Tools for Decoding the Structure-Function Relationships of Biopolymers in Nanotechnology and Glycobiology

by

Venkataramanan Soundararajan
Indian Institute of Technology Guwahati, India

Submitted to the Department of Biological Engineering
in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy
in
Biological Engineering

at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

April 2010

© 2010 Massachusetts Institute of Technology. All rights reserved.

Received by

Venkataramanan Soundararajan
Department of Biological Engineering
April 7, 2010

Certified by

Ram Sasisekharan
Professor of Health Sciences & Technology and Biological Engineering
Thesis Supervisor

Accepted by

Darrell Irvine
Chairman, Committee for Graduate Students
Thesis Committee Members

Alan J. Grodzinsky
Director, Center for Biomedical Engineering
Professor, Department of Biological Engineering
Professor, Department of Mechanical Engineering
Professor, Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology, Cambridge, USA

Rahul Raman
Director, Bioinformatics Core, Consortium for Functional Glycomics
Research Scientist, Harvard-MIT Division of Health Sciences and Technology
Massachusetts Institute of Technology, Cambridge, USA

Ram Sasisekharan
Director, Harvard-MIT Division of Health Sciences and Technology
Edward Hood Taplin Professor of Biological Engineering and HST
Koch Institute for Integrative Cancer Research
Massachusetts Institute of Technology, Cambridge, USA
Abstract

In this thesis, new tools have been developed for decoding structure-function relationships governing complex biopolymers that have emerged as key players in biology, biotechnology, and medicine. Specifically:

(1.) The first part of this thesis addresses the structure-function relationship of synthetic biodegradable plastics that are at the forefront of nanotechnology for spatiotemporally-regulated targeting and sustained release of drugs to treat Cancer and other chronic diseases. A Voxel-based 3-D platform for accurately simulating all phases of polymeric nanoparticle erosion and drug release is introduced. Using the developed Voxel platform, the release of anti-inflammatory and anti-cancer drugs such as doxorubicin and dexamethasone from poly lactic-co-glycolic acid (PLGA) nanoparticles is precisely predicted. The Voxel platform emerges as a powerful and versatile tool for deducing the dynamics in interplay of polymer, drug, and water molecules, thus permitting the rational design of optimal nanotechnology treatments for cancer.

(2.) The second part of this thesis is focused on development of tools to elucidate structure-function relationships of complex polysaccharides (glycans) that specifically interact with proteins to modulate a host of biological processes including growth, development, angiogenesis, cancer, anticoagulation, microbial pathogenesis, and viral infections. First, towards the fine structure determination of complex linear glycans (glycosaminoglycans or GAGs), enzymatic tools are developed for both depolymerizing GAGs at specific linkages and for effectively modulating their functional groups. Specifically, the development and integrated biochemical-structural characterization of the Chondroitinase ABC-II enzyme that depolymerizes dermatan sulfate and chondroitin sulfate GAGs (CSGAGs), and the 6-O-Sulfatase and N-Sulfamidase enzymes that de-sulfate functional groups on heparin and heparan sulfate GAGs (HSGAGs) are described. Second, the interaction of branched glycans with proteins is analyzed using the interplay of Influenza virus surface proteins (mainly Hemagglutinin and Neuraminidase) with host cell surface sialylated glycan receptors as a model system. For this purpose, the novel triple reassortant "Swine Flu" pandemic virus (or 2009 H1N1 virus) is studied. Finally, in order to overcome the challenges facing protein structure prediction in the "Twilight Zone" of low homology that is rampant in glycan-binding protein (lectin) families, a new tool is introduced for modeling the 3-D structure of proteins directly from sequence. Specifically, it is identified that protein core atomic interaction networks (PCAINs) are evolutionarily non-tinkered and fold-conserved, and this finding is utilized towards assignment of folds, structures, and potential glycan substrates to lectin sequences. It is further demonstrated that the developed tool is effective universally; thus emerging as a promising platform for generic protein sequence-to-structure and function mapping in a broad spectrum of biological applications.

Thesis Supervisor: Ram Sasisekharan, Professor of Biological Engineering and Director of Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, USA
Dedicated to VS

for my conversations with Science
Acknowledgments

To my advisor, Dr. Ram Sasisekharan: sincere thanks for encouraging me to think out-of-the-box; for teaching me the value of focus and razor sharp vision while still permitting me to explore the wide vistas of Science; for the timeless conversations on a broad spectrum of scientific and technological topics; for the think-big and can-do attitudes that have been a privilege to learn from; for supporting me through difficult times on personal fronts; and most of all, for providing a rich and vibrant research environment with a fantastic set of individuals where Science is Science in its all-encompassing, limitless, unbounded beauty.

To my committee member, Dr. Rahul Raman: thanks for providing me the handholding that I needed to get started and going with this thesis research; for introducing me to glycobiology and teaching me glycomics tools; for the limitless patience exuded in steering me as I've erred and learnt; for finding time for our conversations despite the several other responsibilities; for showing by example how electrical/computer engineers can apply and excel in Biological Engineering; and for the constant words of wisdom on all aspects of graduate life that have greatly helped me over the last five years.

To my committee member and chair, Dr. Alan Grodzinsky: it has been my fortune to interact and learn from you right from my first two semesters at MIT when I was your student in 20.410 and 20.430. Your interdisciplinary lectures and discussions fusing electrical, mechanical, chemical, and biological engineering have had a profound impact on the way I approach research. I've drawn a lot of inspiration from the remarkable energy and enthusiasm with which you teach, and I hope to imbibe these qualities towards my career in academia. Thanks also for encouraging and supporting me throughout this thesis research.

To S. Raguram: our numerous long discussions on protein science have been deeply inspiring and will remain etched as few of the best times of my graduate life. Thanks for keeping me grounded, while always encouraging me to think beyond, and patiently supporting my flights of imagination with kind words however naive my thoughts. In you, I found a mentor with an identical background and vision, which provided me strength in times of self-doubt and further energy when things went well. I look forward to our building on the solid PCAIN-based foundations we've established.

To Dr. David Eavarone and Dr. Toomas Haller, it was a pleasure working with both of you on the development of the Voxels platform. I'd like to specially thank David for the patient guidance and focus provided during the initial stages of the project.

To Dr. Vikas Prabhakar, Dr. Ishan Capila, Dr. James Myette, Dr. Zachary Shriver, and Dr. Jonathan Behr: sincere thanks for the powerful set of experimental tools and protocols developed as part of the GAG enzyme toolkit, that provided a fertile ground for developing homology-based models to study enzyme structure-function relationships. These tools were immensely beneficial for the progress we made during the first few years of this Thesis research.

To Dr. Karthik Viswanathan: it was an incredible experience learning the fundamentals of experimental molecular biology from you. It was due to your mentorship that I got exposure to hands-on molecular biology, and will always remember those few months fondly. Thanks sincerely, for your exceptional patience in steering me through the several new and fascinating aspects of experimental protocols. Thanks also for the numerous long lunch sessions that helped me better understand graduate life nuances.
To Dr. Aarthi Chandrasekaran: thanks for being an amazing mentor during the many stressful events of my first few years in the lab. Talking to you made it all seem so easy; and for that I am truly grateful.

To Dr. Kannan Tharakaraman: one of the earliest lessons I learnt as a graduate student was "time spent with Kannan is time well spent". Thanks sincerely, for making the last four years an exciting, informative, and enjoyable experience. The two weeks of intense, burning-the-midnight-oil, research frenzy that we undertook together during the outbreak of the 2009 H1N1 "Swine Flu" influenza pandemic virus, culminating in our integrated sequence-structure understanding of the virus and our subsequent submission to Nature Biotechnology, was one of the most fast-paced phases of this Thesis research. For this and the several other brainstorming sessions we've had on various projects, I am most grateful.

To Dr. Vidya Subramanian: thanks for the many wonderful discussions on protein engineering, and I look forward to all the emerging excitement in our horizon.

To Kenneth Warnock, Neel Patel, Stephanie Yaung, Ganesh Chock, and Molly Eileen Cain -- my exceptional UROP students -- sincere thanks for your brilliant hard work. It has been a genuine pleasure working with each of you, and you've all enriched my graduate life in many special ways. Added thanks to Ken -- my first UROP student -- for his remarkable efforts in piecing together the several gigabytes of evolutionary information from the protein fold universe towards building our PCAIN database, as well as the numerous long philosophical discussions on intriguing topics such as the creation-evolution balance, origins of life, computers as man's greatest evolutionary strengths, suppressing evolution of drug-resistant viruses, and decoding protein misfolding in the context of neurodegenerative diseases.

To the wonderful MIT Biological Engineering faculty, a big thanks. To Dr. Forest White, Dr. Ernest Fraenkel, and Dr. Neda Bagheri: it was a genuine pleasure teaching 20.320 with you. Thanks for the incredible opportunity to teach and learn. To Dr. Doug Lauffenburger: special thanks for always being supportive and encouraging of my academic and research pursuits. To Dr. John Essigman: thanks for making learning biochemistry so much fun - I will always remember your lectures fondly.

To my exceptional TAs, Dr. Jared Toettcher, Dr. Mikhail Shapiro, and Dr. Victor Lelyveld: it was an incredible experience learning from the three of you -- thanks for making grad school courses so much fun. To my ever-helpful BE grad seniors, especially Dr. Shan Wu, thanks for helping me find my place in BE, coming from a vastly different academic background.

To Ada Ziolkowski: sincere thanks for your patient and timely help throughout the last five years. Thanks also to Dalia Fares, Mariann Murray, and Linda Bragman, for facilitating the various administrative aspects during the course of this Thesis.

To members of the Sasisekharan group, BE colleagues, MIT friends, IIT friends, and school friends from India: thanks for making the whole ride memorable. To all my teachers right from school to college: none of this would have been remotely possible if not for your patient nurturing and loving guidance.

To my father Mr. K. Soundararajan, mother Mrs. S. Anjana Jayanthi, and grandparents: you have been instrumental for all things good in my life. To my brother Harish Chandra: you have been a phenomenal "co-passenger" in the ride called life, particularly a remarkable
colleague during all our conversations on science and technology; and I wish you the best as you embark on your PhD in Molecular Biology at Trinity/Cambridge/LMB. To my in-laws and family: thanks for all the support and well-wishes.

To my ever-loving wife Barghavi: my past, present, future, and beyond; my here, there, and everywhere; are all what they are, because of you.

To Dr. Viswanathan Sasisekharan, my teacher, mentor, and father-figure, who made me realize that imagination is but the only limit of Science: for all things that have been said and remain unsaid... this Thesis is dedicated to you.

Venkataramanan Soundararajan
April 2010
Cambridge, Massachusetts, USA
# Table of Contents

Abstract ..............................................................................................................................................5

1 Introduction .................................................................................................................................... 15
   1.1 Motivation ................................................................................................................................... 15
   1.2 Biopolymers in nanotechnology ............................................................................................... 16
   1.3 Biopolymers in glycobiology ...................................................................................................... 27
   1.4 Summary and specific objectives ............................................................................................... 47

2 Deducing Structure-Function Relationships of Biopolymers in Cancer Nanotechnology ......53
   2.1 Predicting kinetics of anti-cancer drug release from polymeric nanoparticles .........................54
       Published manuscript ............................................................................................................... 66
   2.2 Harnessing Nanotechnology to target the Cancer Cell Immortality Spectrum ............................71
       Published manuscript ............................................................................................................... 88
       Published manuscript ............................................................................................................... 103

3 Recombinant Enzymatic Tools for Decoding GAG Polymer Fine Structure .........................122
   3.1 Decoding the structure-function relationships of GAG enzymes .............................................123
   3.2 Structure-function relationship of Chondroitinase ABC-II from Proteus Vulgaris .......................124
       Published manuscript ............................................................................................................... 127
   3.3 Structure-function relationship of 6-O-Sulfatase from Flavobacterium Heparinum .....................140
       Published manuscript ............................................................................................................... 144
   3.4 Structure-function relationship of N-Sulfamidase from Flavobacterium Heparinum ....................156
       Published manuscript ............................................................................................................... 159

4 Influenza Virus Surface Proteins Interactions with Host Sialylated Branched Glycans .............172
   4.1 Introduction ............................................................................................................................. 173
       Published manuscript ............................................................................................................... 188
       Supplementary information ...................................................................................................... 192

5 De novo Protein Structure Prediction and Lectin Sequence-to-Function Assignment .............197
   5.1 Pursuit of the Elusive Protein Fold Code: Implications of Ab Initio Mapping of Gene Sequences to Folded Protein Structures ................................................................. 198
   5.2 Anfinsen's Dogma in the Post-Genomics Era: Significance and Implications of the Twilight Zone of Homology .............................................................................................................. 199
   5.3 Nature as a Tinkerer: How Evolutionary Tinkering Challenges Illuminating the Twilight Zone of Homology .................................................................................................................. 200
   5.4 The Twilight Zone of Glycobiology: The Need for, and Challenges Posing, Ab initio GBP Structure Prediction ......................................................................................................................... 203
   5.5 Evolutionary Tinkering as Noise Masking De Novo Deduction of Fold Codes: The Signal-to-Noise Ratio of Molecular Biology .............................................................................................. 209
   5.6 The Core of it All: Shielding from Water as a Factor Mitigating Evolutionary Tinkering in Protein Folds.................................................................................................................................................. 210
   5.7 Mining Protein Contact Maps from the Multi-dimensional Protein Universe .............................. 211
   5.8 Introduction to the (p,w) Landscape of Protein Structural Folds: A Valuable Tool for Homology Modeling in the Twilight Zone ............................................................................................. 212
       Published manuscript ............................................................................................................... 218

Bibliography .......................................................................................................................................232
### List of Figures

1.1 Mechanisms of polymeric matrix degradation .......................................................... 17  
1.2 Modes of polymeric matrix erosion ......................................................................... 18  
1.3 Active targeting of nanoparticles to tumor-specific over-expressed receptors .......... 19  
1.4 The broad-spectrum of Nanotechnology based approaches for targeting Cancer ........ 20  
1.5 Release of agents from drug encapsulated and drug-conjugated polymeric nanoparticles .................................................. 22  
1.6 Synergistic, additive, antagonistic, and unknown interactions between common cancer drugs ... 23  
1.7 Diverse drug release profiles from the same polymer-drug combination of nanoparticles ........................................... 24  
1.8 Classification of glycans........................................................................................... 25  
1.9 Chemical structures of GAG disaccharide repeat units ............................................ 28  
1.10 Known GAG-protein interactions and their biological roles .................................... 30  
1.11 The Non-templated Biosynthesis of GAGs ............................................................ 31  
1.12 Ribbon rendering of chondroitinase B ................................................................. 33  
1.13 Contrast of the endolytic mode of HepAC with exolytic mode of ArthroAC .......... 34  
1.14 Molecular surface rendering of cABCII depicting its structural complex with DS .................................................. 35  
1.15 Sequential degradation and processing of heparin and HSGAGs by GAG enzymes ........................................... 36  
1.16 Depiction of FGF bound to HSGAG ...................................................................... 39  
1.17 Chemical diversity of branched glycans .............................................................. 40  
1.18 Classes of lectins and their glycan specificities .................................................... 41  
1.19 Protein structural topologies for extracellular domains in lectin classes .............. 42  
1.20 Influenza virus HA engagement of cone topology and umbrella topology glycans .... 43  
1.21 Breakdown of homology modeling in the twilight zone of low sequence identity .... 44  

2.1 Mode of clearance and chemical structure of polymers in anti-cancer applications ........ 52  
2.2 The 2-D Pixel-based models contrasted with 3-D Voxel-based whole nanoparticle model ............... 55  
2.3 Contrast of the neighborhoods of pixels (2-D) and voxels (3-D) ................................ 57  
2.4 The limited computational cost associated with 3-D voxel simulation of nanoparticles ........................................... 55  
2.5 Correlation between experimental and model results for drug release from nanoparticles ........................................... 60  
2.6 The Cancer Cell Immortality Spectrum .................................................................. 67  
2.7 The versatile degradation lifetimes of polymeric nanoparticles .............................. 68  
2.8 Structures of receptor molecules that are over-expressed in breast cancer cells ......... 69  
2.9 Estimated levels of over-expressed Breast Cancer receptor molecules .................. 70  
2.10 Structure of Nanocell delivery system ................................................................. 72  
2.11 Simulation of polymeric nanoparticle degradation and erosion with time............... 73  
2.12 Simulation of bulk eroding polymeric nanoparticle degradation and erosion ......... 74  
2.13 Extension of 2-D and 3-D modeling platforms to simulate surface eroding microparticles ........................................... 75  
2.14 Spatial compartmentalization at nanoscales and microscales .............................. 76  
2.15 Simulation of agglomer size distribution with time at fixed agglomeration factor values ........................................... 77  
2.16 Variation of agglomeration factor (AF) on the size of agglomer clusters formed is computed ... 78  
2.17 Simulation of agglomeration with the extended Voxel platform ............................ 79  

3.1 Alignment of the chondroitinase enzymes ............................................................... 119  
3.2 Docking of 14mer C4S (green) to cABCII (gray) highlights collision regions ......... 129  
3.3 Analysis of collision region contributions towards exolytic mode of cABCII ........... 130  
3.4 Ramachandran plot verification of the developed cABCII homology model ............. 132
### 3.5 Structural model of 6-O-sulfatase with docked glucosamine monosaccharide .................................................. 135
### 3.6 Function of HSGAG processing enzymes .................................................................................................................. 149
### 3.7 Known Nitrogen-Sulfur bond cleaving enzymes and their chemical mechanism ......................................................... 150

### 4.1 Influenza Virus Lifecycle .................................................................................................................................................. 166
### 4.2 Influenza Hemagglutinin (HA) and Neuraminidase (NA) serotypes ................................................................................. 167
### 4.3 MALDI-MS glycan profile of Human Bronchial Epithelial (HBE) cells ............................................................................ 168
### 4.4 Cone topology (left) and umbrella topology (right) glycans bind influenza HA .............................................................. 169
### 4.5 The glycan-binding and membrane fusion facilitating sites on Influenza HA ................................................................. 171
### 4.6 Chemical structures of sialic acid (boxed) and popular influenza NA inhibitors ................................................................ 172
### 4.7 Crystal structure rendering of the 1918 pandemic H1N1 NA with Sialic Acid ................................................................ 173
### 4.8 Chemical structures of Adamantane derivatives inhibiting influenza M2 protein ................................................................ 174
### 4.9 The hydrophobic amantadine-binding pocket on Influenza M2 protein ............................................................................. 175
### 4.10 Rimantadine docked to homology model of 2009 H1N1 virus MP2 protein ...................................................................... 176
### 4.11 The glycan-binding and membrane fusion facilitating sites on Influenza HA ................................................................... 171
### 4.12 Influenza Zoonoses ......................................................................................................................................................... 178

### 5.1 Decoding Biological Information Flow in Nature's Lego Set: The Elusive Protein Fold Code ........................................ 189
### 5.2 Estimated Evolutionary Sequence Divergence of Protein Fold Families ........................................................................ 193
### 5.3 The Full-Spectrum of Known Protein Architectures in the Biological Universe .............................................................. 194
### 5.4 Function-driven Evolutionary Tinkering: CDR Loops of Antibody β-Sandwich Folds ......................................................... 196
### 5.5 Poor intra-fold pairwise sequence identities of GBPs ........................................................................................................ 198
### 5.6 Chemical and Structural Diversity of Glycans Binding to β-trefoil Proteins .................................................................. 199
### 5.7 Classification of β-trefoil fold protein domains into GBPs and non-GBPs ........................................................................ 200
### 5.8 Inaccurate structure prediction by current homology modeling tools in the twilight zone ............................................. 201
### 5.9 Depiction of Signal-to-Noise Ratio (SNR) of Molecular Biology ...................................................................................... 202
### 5.10 Protein core/scaffold, surface residues, peptide bonds, and water molecules ............................................................ 203
### 5.11 Protein Contact Maps from architectures in the protein universe .................................................................................. 204
### 5.12 Structure-based multiple sequence alignment of β-trefoil protein fold family ............................................................. 205
### 5.13 The PCAIN-based methodology for protein fold and structure prediction .............................................................. 206-07
### 5.14 Examples of structures obtained with PCAIN approach to homology modeling ......................................................... 208
### 5.15 Structures predicted based on PCAIN methodology for CASP target sequences ......................................................... 209
List of Tables

1.1 Polymeric biomaterials classified according to mode of in vivo clearance ........................................... 15
1.2 Degradation lifetimes of hydrolysis-susceptible polymers ................................................................. 16
1.3 Classes of GAGs and their disaccharide building blocks ..................................................................... 27
1.4 The diversity of protein fold families that constitute GBPs ................................................................. 45

4.1 Available Crystal Structures of Influenza Proteins Compiled from the PDB ........................................ 170

5.1 Diversity of the commonly occurring fold families in Glycan-binding proteins .............................. 197
1. Introduction

1.1. Motivation

Structure is a key determinant of function in biopolymeric molecules. With the exception of the most simple biopolymers however, the structure-function relationships for several of the more complex biopolymers has been challenging to decode [1-3]. This Thesis is motivated by the growing needs from biological, biotechnological, and medical applications, that call for the elucidation of structure-function relationships of complex biopolymers [4-10].

The first part of this Thesis is motivated by the need for uncovering the factors governing hydrolysis kinetics of synthetic biopolymers and their interplay with water and drug molecules [11-16]. The main application is developing tools for integrated "spatiotemporally" regulated targeting of polymeric nanoparticle vehicles for fighting Cancer. Spatiotemporal regulation of polymeric nanoparticle vehicles requires understanding structure-function relationships at both polymeric nanoparticle scales (e.g. for predicting nanoparticle fragmentation) and molecular scales (e.g. for predicting receptor-ligand structural fit) [17-20]. It is envisioned that the developed tools would enable prediction of chemotherapeutic drug release kinetics for a wide variety of FDA approved materials, thus advancing nanotechnology-based cancer treatment.

The second part of this Thesis pertains is motivated by the tremendous biological significance of decoding factors governing the interaction of complex polysaccharides (glycans) with polypeptides (proteins) [21-24]. Glycans are structurally complex and chemically heterogeneous carbohydrates present at the interface of a cell and its environment [25-27]. Glycan-binding proteins (GBPs) specifically recognize these structurally diverse glycans and modulate several biological processes ranging from cell-cell communication, intracellular signaling and catalysis to disease, defense and infection [28-32]. Decoding glycan-protein interactions has been met with several challenges in the past, that this thesis aims to overcome [33-50]. Specifically: (a.) the non-template driven biosynthesis and high information density of glycans complicates elucidation of glycan fine structure; (b.) the fuzzy and often multivalent nature of glycan-protein interactions hinders traditional biomolecular interaction readouts; and (c.) the exceptionally high evolutionary sequence divergence of glycan-binding proteins presents tremendous difficulties for *ab-initio* characterization of the native structural fold and glycan-binding properties directly from sequence. It is envisioned that developing tools that overcome these challenges will permit characterizing biologically salient glycan-protein interactions.
1.2 Biopolymers in Nanotechnology

1.2.1 Biopolymers in medical applications

The 20th century witnessed tremendous progress in the synthesis of polymeric materials derived from biological molecules, termed biopolymers [51-54]. Biopolymers are typically suitable for in vivo use owing to their non-toxic, non-carcinogenic, non-mutagenic, non-allergenic, and non-inflammatory properties [55-60]. Polymers are classified as biodegradable, bioeliminable or permanent, depending on the mode of their in vivo clearance (Table 1.1) [61].

Biodegradable materials are initially in the solid or gel phase and are thereafter broken down into natural metabolites within the body by hydrolytic or enzymatic activity [62-64]. PLGA is an example of a biodegradable material that has been approved for clinical use by the Food and Drug Administration (FDA) [65]. PLGA is completely hydrolyzed into lactic acid and glycolic acid which are components of the Kreb’s cycle and are hence naturally metabolized by the body [66]. The biomaterial revolution of the 20th century was born out of the realization that instability of biodegradable polymers in solution that promotes erosion of their matrices is highly desirable for achieving temporal regulation in biotechnology, pharmaceutical, and medical applications [67-70].

One of the first medical applications of biodegradable polymers -- approved around mid-20th century --- was sutures composed of poly lactic acid (PLA), poly glycolic acid (PGA), and poly lactic-co-glycolic acid (PLGA) [71]. Following this, several other biodegradable polymeric materials such as polydioxanone, polytrimethylene carbonate, polycaprolactone, polyanhydrides, polyorthoesters, and polyphosphazenes have become popular for a multitude of medical applications [72-80]. Some of these other salient medical applications of biodegradable polymers include dental devices, orthopedic fixation devices, cardiovascular devices such as stents and grafts, tissue staples, anastomosis rings for intestinal resection, ligating clips, drug delivery vehicles, in vivo imaging devices, and tissue-engineering scaffolds [81-90].

At the heart of each of these medical applications is the precise regulation in temporal kinetics of polymer degradation and opportunities for molecular/cellular targeting made available by the advent of nanotechnology [91-95]. The next two sections introduce factors governing the temporal regulation of biopolymer degradation kinetics and the spatial targeting of nano-devices to cells of interest within the human body.
<table>
<thead>
<tr>
<th>Clearance mode</th>
<th>Polymer</th>
<th>Chemical structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biodegradable</td>
<td>poly(glycolic acid)</td>
<td>![poly(glycolic acid) structure]</td>
</tr>
<tr>
<td></td>
<td>poly(lactic acid)</td>
<td>![poly(lactic acid) structure]</td>
</tr>
<tr>
<td></td>
<td>poly(lactide-co-glycolide)</td>
<td>![poly(lactide-co-glycolide) structure]</td>
</tr>
<tr>
<td></td>
<td>poly(ε-caprolactone)</td>
<td>![poly(ε-caprolactone) structure]</td>
</tr>
<tr>
<td></td>
<td>poly(malic acid)</td>
<td>![poly(malic acid) structure]</td>
</tr>
<tr>
<td></td>
<td>poly(ethylene carbonate)</td>
<td>![poly(ethylene carbonate) structure]</td>
</tr>
<tr>
<td></td>
<td>poly(propylene carbonate)</td>
<td>![poly(propylene carbonate) structure]</td>
</tr>
<tr>
<td></td>
<td>poly(anhydride)</td>
<td>![poly(anhydride) structure]</td>
</tr>
<tr>
<td></td>
<td>poly(alkyl cyanoacrylate)</td>
<td>![poly(alkyl cyanoacrylate) structure]</td>
</tr>
<tr>
<td>Bioeliminable</td>
<td>poly(ethylene oxide)</td>
<td>![poly(ethylene oxide) structure]</td>
</tr>
<tr>
<td></td>
<td>poly(methyl methacrylate)</td>
<td>![poly(methyl methacrylate) structure]</td>
</tr>
<tr>
<td></td>
<td>poly(hydroxyethyl methacrylate)</td>
<td>![poly(hydroxyethyl methacrylate) structure]</td>
</tr>
<tr>
<td>Permanent/Implant</td>
<td>poly(N-isopropyl acrylamide)</td>
<td>![poly(N-isopropyl acrylamide) structure]</td>
</tr>
</tbody>
</table>
1.2.2 Temporal regulation of polymer degradation kinetics and its biomedical applications

Why is temporal regulation of polymer degradation desirable for biomedical applications? The answer lies in the inherent importance of time as a key moderating factor in biomedical treatments. For example, a biodegradable implant that has been carefully engineered to degrade at the rate of bone healing can serve to optimally transfer load from the implant to the healing bone, thus negating the need for a second surgery to remove the implant [96]. Yet another example that highlights the importance of polymer degradation kinetics is drug delivery. It is well known that the controlled release of drugs is key to achieving maximal therapeutic efficacy [97]. This necessitates factoring the desired polymer degradation kinetics -- and as a consequence the drug release rate -- into the design of therapeutic delivery vehicles.

The duration of therapy for various diseases ranges from a few hours to several months [96]. Application-specific drug delivery devices constituted of biodegradable polymers have been developed across the therapeutic duration spectrum in recent years (Table 1.2) [96]. The characteristic degradation half lives of biodegradable polymers varies from a few minutes to several years based upon the ease of their hydrolysis [97]. Matrices composed of biodegradable polymers degrade by cross-link, side chain, or backbone degradation (Figure 1.1), and the polymer molecular weight plays an important role in determining the average degradation lifetimes [97]. Other characteristics of biodegradable polymers that are known to influence their hydrolysis rate include hydrophobicity, steric effects (local structure, glass transition), microstructure (porosity, phase separation, crystallinity), oligomer solubility, autocatalysis, and pH of the medium [97].

<table>
<thead>
<tr>
<th>Polymer class</th>
<th>Degradation half life</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polyanhydrides</td>
<td>&lt; 10 minutes</td>
</tr>
<tr>
<td>Polyorthoesters</td>
<td>~ 5 hours</td>
</tr>
<tr>
<td>Polymers</td>
<td>~ 20 months</td>
</tr>
<tr>
<td>Polyphosphazenes</td>
<td>~ 5 years</td>
</tr>
<tr>
<td>Polyamides</td>
<td>&gt; 50,000 years</td>
</tr>
</tbody>
</table>
Biodegradable polymeric devices are classified as bulk eroding or surface eroding based on the progress of hydrolysis and degradation (Figure 1.2) [70]. Bulk eroding polymer matrices undergo uniform, instantaneous wetting and degradation happens throughout the bulk of the matrix, whereas water diffuses much more slowly into surface-eroding polymer matrices and degradation hence happens only at the exposed surface of these matrices [70]. The drug release profile associated with surface-eroding polymer matrices is generally mono-phasic whereas bulk eroding polymer matrices display more biphasic release profiles with a distinct initial “burst release” phase [98]. During the burst phase, bulk eroding nanoparticles release a significant volume of their drug payload and this is undesirable for therapeutic applications [18]. However, drug molecules may be chemically conjugated to polymer molecules to ensure non-burst release and such vehicles are known as drug-conjugated polymeric devices [18]. The effects of chemical drug conjugation to polymers is described in a later section pertaining to chemotherapy, where such techniques have been extensively explored for the purpose of sustained drug release [98].

In addition to temporal regulation, the ease of chemical conjugation of biopolymers to other synthetic or organic molecules with highly specific interaction spectrums permits their precise targeting to cells or tissues of interest [97]. This spatial targeting of biopolymers is particularly effective for nanoscale devices owing to the several related advantageous properties of these "small scale devices" [96-98], as discussed in the following section.
1.2.3 Spatial regulation of polymeric device targeting: Emergence of biomedical nanotechnology

A nanometer is one billionth of a meter and devices that are less than hundreds of nanometers in dimension are known as nanoscale devices or nano-devices. These devices are typically smaller than human cells and are able to achieve cellular entry [97]. Such devices can further readily interact with biological macromolecules owing to their comparable dimensions [97]. Nanotechnology, hence, presents an unprecedented insight into the complex regulatory and signaling network of biomolecular interactions that constitute all cellular functions. A broad-spectrum of novel nanotechnology based techniques have been proposed for fighting various diseases, giving rise to the promising field of biomedical nanotechnology [96-98].

The exquisite sensitivity offered by nanotechnology in terms of the ability to monitor sub-cellular compartments and biological molecules in vivo, has presented clinicians with a paradigm shift in the diagnosis of diseases at their most elementary stages [99]. The design of nanoscale devices for detection and diagnosis has motivated a broad-spectrum of advancements in molecular sensing, nano-electromechanical systems (NEMS), nano-fluidics, and ultrasensitive imaging technologies [97]. On the therapeutic front also, nanotechnology is revolutionizing the spatiotemporally regulated delivery of drugs to precisely target cells and molecules of interest [96]. For instance, the shift from macro- to nano-scale drug delivery systems in recent decades has been driven by improvements in polymer formulation technology and evidence that medical applications such as chemotherapy are significantly enhanced by the use of biopolymeric nanoscale delivery vehicles [98].
One of the biggest achievements of nanotechnology in the context of biomedicine is in spatial targeting of nano-devices with drug and imaging agent payloads precisely to the tissues, cells, and molecules of interest [20]. Such a mode of spatial localization of nano-devices within the human body is referred as active targeting and involves the chemical conjugation of targeting ligands such as polysaccharides, antibodies, peptides, nucleic acid aptamers, or other small biomolecules, to the surface of nanodevices [96]. For the purposes of active targeting, biomolecules are carefully screened to identify potential targeting ligands that bind selectively to receptors that are unique to, or over expressed in, the cells of interest (Figure 1.3) [97].

Active targeting is particularly important in the context of tumor therapy for minimizing toxicity and ensuring maximal delivery of chemotherapeutic drugs to cancer cells [20]. Taken together with the concomitant developments on biopolymeric device engineering fronts that permits precise temporal regulation in drug release, as described previously, tremendous opportunities are emerging for integrated spatiotemporally regulated diagnostics and therapeutics for combating cancer on various fronts (Figure 1.4) [96-99]. These emerging opportunities for nanotechnology in targeting cancer are discussed in the following section.

Figure 1.3 Active targeting of nanoparticles to tumor-specific over-expressed receptors
1.2.4 Biodegradable polymeric nano-devices in anti-cancer applications

The development of nanoscale systems for targeted delivery of drugs to cancer cells is a major focus of chemotherapy, primarily for the purpose of toxicity reduction [20]. Nanoscale systems that are smaller than 200 nm in diameter effectively navigate through the leaky tumor microvasculature and aggregate selectively within the tumor interstitial space [62]. They are thereafter contained within the tumor microenvironment owing to the dysfunctional lymphatic drainage [63]. Devices that rely primarily on this enhanced permeation and retention (EPR) for delivering drugs to tumors are known as passively targeted or auto-targeted systems [64]. Active targeting, as defined earlier, involves conjugation of targeting ligands -- that bind selectively to receptors over-expressed in cancerous cells -- to surface of nano-devices (Figure 1.3) [96-98].

Figure 1.4. The broad-spectrum of Nanotechnology based approaches for targeting Cancer
The conjugation of appropriate tumor-specific targeting molecules onto drug-encapsulated nano-devices ensures selective delivery of the agent to cancerous cells and hence minimal toxicity to noncancerous cells [96]. Although passive and active targeting are both helpful for directing nanoscale devices to tumor tissues, the latter mode is associated with greater reduction in chemotherapeutic toxicity, specifically to healthy cells in the immediate neighborhood of tumor cells [97].

Chemical conjugation is also useful in the context of drug and polymer molecules for the purpose of prolonging the release duration (sustained release), as mentioned earlier [18]. Sustained release of agents over several weeks is required for many applications such as monitoring of the tumor microenvironment and anti-cancer drug delivery [100, 101]. The chemical conjugation of drug molecules to polymer fragments prevents rapid dissolution of the drug into the medium [18]. The mechanism governing sustained release from drug-conjugated polymeric nanoparticles is as follows (Figure 1.5). Upon repeated hydrolytic or enzymatic cleavage of the polymer backbones, the fragments become increasingly smaller in size. The dissolution of the drug–polymer conjugates commences only after the molecular weight of these fragments decreases below a certain threshold that is characteristic of the polymer. The dissolved drug molecules then begin diffusing out of the nanoparticle.

An added advantage of drug-conjugated nanoparticles is the absence of the initial burst release phase which, as was discussed earlier, is characteristic of drug-encapsulated polymeric nanoparticles [18]. Burst release in the latter occurs primarily because of the near-instantaneous efflux of medium-exposed drug molecules present in the vicinity of the carrier surface. In the case of drug-conjugated polymeric nanoparticles, the chemical binding of the drug molecules to the polymer backbone prevents instantaneous diffusion of even these medium-exposed drug molecules on the carrier surface.

Drug-conjugated polymeric nanoparticle systems have been successfully used for the sustained delivery of chemotherapeutic agents to several cancer types [18, 96-98]. For example, PLGA-doxorubicin nanoconjugates have been employed to effectively deliver the chemotherapeutic drug doxorubicin to tumors over several weeks [18]. Additionally, the polymers constituting the nanoparticles may be surface functionalized with PEG and tumor-targeting molecules, similar to other nano-devices as described earlier [96-98].
Another unique aspect of cancer therapy is the requirement for drugs combinations owing to the complexity of the disease and failure of single-shot approaches to provide effective treatment [102-122]. Cancer is a complex disease that arises from the dysregulation of several biological networks that are frequently interconnected [105-108]. Consequently, the single drug–single target approach of mono-therapies is often less effective than combination chemotherapies that can synergistically target multiple processes simultaneously [102-122]. This is verified by assessments of numerous successful clinical combination therapies for cancer. Therapeutic synergy involves one or more of the following effects — multiplicative increase of cancer cell death, decreased dosage requirement of each drug, reduced toxicity, and minimized development of drug resistance [102]. The last three effects are a natural consequence of using multiple drugs because the decreased requirement for each drug alleviates issues of toxicity and drug resistance. A compilation of synergistic interactions amongst some of the commonly administered cancer drugs indicates an abundance of mechanistic synergy that can be harnessed by combination therapies (Figure 1.6) [102-122].

However, the effective administration of combination therapies requires very precise spatiotemporal control in the release of each drug owing to the sensitivity associated with functional rewiring of cellular biochemical signaling networks [96]. This in turn calls for the design of sophisticated nanoscale delivery platforms that can carry multiple payloads and disburse them selectively to the cancerous tissues in a spatially and temporally regulated fashion [97-99]. The next section outlines the challenges associated with the design of spatiotemporally regulated polymeric nanoparticle systems and the need for computational nanotechnology platforms for overcoming the challenges towards ushering in an era of novel nano-devices for fighting cancer.
Figure 1.6. Synergistic (■), additive (■), antagonistic (■), and currently unknown (□) interactions between common cancer drugs. The following three letter abbreviations are used for the drugs: Capcitabine (Xeloda®), Cyclophosphamide (Cytoxan®), Docetaxel (Taxotere®), Doxorubicin (Adriamycin®), Epirubicin (Pharmorubicin®), Fluorouracil (5FU), Gemcitabine (Gemzar®), Lapatinib (Tyverb®), Letrozole (Femara®), Methotrexate (MTX), Mitoxantrone, Paclitaxel (Taxol®), Tamoxifen (Nalvodex®), Trastuzumab (Herceptin®).

1.2.5 Challenges in predicting drug release kinetics of polymeric nanoparticles: Emergence of computational nanotechnology

The rapid progression from micro- to nano- scale drug delivery systems in recent years has been in part due to improvements in polymer formulation technology as well as accumulating evidence that medical applications such as chemotherapy are significantly enhanced by the use of nanoscale delivery vehicles [19-21, 69, 123]. However, because drug release profiles differ based on the polymer, drug, and design parameters, the availability of computational models to understand and predict drug release is very valuable for the extension and optimization of existing drug delivery technologies [98]. While several models have been reported in the literature to describe drug release and particle breakdown for macro- and microscale polymer-based delivery vehicles [14-17, 124], there is an absence of such models in the nano- realm that is of great importance to cancer therapy, as outlined earlier. From the studies on polymer degradation, it is clear that there are a number of factors that influence drug release
kinetics including water and drug diffusion, drug dissolution, polymer molecular weight, particle size and geometry, polymer degradation, micro-environment pH changes, autocatalysis, polymer swelling, and more [124-126]. Incorporation of these factors is further complicated by a still incomplete understanding of the extent to which they each influence the kinetics of drug release [98]. As a result of this complex interplay between water, polymer, and drug molecules within the milieu of nanoparticles, the prediction of degradation kinetics of polymeric nanoparticles and rates of drug release has been extremely challenging [98].

However, such prediction of drug release kinetics from polymeric nanoparticles is key to optimizing the design process in terms of time and finances, as well as ensuring the highest therapeutic efficacy and maximal feedback vis-a-vis optimal release rates for different tumor types [17, 69]. Furthermore, a broad-spectrum of drug release rates are observed for the same polymer and anti-cancer drug combination (Figure 1.7), thus suggesting that several intricate factors are involves in determining release rates from polymeric nanoparticles, and necessitating the developing of an effective predictive tool [98].

![Figure 1.7](image.png)

**Figure 1.7.** Extremely diverse drug release profiles from the same combination of polymer (PLGA) and drug (dexamethasone) employed in polymeric nanoparticle based anti-cancer systems necessitates the development of computational models that can accurately predict the kinetics of drug release from polymeric nano-devices.
Computational nanotechnology is quickly emerging as a fertile ground for the development of models to understand theoretical nanosciences and designing new nanoscale technologies for targeting Cancer [127]. Computational nanotechnology refers to the design of models for elucidating the mechanistic underpinnings of nanoscale phenomena. In Chapter 2 of this thesis, a novel platform based on the principles of computational nanotechnology is introduced for overcoming the challenges described herein and provide accurate predictions on the drug release kinetics from polymeric nanoparticles [98]. The efficacy of this platform is demonstrated on a wide variety of polymer, drug, and cancer types, thus validating the utility and versatility of the technique [96, 97].

1.3 Biopolymers in Glycobiology

1.3.1 Glycans and glycan-binding proteins: Biological roles and classification into families

Glycans are structurally and chemically heterogeneous carbohydrates present at the cell-extracellular matrix (ECM) interface [128-141]. There is increasing evidence for the role of glycans in several physiological and pathological processes like cell growth, development, tumorigenesis, infection, anticoagulation, immunity, and cell-cell/cell-ECM communication [142-159]. Glycans can be broadly classified according to their backbone chemical structure into (a.) branched sugars - these are presented as N-linked or O-linked glycosylations on glycoproteins/glycolipids, and (b.) linear sugars - a majority of which are glycosaminoglycans (GAGs) that contain long polymers O-linked to a core protein forming proteoglycan aggregates (Figure 1.8) [160-173]. While glycans modulate biological processes generically by virtue of their mass, shape, charge, or other physical properties, several of their more specific functions are attributed to their recognition by glycan-binding proteins (GBPs) [174-177]. Owing to their ubiquitous presence at the cell-ECM interface, glycans interact with numerous GBPs such as growth factors, cytokines, immune receptors, and enzymes (Figure 1.8) [178-180]. GBPs hence serve to decipher the Glycome of biological systems by recognizing and binding specific glycans from the cell's array of complex polysaccharides [181-185]. The underlying glycan-protein interactions are typically characterized by a rich spectrum of specificities, avidities and affinities, allowing for the precise moderation of several different biological processes [186-198]. GBPs are classified as GAG-binding proteins and lectins [199-205].
Figure 1.8. Glycans are classified as simple and complex, and as linear (GAGs) and branched, as shown. GAGs interact with GAG-binding proteins and are processed by GAG enzymes, whereas Lectins interact with simple and branched glycans. The Consortium for Functional Glycomics (CFG) nomenclature is adopted for depicting glycans. This picture was adapted from: The Essentials of Glycobiology by A. Varki et al., CSH Press, 2004.

The following sections describes each class of glycans and GBPs, as well as the salient biological roles of glycan-protein interactions.

1.3.2 GAGs and biological implications of their interaction with GBPs and GAG enzymes

The Glycosaminoglycan (GAG) family is further classified according to the dissacharide repeat units composed of hexosamine (galactosamine/glucosamine), hexose (galactose) and hexuronic acid (glucuronic acid/iduronic acid) as well as varied patterns of sulfations (2-O-sulfate/6-O-sulfate/4-Osulfate/N-sulfate), acetylations (N-acetyl) and linkage geometries such as β1-3/β1-4/α1-4 (Table 1.3) [206-209]. Of these, the heparin/heparan sulfate GAGs (HSGAGs) and chondroitin sulfate GAGs (CSGAGs) are two most prevalent GAG families [210-214].
Table 1.3 Classes of GAGs and their disaccharide building blocks

<table>
<thead>
<tr>
<th>Category</th>
<th>Disaccharide</th>
<th>U</th>
<th>H</th>
<th>Modifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heparin/Heparan Sulfate (HSGAG)</td>
<td>(U_2X (\alpha\beta1,4)H_{NY,3X,6X (\alpha1,4)})</td>
<td>IdoA/GlcA</td>
<td>Glucosamine</td>
<td>X – sulfated</td>
</tr>
<tr>
<td>Chondroitin Sulfate (CSGAG) / Dermatan Sulfate (DSGAG)</td>
<td>(U_2X (\alpha\beta1,3)H_{NAC,4X,6X (\beta1,4)})</td>
<td>IdoA/GlcA</td>
<td>Galactosamine</td>
<td>X – sulfated</td>
</tr>
<tr>
<td>Keratan</td>
<td>(Gal_6X (\beta1,4)H_{NAC,6X (\beta1,3)})</td>
<td>Gal</td>
<td>Glucosamine</td>
<td>X – sulfated</td>
</tr>
<tr>
<td>Hyaluronic Acid</td>
<td>(GlcA (\beta1,3)H_{NAC (\beta1,4)})</td>
<td>GlcA</td>
<td>Glucosamine</td>
<td>None</td>
</tr>
</tbody>
</table>

Chondroitin sulfate GAGs (CSGAGs) and dermatan sulfate GAGs (DSGAGs) belong to a family of GAGs known as galactosaminoglycans or GalAGs (Figure 1.9) [215-218]. GalAGs are linear polysaccharides of 1-3 linked repeating disaccharide units. The disaccharide units consist of a uronic acid (\(\alpha\)-L-iduronic acid; IdoA or \(\beta\)-D-glucuronic acid; GlcA) linked 1-3 to a \(\beta\)-D-N-acetyl-galactosamine (GalNAc) [219-222]. Each disaccharide unit can additionally possess variations in the form of sulfation at the 2-O and 3-O positions of the uronic acid and 4-O and 6-O positions of the GalNAc [223-227].

Heparin and heparan sulfate glycosaminoglycans (HSGAGs) comprise an important polysaccharide constituent of many proteoglycans [228-230]. These glycans are linear polymers based on the variably repeating disaccharide unit (uronic acid 1-4 glucosamine)_n, where 'n' represents a variably repeating number, typically 10–200 (Figure 1.9) [231-233]. As present in nature, these sugars possess an extensive chemical heterogeneity that is largely attributed to the mosaic arrangement of O- and N-linked sulfates present at different positions along each sugar chain [234-236]. Additional structural variations include the presence of N-linked acetates at the glucosamine C2 position as well as epimerization of the uronic acid C5 carboxylate to form either\(\beta\)-D-glucuronic acid or \(\alpha\)-L-iduronic acid [237-239]. HSGAGs are present as structurally defined binding epitopes on the cell surface and hence also play an important role in microbial pathogenesis [240-242].
Figure 1.9. Chemical structures of GAG disaccharide repeat units

GAGs are present in an environment comprised of a variety of proteins, such as growth factors, cytokines, morphogens, and enzymes (Figure 1.10) [243]. GAGs play a critical role in assembling protein-protein complexes such as growth factor-receptor or enzyme-inhibitor aggregates on the cell surface and in the extracellular matrix that are directly involved in initiating cell signaling events or inhibiting biochemical pathways [244]. Furthermore, extracellular GAGs can sequester proteins and enzymes and present them to the appropriate site for activation [245]. Thus, for a given high-affinity GAG-protein interaction, the positioning of the protein-binding oligosaccharide motifs along the GAG chain determines if an active complex is assembled at the cell surface or an inactive complex is sequestered in the matrix [247].

High-affinity GAG-protein interactions are not the only biologically significant ones (Figure 1.10) [248]. GAGs have been shown to play important roles in maintaining morphogen gradients across a cell or tissue, which has been implicated in developmental processes [249]. Maintaining a gradient in the concentration of growth factors or morphogens involves graded affinities between different GAG sequences and the protein [250]. Thus, the nature of GAG-protein interactions coupled with their sequence diversity enables GAGs to fine-tune (analog modulation) the activity of proteins, unlike DNA-protein interactions, which block expression of a protein, completely activating or inhibiting (digital regulation) protein function [251].
<table>
<thead>
<tr>
<th>GAG binding proteins and their biological roles</th>
<th>GAG oligosaccharide specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cell growth and development</strong></td>
<td></td>
</tr>
<tr>
<td>FGF-HSGAG: FGF-oligorization, assembling FGF-FGFR complexes leading to receptor oligomerization and cell signaling. Cell growth and development, angiogenesis.</td>
<td>FGF-1 – HSGAG: -(I2S-HNS,6S)n – n &gt; 2 for binding &gt; 5 for FGF-mediated cell signaling</td>
</tr>
<tr>
<td></td>
<td>FGF-2 – HSGAG: -(I2S-HNS,6S)n – n &gt; 2 for binding &gt; 5 for FGF-mediated cell signaling. Sulfation at 6-O position is not required for binding but may be required for cell signaling</td>
</tr>
<tr>
<td>HGF/SF-dermatan: hepatocyte regeneration, morphogenesis, cell motility, tumorigenesis and metastasis.</td>
<td>I-HNAc,4S-I-HNAc,4S-I-HNAc,4S-I-HNAc,4S</td>
</tr>
<tr>
<td>Midkine, pleiotrophin-chondroitin: neuronal adhesion, migration, and neurite outgrowth.</td>
<td>-(G-HNAc,4S,6S)n – or -(G2S-HNAc,6S)n –</td>
</tr>
<tr>
<td>Other growth factors/Morphogens: FGFs (1-21), TGFβ, VEGF, PDGF, EGF Amphiiregulin, Betacellulin, Neuregulin, IGF II, Activin, Sonic Hedgehog, Sprouty peptides, Wnts (1-13), BMP-2, 4.</td>
<td></td>
</tr>
<tr>
<td><strong>Anticoagulation and antithrombosis</strong></td>
<td></td>
</tr>
<tr>
<td>AT-III–heparin: enhances factor Xa and IIa inhibition.</td>
<td>-(HNAc,6S-G-HNS,3S,6S-)</td>
</tr>
<tr>
<td>Annexin V–heparin: enhances protein oligomerization.</td>
<td></td>
</tr>
<tr>
<td>HCF II–DS: inhibition of factor IIa and factor IIa-fibrin complex.</td>
<td>I2S-HNAc,4S-I2S-HNAc,4S-I2S-HNAc,4S</td>
</tr>
<tr>
<td>Other factors/proteases: factor Xa, IIa, Thrombomodulin</td>
<td></td>
</tr>
<tr>
<td><strong>Microbial pathogenesis</strong></td>
<td></td>
</tr>
<tr>
<td>HSV-1–heparin</td>
<td>ΔU-HNS-I2S-HNAc-I2S(orG2S)-HNS-I2S-HNH2,3S,6S</td>
</tr>
<tr>
<td>FMDV-heparin</td>
<td>-(I2S-HNS,6S)n –</td>
</tr>
<tr>
<td>VCP-heparin</td>
<td>-(I2S-HNS,6S)n –</td>
</tr>
<tr>
<td><strong>Inflammation/immune response</strong></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1.10. Known GAG-protein interactions and their biological roles. This picture was adapted from: R. Raman et al., Annual Review of Biomedical Engineering, 2006.

In addition to GAG-binding proteins, GAGs also interact with, and are processed by several GAG enzymes [252]. In fact, fundamental to understanding GAG structure-activity relationships is the appreciation that polydispersity of GAG fine structure is not random, but is the end product of a complex and concerted biosynthetic pathway involving numerous modifying enzymes (Figure 1.11) [253-255]. The relative expression levels and specific activities of GAG enzymes are regulated in a cell- and tissue-specific fashion [256]. This programmed diversity of GAG structure ultimately plays out at a functional level, namely through the dynamic regulation of numerous biochemical signaling pathways relating to processes like cell growth, differentiation, death, intercellular communication, adhesion and tissue morphogenesis [257].
In contrast to the complex enzymatic process by which GAGs are synthesized, GAG catabolism is considerably more straightforward, both in the scope of its purpose and the means by which it is carried out at the biochemical level [258]. Both bacterial and mammalian cells produce enzymes that degrade GAGs [259]. In microbes, GAG degradation processes are utilized by some soil bacteria, like Pedobacter heparinus, for metabolizing GAGs, whereas in others, like Streptococcus pneumonia, the degradation of GAGs in the ECM enhances the virulence of the microbe [260]. The bacterial depolymerizing enzymes are lyases that cleave the glycosidic linkage between a hexosamine and an uronic acid through the β-elimination mechanism [261]. The lyase action leaves a 4,5 unsaturated linkage on the nonreducing end uronic acid, which has a characteristic absorbance at 232 nm. The overall sequence of GAG degradation in bacteria involves the depolymerization of GAG chains by the heparinases and chondroitinases down to di- and tetrasaccharide fragments [262]. These small fragments are then acted on by the exolytic sulfatases, such as ΔU 2-O sulfatase, Δ4,5 glucuronidase, N-sulfamidase (HSGAG), GlcNAc 3-O sulfatase, GlcNAc 6-O sulfatase, GalNAc 4-O sulfatase, and GalNAc 6-O sulfatase [263].
The mammalian GAG-degrading enzymes, including heparanases and hyaluronidases, are glycosidases that hydrolyze the glycosidic linkage and retain the epimeric state of the uronic acid in the products formed [264]. The human heparanase I enzyme has been demonstrated to cleave the GlcA-GlcNS,3S,6S linkage within the oligosaccharide motif -GlcNAc,6S-GlcA-GlcNS,3S,6S-IdoA2S-GlcNS,6S-, where the 2-O sulfation on the adjacent IdoA plays a key role in the activity of the enzyme [265]. The GAG catabolic pathway involves depolymerization of the endocytosed GAG chains by heparanases and hyaluronidases and desulfation of the smaller fragments by lysosomal sulfatases, which are exolytic enzymes analogous to the bacterial sulfatases [266]. In addition to the lysosomal exolytic sulfatases, the endolytic sulfatases appear to cleave the 6-O sulfate group from GlcNAc and GlcNS monosaccharides in mature HSGAGs at the cell-ECM interface [267]. These endosulfatases represent a novel family of enzymes that modify extracellular GAG chains [268].

The next few sections focus on the different GAG enzyme families, the spectrum of substrates they process, and the biological functions of these enzymes.

1.3.3 *The chondroitinase family of GAG enzymes*

Chondroitinases are a family of lyases that depolymerize a wide variety of GalAG substrates (Figure 1.9) [269]. Chondroitinases have been classified broadly into three subfamilies. Chondroitinase AC depolymerizes chondroitin-4-sulfate and chondroitin-6-sulfate, whereas chondroitinase B depolymerizes dermatan sulfate as its sole substrate [270-273]. Chondroitinase ABC has the broadest substrate specificity in that it depolymerizes both CS and DS substrates [274].

Chondroitinases have been employed in attempts to promote functional locomotor recovery following trauma to the central nervous system [275]. The application of cABCI at the site of central nervous system injury is believed to prune CS chains from proteoglycans localized to the glial scar [276]. The absence of these CS chains, inhibitory to axon regeneration, facilitates neural outgrowth and reconstruction of damaged tissue. However, the use of chondroitinases as therapeutics is limited because of the lack of availability of pure and contaminant-free enzyme [277-279]. Further, chondroitinase enzymes are often difficult to handle, because of thermal instability and spontaneous proteolysis, as reported by various studies [280].
Chondroitinase AC (cAC) and chondroitinase B (c'B) from the soil bacterium *Pedobacter heparinus* have been characterized extensively in terms of their enzymatic activity and substrate specificity [281].

![Image](image-url)

**Figure 1.12** (A.) Ribbon rendering of chondroitinase B - DS\textsuperscript{hexa} complex; (B.) Molecular surface of chondroitinase B showing the three DS oligosaccharides near the active site. This picture was adapted from: *W. Huang et al. 1999.*

The crystal structure and co-crystal structure of cB with its DS substrate (**Figure 1.12**) together with site-directed mutagenesis of its putative active site residues provided detailed insights into its substrate processing and also revealed a calcium-dependent catalytic activity [282]. The co-crystal structures of cAC with different CS and DS oligosaccharide substrate complexes (**Figure 1.13**) led to the proposal of multiple scenarios in which the active site residues contributed to the catalytic activity of the enzyme [283]. The crystal structure of another cAC from *Arthrhobacter aurescens* and its co-crystal structure with CS substrates (**Figure 1.13**) provided molecular insights into the active site of this enzyme and also its exolytic mode of action compared with the endolytic mode of cAC from *P. heparinus* [284]. Two distinct broad substrate specificity GalAG-degrading chondroitinase ABC lyases, cABCI and cABCII, have been identified in *Proteus vulgaris* [285]. In fact, the ability of the conventional enzyme known as “chondroitinase ABC” to catalyze the complete depolymerization of GalAG substrates to disaccharides is actually the result of the joint action of cABCI and cABCII [286]. Recently, cABCI from *P. vulgaris* was characterized in terms of its active site, the role of divalent cations in processing CS and DS substrates, and its endolytic mode of action (**Figure 1.14**) [287].
Building on these previous studies, in the second part of this thesis, the cloning, recombinant expression, and integrated biochemical-structural characterization of chondroitinase ABC-II (cABCII) from *P. vulgaris* is described. The development of such a contaminant free, recombinant bacterial cABCII will be a valuable tool for the isolation and structural characterization of GAGs, which includes decoding their primary sequence, and also therapies pertaining to CNS recovery after injury [288]. The native forms of chondroitinase ABC purified from soil bacteria have been used in the past for the isolation and characterization of GAGs in these applications [289, 290]. However, the utility of these native enzymes has been impeded by contaminating activities of other enzymes that co-purify with the target enzyme. Cross-contamination of enzyme activities limits the accurate assignment of sequence information obtained from depolymerization of the parent GAG chain [289, 290]. To circumvent these issues, in this thesis, cABCII is cloned, recombinantly expressed, and biochemically as well as structurally characterized in detail, so as to enhance the utility of the developed enzyme tool for fine structural characterization of CSGAGs and DSGAGs.
1.3.4 Heparin and heparan sulfate processing GAG enzymes

Heparinases I, II, and III (hep I, II, and III) have been cloned from _P. heparinus_ and have been recombinantly expressed in _Escherichia coli_ (Figure 1.15) [291-293]. These enzymes have been characterized extensively in terms of the critical amino acids involved in substrate binding and catalysis, involvement of divalent cations in enzymatic activity, substrate specificities, and action pattern toward defined HSGAG substrates [294-296]. Hep I prefers to cleave substrates starting from the nonreducing end (exolytic), and continues to depolymerize (processive) at successive cleavable linkages toward the reducing end. Hep I was shown to cleave -GlcNS,6X-12S- linkages in highly sulfated regions of HSGAG chains. Hep II from _P. heparinus_ has the broadest substrate specificity among the heparinases. Biochemical investigation of this enzyme has revealed two distinct active sites involved in the processing of regions of high sulfation and low sulfation, respectively [297-299]. The action pattern of hep II is distinct from that of hep I. Specifically, studies indicate that hep II processing proceeds through an endolytic and nonrandom mechanism. In addition, hep II releases intermediates, unlike hep I, which processively cleaves its substrate. Hep III has an orthogonal specificity compared to that of hep I, in that it cleaves GlcNAc,6X-G or GlcNS,6X-G linkages in regions of low sulfation in HSGAGs.
Building on the efforts to develop enzymatic tools for HSGAGs, additional exolytic HSGAG-degrading enzymes (Figure 1.15) were cloned from *P. heparinus* and recombinantly expressed [300-302]. The first of these enzymes was the Δ4,5 glycuronidase, which cleaves an unsulfated Δ4,5 unsaturated uronic acid at the nonreducing end of heparinase-depolymerized oligosaccharides [303-305]. Using the ΔU-HNS,6S disaccharide substrate, the conditions for optimal activity of this enzyme have been determined [306]. The second enzyme, the 2-O sulfatase, cleaves the 2-O sulfate group of ΔU. Using a homology modeled structure of *P. heparinus* 2-O sulfatase, the structural basis for exolytic activity and substrate specificity of this enzyme has been recently determined [307-309]. The unsaturated Δ4,5 linkage was shown to be critical for the activity of the enzyme. Furthermore, it was shown that the 2-O sulfatase generated the substrates for the Δ4,5 glycuronidase because the 2-O sulfate group inhibited glycuronidase activity. These enzymes have since been used in tandem with the heparinases for the compositional analysis of heparin [309-311].
In the second part of this thesis, the cloning, recombinant expression, purification, and integrated biochemical-structural characterization of 6-O sulfatase (cleaves 6-O sulfate of glucosamine) [312] and N-sulfamidase (cleaves N-sulfate of the glucosamine) [313] enzymes are described, this illuminating the entire HSGAG degradation pathway (Figure 1.15). Homology modeling (comparative protein modeling) is used to predict the structure of these enzymes in this study, since crystal structures of heparinase, sulfatase, and other enzymes have been extremely challenging to elucidate, whereas high-resolution homologous structures are already available to serve as templates.

Indeed, homology-based modeling tools are emerging as attractive tools for predicting the structure-function relationships of GAG enzymes universally, as outlined in the next section.

1.3.5 Need for determining the structure-function relationships of GBPs and GAG enzymes: Emergence of homology modeling

From the examples of GAG-protein interactions described above, it is clear that the specificity of biomolecular interactions goes beyond the GAG primary sequence. In the heparin-antithrombin-III (AT-III) complex, the specificity is precisely defined at the level of the pentasaccharide sequence [26, 45, 155]. Any modifications to this sequence are detrimental to the binding and activation of AT-III. On the other hand, in the case of growth factor binding, it is more challenging to define specificity at the level of a single oligosaccharide motif like the pentasaccharide. Recently, the preferences of different FGFs for defined HSGAG sequences were studied using a library of octasaccharides [38, 156, 233, 241]. The authors concluded that they were unable to determine a highly specific motif for each FGF similar to the pentasaccharide sequence that binds to AT-Ill. Further, they concluded that there is a significant overlap in the specificities of different FGFs to different HSGAG sequences and challenged the notion of specificity in these interactions. Intuitively, these differences can be rationalized in the context of the biological pathways involving these interactions. The coagulation cascade requires a tight regulation in a short time span, which involves activation or inhibition of enzymes (digital regulation). On the other hand, interactions involving cell growth and development require some level of degeneracy in the GAG sequence space to accommodate longer time frames for the signaling events as well as maintaining graded affinities for generating morphogen gradients.
Going beyond the sequence space into the 3-D structure space, it is possible to define specificity of GAG-protein interactions as the optimal fit of a GAG structure into the GAG-binding site of the protein. This fit is governed by the ionic interactions of the sulfate and carboxylate groups of GAGs with the basic amino acids on the protein as well as the van der Waals contacts made by the GAG chain with the binding site of the protein. The topology and distribution of the basic amino acids of the GAG binding site on the protein influence its specificity in the molecular recognition of GAG sequences. Thus, specificity as defined by structural fit or molecular recognition can be viewed in the 3-D space for GAGs in terms of conformation of the polysaccharide backbone and the orientation of the sulfate groups. It is possible for more than one GAG sequence to satisfy the constraints imposed by the structural fit of the GAG chain to the protein. Thus it is important to understand how the primary GAG sequence translates into a specific GAG conformation and topology (i.e., sulfate and carboxylate groups). Because GAGs are typically known to adopt a linear extended conformation with helical symmetry, it is more manageable to relate their primary sequence to 3-D structure (unlike the challenging problem of protein structure prediction). However, since elucidating GAG-protein interactions requires structures of both GAGs and proteins to be deciphered, and considering the several challenges with obtaining crystal structures of GAG enzymes as outlined earlier, there is an emerging need to develop tools to address protein structure prediction as well.

Further to the prediction of GBP structure, is the elucidation of factors motivating glycan-binding specificity. It is notable that the concept of protein-protein interaction network analysis has revolutionized the field of biology by changing the traditional single-protein-single-function notion to the powerful “systems” approach to biomolecular interaction analysis [314-316]. In this post-genomics period where there is increasing evidence of the role of protein-polysaccharide interactions in modulating key biological processes, there is great focus in adopting “systems” engineering approaches to decode the structure-function relationships mediating glycan-protein interaction specificity. There is much lesser knowledge presently about the factors that are responsible for specificity of glycan-protein interactions as compared to protein-protein and protein-DNA interactions, owing to the distinctively complex information density of polysaccharides and their multi-layered interplay with proteins, as discussed earlier. GAG-protein interactions such as Fibroblast growth factor (FGF) interaction with HSGAGs (Figure 1.16) are also comparatively well characterized through structure-function and site-specificity studies as compared to lectins (that bind to smaller, simple or branched glycans).
Figure 1.16 Depiction of FGF (C = gray; O = red; N = blue) bound to HSGAG (green) with the HSGAG-binding site residues and binding surface on FGF highlighted in pink.

The FGF is a huge family of β-trefoil fold proteins that have been indicated as important modulators of cancer and other diseases [38, 156, 233, 241]. The beta-trefoil fold itself is extensive across GBPs, and is found co-complexed with several diverse glycans (including heparin in the case of FGF). An analysis of each of these protein-glycan complexes is performed towards the latter half of the second part of this thesis that pertains to developing tools for predicting glycan-binding protein structure.

In addition to GAGs that are linear glycans, branched glycans are also key mediators of important biological processes, as mentioned earlier. The next section focuses on branched glycans and the biological functions arising from their interactions with GBPs (lectins).

1.3.5 Branched glycans and their interactions with Lectins

Branched glycans are present as N-linked and O-linked glycosylation on glycoproteins or on glycolipids (Figure 1.17) [317-321]. There are three major classes of N-linked glycans: high-mannose, complex, and hybrid [322-326]. High-mannose glycans are two N-acetylglucosamines with many mannose residues. Complex glycans contain almost any number of the other types of monosaccharides, including more than the original two N-acetylglucosamines. The oligosaccharide chain is attached by oligosaccharyltransferase enzymes to asparagine occurring in the tripeptide sequence Asn-X-Ser or Asn-X-Thr (where X could be any amino acid except proline) [327-331]. This sequence has come to be known as a glycosylation sequon.
O-linked glycosylation may involve the addition of N-acetyl-galactosamine (O-GalNAc type), fucose (O-fucose type), glucose (O-glucose type), N-acetyl-glucosamine (O-GlcNAc type), mannose (O-mannose), or other moieties, to serine or threonine residues by enzymes [332-337]. This may be followed by addition of other monosaccharides. O-linked glycans of each of these types occur in various tissues within the human body and are involved in several important biological functions [338-345].

Glycan-binding proteins (GBPs), also called as Lectins, decode the glycome of biological systems by recognizing and binding specific glycans from the cell’s array of information dense polysaccharides [346]. The structural diversity in GBPs mirrors the chemical heterogeneity, non-template driven biosynthesis and high information density of the ligand glycan molecules [347-349]. GBPs bind to N-linked and O-linked glycans and mediate cell adhesion, trafficking, and signaling events in inflammation and immune responses [350]. GBPs also act as receptors for viruses, bacteria and other microbes that use the glycans attached to cell-surface glycoproteins as ligands for the host-cell GBPs [351].

Figure 1.17 Chemical diversity of branched glycans. This picture is adapted from: The Essentials of Glycobiology, A. Varki et al, CSH Press, 2004.
**Lectin**

<table>
<thead>
<tr>
<th>C-type (includes selectin, collectins, etc.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S-type (galactins)</td>
</tr>
<tr>
<td>P-type (M6PRs)</td>
</tr>
<tr>
<td>I-type (includes Siglec family)</td>
</tr>
<tr>
<td>calnexin, calcierticulin, calmegin</td>
</tr>
<tr>
<td>hyaluronan-binding proteins (CD44, link protein, verscan, aggrecan)</td>
</tr>
<tr>
<td>frog egg lectins</td>
</tr>
<tr>
<td>ERGIC-53 and VIP-36</td>
</tr>
<tr>
<td>S4GGnM receptor</td>
</tr>
<tr>
<td>ganglioside-binding proteins</td>
</tr>
<tr>
<td>sulfogluconorosyl lipid-binding protein</td>
</tr>
<tr>
<td>insect hemolymph lectins</td>
</tr>
<tr>
<td>interleukins I and II</td>
</tr>
</tbody>
</table>

**Glycan specificity**

- highly variable
- β-galactosides and sequence motif poly-N-acetyl-lactosamines
- Man-6-P on high-mannose-type N-glycans
- sialic acids (Siglecs), others
- glucosylated high-mannose-type N-glycans in the ER
- hyaluronan chains
- galactose, sialic acids, heparins
- glycans in the ER-Golgi pathway?
- sulfated GaINAc residues on pituitary glycoprotein hormones
- sialylated glycolipids
- sulfogluconorosyl glycolipids
- sialic acids
- high-mannose oligosaccharides

**Figure 1.18 Classes of lectins and their glycan specificities.** This picture is adapted from: *The Essentials of Glycobiology, A. Varki et al, CSH Press, 2004.*

The glycan binding sites on lectins typically accommodate mono- to tetrasaccharide glycan ligand motifs [352]. The underlying glycan-protein interactions are characterized by a rich spectrum of specificities, avidities and affinities [353, 354]. The interaction between any single glycan binding site and the glycan motif is low affinity with values in the micro- to millimolar range [355]. Most of the physiological glycan-lectin interactions, however, are multivalent involving binding of an ensemble of glycan motifs to multimeric carbohydrate recognition domains (CRDs) formed by association of lectins [356, 357]. Lectins are either expressed as soluble or membrane-bound proteins in the monomeric or multimeric forms with multiple glycan binding sites [358]. Also, lectins can either be dispersed on the cell surface or localized in a microenvironment such as microvilli or clathrin-coated pits [359, 360].

The main classes of lectins include C-type lectins, S-type lectins (Galectins), P-type lectins, and I-type lectins (Siglecs) (Figure 1.18) [361]. These lectins possess diverse organization of protein topologies in their extracellular domains that house the CRDs (Figure 1.19) [362].
Several crystal structures of lectins from various biological sources with their cognate glycan ligands have been elucidated, allowing for an exploration of lectin-glycan interactions at the level of atomic resolution [363-365]. Along these lines, the following principles regarding lectin binding sites have emerged as a consensus from numerous previous studies [366-371]. First, the binding sites are of relatively low affinity and are found in shallow indentations on the surface of the lectins. Second, selectivity is mostly achieved via a combination of hydrogen bonds (involving the hydroxyl groups of the sugars) and by van der Waals' packing of the hydrophobic face of monosaccharide rings against aromatic amino acid side chains. Third, further selectivity may be achieved by additional contacts between the saccharide and the protein, sometimes involving bridging water molecules or divalent cations. Finally, the actual region of contact between the saccharide and the polypeptide typically involves only one to three monosaccharide residues. As a consequence of all of the above, these lectin-binding sites tend to be of relatively low affinity, but of high specificity [372-374]. The ability of such low-affinity sites to mediate biologically relevant interactions in the intact system thus appears to require
multivalency. These unique and important characteristics of lectins make the study of lectin biology both interesting and significant.

One of the biological systems in which lectin binding to branched glycans becomes particularly significant is in the context of influenza virus surface hemagglutinin (HA) binding to the epithelial cell surface branched sialylated glycans (Figure 1.20) [317-324]. The Spanish influenza pandemic of 1918 caused by the H1N1 subtype virus resulted in ≈20 million to 50 million deaths worldwide [325-331]. The emergence of avian influenza H5N1 viruses that are able to infect humans (>300 known cases) and produce a high mortality rate (~200 deaths) has raised serious global health concerns [332-337]. Significantly, adaptation of these viruses for efficient human–human transmission could result in a new influenza pandemic, a public health disaster of tragic proportions [338-341].

The evolution of pandemic viruses involves crossing-over of avian influenza viruses (natural host) to humans and adaptation to the human host for subsequent infection and human–human transmission [342]. This cross-over is believed to involve mutations in HA that switch its glycan receptor preference from cone topology α2-3 sialylated (α2-3) to umbrella topology α2-6 sialylated (α2-6) glycans found abundantly in human upper respiratory epithelia [343]. The human upper respiratory tract α2-6 receptor adaptation of HA is a critical step in permitting the viruses to infect and efficiently replicate in these tissues, leading to rapid human–human transmission [344].

Figure 1.20 Influenza virus Hemagglutinin (HA) engagement of (A.) host surface α2,3 linked sialylated glycans with cone topology; and (B.) host surface α2,6 linked sialylated glycans with umbrella topology. This picture is adapted from: The A. Chandrashekar et al, Nature Biotechnology 2008.
The recent incidence and spread in humans of the 2009 pandemic "swine flu" virus has raised global concerns regarding its virulence, human-human transmission, and anti-viral drug resistance properties [375]. The main cause of the so-called swine flu has been identified as human infection by influenza A viruses of a new H1N1 (hemagglutinin 1, neuraminidase 1) subtype, or '2009 H1N1 strain' [376]. The first cases of human infection were reported in April in the Mexican town of La Gloria in Veracruz; soon after, reported infections occurred in areas of southern California and Texas of USA [377]. In the latter half of this thesis, using a representative 2009 H1N1 strain as our starting point and our intricate understanding of homology modeling and glycan-protein interaction analysis tools, the likely human adaptation, transmissibility, virulence, and resistance characteristics of these viruses are investigated [378].

Our ability to engineer lectins with desired specificity for important therapeutic applications (such as treatment of influenza and other hypermutative RNA viruses) also rests upon deducing and exploiting the factors that dictate the glycan-binding specificity of GBPs [353-361]. Such protein engineering applications require the development of precise tools for deducing the structures and functions of the protein of interest [362-364]. However, elucidation of GBP structure is rife with several challenges on both experimental and computational fronts [363-369]. As outlined earlier, it is not always possible to experimentally determine GBP structure since many of these are transmembrane molecules and there are challenges associated with their expression and crystallization [368-371]. The challenges on the computational front, on the other hand, are a consequence of the extreme evolutionary sequence divergence of lectins (twilight zone) presents hurdles for conventional homology modeling protocols [363-374].

Figure 1.21 Breakdown of homology modeling in the twilight zone of low sequence identity
The next section discusses these challenges facing prediction of GBP structure by homology modeling and the emerging need for development of new tools for this application.

1.3.6 Challenges facing GBP structure prediction: Need for new homology modeling tools

The elucidation of GBP structure is not only fundamental to the characterizing key structure-function relationships like their glycan-recognition specificity, but also for designing novel GBPs for biological or pharmaceutical applications. Given the numerous challenges facing experimental structure determination of GBPs, homology modeling is emerging as a desirable alternative. Homology modeling is an in-silico molecular modeling methodology that builds the structure of 'target' protein based on a 'template' with high sequence identity [379-381]. Although there are plenty of outstanding homology modeling servers and tools, the application of these to structure prediction of GBPs has been limited due to several roadblocks [382-386]. One major limitation is the generally poor sequence identity (even less than 10%) that prevails in GBP fold families, believed to be a consequence of large evolutionary sequence drift caused by the binding of several chemically heterogeneous and structurally diverse glycans to GBPs of the same fold (Figure 1.17). Since homology modeling produces significant inaccuracies in the 'twilight zone' of less than 25% target-template sequence identity (Figure 1.21) -- such as entire folds mispredicted [387-390] -- it is challenging to identify the fold of GBP sequences de novo.

Additionally, glycan-binding proteins are composed of a diverse array of protein folds, that further challenges elucidation of folds from sequence de novo (Table 1.4). Since fold identification is the first towards template selection, target-template alignment, and structural coordinates assignment in homology modeling [391-397], the application of conventional homology modeling tools to GBPs becomes further complicated.

In order to address these challenges and develop a tool for GBP structure prediction, in this Thesis, evolutionarily conserved regions of proteins are examined to decode "non-tinkered" information that is characteristic of each fold family. Such "fold signature" information is then utilized towards accurate identification of protein folds directly from sequence and for the purposes of selecting an optimal template molecule. The latter half of the second part of this thesis focuses on the development of this new tools that can extend the application of homology modeling to the twilight zone of low identity, thereby providing a new platform for high-throughput GBP structural modeling.
Table 1.4 The diversity of protein fold families that constitute GBPs

<table>
<thead>
<tr>
<th>Glycan-binding Proteins (GBPs)</th>
<th>Fold family (α/β/αβ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mannose-binding proteins, Snake coagglutinins, Surfactant proteins, Tetraneclins, Selectins, DC-SIGNs, CD69, CD94</td>
<td>C-type lectin-like (αβ)</td>
</tr>
<tr>
<td>Calnexins, Galectins, Lectin leg-like, Legume lectins, Pentraxins</td>
<td>Concanavalin A-like (β)</td>
</tr>
<tr>
<td>Fibroblast growth factors, Ricin B-like lectins, Agglutinins, Glycosidases, Neurotoxins, Ribosome-inactivating proteins</td>
<td>β-trefoil (β)</td>
</tr>
<tr>
<td>Mannose-binding lectins, Delta-endotoxins, Vitelline membrane outer-layer proteins</td>
<td>β-prism (β)</td>
</tr>
<tr>
<td>E set domains, Fibronectins, Immunoglobulins, Tenascins</td>
<td>β-sandwich (β)</td>
</tr>
<tr>
<td>Antithrombins, Heparin co-factors, Protein C inhibitors</td>
<td>Serpins (αβ)</td>
</tr>
<tr>
<td>Amylases, Chitinases, Glycanases</td>
<td>TIM beta/alpha barrel (αβ)</td>
</tr>
<tr>
<td>Angiogens</td>
<td>RNase A-like (αβ)</td>
</tr>
<tr>
<td>Hevein-like lectins</td>
<td>Knottin (β)</td>
</tr>
<tr>
<td>Receptor associated proteins</td>
<td>RAP domain-like (α)</td>
</tr>
<tr>
<td>Apolipoproteins</td>
<td>Four-helical up-and-down bundle (α)</td>
</tr>
<tr>
<td>Tachylectins, Fucose-specific lectins</td>
<td>β-propeller (β)</td>
</tr>
</tbody>
</table>

1.4 Summary and Specific Objectives

Understanding structure is central to decoding the function of biopolymers. In this Thesis, new tools are introduced for investigating the structure-function relationships of biopolymers in nanotechnology and glycobiology, thus furthering their biological, biotechnological, and medical applications.

Advances in nanotechnology have led to an active interest in developing multifunctional biopolymer-based nanoscale platforms capable of sensing the molecular signatures of tumorigenesis (the birth and development of cancer) and delivering therapeutics specifically to
cancerous cells in order to simultaneously reduce toxicity and increase efficacy. Pioneering contributions to polymer-drug formulation technology have particularly enhanced these efforts by providing for distinct advantages such as sustained release, increased circulation lifetime, specific localization to tumor microenvironments, and biocompatible vehicles. There is growing evidence that (i.) precise targeting of surface-functionalized nanoparticles to over expressed receptors on tumor cells (spatial control), combined with (ii.) rational design of polymer-drug formulations for modulation of factors influencing the release profiles of nanoparticle-encapsulated drugs (temporal control), could together offer dramatic therapeutic benefit. Hence, there is emerging interest in the development of molecular modeling and computational nanotechnology tools for the design of such spatiotemporally-regulated anticancer systems. Specifically, embedding temporal control into the release of anti-cancer drugs is extremely challenging owing to the influenza of several distinct phenomena such as polymer hydrolysis, polymer degradation, polymer erosion, drug dissolution, drug diffusion, nanoparticle fragmentation, and microenvironmental physicochemical changes. Hence, this Thesis introduces a new tool for modeling polymeric nanoparticle degradation and the release of chemotherapeutic agents from such vehicles. The spatiotemporally regulated release of FDA-approved anti-cancer drugs from FDA-approved polymeric systems with the Nanocell technology developed by our group is demonstrated, with compelling results of nanoparticle formulation and tumor shrinkage.

On the glycobiology front, this Thesis builds on the solid foundation of several previous studies from our group that have provided significant impetus to the development and standardization of Glycomics. In this post-genomic era, Glycomics is emerging as an exciting area of research in cellular and molecular biology, with numerous salient roles of glycans in fundamental biological processes being discovered. It is clear from these studies that the biochemical structure of glycans is a chief factor governing the specificity and affinity of their interaction with biologically significant proteins such as morphogens, growth factors, cytokines, chemokines, enzymes, and numerous other signaling molecules. However, decoding structure-function relationships mediating glycan-protein interactions has encountered several unique challenges owing to many factors including high information density, chemical heterogeneity, and lack of amplification tools for glycans; their multivalent and "fuzzy" interaction with GBPs; and the rampant lack of homology between functionally-related GBPs that challenges de novo fold characterization for these proteins. Advancing the field of Glycomics hence requires the
development of new tools to address the key challenges associated with decoding the structure-function relationships governing glycan-protein interactions.

With these as the primary goals, the specific objectives of this thesis are as outlined below. The rest of the thesis is organized as follows. The development of a new whole-nanoparticle modeling platform for predicting drug release kinetics from polymeric nanoparticles and its application to enhance the potency of chemotherapy to different tumors is described in Chapter 2. The engineering and structure-function characterization of the recombinant bacterial enzymatic tools Chondroitinase ABC-II, 6-O-Sulfatase, and N-Sulfamidase for decoding the fine structure of complex linear glycans and facilitating GAG sequencing are described in Chapter 3. In Chapter 4, the structure-function relationships governing interactions of branched glycans with lectins is investigated, specifically in the context of Influenza virus hemagglutinin binding to host cell surface sialylated glycans, using the 2009 H1N1 pandemic influenza virus as a model system. This thesis concludes with Chapter 5 wherein the development of a novel tool for the de novo prediction of protein fold and structure directly from sequence is described, along with applications of this tool for GBP structure prediction towards facilitating the high-throughput decoding of structure-function relationships governing protein-glycan interactions.
Specific objectives of the Thesis

1. Develop a new platform for predicting the kinetics of drug release from polymeric nanoparticles and apply this towards achieving spatiotemporally regulated targeting of anti-cancer drugs to tumor tissues for maximal therapeutic efficacy (Chapter 2).

   - Predict drug release profiles from polymeric nanoparticles constituted of FDA-approved polymer-drug combinations.
   - Investigate the biological significance of temporally regulated release of anti-cancer drugs to tumor tissues.
   - Understand the scope of extending the developed platform towards other spatiotemporally regulated nanotechnology devices for targeting the broad spectrum of diverse attributes that together constitute the "Cancer Cell Immortality Spectrum".

2. Develop and structurally-biochemically characterize recombinant enzymatic tools for elucidating GAG structure and decoding GAG-protein interactions (Chapter 3).

   - Develop and biochemically characterize the structure-function relationship mediating substrate specificity, active site residue functions, mode and mechanism of action, for Chondroitinase ABC-II from *Proteus Vulgaris*.
   - Develop and biochemically characterize the structure-function relationship mediating substrate specificity, active site residue functions, sequence of processing, mode and mechanism of action, for 6-O-Sulfatase from *Flavobacterium Heparinum*.
   - Develop and biochemically characterize the structure-function relationship mediating substrate specificity, active site residue functions, sequence of processing, mode and mechanism of action, for N-Sulfamidase from *Flavobacterium Heparinum*. Also understand the structural basis for novel nitrogen-sulfur bond cleavage by this enzyme as opposed to oxygen-sulfur bond cleavage by all other known sulfatases.
Specific objectives of the Thesis (continued)

3. Understand the structure-function relationship mediating the interaction of branched sialylated host cell surface glycans with influenza virus surface lectins and apply this knowledge to examine the virulence, transmissibility, and anti-viral resistance of the 2009 H1N1 "Swine Flu" influenza virus (Chapter 4).

- Model the hemagglutinin (HA) structure of 2009 H1N1 influenza virus to investigate the human adaptation and receptor binding of the present pandemic.
- Model and analyze the neuraminidase (NA) structure of the 2009 H1N1 influenza virus to elucidate the structure basis for its binding to sialic acid and understand the potential for emergence of viruses resistant to the popular anti-virals tamifu and relenza.
- Model the PB2 and M2 structures and analyze the information content of all other remnant genes of the 2009 H1N1 influenza virus to predict the human-to-human transmission, resistance to adamantane derivative anti-virals, and the overall virulence of the present "swine flu" pandemic.

4. Develop a novel tool for de novo prediction of protein fold and structure directly from its sequence and apply this to shed light on homology modeling in the twilight zone, thus providing for new opportunities to model GBP structure (Chapter 5).

- Analyze evolutionarily non-tinkered regions of proteins to elucidate "fold signatures" that can be used for high-throughput assignment of protein folds to amino acid sequences de novo.
- Understand the biological implications of rampant twilight zone of homology in GBPs and explore opportunities for modeling biologically or medically salient GBPs and understanding their interaction with glycans.
- Investigate the universal applicability of the developed protein structure prediction protocol and explore its utility for illuminating the twilight zone in biologically salient examples from fields such as immunology, protein science, and glycobiology.
2. Deducing Structure-Function Relationships of Biopolymers in Cancer Nanotechnology

Summary

This chapter focuses on developing tools for decoding the factors governing kinetics of drug release from polymeric nanoparticles and applying this towards enhancing the efficacy of chemotherapeutic treatments.

Specifically, a 'Voxel' based whole nanoparticle simulation platform that can account for the complex parameters modulating drug release was developed, to model the kinetics of biopolymer degradation upon hydrolysis and the formation of erosion channels thereof. It was observed that the developed Voxel platform could accurately predict all phases of the experimentally-determined kinetics for polymeric matrix erosion and drug release. The platform was also able to provide insight into the fragmentation of polymeric nanoparticles upon formation of erosion channels, with unprecedented resolution. The platform was further able to incorporate the effects of chemical drug-conjugation to polymer molecules, this providing ample application to anti-cancer sustained drug release kinetics prediction. The Voxel platform hence emerged as a versatile method to delineate the effects of all nanoparticle design parameters quantitatively. The successful deployment of the Voxel method to predict the temporally-regulated release -- of FDA-approved anti-cancer drugs (including doxorubicin and dexamethasone) from nanoparticles constituted of the FDA-approved biopolymer PLGA -- was demonstrated with the Voxel-optimized nanocell platform, with compelling results vis-a-vis tumor volume shrinkage.

These studies resulted in 3 peer reviewed publications in (i.) The Journal of Nanoscience and Nanotechnology; (ii.) Macromolecular Rapid Communications; and (iii.) Handbook of Nanophysics: Nanomedicine and Nanorobotics.
2.1 Predicting kinetics of anti-cancer drug release from polymeric nanoparticles

2.1.1 Need for computational models to understand factors governing biopolymer hydrolysis and nanoparticle erosion

The use of polymeric nanoparticles as drug delivery devices is becoming increasingly prevalent in a variety of biological and therapeutic applications [11]. There is accumulating evidence that chemotherapeutic treatments are significantly enhanced by the temporally regulated release of drugs from polymeric devices [12]. The temporal kinetics of drug release from polymeric devices is dependent on most of the physicochemical properties of the polymeric biomaterials involved [13]. These are typically classified based on their mode of clearance (if at all) from the body (Figure 4a) [14]. Biodegradable polymers are broken down and metabolized within the body, bioeliminable polymers remain structurally intact but are cleared by the renal filtering system (kidneys), and permanent polymers are used for long term in-vivo implants [15]. Of these, there has been great interest in the use of biodegradable polymeric nanoparticles for cancer treatment [16].

![Fig. 2.1 Mode of clearance and structure of polymers in anti-cancer applications.](image)
Despite their widespread use in Cancer diagnostics and therapeutics, the factors influencing the kinetics of drug release from polymeric nanoparticles are still not quantitatively understood [17]. Given that drug release profiles depend on a wide variety of design parameters that moderate the interplay between polymer, drug, and water molecules, this in turn challenges the high-throughput engineering of anti-cancer polymer-based nanoscale devices [18].

Recently, the availability of computational models to understand and predict drug release from macroscale polymer matrices has become popular for the extension and optimization of existing drug delivery technologies [11-20]. However, since the mechanism of drug release from nanoparticles differs from that of macro-particles even when both vehicles have the same polymer-drug composition, direct application of existing methods to nanoparticles is challenging [14-19]. For instance, drug diffusion through the polymer matrix has been shown to be slower in nanoparticles than in microparticles, and this may be due to more dense internal structures as a result of fabrication techniques or the neutralization of autocatalysis-enhanced diffusion phenomenon at the smaller length scales [16-18]. Such nanoscale-unique properties compound the development of models to predict drug release kinetics from polymeric nanoparticles.

Hence, there is a need for a platform for simulation of polymeric nanoparticle erosion and chemotherapeutic drug release thereof.

2.1.2 Utility of PLGA and other biopolymers in design of anti-cancer nanoparticles: Challenges in ab-initio optimization of design parameters

The most frequently used biodegradable polymers for anti-cancer applications include the polyesters poly(lactic acid) and poly(glycolic acid), or PLA and PGA respectively, among several other biocompatible polymer families (Figure 2.1) [398]. As PLA and PGA are introduced, they are eventually hydrolytically degraded and form acidic oligomers and monomers, effectively dropping the pH [17]. The acidic pH contributes further to the hydrolysis, eventually reducing the polymeric network via various degradation mechanisms that allow the encapsulated drug to be released.
By fine-tuning the physicochemical characteristics of the polymer, it is possible to obtain a high level of control over the kinetics of drug release [17]. Polymer size, composition, and chain length have all been shown to have profound effects on the properties of the biomaterial device, and are each easily tailored [11-19]. It is also possible to create co-polymers, or polymers of multiple monomeric species, allowing for a nearly infinite number of customizable attributes [15-17]. These traits allow researchers to "fine tune" properties such as the stability, pH, solubility, and tensile strength of nanoscale polymeric devices. By encapsulating or conjugating different drugs to different polymers, it is possible to create a multitude of temporal release characteristics [17]. In fact, the use of different polymer classes allows for absolutely diverse temporal control, ranging from minutes to months [96-98]. This is highly beneficial, as it allows specific temporal control that can be altered for each drug used in the nano-complex. Such spatiotemporal control features of polymeric nano-devices also permit the creation of novel treatments that would otherwise be impossible. For example, highly toxic drugs that would normally prove fatal may be successfully delivered almost exclusively to tumors, effectively broadening the scope of drugs that can be administered by severely reducing off-target effects. This is evidenced by the fact that nanoparticle-based cancer therapeutics are already showing some promise from clinical trials [96, 97].

However, given the several parameters that influence the kinetics of polymer degradation, it is important to accurately zero in on the design parameters for polymer nanoparticles [98]. For this purpose, the development of predictive computational models is key, as outlined earlier. Here, we introduce a 3-D whole-nanoparticle simulation platform for this very purpose.

2.1.3 Development of a platform for prediction of polymeric nanoparticle erosion and drug release: From pixels to voxels

Monte Carlo models based on 'pixel degradation' have been proposed to simulate polymer erosion and subsequent drug release from microparticles as depicted pictorially herein (Figure 2.2) [14, 70, 126]. The pixel model mechanism as initially proposed is simple, reliable and pragmatic. Although much advancement has emerged since the
initial use of pixels for polymer erosion modeling [15-20], there remains remarkable scope for sophistication. The pixel models have not considered key factors such as average polymer molecular weight, monomer composition, geometry of the particle or physical breakdown of the polymer particle following progressive erosion [98]. Some recent extensions to the pixel model have also made assumptions that decrease the versatility of Monte Carlo simulations such as the use of discrete polymer and drug pixels or linear drug pixel degradation lifetimes [15-20]. The decrease in versatility prevents their use for simulating drug-conjugated polymer particles and other such hybrid vehicles that are becoming increasingly popular in anti-cancer drug delivery [18]. Moreover, with the exception of the most recent Monte Carlo models that account for the complete particle, the earlier models had restricted themselves to just the microenvironment around the particle-medium interface and generalized the results to the system at large. Finally, since the concept of 'pixel volume' does not exist, two dimensional models cannot account for density, morphology and distribution of the pores, polymers and drugs inside the device. These limitations greatly undermines the potential of pixel models and necessitates 3-D modeling (Figure 2.2).

An outstanding motivation to consider the 'third dimension' stems from the observation that pixel models work with the principle of 'neighboring unit effect', that is, the combined influence of immediate neighbors on any polymer matrix pixel of interest. But two dimensional modeling accounts for only eight immediate neighbors whereas three dimensional models could potentially account for all twenty six immediate neighbors as shown pictorially (Figure 2.3). Further, the difference in number of neighbors (NoN) between 2-D and 3-D increases drastically as we consider more distant neighbors (Figure 2.3). The development and application of future Monte Carlo models would require accurate calculation of the neighboring unit effect and this necessitates consideration of all three dimensions.

Voxel is the volume element which represents data on a regular grid in 3-D space and is analogous to pixels that are area elements representing 2-D data (Figure 2.3) [399].
Figure 2.3 The 2-D Pixel-based Monte Carlo models of microparticles contrasted with the 3-D Voxel-based whole nanoparticle model introduced herein

Voxels are particularly attractive and time-optimized for computational nanotechnology applications because the number of voxels required for nanoscale modeling is far lesser than that required at micro- and higher-scales (Figure 2.4). Considering the recent advances in processing power and the relatively small value of voxels to pixels ratio (VPR) for nanoparticles the computational cost associated with voxels modeling is comparable to pixels modeling. Considering all of the above factors, we focused on the development of voxels based three-dimensional spherical model that can predict drug release from polymeric nanoparticles.
Figure 2.3 (Top) Contrast of the neighborhoods of pixels (2-D) and voxels (3-D) shows the 8 versus 26 immediate "Number of Neighbors (NoNs)"; and (Bottom) The drastic increase in progressive number of neighbors from 1-D to 2-D to 3-D is shown.
2.1.4 Predicting release of the anti-cancer drug doxorubicin from spherical PLGA nanoparticles with the Voxel platform

For the purposes of the Voxel model introduced herein, recent findings from nanoscale experiments that revealed that the reduced drug diffusivity in nanoparticles limits the influence of diffusion to the interface between polymer and the external aqueous environment was utilized [11-15]. The model hence simulates drug release after degradation of surrounding polymer and the formation of an erosion channel to the particle surface, given that once the drug diffuses into the exposed erosion channel, the drug is released due to the small distances necessary to travel to the particle surface at the nanoscales. Using this framework, drug release from bulk eroding nanoparticles was modeled based upon a three dimensional Voxel based Monte Carlo simulation of polymer erosion.
In a recent study, Yoo et. al. compared the release of the anti-cancer drug doxorubicin from drug-encapsulated and drug-conjugated PLGA nanoparticles [18] and this data is used here to highlight the efficiency of the Voxel method predictions. A more detailed analysis of the Voxel model predictions for other anti-cancer drugs such as Dexamethasone is presented in the publication following this section. The computed simulation result of cumulative drug release from doxorubicin encapsulated PLGA nanoparticles is shown as a blue line (Figure 2.5 A), along with the corresponding experimental data (green dots). The simulation was then performed for the case of doxorubicin-conjugated PLGA nanoparticles and the correlation between experimental data points (orange dots) and voxel model results (red line) are also shown (Figure 2.5 A). The high degree of correlation between voxel model results and experimental data points validates the model and proves its versatility in simulating both drug encapsulated and drug-conjugated polymer nanoparticles.

The predictive potential of the voxel model was also explored through the variation in molecular weight of the polymer constituting the nanoparticle. This factor is known to play an important role in controlling the rate of drug release, however as discussed earlier, none of the pixel-based models have accounted for polymer molecular weight variance. The data from an earlier study of cumulative drug release from 200nm sized nanoparticles composed of dox-PLGA5005 conjugates (of molecular weight 5000) and dox-PLGA5010 conjugates (of molecular weight 10000) was used to verify the predicted model results. The experimental data for cumulative doxorubicin release from dox-PLGA5005 nanoparticles (blue dots) and dox-PLGA5010 nanoparticles (red dots) is shown (Figure 2.5 B). The voxel model results corresponding to dox-PLGA5005 (green line) and dox-PLGA5010 (orange line) are also shown (Figure 2.5 B). The same figure also displays both the experimental data (gray dots) and model results (black line) for cumulative release from doxorubicin encapsulated PLGA nanoparticles. The predicted model results correlate extremely well with the experimental data for both drug encapsulated and drug-conjugated polymeric nanoparticles of varying PLGA molecular weights, thereby demonstrating the predictive capability of three dimensional voxel modeling.
An analysis of the simulation results shows that cumulative doxorubicin release is nearly bi-phasic for drug encapsulated nanoparticles, with an initial burst release (~50% of total drug volume) that is then followed by a relatively slower release phase. Comparison with cumulative doxorubicin release from drug-conjugated nanoparticles shows that the sustained release profile does not have the initial burst phase, but has an increased rate between 8 to 16 days indicating tri-phasic release. This tri-phasic release is characteristic of PLGA particles that undergo bulk erosion driven by percolation effects and autocatalysis [70]. The burst phase is absent in drug-conjugated nanoparticles owing to the chemical bond between polymer matrix and the drug molecules even in the proximity of the nanosphere surface. This hypothesis was confirmed with the voxels model by analyzing the exposed surface area to volume ratio (AVR) which has been indicated to influence polymer erosion kinetics. The quantization of AVR variation was possible owing to the three dimensional whole-nanosphere nature of the voxels model. The dynamic variation of the exposed surface area and remnant volume of the nanosphere (Figure 2.5 C) and the computed AVR variation (Figure 2.5 D) are also shown. The variation of surface area, volume and AVR for different average polymer molecular weights has also been investigated in these figures. There appears to be substantial increase in AVR dynamics for PLGA5005 nanoparticles as compared to nanoparticles of higher average molecular weights. This predicated result correlates with our earlier analysis of the significance of AVR in determining the burst release and subsequent rate of sustained release. It is interesting to note that physical properties associated with polymer molecules have such a significant impact on nanosphere erosion kinetics.

In order to use computational nanotechnology for investigating hitherto unexplored nanoscale phenomena, the influence of nanoparticle size on drug release and average dynamic size of bulk-eroding nanoparticles was studied. These effects have not been empirically studied owing to the technological bottlenecks associated with nanoscale observation and experimentation that were discussed earlier.
Figure 2.5  (a.) The correlation between experimental data and voxel model results for doxorubicin release from DE and DC nanoparticles is shown. The known parameter values from literature are PLGA50:50 polymer crystallinity = 0, nanoparticle diameter = 300nm, weight average molecular weight = 8020 and degradation rate constant = -0.121/day. The model parameters values obtained for PLGA by curve-fitting are $a = 23$, $b = 3$ and $\beta = -0.5$. (b.) The predictions by the voxels model for cumulative drug release from nanoparticles with PLGA of varying molecular weights are correlated with experimental data. (c.) The dynamics of exposed nanoparticle surface area and remnant volume are computed with the voxels model for varying PLGA molecular weights. (d.) The Area to Volume Ratio (AVR) dynamics is computed using the voxels model for PLGA of varying molecular weights. (e.) The variation of cumulative doxorubicin release with time as a function of nanoparticle size is relatively difficult to calculate empirically. Here this variation is computed using the voxels model for DE nanoparticles. (f.) The variation of cumulative doxorubicin release with time is computed from drug-conjugated polymeric nanoparticles for DC nanoparticles.
The predicted release from doxorubicin encapsulated and doxorubicin-conjugated PLGA nanoparticles of varying sizes (100-700 nm) respectively are shown (Figure 2.5 E-F). Experimental data is only available for 300nm sized nanoparticles and this is also shown (black dots) in the same figures. From these results, one can observe that increase in size of nanoparticle causes general decrease in rate of predicted drug release. This observation may be explained by the increase in average drug diffusion path length with increase in particle size for polymer matrices such as PLGA that erode substantially throughout their bulk. In the case of drug encapsulated nanoparticles (Figure 2.5 E), the magnitude of predicted burst release appears to decrease from 89% for 100nm sized nanoparticles to 22% for 700nm sized microparticle. This correlates with experimental data on microparticles for which burst release was documented to diminish with increase in particle size.

These results of this section provide proof-of-principle demonstration regarding the potential of three dimensional voxel models for analysis and prediction of polymeric nanoparticle erosion kinetics, using the anti-cancer drug doxorubicin and FDA-approved biopolymer PLGA as examples. The voxels computational platform emerges to be reliable as demonstrated by its accuracy in prediction of key nano-material properties. The model’s versatility was also successfully exhibited by precise simulation of drug encapsulated as well as drug-conjugated nanoparticles. The following publication discusses the application of the Voxel platform for simulating the release of another popular anti-inflammatory drug, dexamethasone, thus demonstrating the more broader and potentially universal applicability of this platform for nanoparticle erosion and drug release predictions.
2.1.5 Published manuscript


Specific contributions of this Thesis research to the publication: Conceptualized and developed the Voxel modeling platform; and applied this to compute the optimal design parameters for PLGA nanoparticles laden with cancer drugs for maximal therapeutic efficacy.
A Voxel-Based Monte Carlo Model of Drug Release from Bulk Eroding Nanoparticles

David A. Eavarone¹, Venkataramanan Soundararajan², Toomas Haller², and Ram Sasisekharan¹,²,³,*

¹Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
²Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
³Koch Institute of Cancer Research, Massachusetts Institute of Technology, Cambridge MA 02139, USA

The use of polymeric nanoparticles as drug delivery devices is becoming increasingly prevalent in a variety of therapeutic applications. Despite their widespread clinical use, the factors influencing the release profiles of nanoparticle-encapsulated drugs are still not quantitatively understood. We present here a new, semi-empirical model of drug release from polymeric nanoparticles using a formulation of dexamethasone encapsulated within poly(lactic-co-glycolic acid) to set model parameters. We introduce a three-dimensional voxel-based framework for Monte Carlo simulations that enables direct investigation of the entire spherical nanoparticle during particle degradation and drug release. Due to implementation of this model at the nanoscale, we utilize assumptions that simplify the model while still allowing multi-phase drug release to be simulated with good correlation to experimental results. In the future, emerging mechanistic understandings of nanoparticle drug release may be integrated into this simulation framework to increase predictive power.

Keywords: Nanoparticles, Monte Carlo Model, Drug Release.

The shift from micro- to nano-scale drug delivery systems in recent decades has been driven by improvements in polymer formulation technology and evidence that medical applications such as chemotherapy are significantly enhanced by the use of nanoscale delivery vehicles.¹⁻⁴ Because drug release profiles differ based on the polymer, drug, and design parameters, the availability of computational models to understand and predict drug release is valuable for the extension and optimization of existing drug delivery technologies.

Several models have been reported in the literature to describe drug release and particle breakdown for polymer-based delivery vehicles.⁵⁻⁷ It is clear from these studies that the number of factors influencing drug release is too unwieldy for a single model to incorporate. Such factors include water and drug diffusion, drug dissolution, polymer molecular weight, particle size and geometry, polymer degradation, micro-environment pH changes, autocatalysis, polymer swelling, and more.⁸⁻¹¹ Incorporation of these factors is further complicated by a still incomplete understanding of the extent to which they each influence the kinetics of drug release. Previous modeling approaches have been limited by necessity to incorporating the factors most influential to drug release or polymer breakdown for a particular delivery system.¹⁰ Despite this limitation, computational models have still successfully simulated complex multi-phase drug release profiles from polymeric microparticles.¹²⁻¹⁵ Few studies have extended these models to polymeric nanoparticles, though these are becoming prominent in therapeutic applications.¹⁶

The mechanism of drug release from nanoparticles may differ from that of microparticles even when both vehicles have the same polymer composition. Drug diffusion through the polymer matrix has been shown to be slower in nanoparticles than in microparticles, and this may be due to more dense internal structures as a result of fabrication techniques or the neutralization of autocatalysis-enhanced diffusion phenomenon at the smaller length scales.⁸⁻⁹,¹⁷ For the purposes of our model, we postulate that the reduced drug diffusivity in nanoparticles limits the influence of diffusion in the model to the interface between polymer and the external aqueous environment. That is, we assume that drug release occurs only after degradation of surrounding polymer and the formation of an erosion channel to the particle surface. Once the drug diffuses into the exposed erosion channel, the drug is considered to be released due to the small distances necessary to travel to
the particle surface at the nanoscale. Using this framework, we model drug release from bulk eroding nanoparticles based upon a three dimensional Monte Carlo simulation of polymer erosion.

For the Monte Carlo simulation presented here, entire spherical nanoparticles are represented using voxels. A voxel is the volume element which represents data on a regular grid in three dimensional space and is analogous to a pixel which is the area element representing data in two dimensions. Previous two dimensional models have depicted drug release from a pixel as dependent upon the state of the pixel's immediate 8 neighbors. In this three dimensional voxel-based model, we assess the influence of the 18 neighbors most immediately surrounding the voxel in order to give the best representation of a sphere (Fig. 1). The State Diagram (Fig. 2) describes the simulation algorithm governing voxel dynamic state transitions \( (x_{i,j,k}) \) and voxel drug release. Assuming a homogeneous mixture of drug and polymer, each voxel begins with a voxel dynamic state \( (x_{i,j,k}) \) of 1. The lifetime of individual voxels follows a Poisson's distribution given by:

\[
T_{x_{i,j,k}}(t) = -\ln(s/\lambda)
\]

where \( s \) is a random number and \( 1/\lambda \) is the half life of the polymer system. Expiration of this lifetime changes the voxel dynamic state \( (x_{i,j,k}) \) to zero and the polymer is considered degraded. In order for the polymer to be considered eroded, access to the aqueous environment via an erosion channel from the surface of the nanoparticle is required. Such a channel is formed when at least one of the 18 immediate neighboring voxels is eroded, i.e., any of the neighbor's dynamic voxel state \( (x_{i,j,k}) \) becomes \(-1\).

We formulated poly(lactic-co-glycolic acid) (PLGA) nanoparticles encapsulating the hydrophobic agent dexamethasone as described in Figure 3. Hydrolysable polymers such as PLGA are often used in the formulation of nanoparticles as they combine sustained drug release with complete biodegradability. The polymer half life \( (1/\lambda) \) is approximated at 0.121 days from previous measurements in the literature. In order to run the model simulation, we used the drug release data from this formulation to fit values for two parameters: the number of voxels \((N)\) that compose each nanoparticle and a parameter \((\kappa)\) that integrates the influence of mass transport phenomenon into the simulation. The parameter \( \kappa \) (value between 0 and 1) measures the likelihood of drug release from an intact voxel into an aqueous erosion channel upon exposure, and represents the combined contributions of drug diffusion and dissolution on drug release at this interface. The parameter \( N \) is related to the physical size of the particle and also controls the surface area to volume ratio. Polymeric nanoparticles have a higher surface area to volume ratio than microparticles which may lead to a more pronounced relative initial burst of drug release.

The experimental release profile of dexamethasone from PLGA nanoparticles shows an initial burst release of 40% over the first \( \sim 25 \) hours. A more sustained phase of drug release occurs over the next \( \sim 120 \) hours resulting in complete release of encapsulated drug by 150 hours. The model simulation was a good match to experimental drug release with a coefficient of determination \( (R^{2}) \) value of 0.96 (Fig. 3). We are also able to use our model simulation to investigate physical properties of the voxelated nanosphere during erosion by exploring the fragment size.

<table>
<thead>
<tr>
<th>CLASSIFICATION</th>
<th>PIXEL MODELS (2D)</th>
<th>VOXEL MODELS (3D)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( n = 4 )</td>
<td>( n = 8 )</td>
</tr>
<tr>
<td>( d = \sqrt{1} )</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>( d = \sqrt{2} )</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>( d = \sqrt{3} )</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( \text{CCDF} = \frac{1}{2} \sum_{i=1}^{n} \mathbb{I}(d = \sqrt{i}) ) ( Z )</td>
<td>0.50</td>
<td>1.21</td>
</tr>
<tr>
<td>( \text{SVR} = \frac{1}{2} \sum_{i=1}^{n} \mathbb{I}(d = \sqrt{i}) ) ( 3(1.5)^{i/3} )</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Fig. 1. Geometric implications of pixel and voxel neighborhoods. The distance \( (d) \) between the center of a pixel (in 2D) or voxel (in 3D) and the center of its immediate neighbors is \( \sqrt{1}, \sqrt{2}, \text{or} \sqrt{3} \) units. The center to center distance factor (CCDF) and spherical volume ratio (SVR) are used to calculate the neighborhood \( (n) \) that best approximates a circle (or a sphere) enclosed within the 2D pixel (or 3D voxel) grid. A schematic representation of the neighborhoods is depicted with the pixels/voxels colored black (center), red \( (d = \sqrt{1}) \), blue \( (d = \sqrt{2}) \) and green \( (d = \sqrt{3}) \).
distribution and average size of the nanoparticles over time (Fig. 4(a)). Percolation-driven particle breakdown due to bulk erosion is shown to occur after 150 hours, and so does not appear to play a dominant role in drug release. A plot of nanoparticle surface area/volume with respect to time (Fig. 4(b)) indicates an initial lag prior to the formation of erosion channels. This corresponds well with the lag before the change in phase of experimental drug release from initial burst to sustained release (Fig. 3).

The potential of extending this model to additional drug formulations is explored through variation in the parameter $\kappa$ (Fig. 5). The influence of $\kappa$ is restricted to drug release at the interface of polymer and the external aqueous environment. A $\kappa$ value of 1 corresponds to immediate drug release from an exposed intact voxel into an erosion channel, and the resultant rapid drug release profile is consistent with that of a highly hydrophilic drug. A $\kappa$ value of 0 indicates that drug release from an exposed intact voxel does not occur until the voxel is degraded, and the resultant sustained drug release profile is consistent with that of a polymer-conjugated drug. Values of $\kappa$ between 0 and 1 correspond with intermediate capacities for drug release from an exposed intact voxel, and the resultant multi-phase drug release profiles are consistent with that of a hydrophobic drug as presented here.

Because all processes that influence drug release are not yet quantitatively understood, there is value to the use of semi-empirical computational models able to simulate multi-phase drug release profiles. Use of the model presented here does not require knowledge of physical parameters such as polymer porosity, drug solubility, and drug diffusivity which are often not available or can be difficult to measure. Voxelation of the entire nanoparticle facilitates the extension of this platform to non-symmetrical vehicle formulations as well as non-uniform drug and polymer distributions. Variation of the parameter $\kappa$ demonstrates the versatility of this model for simulating release of hydrophilic, hydrophobic, and polymer-conjugated drugs.

This model was intended as a proof of principle, however in the future emerging mechanistic understandings of
A Voxel-Based Monte Carlo Model of Drug Release from Bulk Eroding Nanoparticles

Fig. 3. Dexamethasone release from PLGA nanoparticles in vitro. Nanoparticles were formulated using a modified emulsion-solvent evaporation technique. Briefly, 25 mg of dexamethasone and 50 mg of PLGA were allowed to dissolve completely in 2.5 mL acetone before addition of 0.5 mL methanol and emulsification into an aqueous 2% PVA solution with sonication. Nanoparticles were separated from solution with ultracentrifugation and were demonstrated to have a mean particle diameter of 115 nm by dynamic light scattering. Drug release was quantified by HPLC after dialysis against PBS at 37 °C with gentle shaking. Simulations of drug release were performed using the C programming language and were performed 100 times with mean values shown. Experimental results are means +/− s.e.m. (n = 3). Parameters: N = 5497 voxels, κ = 0.5.

Fig. 4. Simulation of drug release and nanoparticle fragmentation. (a) Model representation of nanoparticle fragmentation: black scatter represents variability between different simulations, blue line represents average fragment size. (b) Simulation results for surface area (exposed voxel area) and volume (voxel count) during nanoparticle fragmentation.

Fig. 5. Influence of the model parameter κ on drug release. Simulation results for κ values between 0 and 1 demonstrate the capacity of the model for simulating the release of drugs with distinct diffusion and dissolution characteristics. Model parameters: N = 5497 voxels, iterations = 100.

phenomenon influencing nanoparticle drug release may be incorporated to further increase accuracy and strengthen predictive power.

Acknowledgments: The authors thank G. Franzesi and P. Shum for assistance with release experiments, and Dr. A. Chandrasekaran and A. Ziolkowski for helpful comments on the manuscript. The work was supported by NIH GM57073.

References and Notes

5. A. Gopferich, Biomaterials 18, 397 (1997).


Received: 10 August 2009. Accepted: 28 September 2009.
2.2 Harnessing Nanotechnology to target the Cancer Cell Immortality Spectrum

2.2.1 The Cancer Cell Immortality Spectrum

Tumorigenesis is a complex process with multiple pathways of progression, and "normal" cells typically require several mutations in order to overcome regulatory mechanisms and metamorphose into tumorous cells. These mutations may be explained by survival of the fittest, wherein cells that acquire mutations promoting increased cell division rate or longevity maybe more likely to produce daughter cells harboring selective proliferation advantages. Some of the common pro-tumorigenic mutations include the following: mutations that increase function of an oncogene (pro-growth/pro-division/pro-life), mutations that decrease function of a tumor suppressor gene (prodeath/anti-growth), or mutations that decrease function of a surveillance gene (DNA repair/cell cycle control). Individuals carrying one or more congenital mutations in these pathways are predisposed to particularly rapid progression of tumorigenesis. The combined genetic mutations in cancer cells rapidly manifest at the level of tumor tissues and are typified by a combination of characteristics that we refer to as factors I-X of the "Cancer Cell Immortality Spectrum" (Figure 2.6).

![Figure 2.6 The Cancer Cell Immortality Spectrum](image-url)
Fundamental to the cancer cell immortality spectrum are the discussed genetic mutations that promote cell growth and/or reduce cell death, culminating in the formation of tumor predisposed cells (factors I, II, IX). The deregulation of DNA repair and cell cycle control are other key factors that occur throughout tumorigenesis, accelerating its progression via faster mutation and replication rates (factor III). As the tumor tissue grows, it requires additional space, oxygen, and nutrients, necessitating the invasion of local tissue through the degradation of the ECM (factor IV) and the formation of new blood vessels through the production of angiogenic growth factors (factor V). Any anti-growth signals that are produced by the organism during this stage can be countered by mutations in tumor cells that reduce effects of such signals (factor VI). Cancer treatment involves use of chemotherapeutic agents (e.g., cytotoxic drugs) for destroying cancer cells and inhibiting further growth of the tumor tissue. However, tumors are able to acquire mutations that render them resistant to chemotherapy (factor VII). This, combined with their ability to overcome senescence through methods such as over-expression of telomerase, causes cancer cells to become immortal (factor X). As the tumor continues to expand, malignant cells detach from the primary tumor and begin to degrade proteins in the surrounding ECM. This allows cancerous cells to escape from the matrix, enter the blood stream or lymph system, and migrate to other cells via a process termed metastasis (factor VIII). Such multi-drug resistant, immortal, metastatic tumors are responsible for a majority of the present cancer-related deaths.

Given that cancer is motivated by all ten hallmarks of the cancer cell immortality spectrum discussed herein, it becomes critical to adopt a multi-pronged therapeutic approach to tumor treatment. Such combination therapeutics approaches, however, require precise temporal regulation in the release kinetics of the multiple drugs involved in order to achieve maximal ‘‘synergy’’ and optimal therapeutic efficacy in their application (Figure 2.7).
2.2.2 Receptor-ligand structural fit as the basis for spatial targeting of polymeric nanoparticles to tumor cells

The conjugation of appropriate tumor-specific targeting molecules onto drug encapsulated nano-devices ensures selective delivery of the agent to cancerous cells and hence minimal toxicity to non-cancerous cells, as described earlier. Additionally, the potency of nano-devices for cancer treatment is generally enhanced due to the fact that several localization techniques involve binding to tumor over-expressed cell surface receptors that are themselves important for cell growth (Figure 2.8). This is because tumor cell surface receptors are often endocytosed upon binding, thus providing a method for intra-cellular delivery of the nano-complex while simultaneously "competing" with the natural growth signals, thereby inhibiting their biochemical signaling.

As a specific example, folate receptor α (FRα) is over-expressed on the surface of certain cancer cells, but is naturally expressed only on the apical (away from the blood) surface of epithelial cells, making it a unique target for therapy. This has allowed the use of folate-conjugated nanoparticles for both imaging and the delivery of drugs. Once the receptor binds a ligand, such as folate, it is taken into the cell via receptor-mediated endocytosis, escapes from the endosome, and is recycled back to the surface after releasing its ligand within the cell. However, one can imagine a technique in which the ligand remains bound to the receptor after depositing the drug, effectively inactivating or reducing the growth receptor pathway.
Figure 2.8 Structures of receptor molecules that are over-expressed in Breast Cancer cells. (a.) Fibroblast growth factor receptor (FGFR) is over-expressed in sarcoma and breast cancers; (b.) Interleukin-13 (IL-13) receptor is over-expressed in skin and brain cancers; (c.) Vascular endothelial growth factor receptor (VEGFR) is over-expressed in colorectal, ovarian, pancreatic, and breast cancers; (d.) Androgen receptor is over-expressed in prostate cancer; (e.) Insulin-like growth factor-1 receptor (IGF1R) is over-expressed in cervical cancer; (f.) Neuropilin-1 (NRP-1) is over-expressed in breast, colon, prostate, and pituitary cancers; (g.) Epidermal growth factor receptor 2 (EGFR-2) is over-expressed in pancreatic, bladder, cervical, breast, and stomach cancers; and (h.) Estrogen receptor alpha (ERα) is over-expressed in breast and ovarian cancers.

This highlights an important dynamic in drug delivery since any localization technique targeting a growth factor receptor has the unique ability to temporarily disrupt pro-growth in a completely drug-independent manner. Estimated levels of cell surface receptor over-expression in breast cancer is compiled from literature and presented herein to highlight the importance of targeting these receptors in localizing the delivery of chemotherapeutic agents specifically to tumor cells (Figure 2.9). Additionally, several such over-expressed cell surface receptors are known for various cancers and is presented herein along with the incidence of over-expression (Table 2.1).
Breast cancer receptor over expression levels

- Estrogen receptor
- Progesterone receptor
- Epidermal growth factor receptor
- Endothelin-A receptor
- Platelet derived growth factor receptor
- Fibroblast growth factor receptor 2

**Figure 2.9 Estimated levels of over-expressed Breast Cancer receptor molecules**

2.2.3 *Voxel model optimization of the polymeric nanoparticle core of Nanocell delivery systems*

A “Trojan horse” approach using a multi-compartment, multi-functional nanoscale system has recently been proposed to exploit the full potential of nanotechnology for targeting cancers. This strategy is based on a delivery system called the “nanocell” (Figure 2.10), so named because of its visual resemblance to a cell and its nanoscale dimensions. A nuclear polymeric nanoparticle conjugated to a chemotherapeutic agent was encapsulated within a PEG-phospholipid block-copolymer envelope that contained a vascular disruption agent (VDA). Thus, a nanoscale vehicle with two drugs was designed, with customizable release profiles.
Table 2.1 Over-expressed cell surface receptors classified according to incidence of cancer

<table>
<thead>
<tr>
<th>Over-expressed cell surface receptor</th>
<th>Cancer type (incidence)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estrogen receptor alpha (ERα)</td>
<td>Breast (70%), Ovary</td>
</tr>
<tr>
<td>Progesterone receptor (PR)</td>
<td>Breast (64%), Uterus</td>
</tr>
<tr>
<td>Human epidermal growth factor receptor 2 (HER2)</td>
<td>Pancreas (26%), Bladder (44%), Cervix (77%), Breast (30%), Stomach</td>
</tr>
<tr>
<td>Endothelin-A receptor (ET_{AR})</td>
<td>Breast (45.3%), Ovary</td>
</tr>
<tr>
<td>Platelet derived growth factor receptor (PDGFR)</td>
<td>Liver (22.1%), Breast (39.2%)</td>
</tr>
<tr>
<td>Fibroblast growth factor receptor 2 (FGFR-2)</td>
<td>Sarcoma, Breast (5-10%)</td>
</tr>
<tr>
<td>Heparin Sulfate Glycosaminoglycan (HSGAG)</td>
<td></td>
</tr>
<tr>
<td>Neuropilin-1 (NRP-1)</td>
<td>Breast, Colon, Prostate, Pituitary</td>
</tr>
<tr>
<td>Androgen receptor (AR)</td>
<td>Prostate (43%)</td>
</tr>
<tr>
<td>Insulin-like growth factor-1 receptor (IGF_1R)</td>
<td>Cervix</td>
</tr>
<tr>
<td>G-Protein coupled receptor (GPCR)</td>
<td>Brain (57%), Stomach, Prostate (60%)</td>
</tr>
<tr>
<td>Cannabinoid-specific receptors (CBR)</td>
<td>Liver (62%)</td>
</tr>
<tr>
<td>Chemokine Receptor (CXCR4)</td>
<td>Skin</td>
</tr>
<tr>
<td>Vascular endothelial growth factor receptor (VEGFR)</td>
<td>Colorectal, Ovary, Pancreas, Breast</td>
</tr>
<tr>
<td>Interleukin-13 (IL-13) receptor</td>
<td>Skin, Brain</td>
</tr>
</tbody>
</table>

As the nanocells are lodged into the tumor tissue by active or passive targeting, the envelope quickly breaks down, releasing both the conjugated nanoparticle and the free VDA into the tumor. The VDA induces vascular collapse, effectively trapping the nanoparticle-chemotheraputic conjugation within the tumor, allowing the chemotherapeutic agent to undergo sustained release without being able to leave the tumor microenvironment.
As demonstrated in the last section, the Voxel modeling platform is useful for predicting the release of drugs encapsulated and conjugated to polymeric nanoparticles. Hence, the Voxel platform was used to model the degradation and erosion kinetics of PLGA nanoparticles conjugated chemically to the anti-cancer agent of interest. A simulation of the progression in degradation and erosion processes within the polymeric nanoparticle was performed, based on the Voxel model results for this system (Figure 2.11). It can be seen from the simulation results that the polymeric nanoparticle is degraded in bulk throughout the vehicle (including the central core) right from the start of the simulation, but erosion only proceeds from the solvent-exposed surface towards the core with progression of time.
This simulation result is hence in agreement with the Voxel model results that predicted *Bulk Erosion* of PLGA nanoparticles owing to instantaneous diffusion of water molecules and uniform wetting of the entire vehicle, followed by more gradual erosion/fragmentation that accelerated with development of erosion channels from the surface of the vehicle. As mentioned earlier, bulk eroding polymeric matrices such as those composed of PLGA, have been traditionally simulated using 2-D of pixels grids (*left*) with intact polymers (*black*), degraded polymers (*gray*) and eroded pores (*white*) depicted as shown (*Figure 2.12*), but these pixel models cannot account for volumes. On the other hand, the 3-D Voxel-based simulation platform (*right*) is able to include the effect of volumes of the biomaterials, pores, and drugs as depicted herein with intact (*pink/black*) and eroded (*blue*) polymers highlighted (*Figure 2.12*).

Polymer matrices undergo either surface erosion or bulk erosion, depending on their dimension and material properties. Smaller particles (e.g. nanoparticles) that are uniformly wetted by water throughout undergo bulk erosion, while larger particles (e.g. microparticles) that are wetted by water only on their surfaces undergo surface erosion. The rate of erosion is primarily governed by percolation kinetics in the case of bulk erosion but is instead governed by the progress of the eroding surface in the case of surface erosion. Both surface eroding and bulk eroding drug-conjugated matrices have been employed extensively in sustained release therapeutic applications.

The Voxel model platform was also successfully extended to PLGA microparticles that undergo surface erosion, as shown. Similarly to bulk erosion, the
simulations of surface erosion of microparticles was also inclusive of volumes of biomaterials, pores, and drugs, and provided incredible insight into the process of polymeric matrix surface erosion (Figure 2.13). From the voxel platform simulation of microparticle surface erosion (Figure 2.13), it can be seen that, unlike bulk erosion (Figure 2.12) wherein instantaneous and uniform wetting of the vehicle promoted degradation throughout the vehicle (including the core), in the case of surface erosion, degradation by hydrolysis of polymers is initially restricted to the solvent-exposed surface and only with time does it progress towards the core. Such differences in water molecule diffusion, polymer hydrolysis, polymer degradation, nanoparticle erosion, and nanoparticle fragmentation are hence adequately captured by the Voxel model for both surface and bulk eroding polymeric particles.

Figure 2.12 Simulation of bulk eroding polymeric nanoparticle degradation and erosion with the 2-D pixel (left) and 3-D voxel (right) modeling platforms.
Some of the limitations of pixel based models that emerges from the Voxel platform results are noted herein. It is observed that pixels cannot account adequately for material properties that require volume based grid specification, such as dissolution kinetics, conjugation density, and neighborhood-based effects. Hence, for instance, existing pixel based models cannot directly be applied to analyze drug-conjugated polymeric nanoparticles. Some recent extensions to the pixel degradation model, have also assumed discrete polymer and drug pixels to simulate drug release. This would however be physically meaningful only for extremely dense grids and decrease the power of the pixel approach to modeling. Another disadvantage of pixel based polymer modeling is that of over-simplification which leads to a decrease in versatility. An
example of this is the assumption of linear distributions to obtain the lifetimes of discrete drug pixels. The assumptions of symmetry and eight neighboring pixel influence further limits their application. While 2-D pixel models serve as an excellent platform to design more sophisticated and realistic computational models, by themselves they are unsuitable for nano-scale phenomenon analysis, which is best captured using volume dependent parameters. There have also been indications of substantial differences in 2-D versus 3-D modeling in other applications such as tumor vasculature modeling. The group proposed an invasion-percolation 2-D model and have noted that 2-D modeling shows discrepancies in simulation of physical phenomena, but was still preferred because at that time 3-D modeling was too computationally unwieldy. The recent advances in computational processing power prompted revisiting 3-D modeling with the Voxel platform and as demonstrated in the last few sections, this has eliminated the limitations of 2-D modeling outlined herein.

2.2.4 From nano to micro: Adaptation of the Voxel platform for modeling polymeric agglomer-based anti-cancer applications

Along the dimension scale from "nano to micro", the phenomenon of agglomeration, or the coming together of individual nanoscale particles to form clusters (or agglomerers), becomes increasingly important due to the high surface energies involved. A computational platform that integrates the nanoscale voxels platform, which currently pertains to single nanoparticles, with multi-nanoparticle simulations wherein inter-particle forces are accounted for, hence becomes important (Figure 2.14). It is found that such an extension to the Voxel platform enables simulation of a diverse array of polymeric vehicles with spatial compartmentalization including quantum dots, core-shell nanoparticles, and nanomer clusters (at the nanoscales) and agglomerers (at the microscales). By varying the interparticle forces with a suitable agglomeration factor parameter as shown, it was possible to tease apart these forces and compute quantitative estimates of their influence on factors that govern polymeric agglomeration (Figure 2.15).
Furthermore, the size of agglomerates are key to their applications for the treatment of specific cancers. For instance, treatment of lung cancers and associated inflammation and infections with large agglomerates delivered as aerosol formulations has been explored. Such aerosol formulations typically require several hundred polymeric nanoparticles or liposomes to agglomerate together. In most of these scenarios, designing nanoscale agglomerates of smaller sizes (e.g. agglomer of size 2 or 3 nanoparticles) is required to ensure that these are specifically localized to the tumor microenvironment (owing to the dimensions of the pores in the “leaky” tumor microvasculature). In order to facilitate the precise engineering of such nano-complexes for application to tumor treatment, the extended Voxels simulation platform was used to analyze the concentrations of agglomerates of the required size as a function of time as shown (Figure 2.16). Thus, the outlined extensions of the voxels-based 3-D simulation platform allow for integrated modeling of polymeric nano- and micro-particle behavior in the context of their anti-cancer applications.
Quantitative insight into polymeric particle agglomeration kinetics may be derived from the above computations. For instance, from the above simulations, it can be deduced that time = 19 units and agglomeration factor = 0.6 (~ 60% activated PEG density on liposome) appears ideal for obtaining one-on-one binding of PEGylated liposomes. The 'agglomeration factor' has been specifically defined to account for material properties and mechanistic forces to facilitate such computations.
Figure 2.16 Effect of variation of agglomeration factor (AF) on the size of agglomer clusters formed is computed, so that polymer characteristics may be appropriately tailored for desired agglomer sizes.

The results indicate that very fine degrees of precise 1-1 bonding can be achieved by controlling the: a.) temperature and, b.) time of reaction -- factors that may be of value in engineering polymeric vehicles of desired spatial compartmentalization. The entire Monte Carlo simulation of polymer agglomeration (Figure 2.17) provides further insight into the dynamics of individual particles and their association into larger agglomerators.
Figure 2.17 Simulation of agglomeration with the extended Voxel platform
2.6.5 Published manuscripts


*Specific contributions of this Thesis research to the publications:* Developed and applied the Voxel modeling platform to the different anti-cancer polymer-based systems described.
Multifunctional Nanoscale Platforms for Targeting of the Cancer Cell Immortality Spectrum

Venkataramanan Soundararajan, Kenneth Warnock, Ram Sasisekharan*

In the post-genomic era, "omics" platforms and cancer systems biology are greatly advancing our knowledge of the molecular and cellular underpinnings of cancer. In this article, we begin by outlining the factors governing the development of cancer (tumorigenesis) and use this framework to motivate the need for systems-approaches to cancer diagnostics and therapeutics. We review recent efforts to tap into the remarkable potential of nanotechnology for (i) systems-surveillance (or "sensing") of the molecular signatures of tumorigenesis, and (ii) spatiotemporally-regulated delivery (or "targeting") of combination therapeutics to cancer cells. Specifically, we highlight the salient role of polymeric biomaterials and describe the physicochemical characteristics that render them attractive for the design of such nanoscale platforms. We conclude with discussions on the emerging role of macromolecular biophysics and computational nanotechnology in engineering spatiotemporally-regulated anti-cancer systems.

Introduction

Cancer is one of the leading causes of death, accounting for nearly 13% of annual mortality according to the World Health Organization (WHO), and it is estimated that another 23 million people are presently living with cancer around the world. While these numbers are certainly alarming, recent reports ominously predict that these are likely to double within the next two decades. Tumorigenesis, or the development of cancer, is a multistep process that begins with the malignant transformation of innocuous cells owing to a combination of genetic predisposition factors and sustained physical/chemical/biological carcinogenic exposure. Malignant cells are typified by a variety of characteristics, including increased replicative potential, evasion of programmed cell death (apoptosis), self-sufficiency in growth signals, insensitivity to anti-growth signals, sustained blood vessel development (angiogenesis), loss of cellular repair mechanisms, evolution of multi-drug resistance, destruction of adjacent tissues (invasion), non-adherence to universal aging (senescence), and spread to non-neighboring organelles (metastasis). These factors work in synergy to promote the rapid progression of the disease from cells to tissues and organs. Detecting cancer at its earliest stages and adopting a personalized, combination therapeutics approach targeting many of these factors hence becomes critical for effective treatment of the disease.

This in turn necessitates the development of high-sensitivity, multi-functional, diagnostic and therapeutic platforms for ultra-precise measurement and manipulation at molecular, cellular, and tissue levels. Advances in nanotechnology have led to an active interest in developing multifunctional nanoscale platforms capable of sensing the molecular signatures of...
Venkataramanan Soundararajan is currently pursuing his Ph.D. in Biological Engineering at the Massachusetts Institute of Technology, Cambridge, USA. For his PhD thesis research he is developing tools for investigating the principles governing biomolecular interactions, evolution, folding and misfolding, with applications to nanoscale therapeutic design, immunology, allergies and infectious diseases. Prior to starting his Ph.D., he graduated with the B.Tech. in Electronics and Communication Engineering from the Indian Institute of Technology, Guwahati, India. His undergraduate thesis on electrophysiology and medical instrumentation won the Agilent Technologies National Best Thesis Award for 2005. He has also previously been a visiting fellow with the Department of Bioengineering at Christian Medical College, Vellore, India and the Advanced Technology Information Processing Systems Laboratory at the University of Calgary, Canada, where he worked on research projects related to bio-MEMS and cancer biology.

Kenneth Warnock was born in Los Angeles, California in 1987. He is currently a senior majoring in Biology at the Massachusetts Institute of Technology, Cambridge, USA. His research with the Sasisekharan group utilizes computational approaches and bioinformatics tools for better understanding of the molecular mechanisms underlying cellular infection and carcinogenesis. On the infectious diseases front, he is investigating the molecular evolution of viral surface proteins such as dengue glycoprotein-E, influenza hemagglutinin and neuraminidase. On the carcinogenesis front, he is examining the structural basis of p53 mutations potentiating protein misfolding, towards elucidating the role of this phenomena in development and progression of cancer.

Ram Sasisekharan is currently the Edward Hood Taplin Professor of Health Sciences & Technology and Biological Engineering at the Massachusetts Institute of Technology, Cambridge, USA. He is also a faculty member of the David H. Koch Institute for Integrative Cancer Research, the Director of the Harvard-MIT Division of Health Sciences and Technology, and serves on the steering committee of the Consortium for Functional Glycomics. He received his Ph.D. in Medical Sciences in 1992 from Harvard Medical School, his M.S. in Biophysics in 1987 from Harvard University, and his B.S. in Physical Sciences in 1985 from Bangalore University, India. His research on complex poly saccharides has led to over 125 publications and over 50 patents, including the core technologies of several start-up companies. The broad research objectives of his research group at MIT are aimed towards developing tools to study complex carbohydrates, understanding the mechanisms governing the extracellular regulation of cell function, and generating novel pharmacological approaches to modulate cell function for the treatment of diseases. He has won both the Burroughs Wellcome and Beckman Foundation Young Investigator Awards and was the recipient of the 1998, 1999, 2000 and 2001 CapCure Awards from the CapCure Foundation.

tumorigenesis and delivering therapeutics specifically to cancerous cells in order to simultaneously reduce toxicity and increase efficacy.6 Pioneering contributions to polymer-drug formulation technology have particularly enhanced these efforts by providing for distinct advantages such as sustained release, increased circulation lifetime, specific localization to tumor microenvironments, and biocompatible vehicles.12-14 There is growing evidence that (i) precise targeting of surface-functionalized nanoparticles to overexpressed receptors on tumor cells (spatial control), combined with (ii.) rational design of polymer-drug formulations for modulation of factors influencing the release profiles of nanoparticle-encapsulated drugs (temporal control), could together offer dramatic therapeutic benefit.15-17 Hence, there is emerging interest in the development of molecular modeling and computational nanotechnology tools for the design of such spatiotemporally-regulated anticancer systems.18-24

In this article, we provide a perspective on nanoscale polymeric systems for cancer diagnostics and combination therapeutics, and macromolecular modeling platforms for engineering these systems.

Cancer Cell Immortality Spectrum

Tumorigenesis is a complex process with multiple pathways of progression, and “normal” cells typically require several mutations in order to overcome regulatory mechanisms and metamorphose into tumorous cells.5,23 These mutations may be explained by survival of the fittest, wherein cells that acquire mutations promoting increased cell division rate or longevity may be more likely to produce daughter cells harboring selective proliferation advantages.26 Some of the common pro-tumorigenic mutations include the following: mutations that increase function of an oncogene (pro-growth/pro-division/pro-life), mutations that decrease function of a tumor suppressor gene (pro-death/anti-growth), or mutations that decrease function of a surveillance gene (DNA repair/cell cycle control).27 Individuals carrying one or more congenital mutations in these pathways are predisposed to particularly rapid progression of tumorigenesis.9 The combined genetic mutations in cancer cells rapidly manifest at the level of tumor tissues and are typified by a combination of characteristics that we refer to as factors I-X of the “cancer cell immortality spectrum” (Figure 1a). Fundamental to this spectrum are the discussed genetic mutations that promote cell growth and/or reduce cell death, culminating in the formation of tumor predisposed cells (factors I, II, IX). The deregulation of DNA repair and cell cycle control are other key factors that occur throughout tumorigenesis, accelerating its progression via faster mutation and replication rates (factor III).23,28 As the tumor tissue grows, it requires additional space, oxygen, and nutrients,
necessitating the invasion of local tissue through the
degradation of the extracellular matrix (ECM) (factor IV)
and the formation of new blood vessels through the pro-
duction of angiogenic growth factors (factor V). Any anti-
growth signals that are produced by the organism during
this stage can be countered by mutations in cancerous cells

Figure 1. Cancer cell immortality spectrum; hallmark of the multi-pronged systems-approach to nanoscale combination chemotherapy.

a) Contrast of the normal cell cycle that balances cell division and cell death with the ten traits that render immortality to cancer cells.

b) Pictorial depiction of multifunctional anti-tumor nanoparticles that house drug cocktails for combating cancer on several fronts simultaneously.

c) The versatile degradation lifetimes of polymeric nanoparticles taken together with the availability of several anti-cancer drugs allows careful tailoring of temporal release characteristics. The same system with a single drug can also be used for multiple shots as shown.
Cancer treatment generally involves the use of chemotherapeutic agents (e.g., cytotoxic drugs) for destroying cancer cells and inhibiting further growth of the tumor tissue. However, tumors are able to acquire mutations that render them resistant to chemotherapy or other types of drugs (factor VII). As the tumor continues to expand, malignant cells detach from the primary tumor and begin to degrade proteins in the surrounding ECM. This allows cancerous cells to escape from the matrix, enter the blood stream or lymph system, and migrate to other cells via a process termed metastasis (factor VIII). Such multi-drug resistant, immortal, metastatic tumors are responsible for a vast majority of the tumor may be eliminated before the drug eliminates the tumor. As a specific example, about 70% of all ovarian cancers are resistant to the drug paclitaxel, and virtually all cases are resistant following remission, its ability to acquire new mutations and the selective pressure to develop drug resistance will almost guarantee that resistance occurs. Hence, unless the cancer proceeds to complete cure, its ability to acquire new mutations and the selective pressure to develop drug resistance will almost guarantee that resistance occurs.

Cancer Combination Therapeutics

Combination therapeutics refer to the treatment of disease with a variety of drugs and treatments, as opposed to a single drug or treatment. Three significant aspects have prompted scientists to search for treatments that involve drug combinations. The first is the concept of synergy, or the idea that the combined benefit of multiple drugs can be several orders of magnitude greater than the sum of the benefits of the individual drugs. Several different models for synergistic effects have been described and are reviewed extensively elsewhere. The synergistic effects of cancer drugs (Figure 2) have found some success in recent years with promising results from multiple clinical trials. The second major driving force of the multi-drug transition has been the emergence of drug resistance. As anti-cancer drugs are introduced to tumors, this creates a strong selective pressure for the tumor cells to acquire mutations that grants resistance against the drug. While a vast majority of the tumor may be eliminated by the drug, unless the population is completely destroyed, small sets of cells may survive by acquiring mutations that protect them from the drug, reproduce, and create a drug resistant population. This threat is even greater for cancer because most tumors’ DNA repair mechanisms are impaired (factor III), thus providing for opportunities to rapidly acquire new mutations and develop drug resistance before the drug eliminates the tumor. As a specific example, about 70% of all ovarian cancers are resistant to the drug paclitaxel, and virtually all cases are resistant following remission; hence, unless the cancer proceeds to complete cure, its ability to acquire new mutations and the selective pressure to develop drug resistance will almost guarantee that resistance occurs.

Cells are also capable of gaining resistance against a wide range of drugs through a variety of mechanisms, such as increased drug efflux, enzymatic deactivation, and decreased cell permeability. This multi-drug resistance (MDR) creates an enormous challenge, as it renders most current cancer treatment options futile. However, nanotechnology is emerging as an attractive option for countering MDR, as evidenced by the development of nanocarriers capable of simultaneously affecting cellular pH, lowering apoptotic threshold, and introducing resistance modulators.

The third and final factor in the need for a multi-drug approach is due to the inherent complexity of cancer. As described in the previous section, tumorigenesis arises from a complex set of mutations involving multiple
biological systems, making mono-therapy cures less likely solutions\(^{[3,25-29,34]\)} given that a single drug that simultaneously modulates growth, angiogenesis, apoptosis, metastasis, cell signaling, and other key factors of the cancer cell immortality spectrum is very unlikely, a drug cocktail for synergistic targeting of multiple factors provides a more pragmatic form of cancer treatment.

However, there are two significant challenges associated with combination therapeutics – (i) possible antagonistic cross-interaction of drugs\(^{[38,60]}\) and (ii) the creation of novel platforms for the spatiotemporal delivery of combination drug cocktails\(^{[50-58]}\). In order to address the possibility of cross-interactions in combination cocktails, massive efforts are underway to map out the synergistic, antagonistic, and additive interactions between the universal set of drug combinations\(^{[59]}\). While the potential for combination therapeutics is likely to expand as new drug cocktails are tested and approved, the development of novel multifunctional platforms for delivering these drug combinations in a spatiotemporally regulated fashion is still very much in its infancy. With the emergence of nanotechnology, there is great interest in the development of such multifunctional nanoscale platforms for the delivery of synergistic combination therapeutics.

Multifunctional nanoscale platforms for targeting cancer are being engineered for two related applications, namely, timely detection (diagnostics) and treatment (therapeutics). As mentioned earlier, detection of cancer at its earliest stages is essential for effective treatment of the disease, and the ability of nanoscale systems to monitor cellular microenvironments and sub-cellular compartments with high sensitivity provides unprecedented opportunities for the timely recognition of tumor metamorphosis\(^{[6,7]}\). In the following two sections we first review some of the emerging nano-devices for cancer diagnostics, and follow with a discussion of multifunctional nanoscale platforms capable of simultaneously targeting multiple factors of the cancer cell immortality spectrum.

The Nanotechnology Era and its Implications for Cancer Systems-Surveillance

One of the greatest challenges in cancer treatment is being able to accurately diagnose and characterize the multiple stages of cancer in a cost-effective, low-risk, high-accuracy manner. Certain current methods, such as biopsies, are incredibly invasive, while alternatives can be relatively inaccurate\(^{[6]}\). For example, there is a 20% chance that any woman will be incorrectly diagnosed with breast cancer over the course of a decade with yearly mammogram testing\(^{[66]}\). Inaccurate diagnoses can introduce grave risks to patients who are not given treatment for a disease they have (false negative), as well as to patients who are recommended treatment for a disease they do not actually have (false positive). Additionally, an inaccurate diagnosis creates a significant financial burden. Effective and timely detection of cancer requires remarkable finesse in diagnostic capabilities, an attribute that has become increasingly feasible since the emergence of nanotechnology.

A systems-based approach to nanotechnology has the potential to monitor cancer at all stages of the disease, including initial diagnosis, subsequent treatment, and detection of relapse\(^{[61]}\). Several methods for high-resolution, high-accuracy nanoscale cancer detection have been explored and are described exhaustively in our recent chapter\(^{[6]}\). For example, femtomolar quantities of known cell-surface tumor markers may be detected through the use of semi-conductors, such as nanowires (NWs), nanotubes (NTs), and nanocantilevers\(^{[62,63]}\). Similarly, the presence of specific point mutations can be identified through the use of gold nanoparticles capable of monitoring changes in the binding patterns of any protein of interest\(^{[64]}\). Likewise, the versatility of quantum dots (QDs) makes them principal candidates for systems-based cancer detection, as they have already been utilized for imaging cells\(^{[65,66]}\) tracking signaling pathways\(^{[67]}\), identifying cell-cell interactions\(^{[68]}\), and detecting metastasis\(^{[69]}\).

As with any new medical technology, the in vivo use of QDs for cancer detection requires monitoring for ethical, safety, and regulatory issues\(^{[70]}\). Along these lines, the safety associated with several new nanobiotechnology-based products, including in vivo use of QDs for tumor detection, is presently being evaluated by the FDA\(^{[71]}\). Pending their approval, it is possible to envision the creation of a nanoscale platform capable of monitoring tumor progression through each of the factors in the cancer cell immortality spectrum. Such a systems-based diagnostic tool would prove invaluable for both researchers and clinicians by providing a comprehensive framework for studying progression of the disease and treating patients.

Another emerging area of interest in tumor diagnosis is characterization of morphological and biochemical modifications that occur in the extracellular environment (ECM) cells during tumorigenesis, given the increasing evidence for role of the ECM in modulating various stages of cancer\(^{[72-76]}\). For example, glycosaminoglycans (GAGs) are long, linear (unbranched) polysaccharides, abundant in the ECM, that are known to modulate numerous biological processes including angiogenesis, apoptosis, and tumor growth\(^{[72-74]}\). A change in the level of activity of GAG-modifying lyases has been implicated in the progression of various diseases including cancer\(^{[72,77]}\). For instance, the increase of heparanase activity in myeloma patients enables monitoring of heparanase activity as a strategic tumor surveillance option\(^{[77,78]}\). Unfortunately, in vivo detection of heparanase activity is challenging, owing to the fact that lyase activity is typically measured by
monitoring absorbance at 232 nm and most biological materials interfere with this absorbance.\(^{[77-81]}\) However, GAG-conjugated gold nanoparticles have recently been used to shift the absorbance spectra, thus providing an interference-free in vivo assay for tumor detection and surveillance thereafter.\(^{[77]}\)

The success of these nanoscale technologies arises from the ability to easily manipulate the localization properties of the nanoscale complexes. Due to the diversity of localization methods available, including DNA/RNA aptamers,\(^{[82]}\) immunoliposomes,\(^{[83]}\) lipoids,\(^{[84]}\) and receptor-based ligands,\(^{[85]}\) nanotechnology allows an unprecedented level of precision and accuracy for spatial localization (e.g., nuclear, endosomal, lysosomal, cytoplasmic, cell surface, extra cellular, etc). Thus, a high-resolution "detection" nano-component can be combined with a high-precision localization component to offer systems-scale tumor surveillance.

While these nanoscale systems are furthering our ability for widespread cancer surveillance, there is critical need for multifunctional platforms that can enable spatiotemporally regulated delivery of combination therapeutics to the detected cancer tissues. Due to their versatile physicochemical properties, surface-functionalized polymeric nanoparticles are emerging as front-runners in nanoscale anti-cancer platform design.\(^{[6,7]}\) The next section focuses on the localization of surface-functionalized polymeric nanoparticles by biomolecular "targeting" techniques and the physicochemical properties of polymeric biomaterials used in the development of these devices for spatiotemporally regulated anti-cancer combination therapeutics.

**Spatiotemporal Control in Targeted Delivery of Nanoscale Polymeric Devices**

As discussed earlier, the various factors constituting the development of cancer necessitate the creation of multipronged therapeutics capable of "combating" tumors on several different fronts. By designing a platform with the ability to localize several drugs to the same tumor tissue in a spatiotemporally regulated fashion, it is theoretically possible to address each of the traits in the cancer cell immortality spectrum. Unfortunately spatiotemporal regulation is challenging from a design perspective, since the volume of the platform is constrained to a few hundred nanometers so that the device may permeate the "leaky" tumor microvasculature.\(^{[73-75]}\) Although a single platform capable of "full spectrum" anti-cancer therapeutic delivery is yet to be designed, great strides have already been made in the creation of nanoscale multifunctional therapeutics platforms.\(^{[6]}\)

The use of multifunctional nanoscale platforms for cancer treatment is enhanced due to the fact that several localization techniques involve binding to tumor overexpressed cell surface receptors, many of which are important for cell growth (Figure 3). This is because tumor cell surface receptors are often endocytosed upon binding, thus providing a method for intra-cellular delivery of the nanoparticle while simultaneously competing with the natural growth signals, thereby inhibiting their biochemical signaling.\(^{[7]}\)

As a specific example, folate receptor (FRa) is overexpressed on the surface of certain cancer cells, but is naturally expressed only on the apical (away from the blood) surface of epithelial cells,\(^{[86]}\) making it a unique target for therapy. This has allowed the use of folate-conjugated nanoparticles for both imaging and the delivery of drugs.\(^{[87,88]}\) Once the receptor binds a ligand, such as folate, it is taken into the cell via receptor-mediated endocytosis, escapes from the endosome, and is recycled back to the surface after releasing its ligand within the cell.\(^{[89,90]}\) However, one can imagine a technique in which the ligand remains bound to the receptor after depositing the drug, effectively inactivating or reducing the growth receptor pathway. This highlights an important dynamic in drug delivery since any localization technique targeting a growth factor receptor has the unique ability to temporarily disrupt pro-growth in a completely drug-independent manner.

Gold nanoparticles, another common drug delivery platform, also contain inherent properties that have been exploited to treat cancer in a multipronged fashion.\(^{[6]}\) Gold nanoparticles are known to emit heat when exposed to near-infrared laser light, and since cancer cells are highly susceptible to heat, gold nanoparticles can be delivered into cancer to create localized hyperthermia.\(^{[91-93]}\) This technology has since been expanded by combining the nanoparticles with additional drugs, such as the apoptosis-inducing factor TNFa, allowing for a dual treatment with drugs and hyperthermia.\(^{[94,95]}\) By adding a high affinity growth receptor ligand analogue to the nanoparticle, it would be possible to impair the growth receptors and induce hyperthermia without the presence of an actual chemical drug. The use of a single drug encapsulated within gold nanoparticles can hence potentially influence at least three of the components of the cancer cell immortality spectrum, and by combining several other drugs, it would be possible to simultaneously treat additional aspects of the tumor on a system-wide scale.

Although great advances have been made in the utilization of non-drug components of nano-complexes to aid as "adjuvants" in cancer treatment,\(^{[7]}\) an effective generic multifunctional platform must be able to deliver drugs with highly regulated spatiotemporal control. Nanotechnology has recently begun to address this problem, and drug delivery designs using polymeric nanoparticles are becoming increasingly popular. However, by themselves, polymeric nanoparticles typically
release a significant portion of their drug payload in a rapid "burst" manner, severely limiting the propensity for temporal control.6,7 Fortunately, recent advances in nanotechnology have allowed sustained drug release over several weeks, for instance, by conjugating the drug molecule to polymeric nanoparticles.80,90 Likewise, multiple pulses of drug can be delivered by implanting a single polymeric microchip.99 Thus, it is possible to achieve both sustained release (useful for long-term drugs and surveillance molecules) and controlled pulses of drug release (useful for drugs like chemotherapeutics). Taken together with the previously outlined "targeting" aspects, this creates a tunable "spatiotemporal" component to drug delivery with nanoscale polymeric devices.

The temporal kinetics of drug release from polymeric devices is dependent on the physicochemical properties of the polymeric biomaterials involved.17-21 These are typically classified based on their mode of clearance (if at all) from the body (Figure 4a). Biodegradable polymers are broken down and metabolized within the body, bioelimin-
Multifunctional Nanoscale Platforms for Targeting of...

Macromolecular Rapid Communications

Figure 4. Characteristics of nanoscale polymeric biomaterials. a) Classification of polymeric biomaterials as biodegradable, bioeliminable and permanent/implant based on their mode of clearance within the body. b) Chemical structures of popular biocompatible polymers. c) Mechanisms of polymer degradation. d) Degradation lifetimes of biocompatible polymers.

able polymers remain structurally intact but are cleared by the renal filtering system (kidneys), and permanent polymers are used for long term in-vivo implants. Of these, there has been great interest in the use of biodegradable polymeric nanoparticles for cancer treatment.

The most frequently used biodegradable polymers include the polyesters poly(lactic acid) and poly(glycolic acid), or PLA and PGA respectively, among several other biocompatible polymer families (Figure 4b). As PLA and PGA are introduced, they are eventually hydrolytically degraded and form acidic oligomers and monomers, effectively dropping the pH. The acidic pH contributes further to the hydrolysis, eventually reducing the polymeric network via various degradation mechanisms (Figure 4c) that allow the encapsulated drug to be released.

By fine-tuning the physicochemical characteristics of the polymer, it is possible to obtain a high level of control over the kinetics of drug release, polymer size, composition, and chain length have all been shown to have profound effects on the properties of the biomaterial device, and are each easily tailored. It is also possible to create co-polymers, or polymers of multiple monomeric species, allowing for a nearly infinite number of customizable attributes. These traits allow researchers to "fine tune" properties such as the...
As discussed previously, angiogenesis, or the growth of new blood vessels, is an important component of cancer, as it allows oversized tumors to receive additional blood. Vascular disrupting agents (VDAs) comprise a class of cancer drugs that simultaneously disrupt angiogenesis while shutting down the existing vasculature. In theory, this should eliminate the oxygen supply of the tumor, leading to hypoxia and eventual necrosis; in practice, VDAs have been able to slow the growth of tumors, but have not emerged as the “magic bullet” against cancer treatment as was originally expected. Although the individual effectiveness of VDAs has been limited, the combination of VDAs with chemotherapy has produced positive results, with over 10 different VDAs currently in clinical trials. Although simultaneous delivery of VDAs and chemotherapeutics possesses enormous therapeutic potential, the underlying attributes of both classes of drugs raise unique obstacles for effective transport.

The first of three major challenges is that chemotherapeutic agents are traditionally delivered to the tumor via the bloodstream. Since the purpose of VDAs is to effectively shut down the surrounding vasculature, this severely reduces the amount of chemotherapeutic drug that can reach the tumor if it is delivered after vascular disruption has begun. The second problem is that as tumor cells become hypoxic, they over-express hypoxia-inducible factor-1α (HIF-1α), which in turn inhibits the pro-apoptotic effect of chemotherapeutic drugs. Although the combination of VDAs and chemotherapeutics is highly synergistic, it contains an inherent disadvantage. This is because vascular shutdown by VDAs reduces the amount of chemotherapeutic drug that reaches the tumor, increasing the amount of drug required for therapeutic efficacy and escalating systemic toxicity. Finally, the optimal amount of drug delivered and duration of treatment for both classes are completely different. Whereas VDAs typically require long-term administration to prevent angiogenesis, chemotherapy is given in high doses over short treatment cycles, highlighting the need for separate customizable release kinetics.

A “Trojan horse” approach using a multi-compartment, multi-functional nanoscale system has recently been proposed to overcome these challenges. This strategy is based on a delivery system called the “nanocell” (Figure 5), so named because of its visual resemblance to a cell and its nanoscale dimensions. A nuclear polymeric nanoparticle conjugated to a chemotherapeutic agent was encapsulated within a pegylated-phospholipid block-copolymer envelope that already contained a VDA. Thus, an integrated nanoscale platform with two drugs of varying function was designed, each with its own customizable release profile.

As the nanocells are lodged into the tumor tissue by active or passive targeting, the envelope quickly breaks down, releasing both the conjugated nanoparticle and the free VDA into the tumor. The VDA induces vascular collapse, effectively trapping the nanoparticle-chemotherapeutic conjugation within the tumor, allowing the chemotherapeutic agent to undergo sustained release without being able to escape the tumor microenvironment.

To demonstrate the effectiveness of this integrated multifunctional nanoscale platform, the cytotoxic chemotherapeutic drug doxorubicin was conjugated to a PLGA-fabricated nanoparticle. The conjugated nanoparticle was then nucleated inside the VDA drug combrestatin-laden
Nanoscale combination therapeutic device engineering

PEG

Tumor targeting molecule

Inner hydrophilic drug-conjugated polymeric nanoparticle encloses anti-cancer agent for sustained release

Outer hydrophobic lipid layer encloses anti-angiogenic drug

Figure 5. Design of a multifunctional and multi-compartmental nanoscale platform for integrated targeting of manifold factors in the cancer cell immortality spectrum. The nanocell is a multi-compartmental nanoscale system engineered for dual-attack on angiogenesis and cancer cell proliferation. The outer liposomal layer being hydrophobic encloses the anti-angiogenic drug such as chrombetastatin which is released first, followed by the inner polymeric nanoparticle layer releasing the anti-cancer agent from within the blood vessel deprived tumor tissue.

phospholipid block copolymer envelope composed of PEG-DSPE, phosphotidylcholine, and cholesterol (Figure 5a). A tumor-endothelium co-culture bioassay revealed that unlike controls, the nanocells were able to shut down the vasculature surrounding tumor cells in 12 hours and completely ablate the tumor in 30 hours, thus proving the effectiveness of this nanoscale platform for integrated targeting of tumor cells and angiogenesis ex vivo.

The in vivo capacity of the platform was demonstrated by introducing nanocells into mice with either B16/F10 melanomas or Lewis lung carcinoma. Compared to single-drug, no-drug, and non-temporally restrained multi-drug controls, nanocells significantly reduced tumor size, increased apoptosis rates, decreased the amount of HIF1-α and VEGF, reduced the amount of overall system toxicity, and increased the overall survival rates of the mice. These data indicate that the use of nanocells is far more effective than using either drug alone or simultaneously delivering both drugs encapsulated in the same liposome without the polymeric nanoparticle.

Thus, the multi-compartmental and multifunctional nanocell platform (Figure 5) allows for accurate control of both the spatial and temporal targeting, and is also able to overcome the hypoxia-induced resistance to chemotherapeutic drugs. Because the nanoparticle-chemotherapeutic conjugate is trapped within the tumor, the nanocell also significantly alleviates off-target cytotoxic side effects, making it safer than traditional chemotherapy. Additionally, nanocells are able to achieve even greater specificity through the incorporation of cell surface receptors, receptor ligands, antibodies, or aptamers. This extra layer of targeting ensures that the nanocells reaches their appropriate molecular target. This could further limit the already low off-target effects, potentially allowing the use of otherwise highly toxic cancer drugs. Moreover, this would allow an additional level of treatment, as both the polymeric nanoparticle and the targeting component could be used therapeutically in drug-independent manners as previously discussed. Finally, the physicochemical properties of nanocells are easily tailored, allowing for a high degree of customization in delivery rates, drug payloads, and specificity. Since every tumor type reacts differently to each of these parameters, the ability to customize treatments is both beneficial and required for the nanocell to function as a universal multi-drug delivery platform.

As future advancements to the nanocell platform, it is envisioned that strategic drug combinations may be tested as payloads to further the potency of the platform and influence many or all factors of the cancer immortality spectrum. A major challenge to the extension and optimization of the nanocell delivery platform is the still incomplete understanding of the complex interaction of factors relevant to drug release from polymer formulations. As a result, there is a growing need for better modeling and simulation platforms to further the understanding and predictive capabilities of drug release. In the following section, we discuss the recent progress of integrated nanoscale modeling and simulation platforms.

Development of In-Silica Platforms for Computational Nanotechnology and Macromolecular Simulations

Recent work has emphasized the dramatic therapeutic benefits obtained through rational control over the design parameters of polymer-drug formulations. Rational
design of nanoscale polymeric systems for the optimization of delivery strategies requires further understanding of the complex interaction of factors relevant to drug release from polymer formulations. As a result, there is a growing need for better models to understand and predict phenomena such as polymer degradation and nanoparticle erosion. Existing models, however, have been limited in their application to predict drug release from nanoparticles. They typically utilize simplified two-dimensional simulation structures to limit computational cost at the expense of accuracy, and they are additionally structured almost exclusively from microparticle-drug release data.

Recently, a three-dimensional platform has been proposed for modeling and simulating drug release from polymeric nanoparticles (Figure 6a). This model was based on selective incorporation of known nanoscale mechanistic phenomenon pertinent to polymeric devices and parameter integration from material, computational, and pharmaceutical nanotechnologies. As part of this study, a three-dimensional voxel (portmanteau of “volume” and “pixel”) based framework for Monte Carlo simulation was proposed (Figure 6b), that enables direct investigation of the entire spherical nanoparticle during various phases of particle degradation and drug release. Drug release from an experimental formulation of the chemotherapeutic agent dexamethasone encapsulated within poly(lactic-co-glycolic acid) nanoparticles was studied, and good correlation to the simulation results was observed.

This semi-empirical platform was also used to obtain insights into the classical bulk erosion (Figure 6b) and surface erosion (Figure 6c) mechanisms of polymeric matrix degradation and erosion. The 3-D nanoscale platform permitted inclusion of volumes of the polymers, drugs, and pores, and its versatility allowed

---

**Figure 6.** The nanoscale simulation platform developed as part of our multi-pronged approach to targeting the cancer cell immortality spectrum. a) Whole nanoparticle simulation using the 3D platform to study the physicochemical behavior of nanoscale polymer matrices enclosing or conjugated to drug molecules of interest. In this depiction of the platform herein, the progress of polymer degradation (red) and erosion (blue) in a bulk eroding polymeric nanoparticle (gray) is shown. b) Bulk eroding polymeric matrices are traditionally simulated using 2D grids (left) with intact polymers (black), degraded polymers (gray) and eroded pores (white) but these cannot account for volumes, whereas the 3D simulation platform (right) is able to include the effect of volumes of the biomaterials, pores, and drugs as depicted herein with intact (pink/black) and eroded (blue) polymers colored. c) Similarly to bulk erosion, surface erosion of microparticles and nanoparticles is also simulated inclusive of volumes of biomaterials, pores, and drugs, with the 3D simulation platform.
easy incorporation of physicochemical parameters that are important for predictive power. Along the dimension scale from “nano” to “micro”, the phenomenon of agglomeration, or the coming together of individual nanoscale particles to form clusters (or agglomerators), becomes increasingly important due to the high surface energies involved. A computational platform that integrates the nanoscale voxels platform, which currently pertains to single nanoparticles, with multi-nanoparticle simulations wherein interparticle forces are accounted for, hence becomes important (Figure 7a). By varying the interparticle forces with a suitable agglomerate...

Figure 7. From nanomers to agglomerators. a) Spatial compartmentalization at nanoscales (quantum dots, core-shell nanoparticles, and nanomer clusters) and microscales (agglomerators) is depicted pictorially. b) Simulation of agglomer size distribution with time can be computed at fixed agglomeration factor (AF) values to predict time spans when desired agglomer clusters will occur. c) Effect of variation of agglomeration factor (AF) on the size of agglomer clusters formed can be computed as shown and polymer characteristics may be appropriately tailored for desired agglomer sizes.


www.mrc-journal.de
tion factor parameter as shown, it is possible to tease apart these forces and compute quantitative estimates of their influence (Figure 7b). Furthermore, the size of the agglomerates is key to their applications for the treatment of specific cancers.\textsuperscript{[6]}

For instance, treatment of lung cancers and associated inflammation and infections with large agglomerates delivered as aerosol formulations has been explored.\textsuperscript{[118-120]} Such aerosol formulations typically require several hundred polymeric nanoparticles or liposomes to agglomerate together. In most of these scenarios, designing nanoscale agglomerates of smaller sizes (e.g., agglomerate of 2 or 3 nanoparticles) is required to ensure that these are specifically localized to the tumor microenvironment (owing to the dimensions of the pores in the "leaky" tumor microvasculature).\textsuperscript{[6,7,118-121]} In order to facilitate the precise engineering of such nano-complexes for application to tumor treatment, simulation platforms can be used to analyze the concentrations of agglomerates of the required size as a function of time as shown (Figure 7c). Thus, the outlined extensions of the voxel-based 3-D simulation platform will allow for integrated modeling of polymeric nano- and micro-particle behavior in the context of their anti-cancer application.

Given that several phenomena including water and drug diffusion, drug dissolution, drug solubility, particle size and geometry, polymer degradation kinetics, porosity, swelling, average molecular weight of polymer, micro-environmental pH changes, and autocatalysis have all been shown to influence drug release from polymeric devices, there is tremendous value to the development of computational models for predicting multi-phase drug release kinetics. With the advent of more such 3-D computational platforms for engineering nanoscale devices, it is reasonable to expect further progress in our ability to temporally regulate release kinetics. Concomitantly, the virtual explosion of high-resolution 3-D structures for over-expressed cancer cell surface receptors made possible by significant strides in high-throughput x-ray crystallography, structural genomics, and homology-based (comparative) modeling, is continually enabling new opportunities for targeting nanoscale devices to tumor specific signatures. Taken together, the exquisite spatial and temporal regulation of diagnostic and therapeutic agent delivery from nanoscale platforms is likely to present novel methods in the foreseeable future for integrated targeting of the many factors in the cancer cell immortality spectrum.

**Conclusion**

There is an emerging understanding that the technology used to deliver drugs for fighting cancer is as important as the drugs themselves for obtaining maximal therapeutic benefit. The field of nanotechnology is notable for its application to cancer chemotherapy, and pioneering work in nanoparticle-drug formulation has resulted in therapies with longer blood circulation, increased tumor uptake, and enhanced anti-tumor action. However, cancer is a complex disease for which the "magic bullet" single-drug treatment is increasingly appearing unlikely to succeed, and hence, there is great interest in exploration of combination therapeutics for cancer treatment. In this article, we have outlined the molecular underpinnings of tumors using the cancer cell immortality spectrum and motivated the need for a combination therapeutic approach to cancer treatment. We then described emerging nanoscale platforms for both cancer diagnosis and therapeutics, emphasizing the novel approaches that target multiple factors of the cancer cell immortality spectrum. With the development of further such multifunctional nanoscale platforms, it is envisioned that the search for the elusive cancer cure will culminate in potent, cost-effective, minimal side-effect solutions to this dreaded disease.

Acknowledgements: This work was supported by the Singapore-MIT Alliance for Research and Technology (SMART). The authors thank David Weingeist and Luke Robinson for help with literature mining pertaining to clinical combination chemotherapies; Dr. Toomas Haller and Dr. David Bavorone for helpful discussions on polymeric nanoparticles; Barghavi Govindarajan for careful review of this article; and Ada Zielkowksi for administrative help.

Received: August 18, 2009; Revised: October 7, 2009; DOI: 10.1002/marc.200900596

Keywords: cancer; combination therapeutics; nanoparticles; nanotechnology; polymers

Multifunctional Nanoscale Platforms for Targeting of...

Macromolecular Rapid Communications

34.1 Introduction

There exists a delicate balance between life and death within all biological systems. It is estimated that an average of 60 billion cells undergo programmed cell death or apoptosis per day in the adult human body, counterbalancing cell division (Alberts 1983). This equilibrium facilitates the maintenance of an optimal density of cells in healthy tissue. Cancer or tumor is a disproportion that is pathologically characterized by uncontrollable cell division and apoptosis evasion. Tumors are sustained by reservoirs of pro-growth signals and additionally have a distinct insensitivity to anti-growth signals (Hanahan and Weinberg 2000). Tumor tissues thrive on angiogenesis, which is the construction of new blood vessels. These hastily erected blood vessels ensure adequate supply of nutrients and oxygen for the greedy cancer cell colonies (Figg and Folkman 2008). Expanding tumors are typified by an obstinate lack of boundary consciousness and they often impinge on neighboring organelles, disrupting their normal function. Cancer cells, like pathogens, are further capable of metastasizing to non-neighboring organs by hitching onto...
lymphatic or blood vessels. Metastasis, which is characteristic of malignant tumors, permits the rapid colonization of multiple organs with tumor-favoring microenvironments (Harold 2001). Such an irrepressible spread of the disease is usually responsible for the eventual death of the organism.

A number of cancer-causing agents, also known as carcinogens, have been identified (Milman and Weisburger 1994). Sustained exposure to carcinogens causes irreparable genetic mutations, which increases the risk of cancer development. The carcinogenic potential of materials such as asbestos, arsenic, coal tar, heavy metals, tobacco, and dyes as well as environmental risk factors such as exposure to ultraviolet radiation is well characterized. Further, some pathogens have been recognized as cancer-causative agents (Table 34.1). For example, chronic infection with the human papillomavirus (HPV) has been linked with cervical cancer development (Wu 2004). Carcinogenic synergies between multiple agents have also been recorded. For instance, carcinogenic synergy between hepatitis B virus and aflatoxin has been observed to drastically amplify the possibility of sensing technology. Furthermore, miniaturization is also essential for the design of effective anticancer therapeutic systems because the systemic administration of chemotherapy or anticancer chemicals results in high levels of toxicity to healthy systems.

The detection of cancer at its earliest stages is essential for the effective treatment of the disease. The ability to monitor cellular microenvironments and subcellular compartments with high sensitivity is in turn fundamental to the timely recognition of tumorous metamorphosis. This necessitates the miniaturization of sensing technology. Furthermore, miniaturization is also essential for the design of effective anticancer therapeutic systems because the systemic administration of chemotherapy or anticancer chemicals results in high levels of toxicity to healthy tissues. With the advent of nanotechnology, there has been a surge of attempts to "sense" molecular signatures of cancer at its onset and "target" therapeutics to cancer cells.

TABLE 34.1 Pathogens with Suspected Roles in Promoting Carcinogenesis

<table>
<thead>
<tr>
<th>Pathogen</th>
<th>Family</th>
<th>Cancer Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hepatitis B virus (HBV)</td>
<td>Hepadnaviridae</td>
<td>Liver</td>
</tr>
<tr>
<td>Hepatitis C virus (HCV)</td>
<td>Hepadnaviridae</td>
<td>Liver</td>
</tr>
<tr>
<td>Human papilloma virus (HPV)</td>
<td>Papillomaviridae</td>
<td>Cervix, anus, vulva, oropharynx</td>
</tr>
<tr>
<td>Simian virus 40 (SV40)</td>
<td>Polyomaviridae</td>
<td>Lung</td>
</tr>
<tr>
<td>Herpes simplex virus-2 (HSV-2)</td>
<td>Herpesviridae</td>
<td>Cervix</td>
</tr>
<tr>
<td>Epstein-Barr virus (EBV)</td>
<td>Herpesviridae</td>
<td>Nasopharynx, blood</td>
</tr>
<tr>
<td>Human T-lymphotropic virus (HTLV)</td>
<td>Retroviridae</td>
<td>Blood, bone marrow</td>
</tr>
<tr>
<td>Helicobacter pylori</td>
<td>Helicobacteriaceae</td>
<td>Stomach, pancreas</td>
</tr>
<tr>
<td>Salmonella typhi</td>
<td>Enterobacteriaceae</td>
<td>Gallbladder</td>
</tr>
<tr>
<td>Streptococcus bovis</td>
<td>Enterococcaceae</td>
<td>Colonrectum</td>
</tr>
<tr>
<td>Chlamydia pneumoniae</td>
<td>Chlamydiaceae</td>
<td>Lung</td>
</tr>
<tr>
<td>Mycoplasma</td>
<td>Mycoplasmataceae</td>
<td>Stomach, intestine, colon, esophagus</td>
</tr>
</tbody>
</table>

A nanometer is one billionth of a meter and devices that are less than hundreds of nanometers in dimension are known as nanodevices. Nanoscale devices are typically smaller than human cells and are able to achieve cellular entry. Such devices can further readily interact with biological macromolecules owing to their comparable dimensions. Nanotechnology, hence, presents an unprecedented insight into the complex regulatory and signaling network of biomolecular interactions that motivate cancerous transformation of normal cells. Nanotechnology has also provided significant impetus to the development of promising approaches for fighting cancer on various fronts (Figure 34.1).

This chapter describes the full spectrum of nanodevices for molecular diagnostics and therapeutics, specifically highlighting the state-of-the-art nanoscale techniques for targeting cancer. On the diagnostics front, the development of nanoscale systems for the molecular sensing of tumorigenesis fingerprints and for in vivo imaging of the tumor microenvironment is discussed. On the therapeutics front, the development of nanoscale platforms for targeted drug delivery, combination chemotherapy, and sustained drug release is described. Nanoscale systems for other purposes such as multidrug resistance reversal and personalized medicine are then discussed with emphasis on the significance of computational nanotechnology in this post-genomic era. The chapter concludes with a summary of opportunities and challenges involving the translation of nanoscale anticancer devices to the clinical realm.

### 34.2 Nanoscale Molecular Sensing and Early Malignancy Detection

The exquisite sensitivity of nanoscale technology and its ability to monitor subcellular compartments presents clinicians with a paradigm shift in the identification of tumors at elementary stages (Amiji 2006). The design of nanoscale carcinogenesis detection systems is motivated by advancements in molecular sensing, nanoelectromechanical systems (NEMS), nanofluidics, and ultrasensitive imaging technologies. This section showcases these emerging nanoscale opportunities for the timely detection of cancer.

#### 34.2.1 Molecular Combing for Unearthing Genomic Instability

Molecular combing is a prospective high-resolution technology that provides for the linearization and alignment of deoxyribonucleic acid (DNA) molecules thus permitting a thorough analysis of the encoded genetic information. This technique is useful for unearthing mutated genomic domains that may have been missed by conventional sequencing-based screens in cancer patients (Weitzman 2001). Since conventional screenings produce considerable proportion of false negative diagnoses, "cancer combing" is a major step forward toward the development of foolproof procedures for the identification of mutations that arise during the genesis of cancer.
34.2.2 Nanomechanical Cellular Analysis with Atomic Force Microscopy

The atomic force microscope (AFM) can be used to measure the stiffness and elasticity of materials at nanoscale resolutions with high precision. The AFM has been employed to show that lung, breast, and pancreatic tumor cells are all substantially softer than normal cells (Cross et al. 2007). Such studies enable the rational differentiation of cancer cells from noncancerous cells of the same tissue microenvironment, in spite of their general similarity in shape. These studies also contribute to our understanding of the change in mechanical properties underlying each cellular phenotypic event that constitutes carcinogenesis. However, the translation of these in vitro studies to viable in vivo platforms would be paramount to the deployment of cellular elasticity sensing platforms in real-time tumor detection applications.

34.2.3 Tumor Signal Sensing with Nanoelectromechanical Systems

NEMS are nanoscale devices that are constructed out of components such as nanocantilevers, nanosieves, nanowires, and nanochannels. These devices are useful for the collection and analysis of biological signals. NEMS can aid in the timely detection of rare molecular signals associated with malignancy (Cui et al. 2001). For example, nanowire sensor arrays with surface receptors are demonstrated to recognize even femtomolar concentrations of carcinoembryonic antigen and prostate-specific antigen from as few as ten tumor cells (Zheng et al. 2005). This ability to identify tumor molecular markers at small concentrations holds immense potential for the early detection of malignancy.

34.2.4 Tumor Signal Amplification with Superparamagnetic Nanoparticles

Superparamagnetic materials are generally nonmagnetic materials that are composed of small, randomly oriented ferromagnetic clusters. However, when an external magnetic field is applied, thermal fluctuations aid in orienting the clusters resulting in "switching on" of the materials' magnetism (Bean and Livingston 1959). Magnetic resonance imaging or MRI involves scanning materials with externally applied magnetic fields. Superparamagnetic materials are ideal candidates for the amplification of low-amplitude tumor signals owing to their polarization during MRI scans. This property of superparamagnetic nanoparticles has been used for noninvasive illumination of the anatomical contours of brain tumors in the early stages of development (Simberg et al. 2007). Tumor signal amplification with superparamagnetic nanoparticles has also been helpful in locating tumor colonies that may not appear on conventional...
MRI scans owing to the associated poor signal-to-noise (SNR) ratio. This high-sensitivity technology is useful for the timely and accurate detection of cancer.

### 34.2.5 Ultrasensitive Nanoscale Imaging Technology

Conventional optical interferometer systems use resolution-limited, Fourier transform spectroscopy (FTS) to monitor signals from biological samples. Although this provides for faster signal processing and more affordable system design, FTS is unsuitable for nanoscale sensing which by its very nature, demands extremely high resolutions for sensitive scanning. Terahertz spectroscopy systems designed with quantum cascade structures employ far-infrared radiation to extract molecular spectral information with very high resolutions (Ferguson and Zhang 2002). This ultrasensitive imaging technology offers some insight into the native conformations of biomolecules, many of which have collective vibration modes in the terahertz range. One of the applications of this technology is the biomolecular characterization of tumor protein–ligand interactions and this is beneficial for the early detection of cancer (Menik et al. 2002).

The following section examines tumor-specific nanoscale imaging systems.

### 34.3 In Vivo Imaging of the Tumor Microenvironment with Nanoscale Systems

Nanoscale devices are commonly employed for many medical imaging applications (West et al. 2006). Owing to their smaller dimensions compared to the relatively larger pores on the leaky tumor microvasculature, nanoscale devices preferentially accumulate within tumor tissues. This “auto-targeting” of nanoscale devices enables localization of the imaging agent in tumor microenvironments. Nanoscale imaging provides compelling snapshots of the biological mechanisms governing tumor maturation while also enabling effective monitoring of disease spread and assessment of therapeutic efficacy. This section discusses the use of nanoscale systems for in vivo imaging of the tumor microenvironment.

### 34.3.1 Quantum Dots

Quantum dots (QDs) are nanoparticles less than 10 nm in diameter, possessing unique optical, chemical, and electronic characteristics. These include improved signal intensity, narrow emission spectra, good photo-stability, single-wavelength excitation of multiple colors, size-tunable light emission, resistance to photo-bleaching, and ease of surface functionalization (Bruchez and Hotz 2007). Such distinctive properties render QDs as ideal fluorescent probes for the simultaneous illumination of multiple subcellular compartments with high resolution, appreciable sensitivity, and agreeable color contrast. A variety of QD probes have been developed for monitoring the tumor microenvironment. These probes are generally composed of an amphiphilic polymer matrix encapsulated within QDs that are surface-functionalized with tumor-targeting ligands. The powerful combination of ultrasensitivity and ultraprecise targeting enables high-resolution imaging of tumor cells and subcellular structures.

#### 34.3.2 PEGylated Quantum Dots

Polymeric QDs have limited in vivo lifetimes (~minutes). However, they may be optimized for imaging tumors by ensuring adequate circulation time and sustained fluorescence over longer periods. The conjugation of polyethylene glycol (PEG) molecules to the QD polymer coat produces a manifold increase in their lifetime. The resulting fluorescent signal from PEGylated QDs also lasts for several months after injection into the bloodstream. This increased lifetime of PEGylated QDs is due to the presence of a hydration shell around the device that improves its ability to resist pressure (Allen et al. 2002). The hydration sphere in turn is stabilized by a rich network of hydrogen bonds between PEG and water molecules. Furthermore, the excellent structural fit of PEG with the tetrahedral lattice of water molecules enhances the stability of the hydration sphere. Given their greatly increased circulation lifetime, PEGylated QDs are ideal tools for sustained monitoring of tumor cells.

#### 34.3.3 Immunofluorescent Label-Conjugated Quantum Dots

PEGylated QDs can conveniently be conjugated to biomolecules through covalent chemical bonding. These conjugated QDs allow for selective labeling of cell surface receptors, cytoskeleton components, nuclear antigens, and other biomolecules. Immunofluorescent labeling is one technique by which the molecular footprint of tumors may be mapped to aid in the design of appropriate therapeutic systems. For instance, QDs linked with streptavidin and immunoglobulin molecules have been used to selectively label the Her2 receptors on the surface of live breast cancer cells (Wu et al. 2003). Such selective immunofluorescent labeling of tumor biomarkers helps to track the progress of the disease and to evaluate the efficacy of therapeutic regimens.

#### 34.3.4 Bioluminescent Quantum Dots

QDs generally require blue light for efficient excitation. However, the visible frequency range has low tissue-penetration capability and is associated with random excitation of endogenous fluorophores resulting in high levels of background noise (Arnone et al. 2000). Bioluminescent QDs can auto-fluoresce and therefore do not call for any external excitation. Bioluminescence is the phenomenon of light production involving the conversion of biochemical energy to light energy in organisms such as fireflies, glow worms, honey mushrooms, gulper eels, coral,
and vibronaceae. Bioluminescence resonance energy transfer (BRET) is a technology that is prevalently used in a variety of biotechnological applications for harnessing bioluminescence from natural and genetically engineered organisms (Shrestha and Deo 2006). BRET involves the transformation of a naturally occurring fluorophore protein such as luciferase into a bioluminescent probe by conjugating it with QDs. Since fluorophores such as luciferase are of fragile structure, bioluminescent QDs are likely to possess constrained fluorescent activity and highly reduced circulation lifetimes. In order to retain the activity of these autofluorescent devices, synthetic biology techniques are employed to design stable variants of the luciferase gene. Furthermore, genetically engineered fluorophores may be tailored to emit light of shorter wavelength than the natural firefly luciferase protein, thereby synergizing better with the absorption spectrum of QDs. Bioluminescent QDs are powerful tools for probing the anatomical contours of living systems provided the in vivo resistance to degradation by host immune system is enhanced.

### 34.3.5 Cysteine-Coated Quantum Dot Nanocrystals

Coating of QD nanocrystals generally causes variations in their optical properties that adversely affect their pharmacokinetics. However, coating of QD nanocrystals with the sulfur-rich cysteine amino acid renders them bright and stable (Liu et al. 2007). The cysteine coating also prevents the undesirable adsorption of random serum proteins onto the QD surface that typically leads to enlargement and distortion. Furthermore, unbound QDs are rapidly cleared via the kidney’s filtering mechanisms. This suppresses the otherwise rampant background noise contribution of unbound QDs. Cysteine-coated QD nanocrystals are hence particularly useful for the in vivo imaging of microenvironments wherein the tumor signal-to-background noise ratio is generally low.

### 34.3.6 Luminescent Carbon Nanotubes

Carbon nanotubes (CNTs) are cylindrical carbon allotropes of length-to-diameter ratios greater than a million. These nanostructures are characterized by extraordinary strength, chemical inertness, and thermal conductivity (Reich et al. 2004). Hybrid nanostructures of CNTs conjugated to luminescent QDs are used to selectively illuminate cancer cells. An added advantage of these hybrid nanostructures is the additional cavity volume present within the CNT where substantial payloads of anticancer drug can be stored. The successful translation of these luminescent CNTs to clinical trials would require toxicity assays relating to long-term CNT exposure.

### 34.3.7 Nanoscale Polymeric Matrices

Polymeric matrices are highly versatile, biocompatible, biodegradable devices that are extensively used for in vivo imaging applications and a well known example is poly-lactic-co-glycolic acid (LaVan et al. 2003). These devices also offer the advantages of high agent encapsulation efficiency and specific accumulation within tumor sites. Moreover, polymeric matrices can be conveniently surface-conjugated to polyethylene glycol (PEG) molecules for increasing their circulation lifetime and functionalized with tumor-targeting molecules for homing the device specifically to tumor sites (Aubin-Tam and Flamad-Schifferli 2008). These characteristics that render polymeric matrices attractive for in vivo imaging are also responsible for the extensive use of these devices as targeted drug delivery vehicles.

The following section describes the design of targeted nanoscale drug delivery systems with emphasis on the recently discovered biomaterials and technologies that are being utilized for anticancer applications.

### 34.4 Targeting the Nanoscale Delivery of Chemotherapy to Tumors

The development of nanoscale systems for targeted delivery of drugs to cancer cells is a major focus of chemotherapy, primarily for the purpose of toxicity reduction. Nanoscale systems that are smaller than 200nm in diameter effectively navigate through the leaky tumor microvasculature and aggregate selectively within the tumor interstitial space. They are thereafter contained within the tumor microenvironment owing to the dysfunctional lymphatic drainage. Devices that rely primarily on this enhanced permeation and retention (EPR) for delivering drugs to tumors are known as passively targeted or auto-targeted systems. Active targeting, on the other hand, implies that targeting ligands such as polysaccharides, antibodies, peptides, nucleic acid aptamers, or other small biomolecules are conjugated to the surface of the nanodevice. Biomolecules are carefully screened to identify potential targeting ligands that bind selectively to receptors that are unique to, or overexpressed in, cancerous cells. A compilation of overexpressed cell surface receptors in various tumor types is presented (Table 34.2). The conjugation of appropriate tumor-specific targeting molecules onto drug-encapsulated nanodevices ensures selective delivery of the agent to cancerous cells and hence minimal toxicity to noncancerous cells (Figure 34.2). Although passive and active targeting are both helpful for directing nanoscale devices to tumor tissues, the latter mode is associated with greater reduction in chemotherapeutic toxicity, specifically to healthy cells in the immediate neighborhood of tumor cells. This section outlines the advancements in targeting the nanoscale delivery of anticancer drugs.

### 34.4.1 Biodegradable Polymeric Nanoparticles

The duration of therapy for various diseases ranges from a few hours to several months. Application-specific injectable drug delivery devices are being developed across this therapeutic duration spectrum (Chasin and Langer 1990). These devices are typically constituted of polymeric materials that are suitable for in vivo use, that is, nontoxic, noncarcinogenic, nonmutagenic,
TABLE 34.2 Overexpressed Cell Surface Receptors Classified According to Incidence of Cancer Type

<table>
<thead>
<tr>
<th>Overexpressed Cell Surface Receptor</th>
<th>Cancer Type (Incidence If Available)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estrogen receptor alpha (ERα)</td>
<td>Breast (70%), ovary</td>
</tr>
<tr>
<td>Progesterone receptor (PR)</td>
<td>Breast (64%), uterus</td>
</tr>
<tr>
<td>Human epidermal growth factor receptor 2 (HER2)</td>
<td>Pancreas (26%), bladder (44%), stomach</td>
</tr>
<tr>
<td>Endothelin-A receptor (ETAR)</td>
<td>Breast (45.3%), ovary</td>
</tr>
<tr>
<td>Platelet-derived growth factor receptor (PDGF-R)</td>
<td>Liver (22.1%), breast (39.2%)</td>
</tr>
<tr>
<td>Fibroblast growth factor receptor 2 (FGFR-2)</td>
<td>Sarcoma, breast (3%-10%)</td>
</tr>
<tr>
<td>Heparin Sulfate Glycosaminoglycan (HSGAG)</td>
<td>Breast, colon, prostate, pituitary</td>
</tr>
<tr>
<td>Neuropilin-1 (NRP-1)</td>
<td>Brain (57%), stomach, prostate (69%)</td>
</tr>
<tr>
<td>Androgen receptor (AR)</td>
<td>Cervix</td>
</tr>
<tr>
<td>Insulin-like growth factor-1 receptor (IGF,IR)</td>
<td>Liver (62%)</td>
</tr>
<tr>
<td>G-Protein coupled receptor (GPCR)</td>
<td>Skin</td>
</tr>
<tr>
<td>Cannabinoid-specific receptors (CBR)</td>
<td>Rectal, ovary, pancreas, breast</td>
</tr>
<tr>
<td>Chemokine receptor (CXCR4)</td>
<td>Skin</td>
</tr>
<tr>
<td>Vascular endothelial growth factor receptor (VEGFR)</td>
<td>Skin, brain</td>
</tr>
</tbody>
</table>

TABLE 34.2 (Cont.) Biodegradable materials are initially in the solid or gel phase and are thereafter broken down into natural metabolites within the body by hydrolytic or enzymatic activity (Hasirci et al. 2001). PLGA is an example of a biodegradable material that has been approved for clinical use by the Food and Drug Administration (FDA). PLGA is completely hydrolyzed into lactic acid and glycolic acid which are components of the Kreb's cycle and are hence naturally metabolized by the body. The characteristic degradation half-lives of biodegradable polymers varies from a few minutes to several years based upon the ease of hydrolysis (Table 34.4). Matrices composed of biodegradable polymers degrade by cross-link, side chain, or backbone degradation and the polymer molecular weight plays an important role in determining the average degradation lifetimes (Figure 34.3a). Other characteristics of biodegradable polymers that are known to influence their hydrolysis rate include hydrophobicity, steric effects (local structure, glass transition), microstructure (porosity, phase separation, crystallinity), oligomer solubility, autocatalysis, and pH of the medium (Kumar 2007). Biodegradable polymeric devices are classified as bulk or surface eroding based on the hydrolysis and diffusion rates (von Burkersroda et al. 2002). Bulk eroding polymer matrices undergo uniform, instantaneous wetting and degradation happens throughout the bulk of the matrix, whereas water diffuses much more slowly into surface-eroding polymer matrices and degradation hence happens only at the exposed surface of these matrices (Figure 34.3b). The drug release profile associated with surface-eroding polymer matrices is generally monophasic whereas bulk eroding polymer matrices display more biphasic release profiles with a distinct initial "burst release" phase. During the burst phase,
### TABLE 34.3 Polymeric Biomaterials Classified According to Mode of In Vivo Clearance

<table>
<thead>
<tr>
<th>Clearance Mode</th>
<th>Polymer</th>
<th>Chemical Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biodegradable</td>
<td>Poly(glycolic acid)</td>
<td><img src="image" alt="Poly(glycolic acid) chemical structure" /></td>
</tr>
<tr>
<td></td>
<td>Poly(lactic acid)</td>
<td><img src="image" alt="Poly(lactic acid) chemical structure" /></td>
</tr>
<tr>
<td></td>
<td>Poly(lactide-co-glycolide)</td>
<td><img src="image" alt="Poly(lactide-co-glycolide) chemical structure" /></td>
</tr>
<tr>
<td></td>
<td>Poly(e-caprolactone)</td>
<td><img src="image" alt="Poly(e-caprolactone) chemical structure" /></td>
</tr>
<tr>
<td></td>
<td>Poly(malic acid)</td>
<td><img src="image" alt="Poly(malic acid) chemical structure" /></td>
</tr>
<tr>
<td></td>
<td>Poly(ethylene carbonate)</td>
<td><img src="image" alt="Poly(ethylene carbonate) chemical structure" /></td>
</tr>
<tr>
<td></td>
<td>Poly(propylene carbonate)</td>
<td><img src="image" alt="Poly(propylene carbonate) chemical structure" /></td>
</tr>
<tr>
<td></td>
<td>Poly(anhydride)</td>
<td><img src="image" alt="Poly(anhydride) chemical structure" /></td>
</tr>
<tr>
<td>Bioeliminable</td>
<td>Poly(alkyl cyanoacrylate)</td>
<td><img src="image" alt="Poly(alkyl cyanoacrylate) chemical structure" /></td>
</tr>
<tr>
<td></td>
<td>Poly(ethylene oxide)</td>
<td><img src="image" alt="Poly(ethylene oxide) chemical structure" /></td>
</tr>
<tr>
<td>Permanent/implant</td>
<td>Poly(methyl methacrylate)</td>
<td><img src="image" alt="Poly(methyl methacrylate) chemical structure" /></td>
</tr>
<tr>
<td></td>
<td>poly(hydroxyethyl methacrylate)</td>
<td><img src="image" alt="poly(hydroxyethyl methacrylate) chemical structure" /></td>
</tr>
<tr>
<td></td>
<td>Poly(N-isopropyl acrylamide)</td>
<td><img src="image" alt="Poly(N-isopropyl acrylamide) chemical structure" /></td>
</tr>
</tbody>
</table>
bulk eroding nanoparticles release a significant volume of their drug payload and this is undesirable for therapeutic applications (Langer and Folkman 1978). However, drug molecules may be conjugated chemically to polymer molecules to ensure non-burst release.

34.4.2 Liposomes, Micelles, and Polymerosomes

Liposomes are nanoscale spheres composed of one or more bilayer, self-assembled, concentric lipid membranes which can be used to encapsulate various biological molecules. Micelles are self-assembled, spherical lipid monolayers with a hydrophobic core—hydrophilic shell architecture. Polymerosomes are composed of synthetic polymer amphiphiles and are structurally similar to liposomes. Certain characteristics of these lipid-based carriers enable their use for chemotherapeutic drug delivery. These include biodegradability, biocompatibility, drug insulation, ease of surface functionalization, and, encapsulation of drugs with wide-spectrum physicochemical properties (Peer et al. 2007). Furthermore, some of these carriers can be loaded with multiple drug combinations. For example, the inner cavity of liposomes being aqueous is ideal for encapsulating hydrophilic drugs, while the volume between successive lipid membranes is well suited for housing relatively hydrophobic drugs (Figure 34.4a). While these advantages have resulted in extensive deployment of lipid-based nanoscale carriers in therapeutic applications, the synthesis of these carriers remains cumbersome. This challenge is overcome by the use of lipidoid libraries.

34.4.3 Lipidoid Libraries

The time-intensive synthesis of liposomes is a major limitation to the throughput of this class of delivery system. Lipidoid libraries are designed by the rapid and parallel combination of amino molecules with alkyl-akrylates and alkyl-acrylamides. This multiplexed synthesis is capable of producing more than a 1000 different lipid-like structures in an accelerated manner (Akinc et al. 2008). Lipidoids are a viable platform for the targeted intracellular delivery of short strands of ribonucleic acid (RNA). The targeted approach is necessary because direct introduction of RNA into the bloodstream results in its breakdown by the body’s immune system. The RNA may be engineered to block the action of selectively targeted oncogenes, while leaving other cellular mechanisms intact. This exquisite specificity offers a new tool for harnessing the potential of RNA interference (RNAi) which is aimed at suppressing the expression of specific proteins that have been associated with cancer.

---

**TABLE 34.4** Degradation Lifetimes of Hydrolysis-Susceptible Polymers

<table>
<thead>
<tr>
<th>Polymer Class</th>
<th>Degradation Half-Life</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polyanhydrides</td>
<td>&lt;10 min</td>
</tr>
<tr>
<td>Polyorthoesters</td>
<td>~5 h</td>
</tr>
<tr>
<td>Polysters</td>
<td>~20 months</td>
</tr>
<tr>
<td>Polyphosphazenes</td>
<td>~5 years</td>
</tr>
<tr>
<td>Polyamides</td>
<td>&gt;50,000 years</td>
</tr>
</tbody>
</table>

**FIGURE 34.3** (a) Mechanisms of polymer degradation; (b) modes of polymeric matrix erosion.
34.4.4 Immunoliposomes

Antibodies or immunoglobulins are proteins used by the immune system to identify and neutralize bacteria, viruses, and other foreign bodies. The chemical conjugation of monoclonal antibodies to the distal ends of PEGylated liposomes yields immunoliposomes (Figure 34.4b). These are useful both for the intracellular targeting of therapeutic nucleotides aimed at malignant cells as well as for roping in the immune system's arsenal (Kontermann 2006). Moreover, unlike liposomes that are incapable of permeating the blood-brain barrier composed of the brain capillary endothelial wall, immunoliposomes undergo receptor-mediated transcytosis. This has been used to successfully deliver small molecule anticancer drugs such as doxorubicin to brain tumors (Huwyler et al. 1996).

34.4.5 Dendrimers

Dendrimers are synthetic organic macromolecules structured like branched trees (Figure 34.4c). The fabrication of these macromolecules is relatively new to polymer sciences (Tomalia and Frechet 2003). Dendrimers are classified as bioeliminable because they are cleared rapidly from blood by the kidneys owing to their nanoscale (<10 nm) dimensions. They are also water soluble, have well-defined chemical structures, and are completely biocompatible. Dendrimers can also be conveniently conjugated with appropriate targeting molecules with moderate affinity and high avidity. The term "avidity" refers to multivalent collective binding. High avidity of multi-ligand conjugated dendrimer nanovehicles is responsible for manifold increase in their targeting ability (Hong et al. 2007). The moderate affinity of dendrimers is important because high binding affinity of nanocarriers to their targets decreases the probability of their tissue penetration due to "binding-site barrier" effects. Although dendrimers are well suited for targeted drug delivery applications, their mass production is a challenging task.

34.4.6 Natural and Synthetic Viral Capsids

Capsids are protein shells of viruses consisting of several oligomeric units. These are structured as icosahedrons (20 equilateral triangular faces) or other complex geometries and house the viral genetic matter (Figure 34.4d). Like carbon buckyballs, viral capsids self-assemble into near-spherical geometries with high surface area-to-volume ratios (Wagner and Hewlett 2004). The specificity of biomolecular interactions between glycoproteins on the host cell surface and the virus coat dictates cellular entry, implying that certain cells are intrinsically more susceptible to infection with certain vectors. This is the rationale behind cell-specific targeted gene therapy with viral vectors (Table 34.5). Furthermore, vectors can be tailored for tumor-centric...
TABLE 34.5 Tumor Targeting with Viral Capsids and Vectors

<table>
<thead>
<tr>
<th>Viral Family</th>
<th>Type</th>
<th>Capsid</th>
<th>Target Molecules</th>
<th>Tumor Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adenoviridae</td>
<td>DNA</td>
<td>Icosahedral</td>
<td>Urokinase-type plasminogen activator</td>
<td>Skin, prostate</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Matrix metalloproteinases (MMP)</td>
<td></td>
</tr>
<tr>
<td>Herpesviridae</td>
<td>DNA</td>
<td>Icosahedral</td>
<td>Urokinase-type plasminogen activator</td>
<td>Brain</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Matrix metalloproteinases (MMP)</td>
<td></td>
</tr>
<tr>
<td>Poxviridae</td>
<td>DNA</td>
<td>Complex</td>
<td>Urokinase-type plasminogen activator</td>
<td>Liver</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Matrix metalloproteinases (MMP)</td>
<td>Skin</td>
</tr>
<tr>
<td>Reoviridae</td>
<td>RNA</td>
<td>Icosahedral</td>
<td>Matrix metalloproteinases (MMP)</td>
<td>Skin</td>
</tr>
<tr>
<td>Picornaviridae</td>
<td>RNA</td>
<td>Icosahedral</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Rhabdoviridae</td>
<td>RNA</td>
<td>Helical</td>
<td>Matrix metalloproteinases (MMP)</td>
<td>Skin</td>
</tr>
<tr>
<td>Paramyxoviridae</td>
<td>RNA</td>
<td>Helical</td>
<td>Signaling lymphocytic activation molecule</td>
<td>Bladder, skin, ovary, brain</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Membrane cofactor protein (MCP)</td>
<td></td>
</tr>
</tbody>
</table>

34.4.9 Biomimetic Smart Material Nanodevices

Materials may be classified as "regular" or "smart" depending on their ability to adapt to external stimuli. The properties of smart materials can be significantly altered in a controlled manner using specific external stimuli to which the material is sensitive. Regular materials, on the other hand, are able to offer only their characteristic and often limited set of properties (Langer and Tirrell 2004). Most regular materials (e.g., steel) are a sharp contrast to biological materials that show definitive adaptability to environmental stimuli (e.g., plants spreading out to maximize the area exposed to sunlight). Biomimetics refers to the engineering of stimulus-responsive smart systems based upon design principles borrowed from the study of biological materials (Campbell 1995). The ability to quantitatively predict the change in properties of smart materials with varying external stimuli is beneficial for the rationale design of responsive materials for biomedical applications. There are numerous examples of nanoscale “smart” anticancer technologies (Table 34.6). However, many of these technologies are associated with safety concerns relating to deployment of smart materials in vivo. One exception is stimulus-responsive nanogels that are biodegradable and hence safe for in vivo use.

34.4.10 Hydrogels

Hydrogels are water-soluble, cross-linked networks that swell with water inflow (Figure 34.4f). Hydrogels are classified according to the kind of chemical bonding involved. Ease of chemical modification, in situ formability and biodegradability are some of the properties that make hydrogels desirable for drug delivery (Khademhosseini et al. 2006). Moreover hydrogels can be readily tailored with recognition sites such as adhesion or collagenase sequences for a variety of biological applications (Mann et al. 2001). Further, the environmental responsiveness of hydrogels to stimuli such as change in pH, ionic strength, or presence of analytes, is very useful for the regulation of network swelling. This in turn controls the kinetics of drug release from hydrogels. For instance, pH-responsive nanoscale hydrogels or nanogels may be designed to swell specifically within the relatively acidic...
TABLE 34.6 Stimulus-Responsive Smart Materials in Nanoscale Antitumor Applications

<table>
<thead>
<tr>
<th>Material</th>
<th>Stimulus</th>
<th>Variable</th>
<th>Antitumor Applications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Piezoelectric</td>
<td>Force</td>
<td>Voltage</td>
<td>Sensing of tumor cells</td>
</tr>
<tr>
<td>Shape memory alloy/polymer</td>
<td>Temperature</td>
<td>Shape</td>
<td>Controlled drug release</td>
</tr>
<tr>
<td>Magnetic shape memory alloy</td>
<td>Magnetic field</td>
<td>Anti-inflammatory drug delivery</td>
<td></td>
</tr>
<tr>
<td>pH-sensitive polymers</td>
<td>pH</td>
<td>Volume</td>
<td>Targeted drug delivery</td>
</tr>
<tr>
<td>Electro rheostatic</td>
<td>Electric field</td>
<td>Viscosity</td>
<td>Controlled drug release</td>
</tr>
<tr>
<td>Magnetorheostatic</td>
<td>Magnetic field</td>
<td>Tumor cell separation</td>
<td></td>
</tr>
<tr>
<td>Non-Newtonian fluid</td>
<td>Force</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Halochromic</td>
<td>pH</td>
<td></td>
<td>High-contrast imaging</td>
</tr>
<tr>
<td>Electrochromic</td>
<td>Electric field</td>
<td></td>
<td>Subcellular imaging</td>
</tr>
<tr>
<td>Thermochromic</td>
<td>Temperature</td>
<td>Color</td>
<td>Tumor microenvironment monitoring</td>
</tr>
<tr>
<td>Photochromic</td>
<td>Light intensity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Supramolecular assemblies</td>
<td>Enzymes</td>
<td>Entropy</td>
<td>Metastasis and angiogenesis detection</td>
</tr>
<tr>
<td>Chemically cross-linked hydrogels</td>
<td></td>
<td></td>
<td>Agent encapsulation in nanoparticles</td>
</tr>
<tr>
<td>Enzyme-responsive surfaces</td>
<td>Enzymes</td>
<td></td>
<td>Combination chemotherapy</td>
</tr>
</tbody>
</table>

34.5 Nanoscale Systems for Sustained Drug Release and Tumor Monitoring

Sustained release of agents over several weeks is required for many applications such as monitoring of the tumor microenvironment and anticancer drug delivery (Langer and Folkman 1978). The design of nanoscale systems that are tailored for sustained release applications is outlined in this section.

34.5.1 Drug-Conjugated Polymeric Nanoparticles

The chemical conjugation of drug molecules to polymer fragments prevents rapid dissolution of the drug into the medium. The mechanism governing sustained release from drug-conjugated polymeric nanoparticles is as follows (Figure 34.5a). Upon repeated hydrolytic or enzymatic cleavage of the polymer backbones, the fragments become increasingly smaller in size. The dissolution of the drug–polymer conjugates commences only after the molecular weight of these fragments decreases below a certain threshold that is characteristic of the polymer. The dissolved drug molecules then begin diffusing out of the nanoparticle. An added advantage of drug-conjugated nanoparticles is the absence of the initial burst release phase which, as was discussed earlier, is characteristic of drug-encapsulated polymeric nanoparticles. Burst release in the latter occurs primarily because of the near-instantaneous efflux of medium-exposed drug molecules present in the vicinity of the carrier surface (Figure 34.5a). In the case of drug-conjugated polymeric nanoparticles, the chemical binding of the drug molecules to the polymer backbone prevents instantaneous diffusion of even these medium-exposed drug molecules on the carrier surface. Drug-conjugated polymeric nanoparticle systems have been used for the sustained delivery of chemotherapeutic agents. For example, PLGA-doxorubicin nanoconjugates have been employed to effectively deliver the chemotherapeutic drug doxorubicin to tumors over several weeks (Yoo et al. 2000). Additionally, the polymers constituting the nanoparticles may be surface functionalized with PEG and tumor-targeting molecules, similar to other nanodevices. While the use of drug-conjugated polymeric nanoparticles is associated with these multiple advantages, implants of nanoscale devices help in attaining a more locally targeted delivery in the immediate neighborhood of the implantation site (Grayson et al. 2003).

34.5.2 Nanochip Implants

Polymeric multi-reservoir microchips are designed for temporally controlled local release of multiple agents over several weeks (Grayson et al. 2003). By varying the enclosing "gate" characteristics that determine the kinetics of polymer matrix hydrolysis, unique temporal release profiles are obtained for each drug contained within the chip reservoirs. Polymeric nanochips serve as preprogrammed implants for the sustained and controlled release of combination chemotherapy or for simultaneous delivery of drugs and imaging agents for parallel therapy and monitoring. However, tumor sites are not always conducive for procedures such as implantation and it hence becomes important to develop minimally invasive platforms for sustained combination chemotherapy. The following section examines such technologies that are driving nanoscale combination chemotherapy.
34.6 Nanoscale Combination Therapy: Angiogenesis, Metastasis, and Beyond

Cancer is a complex disease that arises from the dysregulation of several biological networks that are frequently interconnected (Keith et al. 2005). Consequently, the single drug-single target approach of monotherapies is often less effective than combination chemotherapies that can synergistically target multiple processes simultaneously. This is verified by assessments of numerous successful clinical combination therapies for cancer (Zimmerman et al. 2007). Therapeutic synergy involves one or more of the following effects—multiplicative increase of cancer cell death, decreased dosage requirement of each drug, reduced toxicity, and minimized development of drug resistance (Chou 2006). The last three effects are a natural consequence of using multiple drugs because the decreased requirement for each drug alleviates issues of toxicity and drug resistance. A compilation of synergistic interactions amongst some of the commonly administered cancer drugs indicates an abundance of mechanistic synergy that can be harnessed by combination therapies (Table 34.7).

However, the effective administration of combination therapies requires very precise spatiotemporal control in the release of each drug owing to the sensitivity associated with functional rewiring of cellular biochemical signaling networks. This in turn calls for the design of sophisticated nanoscale delivery platforms that can carry multiple payloads and disburse them selectively to the cancerous tissues at appropriate times. This section examines such nanoscale platforms for the administration of cancer combination chemotherapy.

34.6.1 Nanocells

Antiangiogenesis drugs curb the growth of new blood vessels thereby starving the cancer colony of oxygen and nutrients. This limits the expansion of tumor tissues. However, delivery of just antiangiogenesis drugs does not suffice to defeat cancer owing to subsequent systemic hypertoxicity (Folkman 1996). Moreover, the destruction of tumor blood supply by antiangiogenesis drugs cuts off potential routes for the delivery of chemotherapeutic agents (Carmeliet and Jain 2000). The design of the nanocell is based upon the principle that the chemotherapeutic agent could be released from within the tumor after its...
TABLE 34.7 Cancer Combination Chemotherapy

<table>
<thead>
<tr>
<th></th>
<th>Cap</th>
<th>Cyc</th>
<th>Doc</th>
<th>Dox</th>
<th>Epi</th>
<th>Flu</th>
<th>Gem</th>
<th>Lap</th>
<th>Let</th>
<th>Met</th>
<th>Mit</th>
<th>Pac</th>
<th>Tam</th>
<th>Tra</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cap</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cyc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Doc</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dox</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Epi</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flu</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gem</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lap</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Let</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mit</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pac</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tam</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tra</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: Synergistic (●), additive (○), and antagonistic (□) interactions between the two dozen most commonly used cancer drugs. The following three letter abbreviations are used for the drugs: Capecitabine (Xeloda®), Cyclophosphamide (Cytoxan®), Docetaxel (Taxotere®), Doxorubicin (Adriamycin®), Endoxifen (Femara®), Fluorouracil (FU), Gemcitabine (Gemzar®), Lapatinib (Tykerb®), Letrozole (Femara®), Methotrexate (MTX), Mitoxantrone, Paclitaxel (Taxol®), Tamoxifen (Nolvadex®), Trastuzumab (Herceptin®).

blood supply is restricted (Sengupta et al. 2005). The nanocell is a dual-compartment device that derives its name from its resemblance to a cell (Figure 34.5b). The device is composed of an outer PEGylated liposomal vesicle enclosing a polymeric nanoparticle in its interior. The antiangiogenesis drug is encapsulated within the lipid layers and the interior polymeric nanoparticle houses the chemotherapeutic agent. Owing to the nanoscale dimensions of this device and the predominance of enhanced permeation and retention (EPR) in the tumor microenvironment as outlined earlier, the nanocells are selectively targeted to tumor tissues. Once the nanocells are lodged within the tumors, the outer layer releases the antiangiogenesis agent thereby cutting off tumor blood supply. The sustained release of the chemotherapeutic agent then commences from within the tumor tissue. Combination chemotherapy with nanocells is more efficient than nanoscale monotherapy and conventional combination chemotherapy in inhibiting tumor growth. The nanocell platform can be further extended to house drugs that target other features of cancer such as drug resistance (Sengupta and Sasisekharan 2007).

34.6.2 Nanoscale Modulation of Tumor Extracellular Matrix

The surface of cancer cells is rich in a plethora of complex polysaccharides such as heparan sulfate glycosaminoglycans (GAGs) that are known to regulate tumorigenesis, tumor progression, neovascularization, and metastasis (Sasisekharan et al. 2002). Furthermore, the extracellular matrix (ECM) is composed of macromolecular networks that restrict liberal transport of materials thereby limiting the efficiency of anticancer systems (Goodman et al. 2007). The ECM and cell surface can be modulated with enzymes such as collagenase that degrades collagen fibers or heparanase that degrades heparan sulfate GAGs. These ECM-modulating enzymes play critical roles in several aspects of tumor biology (Liu and Sasisekharan 2005). The combination of anticancer and antiangiogenic drugs with such enzymes that alter the ECM is hence crucial for maximally synergistic therapeutic benefit. For instance, conjugation of collagenase onto the surface of nanoparticles for site-specific degradation of collagen networks produces manifold increase in nanoparticle mobility and internalization (Goodman et al. 2007).

34.6.3 Chemo-Thermal Combo-Therapy with Targeted Gold Nanoparticles

Targeted thermal therapy generally inhibits expansion of tumor tissues and antitumor drugs are more adept at killing cancerous cells at elevated temperatures (Everts et al. 2006). Gold nanoparticles have the ability to transduce light into heat and this property appears promising for two-pronged chemo-thermal combination therapy. For example, targeted gold nanoparticles can carry drugs as payload and concomitantly transduce light to heat in the tumor microenvironment resulting in more efficient antitumor activity (Visaria et al. 2006).
34.6.4 Combination Gene Therapy for Synergistic Targeting of Multiple Pathways

Monotherapies typically target specific genes involved in specific pathways responsible for motivating cancer. Combination therapies on the other hand target multiple genetic pathways and harness the resulting synergy to produce a multiplicative increase in cancer cell death (Keith et al. 2005). However, the use of nanoscale platforms for controlled delivery is paramount to realizing the full potential of combination gene therapy. For instance, controlled delivery of multiple tumor-suppressor genes with nanoparticle-based systems is more effective than the individual monotherapies in treating cancer (Deng et al. 2007).

34.6.5 Nanoscale Combo-Therapies for Targeting Metastasis and Angiogenesis

Metastasis from cancer colonies are significantly more difficult to detect and treat than the primary tumor itself. However, since newly formed metastatic tumor colonies thrive on the development of fresh blood vessels by angiogenesis, nanoparticle-based systems may be used to target drugs specifically to the sites of maximal blood vessel formation (Murphy et al. 2008). The targeting approach involves sensing the gradients of pro-angiogenetic factors and directing the nanoparticle-based systems toward increasing concentrations of these factors. This integrated approach to treatment of metastasis and angiogenesis with nanoparticle-based systems is more effective than monotherapies in suppressing the spread of malignancy.

34.6.6 Fighting Intractable Tumors with Nanoparticle-Based Combo-Therapies

Cancer cells develop resistance to anticancer agents leading to the formation of intractable tumors. However, pretreatment of intractable tumors with subtherapeutic levels of a potent but highly toxic anticancer agent followed by nanoparticle-based delivery of a second anticancer agent sensitizes tumors cells to the latter (Kano et al. 2007). Nanoscale systems enable the administration of such drug cocktails with precise spatiotemporal targeting.

The following section focuses on nanoscale systems for tumor drug resistance reversal.

34.7 Drug Resistance Reversal in Tumor Cells with Nanotechnology

Multi drug resistance (MDR) is a property of certain pathological bacterial and tumor cells that successfully defy a wide spectrum of potent apoptosis-inflicting drugs (Krishan and Arya 2002). Increased drug efflux, enzymatic drug deactivation, increased membrane permeability, altered drug molecule receptor sites, and creation of compensatory metabolic pathways are some of the mechanisms employed by tumor cells to attain MDR (Figure 34.6a). Nanoparticle-based delivery provides an alternative route to the internalization of drugs since it utilizes a different set of receptors that are not modified by MDR (Figure 34.6b). This section focuses on nanoscale strategies for MDR reversal.

34.7.1 Direct Intracellular Ceramide Delivery with Polymeric Nanoparticles and Liposomes

Ceramide is the pro-apoptotic mediator that is suppressed by the overexpression of glucosyl ceramide synthase (GCS) enzyme in drug-resistant tumor cells. Intravenous injection of ceramide is not feasible owing to high systemic toxicity, necessitating nanoscale-targeted delivery of ceramide directly to the cytoplasm of cancer cells. Nanoparticles that are co-encapsulated with ceramide and an anticancer drug are effective in reversing MDR and killing tumor cells that are normally insensitive to the anticancer drug. For instance, co-encapsulation of ceramide and paclitaxel in polymeric nanoparticles promotes apoptotic signaling in human ovarian cancer cells with MDR and results in nearly 100-fold increase in their chemotherapeutic sensitization (Devalapally et al. 2008). Nanoliposomal formulations of pro-apoptotic ceramide and the anticancer drug sorafenib are also able to completely eliminate breast and skin cancer cell lines in vitro (Tran et al. 2008). While ceramide inhibits MDR, sorafenib attacks angiogenesis and promotes apoptosis. Liposomes and polymeric nanoparticles are hence useful for direct cytoplasmic delivery of MDR reversing agents and pro-apoptotic drugs.

34.7.2 Nanoscale Electrochemotherapy

Electrochemotherapy is a relatively new cancer treatment modality. This form of therapy uses electric pulses to increase the permeability of cancer cell membranes, followed by the delivery of chemotherapeutic agents. For instance, electroporation combined with delivery of bleomycin is effective in increasing the membrane permeability and promoting the internalization of bleomycin molecules into a variety of cancer cell lines in vitro (Gothelf et al. 2003). However, electrochemotherapy as an in vivo MDR-treatment modality will be practicable only upon adapting the electroporation technology on targeted, safe therapeutic platforms (Wagner 2007). The following section discusses the development of integrated platforms for testing a variety of combination drug cocktails.

34.8 Body-on-a-Chip: Systems Approach to Drug Testing on Integrated Platforms

Microfluidics is the study of fluid behavior at microscales. Microfluidic chips have been designed to host a variety of human cells and mimic different body tissues (Shuler and Xu 2007). This effectively bridges in vitro and in vivo testing by simulating full-fledged tissue microenvironments. Such integrated body-on-a-chip platforms allow for the targeted screening
Nanotechnology for Targeting Cancer

and selection of the best-suited drugs instead of the conventional shot-in-the-dark approach. These platforms are useful to test the efficacy and toxicity of preclinical, anticancer drugs, thereby saving time and money associated with conventional drug screening processes. Furthermore, this technology holds promise for "personalized medicine" wherein patient-specific information is used to predetermine drug efficacy and side effects prior to actual administration.

34.9 Theoretical and Computational Nanotechnology

Computational nanotechnology refers to the design of models for elucidating the mechanistic underpinnings of nanoscale phenomena. An understanding of theoretical nanosciences is fundamental to designing new nanoscale technologies for targeting cancer (Rieth and Schommers 2006). This section discusses a few integrated semiempirical models of nanoscale systems for anticancer applications.

34.9.1 Modeling Polymeric Nanoparticle Erosion and Drug Release

The degradation kinetics of polymeric networks within polymer-based nanoscale and microscale delivery systems determines the rate of drug release. Several computational models have been developed to simulate hydrolysis-driven erosion of polymeric matrices (Siepmann and Gopferich 2001). Many of these models employ two-dimensional pixel grids to represent polymeric matrices and associate the pixels with a degradation lifetime that is derived from the characteristic rate constant of hydrolysis for that polymer. Such pixel-based Monte Carlo (MC) simulations
have been successful in predicting the mechanistic underpinings of polymer hydrolysis and have also been used to quantitatively estimate matrix erosion kinetics, porosity changes, microenvironmental pH variations, and drug release rates.

34.9.2 Multiscale Spatiotemporal Modeling of Nanoparticle Penetration in Tumor Spheroids

The tissue-penetration efficiency of nanoparticles increases with the use of collagenase enzyme to modulate the extracellular environment. Semiempirical, multiscale models of nanoparticle penetration into tumor spheroids are able to accurately predict the spatiotemporal distribution profiles of nanoparticles of various dimensions (Goodman et al. 2007). Furthermore, tissue-specific circulation of the agents released from the nanoparticles can also be estimated from these models. Multiscale models provide a platform to bridge molecular-, cellular-, and tissue-level phenomena that prevail in tumor microenvironments. Multiscale models also collectively contribute to our understanding of tumor architectures and how they impinge on the efficiency of nanoparticle-based delivery systems.

34.10 Summary

There is an emerging understanding that a systems approach to targeting cancer will be fundamental to the development of effective diagnostic, therapeutic, and preventive tools (Hartwell et al. 2006). In this post-genomic era, proteomics (Timms 2008), glycomics (Raman et al. 2005) and other "omics" platforms are greatly advancing our knowledge of the molecular and cellular underpinnings of cancer. In order to transform this knowledge into modular, multifunctional, and potent nanoscale devices for fighting cancer, the National Cancer Institute (NCI) has established the "Alliance for nanotechnology in cancer" (http://nano.cancer.gov/). Other countries have similarly embarked on the nano-war against cancer, resulting in a spurt of innovative anticancer approaches. However, seamless integration of anticancer efforts in nanotechnology, bioinformatics, and modern molecular biology for successful clinical application is in its infancy. Cost efficacy, high throughput, standardization in design, development of benchmarks, long-term toxicity studies, and creation of opportunities for synergistic interaction of engineers, biologists, and physicians are some future milestones toward the assimilation of nano-technology into mainstream clinical biomedicine.

Acknowledgments

This work was supported by the Singapore-MIT Alliance for Research and Technology (SMART). The authors thank Dr. V. Sassekharan for insightful discussions, David Weingeist, and Luke Robinson for help with compiling information on clinical combination chemotherapies, David Eavóre for helpful discussions on polymeric nanoparticles, Barghavi Govindarajan for careful review of this chapter, and Ada Ziolkowski for assistance with correspondence.

References


Nanotechnology for Targeting Cancer


3. Recombinant Enzymatic Tools for Decoding Glycosaminoglycan Polymer Fine Structure

Summary

This chapter focuses on development of recombinant bacterial enzymatic tools for decoding the structure-function relationships of glycosaminoglycan (GAG) enzymes. An integrated biochemical-analytical approach was adopted to characterize the structure-function relationships (including saccharide substrate specificity, enzyme active site, mode and mechanism of action) for the following GAG enzymes:

(a.) Chondroitinase ABC-II (cABCII) from Proteus vulgaris;

The development and characterization of this contaminant-free cABCII emerges as a tool to decode biologically significant structure-function relationships of DS and CSGAGs. The results also offer new therapeutic strategies for tissue repair and recovery after injury.

(b.) Heparin/Heiparan sulfate 6-O-sulfatase from Flavobacterium heparinum;

The findings from this study provide a framework that enables the use of 6-O-sulfatase as a tool for HSGAG structure-activity studies and expands our biochemical and structural understanding of this important class of oxygen-sulfur bond cleaving enzymes.

(c.) Heparin/Heparan sulfate N-sulfamidase from Flavobacterium heparinum

The development and characterization of this recombinant, contaminant-free N-sulfamidase enzyme potentiates in vitro reconstruction of the complete and defined exolytic sequence for the heparin and heparan sulfate GAG degradation pathway. This "HSGAG enzyme toolkit" is particularly powerful for application in tandem toward the exo-sequencing of HSGAG-derived oligosaccharides.

The developed enzymes have been integrated into the GAG databases of the Consortium for Functional Glycomics (CFG) for data mining and polysaccharide sequencing applications.

These studies resulted in 3 peer reviewed publications in (i.) Journal of Biological Chemistry (October 2008); (ii.) Journal of Biological Chemistry (September 2009); and (iii.) Journal of Biological Chemistry (September 2009).
3.1 Decoding the structure-function relationships of GAG enzymes

There has been substantial progress in two key areas to address the challenge of elucidating GAG structure-function relationships [43-46]. The first area pertains to the development of sensitive analytical methodologies using mass spectrometry (MS), HPLC, capillary electrophoresis (CE) and NMR techniques. These methodologies have been able to detect and characterize even femtomolar to picomolar amounts of GAG samples. Generally, only this much amount can be isolated and purified from biological sources such as cells and tissues.

The second area pertains to the development of bacterial enzymes as tools to depolymerize GAGs at specific linkages [43-46]. The development of enzymatic tools involves recombinant expression and purification followed by detailed characterization of substrate activity that in turn enables rational engineering of these enzymes with desired specificity. Much of the past efforts in these directions have focused on investigating the structure-activity relationships of bacterial GAG-depolymerizing lyases starting with the heparanases (a family of GAG lyases from *Pedobacter heparinus* that depolymerize heparin and heparan sulfate GAGs or HSGAGs) [47-50]. After successful development of heparinases as tools to characterize HSGAGs, investigations began on another family of lyases - the chondroitinases [43-46].

As described earlier, chondroitinases are classified broadly into three sub-families – chondroitinase-AC (cAC), chondroitinase-B (cB) and chondroitinase-ABC (cABC). cAC depolymerizes chondroitin-4-sulfate (C4S) and chondroitin-6-sulfate (C6S) whereas cB depolymerizes dermatan sulfate as its sole substrate. cABC has the broadest substrate specificity as it depolymerizes both CS and DS substrates. cABC has been employed in attempts to promote functional locomotor recovery following trauma to the central nervous system in a mouse model. The application of cABC I at the site of brain injury is believed to prune CS chains from proteoglycans localized to the glial scar [206]. The absence of these CS chains, inhibitory to axon regeneration, facilitates neural outgrowth and reconstruction of damaged brain tissue.

Recently, starting with a structural framework, our group had synthesized the recombinant Chondroitinase ABC I (cABC I) from *Proteus Vulgaris*, identified its active site, and investigated its structure-function relationship biochemically [267, 268]. Chondroitinase ABC II (cABC II), also from *P.vulgaris* degrades CS, DS, and hyaluronan substrates via an eliminative cleavage reaction resulting in oligosaccharides with D4, 5-unsaturated uronic acid residues at the non-reducing end. An approach similar to the method adopted for characterizing
cABCI is required to elucidate the structure-function relationship of the cABCII enzyme, and this work is described in the first part of this chapter.

Upon successful characterization of the chondroitinases, the sulfatase enzymes that are downstream to lyases (such as heparinases and chondroitinases) were isolated and characterized by our group and others [155, 255, 260-263]. These sulfatases selectively de-sulfate polysaccharide substrates. Recently, our group reported the molecular cloning and recombinant expression of 2-O-sulfatase as a soluble, highly active enzyme from Pedobacter heparinus and concomitant structural analysis was also performed. At the protein level, the bacterial 2-O-sulfatase was found to possess considerable sequence homology to other members of a large sulfatase family, especially within its amino terminus, where a highly conserved sulfatase domain was located. Within this domain, the critical active site cysteine that is predicted to be chemically modified as a formylglycine (in vivo) was identified by sequence homology.

Further, two other proteins were identified by our group that were found to exhibit a kinetic preference for highly sulfated glucosamines within each disaccharide unit, especially those possessing 6-O-sulfate and N-sulfate groups. Upon further investigation two new sulfatases were discovered, namely, the 6-O Sulfatase and the N-Sulfamidase enzymes. Determination of the structure-function relationships by integrated biochemical-structural methods is reported for these two new enzymes in the second and third parts of this chapter.

### 3.2 Structure-function relationship of Chondroitinase ABC-II from Proteus Vulgaris

#### 3.2.1 Brief overview of motivation, methodology, and principal findings

Chondroitin lyases (or chondroitinases) are a family of enzymes that depolymerize chondroitin sulfate (CS) and dermatan sulfate (DS) galactosaminoglycans, which have gained prominence as important players in central nervous system biology [206]. Two distinct chondroitinase ABC enzymes, cABCI and cABCII, were identified in Proteus vulgaris. Recently, cABCI was cloned, recombinantly expressed, and extensively characterized structurally and biochemically by our group [267, 268]. Building on our previous efforts, this study focuses on the recombinant expression, purification, biochemical characterization, and understanding the structure-function relationship of cABCII. Using an efficient system of
Escherichia coli-mediated expression and purification, recombinant cABCII was obtained, and the conditions for its optimal activity were examined.

A homology-based structural model of cABCII and its complexes with CS oligosaccharides was constructed. For this purpose, alignment of cABCI, cABCII, PhcAC and AacAC (Figure 3.1) was computed and revealed that cABCII has generally high sequence identity with cABCI, barring two prominent insertions of large loop regions (green) on both sides of the active site residues (blue, pink, red). The crystal structure of cABCI from P. vulgaris (Protein Data Bank code 1HNO) complexed to DSGAG were obtained from the protein data bank (PDB). In the case of cAC, two distinct structures were available -- one from Arthrobacter aurescens (AacAC; Protein Data Bank code 1RW9) complexed with CSGAG substrate and the other from Pedobacter heparinus (PhcAC; Protein Data Bank code 1CB8) complexed with DSGAG substrate. Both of these share the same structural fold with cABCI and were obtained from the PDB to serve as additional templates for the purpose of homology modeling.

This developed structural model of cABCII provided molecular insights and predictions into the product profile of the enzyme, that could then be readily compared with that of cABCI. The critical active site residues involved in the catalytic activity of cABCII identified based on the structural model were validated experimentally using site-directed mutagenesis and kinetic characterization of the mutants. Based on the developed homology model of cABCII, we predicted an exolytic mode of action for this enzyme, in contrast with the endolytic mode of action of cABCI. Similarly, the structural model of cABCII suggested a preference of smaller substrated owing to a more constrained active site from both sides of the groove. These predictions made from analyzing the homology model of cABCII were then verified experimentally. Specifically, the kinetic parameters and mode of action of the enzyme were characterized using CS and DS substrates. The profile of products formed by action of cABCII on different substrates was compared with product profile of cABCI, and the predicted smaller-substrate preference of cABCII was verified. It was also determined that cABCII was able to efficiently cleave a DS tetrasaccharide that was resistant to cleavage by cABCI, showing that cABCII has more a broad specific spectrum than cABCI. Similarly, the experimental results also validated the predicted exolytic mode of action of cABCII, by showing that cABCII only cleaves substrates at the nonreducing end of the polysaccharide.
The development of this contaminant-free, recombinant cABCII enzyme provides a new tool to decode the biologically important structure-function relationship of CS and DS GalAGs and offers novel therapeutic strategies for recovery after central nervous system injury.

Figure 3.1 Alignment of the chondroitinase enzymes, highlighting the two prominent N-terminal and C-terminal additional residues in cABCII (green) as well as the active site residues (blue, pink, red).

3.2.2 Published manuscript


Specific contributions of this Thesis research to the publication: Developed the homology-based structural model for cABCII and predicted its structure-function relationship in terms of substrate specificity, active site residue functions, mode and mechanism of action, which were all validated subsequently by targeted experiments.
Recombinant Expression, Purification, and Biochemical Characterization of Chondroitinase ABC II from Proteus vulgaris^S

Received for publication, August 26, 2008, and in revised form, September 29, 2008 Published, JBC Papers in Press, October 10, 2008, DOI 10.1074/jbc.M806630200

Vikas Prabhakar^1, Ishan Capila, Venkataramanan Soundararajan, Rahul Raman, and Ram Sasisekharan^2

From the Department of Biological Engineering, Harvard-MIT Division of Health Sciences & Technology, and the Koch Institute for Integrative Cancer Research Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Chondroitin lyases (or chondroitinases) are a family of enzymes that depolymerize chondroitin sulfate (CS) and dermatan sulfate (DS) galactosaminoglycans, which have gained prominence as important players in central nervous system biology. Two distinct chondroitinase ABC enzymes, cABCI and cABCII, were identified in Proteus vulgaris. Recently, cABCI was cloned, recombinantly expressed, and extensively characterized structurally and biochemically. This study focuses on recombinant expression, purification, biochemical characterization, and understanding the structure-function relationship of cABCI. The biochemical parameters for optimal activity and kinetic parameters associated with processing of various CS and DS substrates were determined. The profile of products formed by action of cABCI on different substrates was compared with product profile of cABCI. A homology-based structural model of cABCII and its complexes with CS oligosaccharides was constructed. This structural model provided molecular insights into the experimentally observed differences in the product profile of cABCII as compared with that of cABCI. The critical active site residues involved in the catalytic activity of cABCI identified based on the structural model were validated using site-directed mutagenesis and kinetic characterization of the mutants. The development of such a contaminant-free cABCII enables additional tools to decode the biologically important structure-function relationship of CS and DS galactosaminoglycans and offers novel therapeutic strategies for recovery after central nervous system injury.

Chondroitin sulfate (CS)^3 and dermatan sulfate (DS)^3 belong to a family of glycosaminoglycans known as galactosaminoglycans (GalAGs). GalAGs are linear polysaccharides of 1→4-linked repeating disaccharide units. The disaccharide units consist of a uronic acid (α-L-iduronic acid; IdoA) or β-D-glu- uronic acid; GlcA) linked 1→3 to a β-D-N-acetyl-galactosamine (GalNAc). Each disaccharide unit can additionally possess variations in the form of sulfation at the 2-0 and 3-0 positions of the uronic acid and 4-O and 6-O positions of the GalNAc (1). GalA-glucuronidase chondroitinases have been classified broadly into three subfamilies. Chondroitinase AC depolymerizes chondroitin-4-sulfate (C4S) and chondroitin-6-sulfate (C6S), whereas chondroitinase B depolymerizes dermatan sulfate as its sole substrate (2-9). Chondroitinase ABC has the broadest substrate specificity in that it depolymerizes both CS and DS substrates (10-12). Chondroitinases have been employed in attempts to promote functional locomotor recovery following trauma to the central nervous system (13-15). The application of cABCI at the site of central nervous system injury is believed to prune CS chains from proteoglycans localized to the glial scar. The absence of these CS chains, inhibitory to axon regeneration, facilitates neural outgrowth and reconstruction of damaged tissue. However, the use of chondroitinases as therapeutics is limited because of the lack of availability of pure and contaminant-free enzyme. Further, chondroitinase enzymes are often difficult to handle, because of thermal instability and spontaneous proteolysis, as reported by various groups (12, 16, 17).

Chondroitinase AC (cAC) and chondroitinase B from Pedobacter heparinus have been characterized extensively in terms of their enzymatic activity and substrate specificity. The crystal structure and co-crystal structure of chondroitinase B with its DS substrate together with site-directed mutagenesis of its putative active site residues provided detailed insights into its substrate processing and also revealed a calcium-dependent catalytic activity (3, 6, 8). The co-crystal structures of cAC with different CS and DS oligosaccharide substrate complexes led to the proposal of multiple scenarios in which the active site residues contributed to the catalytic activity of the enzyme (7). The crystal structure of another cAC from Arthrobacter aurescens and its co-crystal structure with CS substrates provided molecular insights into the active site of this enzyme and also its exolytic mode of action compared with the endolytic mode of cAC from P. heparinus (4).

Two distinct broad substrate specificity GalAG-degrading chondroitinase ABC lyases, cABCI and cABCII, have been identified in Proteus vulgaris (12). In fact, the ability of the conventional enzyme known as "chondroitinase ABC" to catalyze...
the complete depolymerization of GaIAG substrates to disaccharides is actually the result of the joint action of cABCI and cABCII. Recently, cABC from *P. vulgaris* was cloned, recombinantly expressed, and characterized biochemically in terms of its active site and the role of divalent cations in processing CS and DS substrates (10, 18).

Building on our previous efforts, the present study describes the cloning, recombinant expression, and biochemical characterization of cABCI and cABCII from *P. vulgaris*. Using an efficient system of *Escherichia coli*-mediated expression and purification, recombinant cABCII was obtained, and the conditions for its optimal activity were examined. The kinetic parameters and mode of action of the enzyme were characterized using CS and DS substrates. In contrast to cABC, which is predominantly an endolytic enzyme, cABCII is an exolytic enzyme that cleaves the substrate only at the nonreducing end of the polysaccharide. Furthermore, cABCII was able to efficiently cleave a DS tetrasaccharide that was resistant to cleavage by cABC (10).

The structural basis for the role of the active site residues in enzymatic activity and exolytic processing of cABCII was further elucidated using homology-based structural models of the enzyme-substrate complexes constructed using the cABC I and cAC crystal structures (4, 7, 11) as templates. The establishment of a contaminant-free recombinant cABCII and detailed characterization of its structure-function relationship enables using this enzyme in a variety of biotechnological applications that include sequencing biologically important GaIAG motifs and augmenting therapeutic strategies to reverse central nervous system deficit.

**EXPERIMENTAL PROCEDURES**

**Isolating Chondroitinase ABC II from *P. vulgaris*—Genomic DNA was isolated from cultures of *P. vulgaris* (ATCC 6896) using a DNeasy purification kit (Qiagen). The primers were designed based on the available sequence of the gene for both the full-length and mature versions (19). Forward primers were designed so as to incorporate an Ndel restriction site; the reverse primer was designed to incorporate BamHI and XhoI restriction sites. This allowed cloning into a pET-28a vector (Novagen). The primers for cloning cABCII had the sequences: 5'-CATATGCTAATAAAAAACCTTATAGCC-3' (forward primer for the full-length gene), 5'-CATATGTTACCCCACTCTGTCCTCAGAACGC-3' (forward primer for the truncated gene-excluding signal sequence), and 5'-GGATCCTCGAGTTACTAATACTAATAAAGATAGCC-3' (reverse primer). It should be noted that for the truncated gene an additional methionine was introduced into the primer sequence to allow for translation of the protein product. This occurs as an increment in the numbering of the residues by one for the final protein product thus produced. PCR was run using genomic DNA as template with an extension time of 3 min. The PCR product was ligated into the pET-28a vector using the TOPO TA cloning kit (Invitrogen) and transformed into TOP10 *E. coli* cells. Plasmid DNA was isolated, and the cABCII gene was excised by exploiting the Ndel and XhoI restriction sites. The excised gene was ligated into similarly digested pET28a. These ligation products were transformed into DH5α *E. coli* cells. Plasmid DNA isolated from the colonies was screened by restriction digestion for incorporation of the cABCII gene. Sequencing was also undertaken to confirm incorporation of the gene. *E. coli* cells (BL21(D3)) were transformed with plasmid DNA for expression.

**Protein Preparation**—Recombinant cABCII was expressed in *E. coli* using an adapted version of a previous approach (5, 18). The pET28a expression system contains an inducible T7 promoter, as well as an N-terminal six-histidine tag for facile purification. Cultures of *Luria-Bertani* broth containing kanamycin were inoculated, induced with 1 mM isopropyl-β-D-thiogalactopyranoside in mid-log phase (A600 = 0.8), and incubated at room temperature overnight. Centrifugation was used to harvest the cells, and the supernatant was discarded. The cell pellet was kept on ice and resuspended in 50 mM Tris, 250 mM NaCl, 10 mM imidazole, pH 7.9 (binding buffer), and then lysed by sonication. Soluble protein was collected by centrifugation at 15,000 × g for 15 min at 4°C. The soluble lysate was sequentially filtered through a 0.8-μm membrane and then a 0.45-μm membrane. A 5-ml Hi-Trap Metal Chelate column (GE Healthcare) was prepared by charging with 200 mM NiSO₄ and treatment with binding buffer. The protein was loaded onto the column, washed with a buffer containing 100 mM Tris, 250 mM NaCl, and 50 mM imidazole, and eluted into a similar buffer with increased imidazole (250 mM). The six-histidine tag was removed using a thrombin capture kit (Novagen) as previously described (20). The presence and purity of the proteins was assessed by standard methods using SDS-polyacrylamide gel electrophoresis. Protein concentration was measured using the Bradford assay (Bio-Rad) with bovine serum albumin (Sigma) as a standard.

**Site-directed Mutagenesis**—A QuickChange site-directed mutagenesis kit (Stratagene) was used with plasmid DNA template to induce mutations in the cABCII clone. As previously described (18), plasmid denaturation and annealing of custom-crafted oligonucleotides primers were used to introduce mutations. The primers were designed as follows (all primers are read in the 5'→3' orientation): the 5' 123T primer had the sequence GAA GGT GAA AAA AAT AAT GAA (Novagen). The primers for cloning cABCII had the sequences: 5'-CATATGCTGACTTCTGAC TTTCATGAA GC-3' (forward primer for the full-length gene), 5'-GGATCCTCGAGTTACTTAAATTAAACGAGTCG-3' (reverse primer). It should be noted that for the truncated gene an additional methionine was introduced into the primer sequence to allow for translation of the protein product. This occurs as an increment in the numbering of the residues by one for the final protein product thus produced. PCR was run using genomic DNA as template with an extension time of 3 min. The PCR product was ligated into the pCR 4-TOPO vector using the TOPO TA cloning kit (Invitrogen) and transformed into TOP10 *E. coli* cells. Plasmid DNA was isolated, and the cABCII gene was excised by exploiting the Ndel and XhoI restriction sites. The excised gene was ligated into similarly digested pET28a. These ligation products were transformed into DH5α *E. coli* cells. Plasmid DNA isolated from the colonies was screened by restriction digestion for incorporation of the cABCII gene. Sequencing was also undertaken to confirm incorporation of the gene. *E. coli* cells (BL21(D3)) were transformed with plasmid DNA for expression.
Chondroitinase ABC II from P. vulgaris

Qiagen miniprep kit. Each clone was sequenced to confirm the presence of the desired mutation. Plasmid DNA was used to transform BL21 (DE3) E. coli. In addition to the mutants described above, other mutations in residues believed to be important for enzyme activity were also made. The primer sequences for each of the mutants are listed below. The H44A mutant primers have the sequences 5'-CGA GGA AGT GGT TAT CAA ATT ATT ACT CCT GCT GTT GGT TAC CAA ACC-3' and 5'-GTT TTT GTA ACC AAC GC CAT GGT AA AAT TTG ATA ACC ACT TCC TCG-3'. The H453A mutant primers have the sequences 5'-CT GAT GGT TCT ATT TTG GCC CAT TCA CAA CAT AAC CCC GC-3' and 5'-GC GGG GTA ATGGTTTGAATG AGC GTG AAA AAT AGA ACCATCAG A-3'. The H457A mutant primers have the sequences 5'-CCT GAT GGT GCT CAT TCA CAA CAT TAC CCC GC-3' and 5'-GC GGG GTA ATGGTTTGAATG AGC GTG AAA AAT AGA ACCATCAG A-3'. The H546A mutant primers have the sequences 5'-TTT CCC TAC CCC GCT ACC AAC AAG ATA ACG GCT-3' and 5'-GTT TTG GTA ACC AAC AGC AGT AAT AAT TTG ATA ACC ACT TCC TCG-3'. The H563A mutant primers have the sequences 5'-AAC GGA AGT GGT TAT CAA ATT ATT ACT CCT GCT GTT GGT TAC CAA ACC-3' and 5'-AGC ATA AGC GGG GTA ATGGTTTGAATG AGC GTG AAA AAT AGA ACCATCAG A-3'. The Y461A mutant primers have the sequences 5'-CGA GGA AGT GGT TAT CAA ATT ATT ACT CCT GCT GTT GGT TAC CAA ACC-3' and 5'-GTT TTT GTA ACC AAC AGC AGT AAT AAT TTG ATA ACC ACT TCC TCG-3'.

Composition Analysis of Products from cABCII Processing of CS and DS Substrates—To investigate the composition of the final products of cABCII digestion, capillary electrophoresis was performed as previously described (6). Substrates included C6S from shark cartilage (Sigma), DS from porcine intestinal mucosa (Sigma), and CS from shrengen notchord (Seikagaku). Exhauisive overnight digestions of substrate by cABCII were analyzed using a Hewlett Packard three-dimensional capillary electrophoresis instrument with an extended path length cell. A voltage of 30 kV was applied using reverse polarity. Oligosaccharides were injected into the capillary using hydrodynamic pressure. They were detected using an ultraviolet detector set to 232 nm.

Biochemical Characterization of Chondroitinase ABC II Activity—Substrates (C6S and DS) were dissolved at 1 mg/ml concentration in various buffers to determine the relative effects of pH, temperature, ionic strength, and sodium acetate concentration on enzyme activity, as previously described (18). Chondroitin from shark cartilage (Seikagaku), hyaluronan from human umbilical cord (Sigma), heparin (Celsius), heparan sulfate (Celsius), and keratan sulfate (Sigma) were also used in these studies. For activity experiments, 2 μl of enzyme was placed in 248 μl of 50 mM Tris/HCl, pH 8.0, and reacted with 1 mg/ml substrate at 37 °C (0.25 mg/ml for hyaluronan). Product formation was monitored as an increase in absorbance at 232 nm as a function of time in a SpectraMax 190 (Molecular Devices) 96-well quartz format. For kinetic assays, 1 μl of enzyme (0.2–1.0 μg/μl) was added to 249 μl of a solution containing a GalAG substrate. Substrate concentration ranged from 0.1 to 5 mg/ml. Product formation was monitored by measuring the absorbance at 232 nm every 2 s. Kinetic parameters were determined using the initial rate of product formation and calculated as previously described based on Michaelis-Menten and Hanes techniques (18).

Modeling the Theoretical cABCII-Substrate Structural Complex—The crystal structure of cABCII from P. vulgaris (Protein Data Bank code 1H0N) was used as a template to obtain the model of cABCII. Initial inspection of the sequence alignment between cABCII and cABCII revealed that cABCII had multiple insertions of large loop regions as compared with cABCII. In addition to cABCII, the crystal structures of distinct cACs (which share the same structural fold with cABCII) from P. heparinus (PhaAC; Protein Data Bank code 1C8B) and from A. aurescens (AacAC; Protein Data Bank code 1RW9) were also used to model the loop regions in cABCII that aligned with either of the cACs. The structural superimposition of cABCIII and these distinct cACs was obtained using combinatorial extension-Monte Carlo (CE-MC) multiple structural alignment tool (21) (supplemental Fig. S1). A homology-based structural model of cABCII was generated using the homology module of InsightII v2005 (Accelys, San Diego, CA). The deletions in the modeled structure were closed by constrained minimization upon holding most of the structure rigid, except for regions close to the deletion site. This was followed by 200 iterations of steepest descent and 500 iterations of conjugate gradient minimization without including charges. The loops and side chains of all of the residues were then allowed to move freely by performing 500 iterations of steepest descent minimization. The refined structure was then subjected to 500 iterations of steepest descent minimization without including charges and 500 iterations of conjugate gradient minimization including charges to obtain the final predicted model of the cABCII enzyme. The final model of cABCII was validated using Whatif web-based interfaces and the Ramachandran plot explorer (supplemental Fig. S2).

A CS Tetrasaccharide Substrate, (Glca-GalNac,4S)2, was docked into the putative active site of cABCII using the following approach. The SuperPose version 1.0 server (22) was used to superimpose the co-crystal structure of cAC-CS tetrasaccharide complex (Protein Data Bank code 1HMW) with the modeled cABCII structure. The CS tetrasaccharide in this cAC complex had a uronic acid with a Δ4,5 unsaturated linkage at the nonreducing end. The starting model of the CS tetrasaccharide in the cACBII active site was therefore derived from the coordinates of a CS hexasaccharide (Protein Data Bank code 1C4S), which was superimposed on the CS substrate in the cAC co-crystal structure. The enzyme-substrate complex was subject to minimization without charges with 400 steps of steepest descent and 600 steps of conjugate-gradient methods. This was followed by another 500 steps, each of steepest descent and conjugate-gradient methods with charges. To constrain the ring torsion angles to maintain the ring conformation of the tetrasaccharides during the process of energy minimization, a force constant of 7000 kcal/mol was utilized. To evaluate the exolytic versus endolytic propensity of cABCII in comparison
with that of cABCI, an octasaccharide C4S substrate, (GlcA-GalNAc,4S)4 was also docked in a similar fashion (described above) into putative active sites of cABCI and cABCII, respectively. The octasaccharide was generated from the coordinates of a C4S hexasaccharide (Protein Data Bank code 1C4S) by adding another, GlcA-GalNAc,4S, to the reducing end of the hexasaccharide.

The viewer, builder, and discover modules of InsightII v2005 (Accelrys) were used for the visualization, structure building, and energy minimization, respectively. The AMBER force field (Amber95) provided with the Discover module was used to assign the potentials for both the enzyme and substrate. The AMBER force field (Accelrys) were used for the visualization, structure building, and energy minimization, respectively. The AMBER force field-based simulations according to the specifications in the InsightII manual.

RESULTS

Cloning, Expression, and Purification of cABCII—Two versions of the cABCII gene were cloned from P. vulgaris DNA: a full-length version and a truncated version that corresponds to the mature form of the enzyme, that is, without its putative leader sequence. These cABCII transcripts are ~3 kb in length—rather large, but still at an appropriate size to tolerate amplification via standard polymerase chain reaction techniques. Following cloning into a TOPO vector, the PCR product was subcloned into the pET28a expression vector. This facilitates E. coli-mediated uptake of the transcript and, ultimately, expression of the protein. Chondroitinase ABC II was expressed in E. coli as described under "Experimental Procedures." Purification over a charged Ni2+ resin was possible because of the incorporation of an N-terminal His6 tag. was induced in the log phase by the addition of isopropyl-β-D-thiogalactopyranoside. Purification of cABCII generated in excess of 50 mg of protein/500 ml of culture. SDS/PAGE analysis (Fig. 1A) confirmed the presence of highly pure cABCII at ~100 kDa, consistent with previously reported masses of the enzyme (12). Expression of the full-length cABCII clone generated a protein largely present in the insoluble fraction. The yield of soluble enzyme was greatly improved by the engineered removal of the hydrophobic N-terminal signal sequence. We then turned our attention to the truncated clones, both the original sequence and the transcript that underwent our sequential mutagenesis resolution. The recombinant protein with the sequence discrepancies was unable to effectively process Ga1AG substrates (Fig. 1B). Expression of the full-length cABCII clone generated a protein product with reinvigorated functionality as demonstrated by the observed cleavage of chondroitin-6-sulfate (Fig. 1B).

Biochemical Conditions for Optimal Enzyme Activity—After establishing the active recombinant cABCII, the reaction conditions for optimal cleavage of Ga1AGs were investigated. These reaction parameters included temperature, pH, ionic strength, and buffer system. For C6S substrate, cABCII demonstrated maximal processing at 37 °C, with a greater than 50% decline in activity at 42 °C (Fig. 2A). Chondroitinase ABC II similarly acted on DS substrate maximally at 37 °C, with a 50% drop at 42 °C (Fig. 2A). Activity against both substrates fell dramatically in excess of 45 °C. For both C6S and DS substrates, 37 °C was chosen as the optimal temperature for biochemical experiments.

A Tris buffer system was chosen for biochemical experiments because it permitted greater activity relative to phosphate buffer (data not shown). The recombinant enzyme demonstrated maximal activity at pH 8.0 for C6S. For DS, maximal activity occurred in the range from pH 8.0 to 8.5 (Fig. 2B).
Chondroitinase ABC II from P. vulgaris

![Graph A](image)

![Graph B](image)

![Graph C](image)

![Graph D](image)


**TABLE 1**

Specific activity of recombinant chondroitinase ABC II on glycosaminoglycan substrates

<table>
<thead>
<tr>
<th>Substrate</th>
<th>Specific activity (mU/mg protein)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chondroitin-6-sulfate</td>
<td>39,000</td>
</tr>
<tr>
<td>Chondroitin-4-sulfate</td>
<td>18,000</td>
</tr>
<tr>
<td>Dermatan sulfate</td>
<td>17,000</td>
</tr>
<tr>
<td>Chondroitin</td>
<td>5,400</td>
</tr>
<tr>
<td>Chondroitin sulfate D</td>
<td>4,900</td>
</tr>
<tr>
<td>Chondroitin sulfate E</td>
<td>4,900</td>
</tr>
<tr>
<td>Hyaluronan</td>
<td>ND</td>
</tr>
<tr>
<td>Keratan sulfate</td>
<td>ND</td>
</tr>
</tbody>
</table>

**TABLE 2**

Kinetic analysis of chondroitinase ABC II with various substrates

<table>
<thead>
<tr>
<th>Substrate</th>
<th>( k_{\text{cat}} ) (( \mu \text{M}^{-1} \text{min}^{-1} ))</th>
<th>( k_{\text{cat}/K_{M}} ) (( \mu \text{M}^{-1} \text{min}^{-1} ))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chondroitin-6-sulfate</td>
<td>9.8 ± 2.1</td>
<td>3800 ± 1130</td>
</tr>
<tr>
<td>Dermatan sulfate</td>
<td>19.2 ± 2.6</td>
<td>1150 ± 26</td>
</tr>
<tr>
<td>Chondroitin-4-sulfate</td>
<td>16.1 ± 4.2</td>
<td>1000 ± 60</td>
</tr>
</tbody>
</table>

Chondroitinase ABC II processing of GaIAG substrates is greatly curtailed below pH 7.5 and above pH 8.8. For purposes of simplicity, pH 8.0 was taken as the optimal pH for biochemical studies for both C6S and DS in a Tris buffer system.

For cABCII, ionic strength proved to be an important determinant in the processing of GaIAG substrates (18). In the case of cABCII, ~100 mM NaCl reduced processing by 50% for both C6S and DS substrates (Fig. 2C). Processing of C6S and DS is virtually eliminated in excess of 250 mM NaCl. Furthermore, the addition of ~50 mM sodium acetate for C6S and 100 mM sodium acetate for DS activates cABCII-mediated processing (Fig. 2D). Depolymerization of GaIAG substrate by cABCII is nearly completely inhibited at 500 mM sodium acetate. Therefore, unlike cABCII, which required ~50 mM of NaCl for C6S processing (and 125 mM for DS) (18), for cABCII the presence of NaCl in the buffer actually decreases the activity.

**Galactosaminoglycan Processing by cABCII**—The specific activity of the recombinant cABCII was tested against a full panel of glycosaminoglycan substrates (Table 1). The results suggest that cABCII is most proficient in degrading C6S, C4S, and DS substrates. It was not possible to detect reaction progression for hyaluronan, possibly because of low cleavage rate and nonoptimal pH for this substrate. As expected, other glycosaminoglycan families, like heparin and heparan sulfate (which contain glucosamine instead of galactosamine), were not processed by cABCII. Comparison of cABCII activity with that of cABCII (18) shows that cABCII more efficiently processes CS (~10-fold) and DS (more than 7-fold) than cABCII. Kinetic parameters for recombinant cABCII acting on C6S, DS, and C4S are summarized in Table 2. These measurements were obtained from the initial reaction rates against each substrate using both Michaelis-Menten and Hanes analysis. The turnover numbers for cABCII were experimentally determined to be 1300, 1150, and 1000 min⁻¹ for C6S, DS, and C4S, respectively. The catalytic efficiency was highest against C6S substrate (132 μM⁻¹ min⁻¹) and comparable for DS and C4S (both 60 μM⁻¹ min⁻¹).

The depolymerization of GaIAG substrates by cABCII was further scrutinized with capillary electrophoresis. These studies allow for the characterization of the final products of ABCII digestion following a 20-h exhaustive reaction at 37°C and therefore represent an end point assay for cABCII activity. For all of the substrates examined, the product profile contains an overwhelming proportion of disaccharide products (Fig. 3). For C6S, the dominant product is a 6-O-sulfated disaccharide. For C4S, the major product is a 4-O-sulfated disaccharide. With DS, a mixture of disaccharide products includes two mono-sulfated species, the 4-O-sulfated DS disaccharide and the 6-O-sulfated DS disaccharide, and a doubly sulfated disaccharide, the 4-O- and 6-O-di-sulfated disaccharide. It should also be noted that the amount of disaccharide released in each of these end point assays is considerably less (as measured by the double bonds generated, i.e., 1.19±2 mM) than those released by cABCII processing of these substrates. Comparison of...
the ADi6S peak in electrophoretogram AUA-GalNAc6S, ADi4S6S, AUA-GalNAc4S6S. Impurities in commercial sub-

processing of the addition of more enzyme and longer digestion period, whereas cABCII of Cisthe processing of

FIGURE 3. Product profile analysis of recombinant chondroitinase ABC II. Product profiles for cABCI action on C6S (A), C4S (B), and DS (C). Shown in inset of C is the processing of DS by cABCI. Chondroitinase ABCII is unable to cleave DS tetrasaccharides fragments (indicated by Tetra in the peak label) despite the addition of more enzyme and longer digestion period, whereas cABCII processing of DS results only in disaccharides. ADi4S, AUA-GalNAc4S, ADi6S, AUA-GalNAc6S, ADi45S, AUA-GalNAc45S. Impurities in commercial substrate preparations result in the ADi4S peak in electrophoretogram (A) and the ADi6S peak in electrophoretogram (B).

Structural Investigation of the Active Site Residues and Mode of Action of cABCII—The crystal structures of cABCI, PhcAC, and AacAC were used as templates to obtain a homology-based structural model of cABCI and its complex with the substrate (Fig. 4). The resulting model therefore adopted a multi-domain structure (similar to that of cABCI and the cACs) comprising an N-terminal B-domain with a jellyroll fold, the catalytic a-helix domain (incomplete toroid (a/3)-fold), and a C-terminal anti-

parallel B-sheet domain. To contrast the active site and positioning of the substrate within cABCI, the obtained model was superimposed on the enzyme-substrate co-crystal structures of the cAC enzymes (see "Experimental Procedures"). This superimposition showed that the critical active site residues were conserved and aligned spatially (supplemental Fig. S1). Comparison of cABC structures with that of cACs showed that the substrate binding and the active site of the cACs formed a wider groove to accommodate a broad range of CS and DS substrates. The notable differences in the structures of cABCI and cABCII were in two specific loop regions Ile121-Thr124 and Gly314-Leu315 that were present in cABCI and not cABCII (supplemental Fig. S1). The presence of extra loop regions in

AacAC compared with that of PhcAC has been implicated to account for the predominant exolytic activity of AacAC as against the endolytic activity of PhcAC (4). To better understand the role of the loop regions in cABCI in substrate processing, an octasaccharide of C4S was docked into the putative active site of cABCI. Assuming that the GalNAc4S-GlcA cleavable linkage occupied subsites -1 and +1, respectively, the sugars on the nonreducing end of the cleavable linkage occupied -2, -3, and -4 positions, whereas the sugars on the reducing end occupied +2, +3, and +4 positions.

In the docked model of the cABCI-octasaccharide complex, the sugars at the nonreducing end -3 and -4 positions had unfavorable steric contacts with the Ile121-Thr124 loop region (Fig. 5). Also, it appeared that for a long oligosaccharide (octa and higher) to be accommodated into the active site of cABCI, the chain would have to bend out of the active site groove. The cABCI active site groove, on the other hand, is open at both reducing and nonreducing ends of the substrate and hence is readily able to accommodate longer oligosaccharide substrates in the active site. Together the above observations offer an explanation based on the structural model for the ability of cABCI to process smaller oligosaccharides such as tetrasaccharides more efficiently than cABCII. Further, these observations also support the notion that cABCI processes its substrates in a predominantly exolytic fashion starting from the nonreducing end.

To understand the interactions of cABCI with its substrates, a C4S tetrasaccharide was docked into the putative active site of cABCI (see "Experimental Procedures"). It has been shown previously that a tetrad of residues including Arg, His, Tyr, and Glu play a critical role in the lyase activity of cABCI and the cACs. The analogous tetrad of residues in cABCI (that struc-
Chondroitinase ABC II from P. vulgaris

residues His\(^{514}\) and His\(^{519}\) are proximal to the 4-sulfate group of the GlcNAc4S residue at the \(-1\) subsite.

Site-directed Mutagenesis and Kinetic Characterization of the Putative Active Site Residues—

Based on the above structural model of cABCII-CS tetrasaccharide complex, a putative catalytic tetrad, His\(^{455}\), Tyr\(^{460}\), Arg\(^{518}\), and Glu\(^{608}\), and the other residues in the active site that are positioned to interact with the substrate such as His\(^{545}\), His\(^{546}\), and His\(^{546}\) were identified. To probe the contribution of these residues to enzymatic activity, they were mutated to alanines, and the resulting enzyme products were assayed for activity on chondroitin-6-sulfate and dermatan sulfate substrates (Table 3). The results show that mutants H453A, Y460A, R513A, and E608A all show no detectable activity on either C6S or DS, consistent with their designation as the critical catalytic tetrad residue required for enzyme activity. H452A shows similar \(K_m\) values to the wild type enzyme but a greatly (100-fold) reduced catalytic efficiency. On the other hand, mutation of His\(^{546}\) to alanine has limited effect on overall catalytic efficiency (≈3-fold reduction), indicating that this residue likely does not play an important role in catalysis. Surprisingly, mutation of His\(^{545}\) to alanine also yields an enzyme that shows no detectable activity on C6S and DS, suggesting that His\(^{545}\) plays a critical role in the enzymatic activity in addition to the catalytic tetrad.

**DISCUSSION**

This report is the first to describe the expression and characterization of a stable, highly active, contaminant-free recombinant chondroitinase ABC II from P. vulgaris. The sequence anomalies in the original clone resulted in a catalytically inactive enzyme, although these anomalies were not in the critical active site residues (Fig. 1B). It is therefore likely for these sequence differences to affect the overall stability of the enzyme, which in turn results in an inactive enzyme. Fixing these anomalies resulted in a fully active recombinant enzyme that efficiently processed both CS and DS substrates. This recombinant cABCII was examined structurally and biochemically, including reaction conditions to maximize enzyme efficacy, the product profile following digestion of GalAG substrates, kinetic analysis, substrate specificity, mode of action analysis, and structural insights into the active site-substrate interactions. Comparison of substrate processing of cABCII with that of cABC I showed that cABC I does cleave GalAG substrates at superior rates to cABC II; however, demethylolation mediated by cABC II proceeds by a course distinct from cABC I. Chondroitinase ABC I seems to prefer longer chain substrates

![FIGURES: Structural rationale for exolytic action of cABCII versus endolytic action of cABC I. Shown in the figure is the Connolly surface rendering of cABC I (top in pink) and cABC II (bottom in gray) with a C4S octasaccharide (shown as a stick model) docked into their putative active sites (generated using PyMol). Shown on the left and right is the active site groove seen from the nonreducing and reducing end of the octasaccharide respectively. Note that the octasaccharide is readily accommodated in the active site of cABC I. On the other hand, the access for the octasaccharide appears to be constricted toward the active site of cABC II.](http://www.jbc.org/content/suppl/2008/03/17/M600683200.DC1.html)

**TABLE 3**

<table>
<thead>
<tr>
<th>Enzyme</th>
<th>Chondroitin 6-sulfate</th>
<th>Dermatan sulfate</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(K_m) (\mu \text{m}^{-1})</td>
<td>(K_m) (\mu \text{m}^{-1})</td>
</tr>
<tr>
<td>cABCII</td>
<td>9.8 ± 2.1</td>
<td>1500 ± 113</td>
</tr>
<tr>
<td>H453A</td>
<td>NDA</td>
<td>NDA</td>
</tr>
<tr>
<td>H452A</td>
<td>7.0 ± 1.5</td>
<td>101 ± 1.9</td>
</tr>
<tr>
<td>H513A</td>
<td>NDA</td>
<td>NDA</td>
</tr>
<tr>
<td>H608A</td>
<td>10.4 ± 2.4</td>
<td>493 ± 51</td>
</tr>
<tr>
<td>Y460A</td>
<td>NDA</td>
<td>NDA</td>
</tr>
<tr>
<td>E608A</td>
<td>NDA</td>
<td>NDA</td>
</tr>
</tbody>
</table>

Structurally aligned with cABC I and cAcs are Arg\(^{513}\), His\(^{545}\), Tyr\(^{460}\), and Glu\(^{608}\) (Fig. 6). Although Glu\(^{608}\) does appear to be directly involved in the contact with the sugar, it is positioned to make critical hydrogen bonding interactions with His\(^{455}\) and Arg\(^{518}\), thus positioning these residues for their activity (Fig. 6). In addition to these critical active site residues, there are several other residues that are proximal to the CS substrate and hence could play a key role in positioning the substrate for enzyme action. The residue His\(^{452}\) proximal to His\(^{453}\) is on the opposite side of Arg\(^{513}\) in the base of the groove and is positioned to interact with a carboxyl group of the GlcA in the cleavage site. Residues His\(^{545}\) and Tyr\(^{460}\) on the top side of the groove (opposite to the catalytic tetrad at the base) are positioned to interact with GlcA and GalNAc4S in the +1 and +2 subsites and hence could play a critical role in anchoring the substrate in the active site. The

![Supplemental Material can be found at http://www.jbc.org/content/suppl/2008/03/17/M600683200.DC1.html](http://www.jbc.org/content/suppl/2008/03/17/M600683200.DC1.html)
and cleaves in a predominantly endolytic fashion (10, 12). Chondroitinase ABC II, on the other hand, appears to cleave shorter oligosaccharide substrates more efficiently in a predominantly exolytic fashion, resulting only in disaccharide products. The presence of these distinct cABCs offers the bacteria the ability to rapidly process GalAG substrates where cABCI would cleave the naturally occurring long GalAG polysaccharides and cABCII would then act on the smaller oligosaccharide fragments from cABCI processing to generate disaccharides that can be readily utilized for bacterial metabolism. The proposed exolytic mechanism of cABCII and its ability to efficiently process smaller GalAG substrates is also supported by the structural model of the cABCII-C4S octasaccharide complex. In this model the two extra loop regions, Ile294-298-Thr299 and Gly299-Leu997, in cABCII appear to constrict the active site groove specifically on the non-reducing side of the substrate. This groove constriction restricts the access of internal cleavable linkages in a long GalAG chain and thus points to a predominant exolytic mode of action for cABCII. The structural basis for differences in the activity of cABCI and cABCII is similar to the framework proposed for endolytic action of PhcAC versus exolytic action of AacAC (4, 9).

A combination of the theoretical structural model of the enzyme-C4S tetrasaccharide substrate complex and site-directed mutagenesis enabled the identification of critical residues involved in the enzymatic activity. This combined analysis further suggests the likely role of these active site residues in catalytic action of the enzyme. A tetrad of residues including His653, Tyr656, Arg913, and Glu908 in cABCI was structurally conserved in cABCII and the cACs. The corresponding tetrad in cABCI has been shown to be critical for enzyme activity (10). The mutation of any of these four critical residues to Ala resulted in a reduction of the enzyme activity. This combined analysis further suggests the likely role of these active site residues in catalytic action of the enzyme.

FIGURE 6. Critical residues involved in catalytic action of cABCI. A shows stereo view of a C4S tetrasaccharide (carbon, cyan; oxygen, red; nitrogen, blue; sulfur, yellow) substrate docked into the active site groove of the structural model of cABCI. The groove is shown as a cartoon model generated using PyMol, and the residues that are positioned to interact with the substrate are also shown in the following colors: His, Arg, Lys (blue); Tyr (purple); and Asp, Glu (red). B shows a two-dimensional schematic of the chemical structure of the C4S tetrasaccharide and its interactions with the critical residues in the active site. The sugars are numbered -2, -1, +1, and +2 from nonreducing to reducing end of the C4S substrate where the cleavage occurs between GaINAc-GlcA in the -1 and +1 sites, respectively. The tetrad of residues His513, Tyr516, Arg519, and Glu608 are conserved in cABCI, PhcAC, and AacAC and have been shown to play a critical role in the catalytic activity of these enzymes.
Chondroitinase ABC II from P. vulgaris

tioned earlier, Glu604 is not directly involved in interactions with the substrate, but it is positioned to interact with both His493 and Arg513 via hydrogen bonding. Hence mutation of Glu604 to Ala disrupts these interactions and is therefore unfavorable for the optimal positioning of these critical residues for catalytic activity. Earlier studies (2, 10, 24) implicate distinct interactions of cABCI with CS and DS substrates. The same scenario holds good for cABCII because of the structurally conserved active site tetrad. Therefore, the proposed roles of the tetrad in catalytic activity of cABCII could potentially be interchanged to accommodate a broad range of CS and DS substrates.

In addition to the tetrad, it was surprising to note that the H343A mutant completely lost the catalytic activity toward both CS and DS substrates (Table 3). This residue is positioned on the top side of the active site groove opposite the tetrad, which is at the base of the groove (Fig. 6). In the structural model of cABCII described in this study, His493 is not as proximal as the tetrad to the substrate. However, the model developed in this study is based primarily on the uncomplexed cABCI enzyme. Therefore, it is possible that in the presence of the substrate the active site groove could become more “closed,” which would position His493 proximal to the CS atom of the GlcA where it could play a critical role in either neutralizing the carboxylate group or stabilizing the CS carbanion transition state.

This understanding of the structure and mechanism of action of cABCII extends our understanding of precisely how these lyases function and how various structural features contribute to the depolymerization process. The distinct substrate processing ability of cABCII enables its use as an additional valuable resource in technologies directed at determining the fine structural elements of biologically relevant GaIAGs. These enzymes may further be useful directly in strategies to interfere with GalAG function in vivo, for example, in neural regeneration therapies and other such biomedical applications.

REFERENCES


Supplemental Material can be found at http://www.pjbc.org/content/jbc/284/2/135.full.html
3.2.3 Understanding chondroitinase ABC II preference for smaller GAG substrates and exolytic mode of action by glycan-docking

Docking of 14mer C4S polysaccharide to the developed homology-based cABCII structural model was performed to further understand the enzyme’s preference for smaller GAG substrates and its exolytic mode of action. Upon successive minimizations of the enzyme-substrate complex, the molecular surface was plotted (gray) and two distinct collision regions (denoted CR-1 and CR-2 respectively) were identified on the cABCII structure that impinge on the long polysaccharide substrate (Figure 3.2). The CR-1 region was identified as residues 231-243 (reducing end) and the CR-2 region was identified as residues 910-924 (non-reducing end). These two collision regions together render a small-substrate, exolytic preference to cABCII. These docking results are in agreement with the experimental results presented in the publication preceding this section.

Figure 3.2 Docking of 14mer C4S (green) to cABCII (gray) highlights collision regions (red) that make cABCII exolytic with preference for smaller polysaccharide substrates.
Figure 3.3 Analysis of collision region contributions towards exolytic mode by deletion of CR-1 (top row; orange), CR-2 (middle row; blue), and CR-1+CR-2 (bottom row; pink)
The first mutant cABCII structure (Figure 3.3; orange) was generated by deletion of residues 224-241 from CR-1 upon which multiple rounds of energy minimizations were performed on the mutant enzyme-substrate complex. Although this mutant structure was observed to be sterically unhindered in the reducing end (where CR-1 originally existed), there was still remnant steric hindrance on the non-reducing end (proximal to CR-2). Therefore, it appears that CR-1 contributes lesser towards the exolytic mode of action and small-substrate preference of cABCII. The second mutant cABC-II structure (Figure 3.3; blue) was generated by deletion of residues 912-937 from CR-2 upon which multiple rounds of energy minimizations were performed on the mutant enzyme-substrate complex. This mutant structure is unhindered on the non-reducing end (where CR-2 originally existed) and does not impinge on residues 233-242 of CR-1. This implies that CR-2 is the major contributor to the experimentally observed exolytic mode and small-substrate preference of cABCII. A third mutant cABCII structure (Figure 3.3; pink) was generated by deletion of both residues 224-241 from CR-1 and residues 912-937 from CR-2, as a frame of reference. Energy minimizations were subsequently performed on the mutant enzyme-substrate complex. From an observation of this molecular surface, it is clear that there are no steric hindrances on either the reducing or the non-reducing end and this in turn implies that in the absence of CR-1 and CR-2, cABCII mutants may be endolytic similar to the experimentally observed endolytic mode of cABCI.

3.2.4 Additional verification of the cABCII homology model with the \((\varphi, \psi)\) plot

The homology-based structural model of cABCII and all mutants were verified with the \((\varphi, \psi)\) Plot. The \((\varphi, \psi)\) plots for cABCII structural model and cABCI crystal structures computed are shown herein (Figure 3.4). These \((\varphi, \psi)\) plots show that the conformations of most of the residues in both cABCI and cABCII are within the allowed regions. To check whether the conformation of the critical active site residues in cABCII are within the allowed region (Figure 3.4), the Ramachandran Plot explorer 1.0 software (http://boscoh.com/ramaplot/) was used to analyze the conformational maps (shown in bottom-left) of these residues (shown in bottom-right). This analysis shows that the conformation of the active site residues in the cABCII structural model fall in the permitted regions of the \((\varphi, \psi)\) plot, thus providing additional verification for the developed homology model of this enzyme.
Having thus deduced the structure-function relationship of the CSGAG and DSGAG cleaving chondroitinase ABC II enzyme from *Proteus vulgaris*, we next proceeded to clone, recombinantly express, and biochemically-structurally characterize the heparin and HSGAG processing 6-O-sulfatase and N-sulfamidase enzymes from *Flavobacterium heparinum*, as described in the next two sections of this chapter.
3.3 Structure-function relationship of 6-O-Sulfatase from Flavobacterium Heparinum

3.3.1 Brief overview of motivation, methodology, and principal findings

Sequential HSGAG degradation has been demonstrated in several microorganisms, which depend on these sulfated polysaccharides not only as a carbon source but often as a means of scavenging inorganic sulfate [218, 255, 266]. The Gram-negative soil bacterium Flavobacterium heparinum (a.k.a. Pedobacter heparinus) is an excellent example of this process. As such this organism has proven to be a rich biological source for the isolation and molecular cloning of several GAG degrading enzymes [21, 25]. Like the enzymes of the lysosomal pathway, many of the flavobacterial enzymes possess well defined substrate specificity.

Given the manifold biological applications of other HSGAG-degrading enzymes we had previously derived from F. heparinum, including the heparinas as well as the 4,5-glycuronidase and the 2-O-sulfatase [21, 25, 218, 255, 266], we reasoned that cloning and characterization of additional downstream sulfatases would enable the development of important tools for more specifically investigating HSGAG structure. This is especially relevant given extensive experimental evidence that points to the regulated expression of endolytic HSGAG desulfating enzymes (especially 6-O-sulfatase) and their secretion into the extracellular matrix as a mechanism of modulating GAG structures critical for protein interactions. Thus, the use of HSGAG degrading and desulfating enzymes as analytical tools is central to unlocking the structural basis of HSGAG function and their potential use in generating structure-specific, bioactive glycans for therapeutic applications [35]. To utilize these enzymes correctly and efficiently requires not only a detailed understanding of the biochemistry of these enzymes; it also requires their ample availability for in vitro use. Both of these criteria already have been satisfied in structure-function studies of several HSGAG-related degrading enzymes, such as heparinase I, heparinase II, a unique unsaturated glycuronidase, and the 2-O-sulfatase [21, 25, 218, 255, 266].

Recently, our group cloned and identified an additional sulfatases downstream to these enzymes, namely the 6-O-sulfatase. The enzyme was cloned from the gram-negative soil bacterium Flavobacterium heparinum, recombinantly expressed in Escherichia coli in a soluble, active form, and identified as a specific HSGAG sulfatase. So as to investigate the mechanism of action of the enzyme through biochemical and structural studies, a homology based structural model of -O-sulfatase was developed. For this purpose, ClustalW was used to perform multiple
sequence alignment of 6-O-sulfatase with bacterial and lysosomal enzymes from the highly conserved sulfatase family. The crystal structures of human arylsulfatase A (PDB code: 1AUUK), human arylsulfatase B (PDB code: 1FSU) and P. aeruginosa arylsulfatase B (PDB code: 1HDH) were obtained and structure-based multiple superposition was performed using CE-MC and SuperPose tools. Human arylsulfatase A was used as a template for homology-based modeling with Insight II molecular simulations package (Accelrys, San Diego, CA) and the preliminary coordinates of 6-O-sulfatase were obtained. The AMBER force field was used to assign the potentials for the structure. The homology module of Insight II was used to model the loop regions of the structure and discover module was used to close the deletions in the model, by constraining most of the structure while permitting the regions in the proximity of the deletion site to move freely during multiple iterations of steepest descent and conjugate gradient minimization. The loop regions and side-chains of all residues were then allowed to move freely while performing further iterations of steepest descent minimization and the refined structure was finally subjected to more iterations of steepest descent minimization without including charges and conjugate gradient minimization including charges to obtain the final predicted model of 6-O-Sulfatase. The structures of GlcNS, GlcNS,6S and GlcNS,3S,6S monosaccharides were then obtained into the builder module of Insight II by appropriate modifications to the coordinates of heparin structure (1HPN) from the PDB. Owing to these monosaccharides being the possible natural substrates for the 6-O-sulfatase enzyme, these were docked to the enzyme model for molecular analysis. In order to further verify the substrate specificity of the enzyme, the coordinates of chondroitin sulfate were obtained from its crystal structure (PDB code 1C4S) and imported into the builder module of Insight II to generate the structural models of the GalNAc,6S and GalNAc,6S,4S monosaccharides. The initial orientation of the different substrates in the groove of the enzyme active site was facilitated by the highly conserved position of the cleavable sulfate group relative to Oy1 of the geminal diol as observed from the crystal structures of the sulfatase enzyme family. The potentials for all structures were assigned using the AMBER force field modified to include carbohydrates with sulfate and sulfamate groups. Following this, the enzyme-substrate complexes were subjected to optimization with minimization. In order to constrain the ring torsion angles to maintain the ring conformation of the monosaccharides during the process of minimization, a force constant was utilized. The final homology-based model structural model of the 6-O-sulfatase enzyme-substrate complexes were this obtained for further analysis (Figure 3.5).
As detailed in the manuscript that follows, all predictions made from analysis of the homology-based molecular modeling studies regarding the substrate specificity of the *F. heparinum* 6-O-sulfatase enzyme, its mode of action, kinetic properties, active-site topology, substrate specificity, role of divalent metal ions, and key enzyme-substrate interactions, were verified by targeted biochemical and molecular biology experiments. Taken together, the structural and biochemical studies indicate that 6-O-sulfatase is a predominantly exolytic enzyme that specifically acts on N-sulfated or N-acetylated 6-O-sulfated glucosamines present at the non-reducing end of HSGAG oligosaccharide substrates. This requirement for the N-acetyl or N-sulfo groups on the glucosamine substrate can be explained through eliciting favorable interactions with key residues within the active site of the enzyme, using the homology model developed. These findings provide a framework that enables the use of 6-O-sulfatase as a tool for HSGAG structure-activity studies as well as expand the horizons of biochemical-structural understanding of this key class of oxygen-sulfur bond cleaving enzymes.
3.3.2 Published manuscript


*Specific contributions of this Thesis research to the publication: Developed the homology-based structural model for 6-O-sulfatase and predicted its structure-function relationship in terms of substrate specificity, active site residue functions, role of divalent metal ions, mode and mechanism of action, which were all validated subsequently by targeted experiments.*
Heparin/Heparan Sulfate 6-O-Sulfatase from Flavobacterium heparinum

INTEGRATED STRUCTURAL AND BIOCHEMICAL INVESTIGATION OF ENZYME ACTIVE SITE AND SUBSTRATE SPECIFICITY*

Received for publication, August 7, 2009. Published, JBC Papers in Press, September 2, 2009. DOI: 10.1074/jbc.M109.053801

James R. Myette1, Venkataramanan Soundararajan, Zachary Shriver, Rahul Raman, and Ram Sasisekharan2

From the Harvard-Massachusetts Institute of Technology Division of Health Sciences and Technology, Koch Institute for Integrative Cancer Research, and Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139

Heparin and heparan sulfate glycosaminoglycans (HSGAGs) comprise a chemically heterogeneous class of sulfated polysaccharides. The development of structure-activity relationships for this class of polysaccharides requires the identification and characterization of degrading enzymes with defined substrate specificity and enzymatic activity. Toward this end, we report here the molecular cloning and extensive structure-function analysis of a 6-O-sulfatase from the Gram-negative bacterium Flavobacterium heparinum. In addition, we report the recombinant expression of this enzyme in Escherichia coli in a soluble, active form and identify it as a specific HSGAG sulfatase. We further define the mechanism of action of the enzyme through biochemical and structural studies. Through the use of defined substrates, we investigate the kinetic properties of the enzyme. This analysis was supplemented by homology-based molecular modeling studies that sought to rationalize the substrate specificity of the enzyme and mode of action through an analysis of the active-site topology of the enzyme including identifying key enzyme-substrate interactions and assigning key amino acids within the active site of the enzyme. Taken together, our structural and biochemical studies indicate that 6-O-sulfatase is a predominantly exolytic enzyme that specifically acts on N-sulfated or N-acetylated 6-O-sulfated 6-O-sulfated glucosamines present at the non-reducing end of HSGAG oligosaccharide substrates. This requirement for the N-acetyl or N-sulfos groups on the glucosamine substrate can be explained through eliciting favorable interactions with key residues within the active site of the enzyme. These findings provide a framework that enables the use of 6-O-sulfatase as a tool for HSGAG structure-activity studies as well as expand our biochemical and structural understanding of this important class of enzymes.

Heparin sulfate glycosaminoglycans (HSGAGs)3 comprise an important polysaccharide constituent of many proteoglycans (see Ref. 1, for a review). These glycans are linear polymers based on the variably repeating disaccharide unit (uronic acid 1→4 glucosamine)n, where n represents a variably repeating number (typically 10–200). As present in nature, these sugars possess an extensive chemical heterogeneity that is largely attributed to the mosaic arrangement of O- and N-linked sulfates present at different positions along each sugar chain (2, 3). Additional structural variations include the presence of N-linked acetates at the glucosamine C2 position as well as epimerization of the uronic acid C5 carboxylate to form either β-D-glucuronic acid or α-L-iduronic acid. Fundamental to understanding HSGAG structure-activity relationships is the appreciation that polydispersity of the glycans fine structure is not random. Instead, it is the end product of a complex and concerted biosynthetic pathway involving numerous modifying enzymes, whose relative expression levels and specific activities are regulated in a cell- and tissue-specific fashion. This programmed diversity of HSGAG structure (4) ultimately plays out at a functional level, namely through the dynamic regulation of numerous biochemical signaling pathways (2) relating to such processes as cell growth and differentiation (5), cell death (6, 7), intercellular communication, adhesion and tissue morphogenesis (8). HSGAGs are present as structurally defined binding epitopes on the cell surface and hence also play an important role in microbial pathogenesis (9, 10).

In contrast to the complex enzymatic process by which these polysaccharides are made, it appears that their catabolism is considerably more straightforward, both in the scope of its purpose and the means by which it is carried out at the biochemical level. Sequential HSGAG degradation has been demonstrated in several microorganisms (11–13), which depend on these sulfated polysaccharides not only as a carbon source but often as a means of scavenging inorganic sulfate (14). The Gram-negative soil bacterium Flavobacterium heparinum (a.k.a. Pedobacter heparinum) is an excellent example of this process. As such this organism has proven to be a rich biological source for the isolation and molecular cloning of several GAG degrading enzymes (15, 16). Like the enzymes of the lysosomal pathway, many of the flavobacterial enzymes possess well defined substrate specificity. Given the utility of other HSGAG-degrading enzymes, whose relative expression levels and specific activities are defined on the cell surface and hence also play an important role in microbial pathogenesis (9, 10).

*This work was supported, in whole or in part, by National Institutes of Health Grant GM57073.
1 Present Address: Momenta Pharmaceuticals, 675 West Kendall St., Cambridge, MA 02142.
2 To whom correspondence should be addressed: 77 Massachusetts Ave., Rm. 16-561, Cambridge, MA 02139. Fax: 617-258-9409; E-mail: rams@mit.edu.
3 The abbreviations used are: HSGAG, heparin/heparan sulfate glycosaminoglycan; AT, ΔΔ3-unaturated uronic acid; GlcNAc, N-, 6-O-disulfated α-D-glucosamine; GlcN, N-, 6-O, 3-O-trisulfated α-D-glucosamine; GlcNAc, N-acetylated, 6-O-sulfated α-D-glucosamine; GalNAc, N-acetylated, 6-O-sulfated β-D-galactosamine; 4MU, 4-methylumbelliferone; CE, capillary electrophoresis; LC/MSD, liquid chromatography/mass selective detector; MES, 4-morpholinethanesulfonic acid; MOPS, 3-[N-morpholino]propanesulfonic acid; ESI, electrospray ionization; MS, mass spectrometry; ORF, open reading frame; PDB, Protein data bank.
enzymes derived from *F. heparinum*, including the heparinases as well as the ΔΔ4,5-glycuronidase (17) and the 2-O-sulfatase (18), we reasoned that cloning and characterization of additional sulfatases would enable the development of important tools for investigating HSGAG structure. This is especially relevant given extensive experimental evidence that points to the regulated expression of endolytic HSGAG desulfatizing enzymes (especially 6-O-sulfatase) and their secretion into the extracellular matrix as a mechanism of modulating (especially 6-O-sulfatase) and their secretion into the extracellular matrix as a mechanism of modulating (19, 20). Therefore, the use of HSGAG degrading enzymes as analytical tools is central to unlocking the structural basis of HSGAG function and their potential use in generating structure-specific, bioactive glycans for therapeutic applications (21). To utilize these enzymes correctly and efficiently requires not only a detailed understanding of the biochemistry of these enzymes; it also requires their ample availability for *in vitro* use. Both of these criteria already have been satisfied in structure-function studies of several HSGAG-related degrading enzymes, such as heparinase I (22), heparinase II (23), a unique unsaturated glycuronidase (17), and the 2-O-sulfatase (24). We have been able to clone and identify two additional sulfatases downstream to these enzymes, namely the 6-O-sulfatase and N-sulfamidase.

In this study, we report the molecular cloning and recombinant expression in *Escherichia coli* of the *F. heparinum* sulfatase, namely the glucosamine 6-O-sulfatase, and in the accompanying manuscript (39) we report detailed studies on the N-sulfamidase as well as a model for how the 2-O- and 6-O-sulfatases, and N-sulfamidase work in concert. In both studies, we also present detailed molecular structural modeling and biochemical analyses of the recombinant enzyme as it relates to substrate specificity, optimal reaction conditions, and the role of divalent metal ions in the action of the enzyme. Taken together with our previous studies regarding other members of the heparinase-sulfatase family of enzymes, we have now unraveled the complete framework for HSGAG degradation in *F. heparinum*.

**EXPERIMENTAL PROCEDURES**

**Reagents**—Fluorescent glucopyranoside substrates 4-methylumbelliferyl-α/β-D-glucopyranoside (4-MUG5-Glc and 4-MUG5-Glc) were purchased from EMD Biosciences, Inc. (San Diego, CA). 6-O-Sulfated fluorogenic glycopyranoside derivatives were obtained through Toronto Research Chemicals (Toronto, Canada). Glucosamine and galactosamine monosacharides, and arylsulfate substrates 4-catechol-sulfate and 4-methylumbelliferyl-sulfate were purchased from Sigma. Exo-glucosidases were purchased from MP Biomedicals (Irvine, CA). Materials for genomic library construction and screening were obtained from Stratagene (La Jolla, CA). Reagents for site-directed mutagenesis were also obtained from Stratagene. PCR enzymes, TOP10 chemically competent cells, and oligonucleotide primers were obtained from Invitrogen. Additional molecular cloning reagents were purchased from New England Biolabs (Beverly, MA) or the manufacturers listed.

**Molecular Cloning of Flavobacterial 6-O-Sulfatase**—The flavobacterial sulfatase gene was cloned by PCR from a λZAPII flavobacterial genomic library originally screened using DNA hybridization probes specific to the 2-O-sulfatase (18). Library construction, hybridization screening, and phage excision were as described. An overlapping clone was expanded by chromosomal walking and restriction mapping using the Lambda DASH II genomic cloning kit (Stratagene) for the ligation of size-fractionated genomic DNA (generated by partial Sau3AI digestion). 2-O-Sulfatase-positive clones from an amplified library were plaque purified through three successive rounds, and the DNA were purified from a high titer lysate using standard techniques. For DNA sequencing, recombinant phage DNA was subcloned into pBluescript SK-·. The coding sequence of the putative sulfatase gene (described throughout this paper as orfb) was identified by the canonical Pfam (25) sulfatase family identifier (CXPRXXXTXS/TG) and subsequently PCR amplified using the following primer set for orfb (6-O-sulfatase), 5'-GAA TTC ATA TCG GTA AAT TGA AAT TAA TTT TA-3' (forward) and 5'-GGA TCC TCG AGT ATG TAA GGA GCC TCA GTT GGA TCA GT-3' (reverse). The amplified gene was subcloned into the T7-based bacterial expression vector pET28a (Novagen) as an NdeI-Xhol cassette (restriction sites underlined). Cloning as such allowed the gene to be expressed as an NH2-terminal His6 fusion with an intervening thrombin cleavage site for facile removal of this tag following protein purification.

**Bacterial Expression and Protein Purification**—Recombinant protein expression in *E. coli* strain BL21(DE3) and one-step affinity purification by nickel chelation chromatography were as described for the 2-O-sulfatase (18). The prediction of the NH2-terminal signal sequence and putative cleavage site for the protein was made by the computational method of von Heijne (26). Engineering and expression of the truncated protein (minus signal sequence) was as described above for the full-length gene with the exception of substituting the 5' primer that was used in the original PCR amplification step. The internal primer included 5'-CTC AGA CAT ATG TCT GCC CAG CAG CCA A-3' for orfb with the NdeI site underlined. As such, the orfb gene sequence begins at Met-18. Removal of the His6 tag was achieved by site-specific protease cleavage using the thrombin cleavage capture kit (Novagen). Proteolysis conditions were generally as described for other recombinantly expressed flavobacterial heparin-degrading enzymes (17, 18). Following concentration of orfb by ultrafiltration, the cleaved protein was dialyzed against 4 liters of 50 mm Tris, pH 7.5, and 0.1 m NaCl, 4 °C, overnight using a 3-ml slide-a-lyzer cassette with a 10,000 molecular weight cutoff (Pierce).

Final protein concentration was determined colorimetrically using the Bradford assay (Bio-Rad) and confirmed by UV absorption spectroscopy using theoretical molar extinction coefficients (εmax) of 94,730 m-1 (61,572 Da) for the NH2-terminal truncated orfb (6-O-sulfatase). This value was calculated for the thrombin-cleaved protein lacking a His6 purification tag. The enzyme was stored at 4 °C at a concentration of ~10 mg/ml and full enzyme activity was retained for several months under these conditions.

**Site-directed Mutagenesis of Putative 6-O-Sulfatase Active Site Residues**—Based on homology modeling of the 6-O-sulfatase (described below), three residues were initially chosen for
were mutated to alanines. Site-directed mutagenesis was completed by thermal cycling using the QuikChange™ method (Stratagene) generally as recommended by the manufacturer. The same pET28 (His6) vector used for bacterial expression of the wild type 6-O-sulfatase was likewise used as the template for linear amplification. The following oligonucleotide primer pairs (forward and reverse) were used: 5'-CGT TTT TTG TAC CAA TGC AGT Tgc GGc GCC ATC CAG GGC TAC-3', 5'-GTA CTG GTG CAT GCC CTG GAT GCC CCG gcA ACT GCA TTT GTC GTT ACA AAA CAG-3' (C80A); 5'-C TTG AAA AAG GAC CAT GCt AAA CCC TTT CTG ATG ATT TAC-3', 5'-G TAA ATC ATC AGA AAG GTG TTA gcA TGG TCC CTT TTT TCA AGG-3' (D186A); 5'-CC ATT ATT GTC TAT ACT TCC GTG CAG GCC TTT TAT TTT GGT G3-5', 5'-CA CCC AAA TAA AAg gCt CTC TGA GcG GAA GTA TAG ACA ATA ATG G3-5'. Base pair changes are noted in lowercase, boldface type.

mutagenesis: Cys-80, Asp-186, and Asp-374. All three residues were mutated to alanines. Site-directed mutagenesis was completed by thermal cycling using the QuikChange™ method (Stratagene) generally as recommended by the manufacturer. The same pET28 (His6) vector used for bacterial expression of the wild type 6-O-sulfatase was likewise used as the template for linear amplification. The following oligonucleotide primer pairs (forward and reverse) were used: 5'-CGT TTT TTG TAC CAA TGC AGT Tgc GGc GCC ATC CAG GGC TAC-3', 5'-GTA CTG GTG CAT GCC CTG GAT GCC CCG gcA ACT GCA TTT GTC GTT ACA AAA CAG-3' (C80A); 5'-C TTG AAA AAG GAC CAT GCt AAA CCC TTT CTG ATG ATT TAC-3', 5'-G TAA ATC ATC AGA AAG GTG TTA gcA TGG TCC CTT TTT TCA AGG-3' (D186A); 5'-CC ATT ATT GTC TAT ACT TCC GTG CAG GCC TTT TAT TTT GGT G3-5', 5'-CA CCC AAA TAA AAg gCt CTC TGA GcG GAA GTA TAG ACA ATA ATG G3-5'. Base pair changes are noted in lowercase, boldface type. Parental DNA was restricted using DpnI. Mutated DNA was transformed into XL Blue chemically competent cells. Targeted mutations were confirmed by direct DNA sequencing of both strands.

 Arylsulfatase Assay—Arylsulfatase activity was measured independently using two chromogenic substrates, 4-catechol sulfate and 4-methylumbelliferyl sulfate, colorimetric activity was determined spectroscopically at 515 nm. The fluorimetric arylsulfatase assay using 4-methylumbelliferone was measured at this alkaline pH and dilution as described. The ability of the enzyme to desulfate unsaturated heparin and chondroitin disaccharides was assessed essentially as described for CE-based compositional analyses of enzymatically generated glycosaminoglycan di- and tetrasaccharides (29). For these studies, the following disaccharide substrates were tested: AUGCNa6s, AUGCNa6s, AUGCNa8s, AUGCNa10s, and AUGCNa12s. Reactions included 500 µM substrate, 10 µM enzyme, 50 mM sodium acetate, pH 6.5, and ±2 mM CaCl₂ in a 20-µl reaction volume.

Coupled Enzyme Assay for the Determination of Biochemical Reaction Conditions and Steady-state Kinetics—Indirect measurement of enzyme activity was also made using a fluorimetrically based plate assay in which the prerequisite desulfation of the appropriate glucopyranoside 1→4-methylbifurilose substrate by the 6-O-sulfatase was coupled to the glucosidase-mediated hydrolysis of the stereo-specific 1→4-glycosidic linkage between the pyranose ring and the adjoining fluorophore. Release of the free fluorophore (4-MU) was monitored spectrophotometrically as described above for the arylsulfatase assay using 4-MU sulfate. Hydrolysis of the substrate 4-MU-β-D-GlcNAc6s at the 6-OH position was coupled to β-glucosidase purified from sweet almonds (MP Biomedicals, catalogue number 195197). The efficacy of the coupled assay was contingent on the intrinsically poor ability that the glucosidase possesses for hydrolyzing the glycosidic bond when the adjoining glucosamine is modified by a sulfate. The presumption of the sulfatase activity being the rate-limiting step was established experimentally. Reaction conditions were optimized to satisfy three criteria: 1) linear readout of the fluorescent signal that was directly proportional to sulfatase activity; 2) quantitative release of 4-MU by glucosidase activity under the biochemical conditions examined; and 3) negligible fluorescent quenching of free chromophore.

The standard reaction conditions for the assay included 2 µM recombinant enzyme, 50 mM sodium acetate buffer, pH 5.5, and 5 mM CaCl₂ in a 20-µl reaction volume. The 4-MU-GlcNAc6s substrate concentration was varied from 0.1 to 2 mM. Each well of a microtiter plate (prechilled on ice) was treated with 2 µl of
Heparin/Heparan Sulfate 6-O-Sulfatase from F. heparinum

due to. The enzyme, followed by gentle vortexing of the plate and spin
down of the well contents for 1 min at 500 × g and 4 °C. The
assay was initiated by transferring the 96-well plate to a heating
block preequilibrated at 30 °C. The incubation of the 6-O-sulfat-
ase enzyme was carried out at 30 °C for 20 min, after which
the enzyme activity was inactivated by heat denaturation
(95 °C, 10 min). For the glucosidase enzyme, the microtiter plate
was once again chilled on ice and 40 units of β-glucosidase was
added to each well. These were followed by vortex of the plate for
mixing and spin down at 500 × g for 1 min at 4 °C. The contents
were then transferred to a heating block preequilibrated at
37 °C. Incubation proceeded for 60 min prior to being quenched with
200 μL of 0.5 M Na₂CO₃, pH 10.7. Reactions were transferred to a black
96-well, flat-bottom fluoroimmunoassay plate and fluorescence
measured as described above for the detection of free 4-MU. Flu-
orescent signal was adjusted to background (minus sulfatase con-
trol). For β-glucosidase, this background hydrolysis was somewhat
dependent on the initial 4-MU-GlcNAc6O concentration, but was
typically less than 10%. More conversion of product was extrapo-
lated from a standard curve generated from various concentra-
tions of 4-MU from 0 to 300 μM. Michaelis-Menten kinetics was
extrapolated from Vₗ versus substrate concentration plots fit by
non-linear regression to pseudo-first order kinetics and all
obtained data represent the mean of three experimental trials.

The coupled enzyme assay was also used for the assessment of
enzyme activity for select site-directed mutants. Enzyme
activity was measured kinetically using a single saturating con-
centration (2 mM) of fluorescently labeled substrate. Results are
reported as % activity relative to the wild type enzyme.

*Compositional Analyses of Sulfate-treated Heparin*—20 μg
of heparin was preincubulated with 10 μM 6-O-sulfatase for 8 h at
30 °C in a 20-μL reaction volume that included 25 mM sodium
acetate, pH 7.0, and 2 mM calcium acetate, pH 7.0. Following
this preincubcation, the enzyme was inactivated by heat dena-
turation at 95 °C for 10 min, and heparin was exhaustively
digested overnight at 37 °C by the addition of 2 μL of a concen-
trated enzyme mixture containing heparinase I and III. Subse-
quent CE-based compositional analyses of heparinase-derived
disaccharides were completed as described (17).

*Electrospray Ionization (ESI)-Mass Spectrometry of Sulfated
  Glucosamine Monosaccharides*—Electrospray ionization-mass
spectrometry was performed in the negative ion mode using an
Agilent 1100 Series VI. LC/ESI/MS/MS trap. For simplicity, the
samples were prepared by adding MeOH directly to the enzymatic
reaction mixtures without purification, and directly injected
into the source of the mass spectrometer using a syringe pump
at a rate of ~8 μL/min. The SPS function of the software (LC/MSD
Trap Software 4.1 Build 143, MSD Trap Control Version 5.0 Build 65)
was used to tune the instrument, with the target mass
set to the mass of the substrate, the sample stability set to
50%, and the drive level set to 100%. Data were acquired over
the scan range of m/z 100–2200 by accumulating 30,000 ions
per scan. Capillary voltage was set to 3000 V. Nitrogen was used
as the drying gas, whereas helium was used as the nebulizing
gas, with flow rates of 5 and 15 liters/min, respectively. In
each case a minimum of 10 spectra were averaged. Substrate spe-
cificity was determined using unlabeled glucosamide monosac-
charide substrates with various substitutions at the 2-N, 3-OH,
and 6-OH positions. Reactions were carried out with 2.5 mM
substrate, 5 mM CaCl₂, 1 μM enzyme, and 50 mM acetate buffer
at pH 5.5 and 37 °C. Reactions were quenched by diluting the
samples 1:4 in MeOH. The carrier solvent was H₂O:MeOH (1:4,
v/v). In the experiments, 4-MU-GlcNS was added prior to injection as the internal standard to monitor ionization and
mass accuracy in source and trap.

*Homology Modeling of 6-O-Sulfatase and Docking of
  Substrates—ClustalW (33)* was used to perform multiple
sequence alignment of 6-O-sulfatase with bacterial and lysoso-
mal enzymes from the highly conserved sulfatase family (fig. 1).
The crystal structures of human arylsulfatase A (PDB code
1AU), human arylsulfatase B (PDB code 1FSU), and P. aerugi-
nosa arylsulfatase B (PDB code 1HDH) were obtained and
structure-based multiple superposition was performed using
CE-MC (30) and SuperPose (31) tools. A homology-based
structural model of the 6-O-sulfatase was generated using the
Homology module of InsightII version 2005 (Accelrys, San
Diego, CA) using human arylsulfatase A as a template. The
deletions in the modeled structure were closed by constrained
minimization upon holding most of the structure rigid, except
for regions close to the deletion site during 300 iterations of
steepest descent and 500 iterations of conjugate gradient mini-
imization without including charges. The loops and side chains
of all residues were then allowed to move freely by performing
500 iterations of steepest descent minimization. The refined
structure was then subjected to 500 iterations of steepest
descent minimization without including charges and 500 itera-
tions of conjugate gradient minimization including charges to
obtain the final predicted model of the enzyme.

The structures of GlcNS, GlcNS,6S, and GlcNS,6S,6S monosac-
charides were obtained by appropriate modifications to the
coordinates of the heparin structure (PDB 1HPS). These mono-
saccharide substrates were docked to the enzyme. To further
verify the substrate specificity of the enzyme, the coordinates
of chondroitin sulfate were obtained from its crystal structure
(PDB code 1CAS) to generate the structural models of the mono-
saccharides GalNAc6S and GalNAc6S,6S. The initial orientation
of the different substrates in the groove of the enzyme active
site was facilitated by the highly conserved position of the
cleav-
able sulfate group relative to Ogly-1 of the geminal diol as
observed from the crystal structures of the sulfatase enzyme
family. The glycosidic and exocyclic torsion angles were then
adjusted manually upon fixing the 6-O-sulfate group to elimi-
nate unfavorable steric contacts with the amino acids of the
enzyme active site. Following this, the enzyme-substrate com-
plexes were subjected to optimization with 400 steps of steepest
descent followed by 500 steps of Newton-Raphson minimiza-
tion including charges. To constrain the ring torsion angles to
maintain the ring conformation of the monosaccharides during
the process of energy minimization, a force constant of 7000
kcal/mol was utilized. Most of the enzyme was held rigidly,
whereas regions constituting the active site were allowed to
move freely.

The Viewer, Builder, and Discover modules of InsightII ver-
sion 2005 (Accelrys) were used for visualization, structure
building, and energy minimization, respectively. The potentials
for all structures were assigned using the AMBER force field modified to include carbohydrates with sulfate and sulfamate groups. The AMBER force field (Amber95) provided with the Discover module was used to assign the potentials for both the enzyme and substrate. The parameters for sulfates and sulfamate groups in glycosaminoglycans described previously (32) were incorporated into this force field to assign the potentials the genetically engineered removal of the amino-terminal signal sequence of the protein. Exclusion of this domain had little deleterious effect on the specific activity of the enzyme (data not shown). At the same time, replacement of this NH$_2$-terminal peptide with a histidine (His$_6$) tag facilitated purification of the recombinant protein in essentially a single chromatographic step to greater than 80% purity (data not shown). Sub-

**RESULTS**

**Molecular Cloning and Recombinant Expression of F. heparinum Sulfatase Gene**—The sulfatase gene described in this study was first identified through the screening of a genomic library with hybridization probes directed toward the flavobacterial 2-O-sulfatase (18). An overlapping phagemid clone identified during this process was expanded by chromosomal walking and restriction mapping. Sequence analyses of this genomic region revealed a sizeable open reading frame of 1647 base pairs (described hereafter as orfb). The gene sequence putatively encodes a protein of 545 amino acids in length starting at the initiating Met. The sequence does not possess an obvious Shine-Delgarno ribosomal binding site within 10 nucleotides of the initiating ATG codon. A closer examination of the sequence at the protein level revealed several important features. The flavobacterial orfb possesses an NH$_2$-terminal hydrophobic signal peptide and the corresponding cleavage site sequence predicted by the von Heijne method for Gram-negative bacteria (26). The orfb gene product is characterized by a theoretical pl of 8.6 and also contains the canonical sulfatase domain as described by the Protein Family (PFAM) identifier PF000884 (see “Discussion”). The putative sulfatase gene was robustly expressed in E. coli as a soluble enzyme. Satisfactory expression of the soluble enzyme required

**FIGURE 1. Multiple sequence alignment of the sulfatases using ClustalW.** The flavobacterial 6-sulfatase enzyme is a member of a large sulfatase family. The putative active site is boxed, with the critically modified Cys-80 noted by an asterisk. Invariant residues are shaded in dark gray, those with partial identity in light gray, and conservative substitutions in charcoal. Multiple sequence alignment was generated by ClustalW (1) using only select sequences incited by a BLAST search of the protein data base. Enzymes with the following GenBank accession numbers are abbreviated as follows: F. heparinum 6-O-sulfatase (FH6S); F. heparinum 2-O-sulfatase (FH2S); Pseudomonas aeruginosa arylsulfatase (PAR; GenBank code CAAB8421); human N-acetylgalactosamine-6-sulfate sulfatase or chondrinin 6-sulfatase (iGal65; GenBank code AAC51350); human cerebroside-3-sulfate sulfatase or arylsulfatase A (HARSA; GenBank code AAC51350); human N-acetylgalactosamine-4-sulfate sulfatase or arylsulfatase B (HARSB; GenBank code AAC51350).
Heparin/Heparan Sulfate 6-O-Sulfatase from F. heparinum

sequent thrombin cleavage of the histidine tag was carried out as described under "Experimental Procedures." These ΔNH₂-terminal truncations (lacking both the native signal sequence and NH₂-His₆ tag) were used in all subsequent biochemical characterizations of the sulfatase. The apparent molecular mass of the recombinant protein based on SDS-PAGE was consistent with its theoretical molecular mass calculated from the amino acid composition (orfB gene product, 61,572 Da).

Biochemical Characterization of Recombinant HSGAG 6-O-Sulfatase: Preliminary Determination of Monosaccharide Substrate Specificity—As a first step in biochemical characterization of the orfB sulfatase, we examined whether this enzyme, similar to the previously characterized 2-O-sulfatase (18), can function as a generic arylsulfatase. Both enzyme activities were tested against 4-catechol sulfate and 4-MU sulfate, two different aromatic sulfate esters commonly used as substrates to make this assessment. The 2-O-sulfatase exhibited an appreciable level of hydrolytic activity relative to a known arylsulfatase from A. aerogenes, which served as a positive control (data not shown). The orfB sulfatase partly hydrolyzed the 4-MU sulfate at a much slower but discernible rate. To further test our prediction that the orfB sulfatase acts on carbohydrates, in particular, heparin and heparan sulfate, we used a modified substrate wherein the sulfated hexosamine was linked 1→4 (α or β) to 4-MU. The presence of this chromophore allowed us to directly monitor desulfation of the monosaccharide by capillary electrophoresis. Four monosulfated substrates were tested, all of which were commercially available. These included the two "heparin" monosaccharides 4-MU-GlcNS₆S and 4-MU-GlcN₆S in addition to the 6-O-sulfated galactose sugars 4-MU-Gal₆S and 4-MU-GalN₆S (corresponding to the monosaccharide constituents of keratan sulfate and chondroitin/dermatan sulfate, respectively). In this analysis, the orfB sulfatase was found to be specific for the 6-O-sulfated glucosamine (Fig. 2A) and did not act upon either of the two 6-O-sulfated galactose sugars. We further investigated the substrate specificity of this enzyme by examining the influence of various substitutions at the 2-amino, 3-OH, and 6-OH positions of the glucosamine. In these experiments, desulfation of non-derivatized monosaccharide substrates was detected and quantified by ESI-mass spectrometry. In this analysis, the enzyme acted specifically on the 6-O-sulfate position and required a substituted amine (acetate or sulfate) at the 2-amino position. A comparative kinetic analysis of the two corresponding substrates (Glc₆N₆S₆S versus Glc₆N₆S) indicated only a modest preference of the enzyme for the monosulfated substrate (Fig. 2B) based on initial kinetic profiles. The enzyme was also completely inhibited by the presence of a 3-O-sulfate group (data not shown). Based on these findings, we defined the orfB as a heparin/heparan sulfate 6-O-sulfatase. To investigate these experimental observations in greater detail, we proceeded to analyze the purified recombinant enzyme using a combination of more detailed biochemical and structural studies.

The CE-based assay was also used to determine the pH optimum for the enzyme. The 6-O-sulfatase enzyme exhibited slightly acidic pH optima (between 5.5 and 6.5) and showed higher activity in acetate buffer when compared with sulfonate.
buffers, such as MES and MOPS when examined over this same pH range.

Optimization of in Vitro Reaction Conditions—Having identified a suitable chromogenic substrate for the recombinant sulfatase, we also used this substrate to further develop a fluorescence-based plate assay as the means to define the optimal in vitro reaction conditions for 6-0-sulfatase. We investigated parameters such as ionic strength and the effect of divalent metal ions as well as steady-state enzyme kinetics. We chose a coupled enzyme assay in which the recombinant sulfatase served as the primary (product limiting) enzyme and β-glucosidase served as the secondary enzyme. Use of this second enzyme permitted the indirect detection of relative sulfatase activity by means of the stoichiometric release of free 4-MU, which served as the fluorescent signal. This coupled assay for the 6-0-sulfatase was validated in control experiments demonstrating only modest hydrolysis by the glucosidase of the \( \beta_{1-4} \) mucosidic linkage of the sulfated glucosamine. The 6-0-sulfatase was sensitive to increasing ionic strength as measured by the addition of NaCl and 50% inhibition was observed at \( 200 \text{ mM} \) NaCl with less than 20% activity remaining at 1 mM NaCl relative to the zero NaCl control. Moreover, 6-0-sulfatase was inhibited by the addition of sulfate or phosphate, and between these anions, phosphate was clearly a more effective inhibitor with 50% inhibition of 6-0-sulfatase activity observed at \( 2 \text{ mM} \) \( \text{PO}_4^2- \) compared with \( 20 \text{ mM} \) \( \text{SO}_4^2- \).

Structural Investigation of Substrate Binding and Enzyme Action of the 6-0-Sulfatase—Having defined the basic biocatalytic context of the 6-0-sulfatase, we sought to provide a structural context to our findings. A homology based structural model of 6-0-sulfatase was constructed. A representative disulfated monosaccharide GlcNS,6S, was docked into the putative active site of the enzyme to investigate the molecular interactions between the enzyme and the substrate (Fig. 3). The critical active site residues at the base of the sugar binding pocket involved in the positioning of the sulfate group and the sulfate-ester bond cleavage are structurally conserved among the different sulfatases (highlighted in bold in Table 1). The functional roles of these conserved residues in the sulfate-ester hydrolysis mechanism were assigned (Table 2) based on the analysis of the enzyme-substrate complex as well as on the assigned roles of their structurally conserved residues in the other sulfatases.

Apart from the conserved residues, the other residues in the proximity of the active site are Asp-186, Lys-187, Pro-188, Phe-189, Thr-300, Trp-307, Tyr-378, and His-507. These residues are not structurally conserved in the other sulfatases and therefore they potentially play a unique role in the substrate specificity of the 6-0-sulfatase. Indeed, examination of the crystal structures for other, non-\( \text{HSGAG} \) sulfatases, indicates that there is no equivalent for these residues (Table 1). Furthermore, many of these residues are present as a part of the additional loop regions in the 6-0-sulfatase as compared with the other sulfatases. The presence of these additional loop regions marks the substrate-binding groove of 6-0-sulfatase more constricted as compared with the other sulfatases. This narrow substrate binding groove taken together with these additional residues are likely to restrict the processing of GlcNS,6S, sugar present either as a monosaccharide or at the non-reducing end of a short oligosaccharide (exolytic processing of substrate).

It was observed that Phe-189 and Tyr-378 are positioned to stack the pyranose sugar ring and thus lock the substrate in the active site. Asp-186 is positioned to make favorable contacts with the amine group in both GlcNS,6S and GlcNAcNS,6S substrates. Due to its role in substrate binding, mutation of Asp-186 to alanine resulted in a modest decrease in 6-0-sulfatase activity consistent with the fact that multiple contacts are important for orientation of substrate and stabilization of its binding. His-507 on the other hand interacts with the equatorial 4-OH of the Glc pyranose sugar. The critical conserved residues are Asp-186, Lys-187, Pro-188, Phe-189, Thr-300, Trp-307, Tyr-378, and His-507. These residues are not structurally conserved in the other sulfatases and therefore they potentially play a unique role in the substrate specificity of the 6-0-sulfatase. Indeed, examination of the crystal structures for other, non-\( \text{HSGAG} \) sulfatases, indicates that there is no equivalent for these residues (Table 1). Furthermore, many of these residues are present as a part of the additional loop regions in the 6-0-sulfatase as compared with the other sulfatases. The presence of these additional loop regions marks the substrate-binding groove of 6-0-sulfatase more constricted as compared with the other sulfatases. This narrow substrate binding groove taken together with these additional residues are likely to restrict the processing of GlcNS,6S, sugar present either as a monosaccharide or at the non-reducing end of a short oligosaccharide (exolytic processing of substrate).

It was observed that Phe-189 and Tyr-378 are positioned to stack the pyranose sugar ring and thus lock the substrate in the active site. Asp-186 is positioned to make favorable contacts with the amine group in both GlcNS,6S and GlcNAcNS,6S substrates. Due to its role in substrate binding, mutation of Asp-186 to alanine resulted in a modest decrease in 6-0-sulfatase activity consistent with the fact that multiple contacts are important for orientation of substrate and stabilization of its binding. His-507 on the other hand interacts with the equatorial 4-OH of the Glc pyranose sugar. The critical conserved residues are Asp-186, Lys-187, Pro-188, Phe-189, Thr-300, Trp-307, Tyr-378, and His-507. These residues are not structurally conserved in the other sulfatases and therefore they potentially play a unique role in the substrate specificity of the 6-0-sulfatase. Indeed, examination of the crystal structures for other, non-\( \text{HSGAG} \) sulfatases, indicates that there is no equivalent for these residues (Table 1). Furthermore, many of these residues are present as a part of the additional loop regions in the 6-0-sulfatase as compared with the other sulfatases. The presence of these additional loop regions marks the substrate-binding groove of 6-0-sulfatase more constricted as compared with the other sulfatases. This narrow substrate binding groove taken together with these additional residues are likely to restrict the processing of GlcNS,6S, sugar present either as a monosaccharide or at the non-reducing end of a short oligosaccharide (exolytic processing of substrate).
TABLE 1
Structure-based comparison of sulfatase active site residues

<table>
<thead>
<tr>
<th>6-O-Sulfatase (F. heparinum)</th>
<th>2-O-Sulfatase (F. heparinum)</th>
<th>Arylsulfatase A (human)</th>
<th>Arylsulfatase B (human)</th>
<th>Arylsulfatase (P. aeruginosa)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cys-80</td>
<td>Cys-82</td>
<td>Cys-69</td>
<td>Cys-91</td>
<td>Cys-31</td>
</tr>
<tr>
<td>Arg-94</td>
<td>Arg-96</td>
<td>Arg-73</td>
<td>Arg-95</td>
<td>Arg-25</td>
</tr>
<tr>
<td>Lys-128</td>
<td>Lys-134</td>
<td>Lys-122</td>
<td>Lys-145</td>
<td>Lys-113</td>
</tr>
<tr>
<td>His-130</td>
<td>His-136</td>
<td>His-125</td>
<td>His-147</td>
<td>His-115</td>
</tr>
<tr>
<td>Lys-387</td>
<td>Lys-398</td>
<td>Lys-302</td>
<td>Lys-318</td>
<td>Lys-375</td>
</tr>
<tr>
<td>Glu-396</td>
<td>Glu-397</td>
<td>His-229</td>
<td>His-242</td>
<td>His-211</td>
</tr>
<tr>
<td>Arg-40</td>
<td>Arg-42</td>
<td>Asp-23</td>
<td>Asp-53</td>
<td>Asp-13</td>
</tr>
<tr>
<td>Asp-41</td>
<td>Glu-43</td>
<td>Asp-30</td>
<td>Asp-54</td>
<td>Asp-14</td>
</tr>
<tr>
<td>Asp-374</td>
<td>Asp-309</td>
<td>Asp-301</td>
<td>Asp-317</td>
<td></td>
</tr>
<tr>
<td>Glu-375</td>
<td>His-296</td>
<td>Asp-202</td>
<td>Asp-301</td>
<td></td>
</tr>
<tr>
<td>Asp-186*</td>
<td>Asp-187*</td>
<td>Asp-188</td>
<td>Asp-189</td>
<td></td>
</tr>
<tr>
<td>Lys-157*</td>
<td>Lys-158*</td>
<td>Lys-159</td>
<td>Lys-160</td>
<td></td>
</tr>
<tr>
<td>Thr-300*</td>
<td>Thr-301*</td>
<td>Thr-302</td>
<td>Thr-303</td>
<td></td>
</tr>
<tr>
<td>Thr-378*</td>
<td>Thr-379*</td>
<td>Thr-380</td>
<td>Thr-381</td>
<td></td>
</tr>
<tr>
<td>His-507*</td>
<td>His-508*</td>
<td>His-509</td>
<td>His-510</td>
<td></td>
</tr>
<tr>
<td>Trp-307*</td>
<td>Trp-308*</td>
<td>Trp-309</td>
<td>Trp-310</td>
<td></td>
</tr>
</tbody>
</table>

*Regions obtained using the SuperPose tool (25). Residues that are conserved are displayed in boldface and non-conserved residues are listed in normal font. Regions of deletion are marked with the minus sign. Amino acids of 6-O-sulfatase that could potentially be involved in substrate binding are denoted with the asterisk.

TABLE 2
Functional assignment of 6-O-sulfatase active site residues

<table>
<thead>
<tr>
<th>Amino acids</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sulfate ester hydrolysis FGly-80</td>
<td>Cys-80 in the active enzyme is modified to FGly and the FGly is further hydrated to form an aldehyde hydrate comprised of Oγ-1 and Oγ-2 atoms</td>
</tr>
<tr>
<td>FGly-80</td>
<td>Stabilizes the hydrated FGly by interaction with Oγ-1. His-130 is well positioned also for proton abstraction from Oγ-2 after the catalytic process for elimination of sulfate and regeneration of geminal diole</td>
</tr>
<tr>
<td>FGly-80</td>
<td>Positioned to interact with oxygen atoms of the 6-O-sulfate group to enhance electron density withdrawal from sulfate, hence contributing to the electrophilicity increase of the sulfur center. Also, Lys-387 is well positioned to protonate the oxygen atom of the leaving substrate</td>
</tr>
<tr>
<td>Lys-128, Lys-387, Glu-396</td>
<td>Positioned to coordinate with a divalent metal ion such as Ca²⁺. Asp-374 also could donate proton and enhance nucleophilicity of Oγ-1. Thr-300 being weakly acidic could also participate instead of Glu-375 in metal ion coordination</td>
</tr>
<tr>
<td>Lys-128, Lys-387, Glu-396</td>
<td>Positioned to stack with the pyrrole ring of the Glc sugar</td>
</tr>
<tr>
<td>Lys-507, Asp-186, Tyr-378</td>
<td>Positioned to interact with the acetyl group of the GlcNAc sugar</td>
</tr>
</tbody>
</table>

for the observed kinetic preference of the monosulfated GlcNAc6s substrate. Furthermore, the position of the 3-OH is such that a glucosamine sugar containing a 3-O-sulfate group (such as Glc3sNAc6s,6s) would have unfavorable steric hindrance with the Trp-307 residue.

TABLE 3
Assessment of relative activity of 6-O-sulfatase active site mutants

<table>
<thead>
<tr>
<th>Mutant</th>
<th>Classification</th>
<th>% Activity After 0.5 h</th>
<th>% Activity After 5 h</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild type</td>
<td>Control</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Asp-374→Ala</td>
<td>Active site (FGly)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Asp-186→Ala</td>
<td>Ca²⁺ coordination</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Asp-186→Ala</td>
<td>Interaction with 2⁺</td>
<td>53</td>
<td>92</td>
</tr>
</tbody>
</table>

*Activity was measured using the coupled enzyme assay under maximum (saturating substrate) conditions as described under "Experimental Procedures."

TABLE 4
Steady state kinetic parameters for 6-O-sulfatase using 4-MU monosaccharide substrates

<table>
<thead>
<tr>
<th>Addition</th>
<th>K_m (µM)</th>
<th>kcat/k_m (X10²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>2.5</td>
<td>146</td>
</tr>
<tr>
<td>5 mM Ca²⁺</td>
<td>3.3</td>
<td>217</td>
</tr>
<tr>
<td>5 mM Ca²⁺ + 1 mM EDTA</td>
<td>6.8</td>
<td>327</td>
</tr>
</tbody>
</table>

To understand the specificity of this enzyme toward α-D-glucosamine sugar in HSGAGs versus β-D-galactosamine sugar in chondroitin and dermatan sulfate, a GalβNAc6s sugar was docked into the active site (Fig. 4). The ring stacking constraints imposed by hydrophobic residues Phe-189 and Tyr-378 caused the pyranose ring of GalβNAc6s to coincide with that of GlcβNAc6s sugar. However, axial orientation of the 4-OH in GalβNAc6s (as compared with its equatorial orientation in GlcβNAc6s) causes the 6-O-sulfate group to move away from the active site. If the 6-O-sulfate group of GalβNAc6s has to coincide with that of GlcβNAc6s sugar, then the C6-C5 bond will be cis relative to the C4-O4 bond, which is not energetically favorable given the unfavorably close proximity of O6 and O4 atoms in this conformation. These observations provide structural insights into the specificity of the 6-O-sulfatase enzyme toward HSGAG oligosaccharide fragments comprising the GlcβNAc6s6s sugar in the reducing end.

Role of Divalent Metal Ions in Activity of 6-O-Sulfatase—Our next step was to investigate the interactions of the active site residues with a divalent metal ion such as Ca²⁺ that has been shown to play a critical role in sulfatase activity of many arylsulfatases (33, 35). In each of the arylsulfatases, the metal ion coordinates with the oxygen atoms of the sulfate group of the respective substrate. Additionally, a cluster of four highly conserved acidic amino acids have been observed to coordinate with this divalent metal ion. The corresponding metal ion coordinating amino acids in the 6-O-sulfatase are identified based on the structural model as Asp-40, Asp-41, Asp-374, and Glu-375 (Table 1). These four amino acids are well positioned to coordinate with the Ca²⁺ ion (Fig. 3). The other amino acid present near the vicinity of this tetrad is Thr-301, which is also highly conserved and could participate in metal ion coordination. Based on these observations, we predict that Ca²⁺ ions would substantially influence the activity of 6-O-sulfatase and that there would be a reduction in the activity in the absence of Ca²⁺ ions.

We set out to test this hypothesis by examining the effect of Ca²⁺ and other divalent metal ions on enzyme activity. 6-O-Sulfatase activity was indeed found to be enhanced 2–3-fold (Table 4) by the presence of calcium in a concentration-de-
Sulfatase. In an attempt to determine the inhibition in the presence of a calcium-specific chelator we measured the potential for enzyme inhibition in the presence of a calcium-specific chelator (EGTA) but found that it had no appreciable effect on the specific activity of 6-O-sulfatase. In an attempt to determine the mechanism by which calcium exerts its effect on the enzyme we followed up these metal ion experiments by measuring the effect of calcium on enzyme steady-state kinetics (Fig. 5B). Consistent with our previous results, the initial rate of enzyme activity was affected by calcium in a concentration-dependent fashion and was largely manifested as a $k_{\text{cat}}$ effect (Table 4). The likely role of Asp-374 in coordinating calcium as predicted by the model was confirmed by site-directed mutagenesis (Table 3). Thus, through a combination of structural analysis and biochemical experiments; we were able to determine a role for Ca$^{2+}$ ion coordination in the enzyme activity of 6-O-sulfatase.

Experimental Validation of Exolytic Mode of Action by 6-O-Sulfatase—The structural analysis performed on our model indicates that the 6-O-sulfatase should be predominantly exolytic, given the highly specific orientation of the active-site residues. To verify this hypothesis, we performed experiments to address the possibility of this enzyme acting exolytically/endolytically on heparin or heparan sulfate-derived oligosaccharides. Of note is the fact that in the natural degradation pathway, 6-O-sulfatase would likely act downstream of the heparinases and possibly the A4,5-glycuronidase (17). Thus, these HSGAG oligosaccharide substrates could possess either an uronic acid (even number of saccharide units) or a hexosamine (odd number oligosaccharide) at the non-reducing end. In the former case, the uronic acid would likely be unsaturated due to the preceding action of heparin lyases, which cleave the GAG chain through a $\beta$-eliminative catalytic mechanism. In the latter case, loss of the unsaturated hexuronic acid would result from the hydrolytic action of the glycuronidase. To address this important issue, the 6-O-sulfatase enzyme was initially tested against a panel of unsaturated heparan disaccharides such as $\Delta U_{2,3}H_{NAC,6S}$ and $\Delta U_{2,3}H_{S6S}$. For these experiments, standard reaction conditions were chosen as defined in the monosaccharide studies. None of the unsaturated disaccharides were desulfated by 6-O-sulfatase (data not shown). The inability of this enzyme to do so was confirmed in a related experiment in which all possible heparin disaccharides were first generated by pre-treating heparin with heparinases I and II prior to adding the sulfatases to the same reaction tube. The converse experiment was also conducted in which unfractionated heparin was preincubated with the 6-O-sulfatase for an extended period of time (8 h) followed by the addition of heparinases I and III. In this particular sequence, sulfatase pretreatment had no effect on the compositional profile of the heparinase-derived cleavage products (data not shown). These experiments categorize the 6-O-sulfatase as an obligatorily exolytic enzyme, confirming the hypothesis derived from observation of the active-site topology.

To further verify the mode of 6-O-sulfatase processing, we proceeded to examine the possibility of 6-O-sulfatase acting on the non-re-

![Figure 4](image-url)
Heparin/Heparan Sulfate 6-O-Sulfatase from F. heparinum

**A. Add 2-O sulfatase**

<table>
<thead>
<tr>
<th>Mass (m/z)</th>
<th>% Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>2800</td>
<td>0</td>
</tr>
<tr>
<td>3140</td>
<td>0</td>
</tr>
<tr>
<td>3480</td>
<td>2</td>
</tr>
<tr>
<td>3820</td>
<td>0</td>
</tr>
<tr>
<td>4160</td>
<td>100</td>
</tr>
</tbody>
</table>

**B. Add A4,5 glycuronidase**

<table>
<thead>
<tr>
<th>Mass (m/z)</th>
<th>% Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>4129.85</td>
<td>0</td>
</tr>
<tr>
<td>4128.85</td>
<td>16.6E+4</td>
</tr>
<tr>
<td>4128.63</td>
<td>11073.5</td>
</tr>
</tbody>
</table>

**C. Add 6-O sulfatase**

<table>
<thead>
<tr>
<th>Mass (m/z)</th>
<th>% Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>4286.0</td>
<td>100</td>
</tr>
<tr>
<td>4050.25</td>
<td>833.25</td>
</tr>
</tbody>
</table>

FIGURE 6. Sequential degradation of an HSGAG tetrasaccharide using recombinantly expressed flavobacterial enzymes. The ability of the 6-O-sulfatase to hydrolyze the non-reducing end of an oligosaccharide is demonstrated in the context of exo-sequencing the heparin-derived tetrasaccharides ΔH_HNS,6S12SHNS,6S (shown here) and structurally related ΔH_HNS,6S12SHNS,6S lacking a 2-O-sulfate at the internal iduronic acid position (data not shown). Sequential treatment of the HSGAG tetrasaccharide was physically assessed after each enzymatic step (A-C) by matrix-assisted laser desorption ionization-MS. Masses listed in each panel represent either peptide alone (~3216 Da) or oligosaccharide-peptide complex. The net mass of the oligosaccharide is listed in parentheses. A, subsequent addition of the 6-O-sulfatase; B, subsequent addition of the A4,5-glycuronidase; C, subsequent addition of the 2-O-sulfatase (note loss of a sulfate represented by a shift in net molecular mass from ~915 to ~835 Da).

In this study, we describe the cloning, biochemical characterization, molecular modeling, and structure-function analysis of a sulfatase gene from the F. heparinum genome. A BLASTP sequence homology search of the flavobacterial gene against the protein data base unambiguously identified the gene product as a member of a large sulfatase family. The 6-O-sulfatase protein sequence possesses the signature Pfam sulfatase motif C/SXPA/KXXX/S/G as well as the highly conserved sequence L/TG (at the +9 through +11 positions relative to this motif). The two flavobacterial sulfatases that we have now cloned (i.e., the 2-O-sulfatase and orfB gene product) share only a limited overall homology to one another. The flavobacterial sulfatase shows a strong sequence homology (greater than 50%) to the mucin-desulfating sulfatase encoded by the enteric bacterium Prevotella strain RS2 (MdSAs gene). In addition to mucin, this particular enzyme is specific for free N-acetylgalactosamine-6-O-sulfate (36). Our previous studies with 2-O-sulfatase and other enzymes from the same producing end of saturated, odd-numbered oligosaccharides. We addressed this using a combination of two structurally related sulfated trisaccharides, HNS,6S12SHNS,6S and HNS,6S12SHNS,6S. Each of these trisaccharides was generated from the corresponding heparin-derived tetrasaccharide ΔU_HNS,6S12SHNS,6S (Fig. 6, data not shown for ΔU_HNS,6S12SHNS,6S) by the tandem use of the 2-O-sulfatase (Fig. 6A) and the A4,5-glycuronidase (Fig. 6B) prior to the addition of 6-O-sulfatase. Desulfation was followed by matrix-assisted laser desorption ionization-MS. The results presented for exo-lytic desulfation of oligosaccharides at the non-reducing end are consistent with the substrate specificity data pertaining to desulfation of monosaccharide substrates. Taken together, our structural and biochemical results indicate that 6-O-sulfatase is a highly specific, exo-lytic enzyme that acts on the non-reducing end of GlcNS,6S or GlcNAc,6S of saturated, odd-numbered oligosaccharide substrates and that its activity is enhanced by the presence of divalent Ca²⁺ ions.

**DISCUSSION**

In this study, we describe the cloning, biochemical characterization, molecular modeling, and structure-function analysis of a sulfatase gene from the F. heparinum genome. A BLASTP sequence homology search of the flavobacterial gene against the protein data base unambiguously identified the gene product as a member of a large sulfatase family. The 6-O-sulfatase protein sequence possesses the signature Pfam sulfatase motif C/SXPA/KXXX/S/G as well as the highly conserved sequence L/TG (at the +9 through +11 positions relative to this motif). The two flavobacterial sulfatases that we have now cloned (i.e., the 2-O-sulfatase and orfB gene product) share only a limited overall homology to one another. The flavobacterial sulfatase shows a strong sequence homology (greater than 50%) to the mucin-desulfating sulfatase encoded by the enteric bacterium Prevotella strain RS2 (MdSAs gene). In addition to mucin, this particular enzyme is specific for free N-acetylgalactosamine-6-O-sulfate (36). Our previous studies with 2-O-sulfatase and other enzymes from the same...
system have confirmed the presence of a cysteine-specific active site (Cys-80) and it is at this conserved cysteine (and not serine) that the critical co- or post-translational oxidation to an α-C-α-formylglycine occurs (37, 38). This obligatory requirement for the covalently modified cysteine has also been demonstrated by site-directed mutagenesis; mutation of Cys-80 to Ala completely abolished 6-0-sulfatase activity (Table 4).

Beyond the predicted function as inferred from structural homology to other known sulfatases, we set out to empirically confirm its putative function, first, by examining the ability of the enzyme to act as a so-called "arylsulfatase," second, to act within the context of HSGAG degradation and finally, by performing an exhaustive atomic analysis of the enzyme structural model active-site topology. We also experimentally validated the role of key active site residues by mutating Cys-80 → Ala or Asp-374 → Ala (as described previously in Ref. 24). These mutants showed no activity (Table 4) in comparison to the wild type enzyme and hence validate the critical role of these amino acids in the catalytic activity of the enzyme.

Although the 6-0-sulfatase was found to be a poor "generic" arylsulfatase on the basis of a rather nonspecific but commonly used biochemical screen, our results obtained by more structurally directed monosaccharide substrates have unequivocally confirmed that we had indeed cloned the heparan N-acetylglucosamine-6-0-sulfatase. These results demonstrate the exclusivity of the recombinant enzyme in terms of the singular position of the sulfate that is hydrolyzed. Moreover, this work goes beyond this basic description and identifies putatively the important structural determinants of enzyme specificity. In particular, the results presented identify the critical spatial orientation of the C4 hydroxyl as an additional structural determinant of substrate specificity, thus making the two flavobacterial sulfatases uniquely heparin/heparan sulfate-specific in this physiological context. These results demonstrate the exclusively directed monosaccharide substrates have unequivocally defined important parameters for enzymatic activity, as a catalytic base to confer hydrolytic activity, even in the presence of a divalent metal ion (e.g. Asp-374). This putative structure-activity assignment to Asp-374 is supported by our mutagenesis studies.

Pivotal in the potential use of the 6-0-sulfatase enzyme for controlled desulfation of heparin/heparan sulfate oligosaccharides is the question of its endolytic versus exolytic potential. By definition, the former mode of action would predict their ability to hydrolyze internally located sulfates within either a disaccharide or oligosaccharide chain. Our combined structural modeling and biochemical characterization results argue against any endolytic potential for this enzyme. On the other hand, an exolytic mode of action would necessarily require this enzyme to sequentially follow 4,5-glycuronidase hydrolysis of terminal uronic acids if, in fact, it is to act on the non-reducing end of these saccharides. The data presented here confirm this prediction, i.e., by demonstrating the ability of the enzyme to hydrolyze the non-reducing end of heparin-derived oligosaccharides, albeit with certain structural constraints, on the non-reducing end of the saccharide, namely a requirement of direct access to a sulfated hexosamine that is not sterically hindered by the presence of an intervening uronate.

The molecular modeling of 6-0-sulfatase, together with the biochemical characterization described herein, provide an insightful understanding of important structure-function relationships. Taken together with our understanding of the substrate specificity of the other flavobacterial sulfatases (including the N-sulfamidase described in the accompanying paper (39)), this work also provides a practical framework toward the use of these enzymes and discrete analytical tools for elucidating the HSGAG fine structure and generating HSGAG-derived oligosaccharides of defined length and structure.

REFERENCES

Heparin/Heparan Sulfate 6-O-Sulfatase from F. heparinum


3.4 Structure-function relationship of N-Sulfamidase from Flavobacterium Heparinum

3.4.1 Brief overview of motivation, methodology, and principal findings

The N-sulfate group is characteristic and unique to heparin sulfate glycosaminoglycans (HSGAGs) [313]. As mentioned previously, HSGAG degradation follows an obligatory sequence of depolymerization steps involving multiple enzymes acting in tandem to cleave the HSGAG chain [312]. Most of these enzymes act exolytically, but may differ in the extent of their processivity, e.g. when comparing functionally similar enzymes between eukaryotic lysosomal and bacterial systems. As mentioned previously, prior work in our laboratory has led to the molecular cloning and biochemical characterization of several HSGAG-degrading enzymes derived from the gram-negative soil bacterium *Flavobacterium heparinum* (also called *Pedobacter heparinus*) [21, 49, 252, 254]. Although earlier work described the enzymology of the heparin lyases (or heparinases), more recent work has focused on downstream enzymes such as an unsaturated glucuronyl hydrolase (Δ4,5-glycuronidase), and the 2-O- and 6-O-sulfatases (Figure 3.6). By their very nature, these latter enzymes provide an attractive subject of study and application due to the following reasons: (1.) their respective unique substrate specificities, (2.) their relevance as tools to probe multiple biological processes, and (3.) a solid biochemical foundation of their structure-function relationships that include several protein structures and a detailed understanding of their common catalytic mechanism. Taken together, these factors make the enzymes a useful “tool kit” for the characterization of heparin and heparan sulfate GAGs.

**Figure 3.6 Function of HSGAG processing enzymes**
In the paper that follows this section, the cloning, recombinant expression in *E. coli*, substrate specificity, metal ion requirements, mode of action, sequence of processivity, optimal reaction conditions, kinetics, and an integrated biochemical-structural characterization of the N-sulfamidase enzyme mechanism from *F. heparinum* are described. The N-sulfamidase enzyme is unique because it hydrolyzes the sulfamate linkage (N–S) and not a sulfoester as is described for the other HSGAG sulfohydrolases (Figure 3.6). In fact, N-sulfamidase is one of only two known enzymes that are able to cleave the nitrogen-sulfur bond (Figure 3.7). So as to elucidate the mechanism of N-sulfamidase action, a homology model of the N-sulfamidase enzyme was developed as outlined in the enclosed manuscript. Based on structure-function analysis of the derived model, with methodologies very similar to that outlined for 6-O-sulfatase in the previous section, a possible mechanism for N–S bond cleavage by N-sulfamidase and the critical amino acids involved in catalytic activity and substrate specificity of the enzyme were predicted.

Taken together with the reported substrate specificities for the previously characterized *F. heparinum* upstream heparin lyases, and the downstream enzymes including the unsaturated glucuronyl hydrolase, the 2-O-sulfatase, and the 6-O-sulfatase (described in the preceding paper), we are now able to reconstruct in vitro, the complete and defined exolytic sequence for the heparin and heparan sulfate GAG degradation pathway of *F. heparinum*, facilitating the tandem application of this "enzyme tool-kit" towards sequencing of HSGAG oligosaccharides.

Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB)

Enzyme Nomenclature. Recommendations

**EC 3.10.1**

*Acting on sulfur-nitrogen bonds*

- **EC 3.10.1.1** *N*-sulfoglucosamine sulfohydrolase
- **EC 3.10.1.2** cyclamate sulfohydrolase

\[
\text{ROSO}_3^- + \text{H}_2\text{O} \rightarrow \text{ROH} + \text{HSO}_4^- \\
\text{RN(H)SO}_3^- + \text{H}_2\text{O} \rightarrow \text{RNH}_2 + \text{HSO}_4^-
\]

*Figure 3.7 Known Nitrogen-Sulfur bond cleaving enzymes and their chemical mechanism*
3.4.2 Published manuscript


Specific contributions of this Thesis research to the publication: Developed the homology-based structural model for N-sulfamidase and predicted its structure-function relationship in terms of substrate specificity, active site residue functions, role of divalent metal ions, mode and mechanism of action, which were all validated subsequently by targeted experiments. Also, proposed mechanisms of enzymatic nitrogen-sulfur bond cleavage; which at the time of this study, had previously not been discussed in the literature for any known enzyme.
Heparin/Heparan Sulfate N-Sulfamidase from Flavobacterium heparinum

STRUCTURAL AND BIOCHEMICAL INVESTIGATION OF CATALYTIC NITROGEN-SULFUR BOND CLEAVAGE

Received for publication, August 7, 2009. Published, JBC Papers in Press, September 2, 2009, DOI 10.1074/jbc.M109.053835

James R. Myette,1* Venkataramanan Soundararajan, Jonathan Behr, Zachary Shriver, Rahul Raman, and Ram Sasishekar

From the Harvard-Massachusetts Institute of Technology Division of Health Sciences and Technology, Koch Institute for Integrative Cancer Research, and Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

Sulfated polysaccharides such as heparin and heparan sulfate glycosaminoglycans (HSGAGs) are chemically and structurally heterogeneous biopolymers that function as key regulators of numerous biological functions. The elucidation of HSGAG fine structure is fundamental to understanding their functional diversity, and this is facilitated by the use of select degrading enzymes of defined substrate specificity. Our previous studies have reported the cloning, characterization, recombinant expression, and biochemical characterization of several HSGAG-degrading enzymes derived from Flavobacterium heparinum (Pedobacter heparinus). Although earlier work described the enzymology of the heparinases (14–17), more recent work has focused on downstream enzymes such as an unsaturated glucuronidase (Δ4,5-glycuronidase) and the 2-O- and 6-O-sulfatases (18–20). By their very nature, these latter enzymes provide an attractive subject of study and application given the following: 1) their respective substrate specificities, 2) their relevance as tools to probe multiple biological processes, and 3) a solid biochemical foundation of their structure-function relationships that include several common catalytic mechanisms (23). Taken together, these factors make the enzymes a useful "tool kit" for the structural characterization of heparin and heparan sulfate.

The *N*-sulfate group is characteristic and unique to heparin sulfate glycosaminoglycans (HSGAGs), which are commonly occurring polysaccharides predominant in proteoglycans (1). HSGAGs are linear polymers with variable repeating disaccharide units and diverse chemical heterogeneity due to the variable positions of O- and N-linked sulfates (2, 3). Other factors contributing to the structural diversity of HSGAGs include the presence of N-linked acetates and possible epimerization at the C-5 carbonylates. The structure-function relationship of this diversity plays out in the dynamic regulation by HSGAGs of various signaling pathways (4), including cell death (5, 6), intercellular communication, cell growth and differentiation (7), and adhesion and tissue morphogenesis (8). Microbial pathogenesis has also been shown to depend upon the HSGAGs that are present as structurally defined binding epitopes on cell surfaces and as part of the extracellular matrix (9, 10).

GAG degradation follows an obligatory sequence of polymerization steps involving multiple enzymes acting in tandem to cleave the HSGAG chain. Most of these enzymes act exotically (11, 12) but may differ in the extent of their processivity, e.g. when comparing functionally similar enzymes between eukaryotic lysosomal and bacterial systems (13). Previously cited work in our laboratory has led to the molecular cloning and biochemical characterization of several HSGAG-degrading enzymes derived from the Gram-negative soil bacterium Flavobacterium heparinum (Pedobacter heparinus). Although earlier work described the enzymology of the heparinases (14–17), more recent work has focused on downstream enzymes such as an unsaturated glucuronidase (Δ4,5-glycuronidase) and the 2-O- and 6-O-sulfatases (18–20). By their very nature, these latter enzymes provide an attractive subject of study and application given the following:

1) their respective substrate specificities,
2) their relevance as tools to probe multiple biological processes, and
3) a solid biochemical foundation of their structure-function relationships that include several protein structures (21, 22) and a detailed understanding of their common catalytic mechanism (23). Taken together, these factors make the enzymes a useful "tool kit" for the structural characterization of heparin and heparan sulfate.

Received, in whole or in part, by National Institutes of Health Grant GM57073.
1 Both authors contributed equally to this work.
2 Present address: Momenta Pharmaceuticals, 675 West Kendall St., Cambridge, MA 02139.
3 To whom correspondence should be addressed: 77 Massachusetts Ave., Rm. 16-561, Cambridge, MA 02139. Fax: 617-258-9409; E-mail: rams@mitedu.
4 The abbreviations used are: HSGAG, heparin/heparan sulfate glycosaminoglycan; 2,4,5 unsaturated uronic acid; A4,5-glycuronidase; GlcNAc, N-acetylated α-D-glucosamine; GalNAc, N-acetylated α-D-galactosamine; 6-O-sulfated; 4-MU, 4-methylumbelliferyl 4-amino-phenyl-β-N-acetylglucosaminide; L-α-formylglycine; MOPS, 3-(N-morpholino)propane sulfonic acid; MALDI, matrix-assisted laser desorption/ionization; AB, open reading frame; CE, capillary electrophoresis.

DOI 10.1074/jbc.M109.053835

159
Heparin/Heparan Sulfate Glycosaminoglycan from *F. heparinum*

Toward this end, we describe in the preceding paper (19) the cloning, recombinant expression, structural and biochemical characterization of enzymatic action, and substrate specificity of the *F. heparinum* 6-0-sulfatase enzyme. In this paper, we extend this work through the cloning of the *N*-sulfamidase gene from *F. heparinum* and the bacterial expression of the gene product. A biochemical and structural characterization of the enzyme is also described, including a determination of the substrate specificity and metal ion requirements for optimal enzyme activity. This sulfamidase enzyme is unique because it hydrolyzes the sulfamate linkage (18), hydrolyzes the sulfamate linkage and the critical amino acids involved in both a homology-based structural model of the *N*-sulfamidase enzyme. This sulfamidase enzyme is unique because it hydrolyzes the sulfamate linkage (18) and not a sulfester as is described for the other HSGAG sulfohydrolases. Using a homology-based structural model of the *N*-sulfamidase enzyme, we have outlined a possible mechanism for the *N*-S bond cleavage and the critical amino acids involved in both the catalytic activity and substrate specificity of the enzyme. Finally, taken together with the reported substrate specificities for the previously characterized *F. heparinum* 2-0-sulfatase, the 6-0-sulfatase (described in the preceding paper, see Ref. 19) and the unsaturated glucuronidase, we are now able to reconstruct *in vitro* the complete and defined exocytic sequence for the heparin and heparan sulfate degradation pathway of *F. heparinum* and apply these enzymes in tandem toward the exo-sequencing of HSGAG-derived oligosaccharides.

**EXPERIMENTAL PROCEDURES**

**Reagents**—Glucosamine and galactosamine monosaccharides and aryl sulfate substrates 4-catechol sulfate and 4-methylumbelliferyl sulfate were purchased from Sigma. Exo-glucosidases were purchased from MP Biomedical (Irvine, CA). *N*-Sulfated fluorogenic glycopyranoside derivatives were obtained through Toronto Research Chemicals (Toronto, Ontario, Canada). Fluorescent glucopyranoside substrates 4-methylumbelliferyl-α/β-d-glucopyranoside (4-MU-α-d-Glc and 4-MU-β-d-Glc) were purchased from EMD Biosciences, Inc. (San Diego). PCR enzymes, TOPO10 chemically competent cells, and oligonucleotide primers were obtained from Invitrogen. Materials for genomic library construction and screening and site-directed mutagenesis were obtained from Stratagene (La Jolla, CA). Additional reagents for molecular cloning were purchased from New England Biolabs (Beverly, MA) or the listed manufacturers.

**Molecular Cloning of Flavobacterial N-Sulfamidase**—PCR from a λZAPII flavobacterial genomic library originally screened using DNA hybridization probes specific to the 2-O-sulfatase (18) and subsequently used for the 6-O-sulfatase (19) were used to clone the sulfamidase gene. Library construction, hybridization screening, and phage excision were performed as outlined. Two overlapping clones were expanded by chromosomal walking and Southern restriction mapping using the Lambda DASH II genomic cloning kit (Stratagene). This was used for the ligation of size-fractionated genomic DNA generated by partial digestion of Sau3AI. The DNA was purified from a high titer lysate using standard techniques, and three successive rounds of plaque purification were performed on 2-O-sulfatase-positive clones from an amplified library. Recombinant phage DNA was subcloned into pBluescript SK+ for sequencing the DNA. The putative *N*-sulfamidase gene is described throughout this paper as orfc. The coding sequence of orfc was identified by the canonical Protein Families Database (24) sulfatase family identifier (CXPRXXXX(S/T)G) and subsequently PCR-amplified using the primer set given by 5'-TCT AGA CAT ATG AAA TTT AAC AAA TTT TCT TTC-3' (forward) and 5'-GGC TCT TCG AGT TAC TTT AAA TAA TTG TAA CTA GAA T-3' (reverse). Subcloning of the amplified gene into T7-based bacterial expression vector pET28a (Novagen) was performed as an Ndel-XhoI cassettes described for the 6-O-sulfatase (19).

**Bacterial Expression and Protein Purification**—Similar to the setup in the 2-O-sulfatase (18) and 6-O-sulfatase (19) expression systems that we had designed earlier, the *Escherichia coli* strain BL21(DE3) and one-step affinity purification by nickel chelation chromatography likewise were used. Von Hejne computational method was used to make the predictions of the NH2-terminal signal sequence and putative cleavage site for the protein (25). The engineering and expression of the truncated protein lacking the signal sequence were as described for the complete gene with the exceptions of substituting the internal 5′-primer with 5′-TCT AGA CAT ATG TCC TGC ACT TCG CCG GAA-3′ (Ndel site underlined) and beginning the orfc gene sequence at Ser-21. Site-specific protease cleavage using the thrombin cleavage capture kit (Novagen) was performed to remove the His8 tag with proteolysis conditions as described for other recombinantly expressed flavobacterial heparin-degrading enzymes. Ultrafiltration was used to concentrate the enzymes, and cleaved proteins were dialyzed against 4 liters of 50 mM Tris, pH 7.5, and 0.1 mM NaCl, 4°C overnight using a 3-mL Slide-a-Lyzer cassette with a 10,000 molecular weight cutoff (Pierce). The Bradford assay (Bio-Rad) was used to determine final protein concentrations and was confirmed by UV absorption spectroscopy with theoretical molar extinction coefficients (ε280) of 86,340 (53,193 Da) for the NH2-terminally truncated orfc (*N*-sulfamidase). This value was calculated for the thrombin-cleaved protein lacking the His8 purification tag. To retain full enzyme activity for months, the enzyme was stored at 4°C and 10 mg/ml concentration.

**Site-directed Mutagenesis of Putative N-Sulfamidase Active Site Residues**—Based on the homology modeling of the *N*-sulfamidase (described below), three residues were initially chosen for mutagenesis: Cys-80, Asp-40, and Lys-158. Site-directed mutagenesis was completed by thermal cycling using the QuickChange™ method (Stratagene, La Jolla, CA) as described in the accompanying paper (19) and using the following oligonucleotide primer pairs: 5′-AT ATC CTG ATG ATC ATG TCC GCT AAC CAA TCC TGG AAC CAC G-3′, and 5′-CTG TGG TTC CAG GAT TGG TTA CGG GAC ATG ATC ATG AGG ATA T-3′ (D90A); 5′-CTT TTC TGC AGT TCA CCT TCC gtg ACg CCG CCC GCA AGG GCT G-3′ and 5′-CA GCC CCT GCG GGT GCA GGG GAT GAA GGT CAA AAA G-3′ (C80A); 5′-G TTC AAA AGT TTT GGC GGA CTT TTA gca gat AAA AAA GAA GAT GGG CTT CCT-3′ and 5′-CAG GGA CCT TCT TTT TTA TCT gcg AAA AAT GGC CCA AAA CTT TTA AAC-3′ (K158A). Base pair changes are noted in lowercase boldface type. The same pET28 (His8) vector used for bacterial expression of the wild type *N*-sulfamidase-sulfatase.
were confirmed. Ase was likewise used as the template for linear amplification. Parental DNA was restricted using DpnI. Targeted mutations were confirmed by direct DNA sequencing of both strands.

**Arylsulfatase, N-Sulfamidase, and Coupled Enzyme Assays**

**Biochemical Protocols and Kinetic Studies**—The 4-catechol sulfate and 4-methylumbelliferol sulfate were two chromogenic substrates used to independently measure arylsulfatase activity. The catechol substrate assay was conducted as described (26). Fluorimetric arylsulfatase assay using 4-methylumbelliferol sulfate was also generally as described previously (27), and detection of fluorescent methylumbelliferone was measured using a SpectraMax microtiter plate reader (GE Healthcare) set at excitation and emission wavelengths of 360 and 440 nm, respectively. Fluorescence intensity was corrected against background (minus enzyme control). In both assays, 0.5 unit of arylsulfatase from *Aerobacter aerogenes* was the positive control.

For the pilot N-sulfamidase assay, initial assessment of substrate specificity and pH optima was made using a capillary electrophoresis-based assay for detection of desulfated products. The fluorescently labeled monosulfated glucosamine and galactopyranosides used as substrates to test enzyme activity included 4-MU-GlcNAc6S, 4-MU-GlcNS, 4-MU-GalNaS6S, and 4-MU-GalNS. Standard reactions included 1 mM substrate, 1–10 μM enzyme, 50 mM sodium acetate, pH 5.5–6.5, and 5 mM CaCl₂ in a 20-μl reaction volume. Exhaustive reactions involving overnight incubations at 30 °C were used for pilot experiments. N-Sulfamidase was inactivated by heat denaturation at 95 °C for 10 min, followed by a 10-fold dilution into water, and the products were resolved by capillary electrophoresis using a 25-cm long, 75-μm inner diameter fused silica capillary (Agilent Technologies). Substrate desulfation was measured as a percentage of substrate depletion relative to a minus enzyme control as monitored by the loss of UV absorbance at 315 nm detected at ~4 min. A standard capillary electrophoresis buffer included 50 mM Tris and 10 mM dextran sulfate (average molecular mass of 10,000 Da) adjusted to pH 2.0 with phosphoric acid. The effect of pH was measured by capillary electrophoresis using the following three sets of buffers with overlapping pH ranging from 4.5 to 8.0: 50 mM sodium citrate at 4.5, 5.0, and 5.5; 50 mM MES at 5.5, 6.0, 6.5, and 7.0; and 50 mM MOPS at 6.5, 7.0, 7.5, and 8.0. For N-sulfamidase, the reactions included 4-MU-GlcNS, 50 mM buffer, and 5 mM CaCl₂ in a 20-μl reaction volume, and this assay was initiated by addition of 2 μl of 10× enzyme stock to 18 μl of preheated reaction mixture. All the reactions were performed at 30 °C for 30 min each. The ability of N-sulfamidase to desulfate unsaturated heparin and chondroitin disaccharides was assessed essentially as described for CE-based compositional analyses of enzymatically generated glycosaminoglycan di- and tetrasaccharides (28).

A fluorimetrically based plate assay was used for indirect measurement of enzyme activity, in which the N-sulfamidase-driven desulfation of the glucopyranoside 1→4 methylumbelliferylone substrate was coupled to the glucosidase-controlled hydrolysis of the stereospecific 1→4 glycosidic linkage between the pyranose ring and the adjoining fluorophore. Release of the 4-MU was monitored spectroscopically as described above for the arylsulfatase assay. N-Sulfamidase hydrolysis of 4-MU-α-D-GlcNS at the 2-amino position was coupled to α-glucosidase (catalogue number 153487, MP Biomedicals), and the efficacy of this coupled assay was contingent on the intrinsically poor ability of the enzyme for hydrolyzing the glycosidic bond when the adjoining glucosamine is modified by a sulfate.

The coupled N-sulfamidase assay was generally described for the 6-O-sulfatase (19) but with the following modifications: 4-MU-GlcNS as substrate, 50 mM sodium acetate at pH 6.0 (instead of 5.5), and 1 μM enzyme. For the second enzyme step, 5 units of α-glucosidase were added. Enzyme incubation was carried out for 22 h at 37 °C. The obvious difference in enzyme efficiencies between α-glucosidase versus β-glucosidase is reflected in the substantially longer incubation times required for the α-glucosidase to quantitatively hydrolyze the glycosidic α1→4 linkage between the fluorophore and the desulfated glucosamine. All other reaction conditions were as described for the coupled 6-O-sulfatase/β-glucosidase assay (19). Michaelis-Menten kinetics were extrapolated from Vₜ versus substrate concentration plots fit by nonlinear regression to pseudo first-order kinetics, and all obtained data represent mean of three experimental trials.

The coupled enzyme assay was also used for the assessment of enzyme activity for select site-directed mutants. Enzyme activity was measured kinetically using a single saturating concentration (2 mM) of fluorescently labeled substrate. Results are reported as % activity relative to the wild type enzyme.

**Compositional Analyses of Sulfamidase-treated Heparin**—10 μM N-sulfamidase was preincubated with 20 μg of heparin for 8 h at 30 °C in a 20-μl reaction, which included 25 mM sodium acetate, pH 7.0, and 2 mM calcium acetate, pH 7.0. The enzyme was then inactivated by heat denaturation at 95 °C for 10 min, and the heparin was completely digested at 37 °C by the addition of 2 μl of concentrated enzyme mixture containing heparinase I and III. The CE-based compositional analyses of heparinase-derived disaccharides were performed as described (29).

**Sequential Degradation of Heparin Oligosaccharide by Flavobacterial Exo-enzymes**—The purified penta- and hexasulfated tetrasaccharides δU(2H)₅N₅S₅S = 2H₅N₅S₅S₅S₅ were a gift from Dr. I. Capila (Momenta Pharmaceuticals, Inc.). The following enzyme sequence was performed: 2-O-sulfatase, δ4,5-glycuronidase, 6-O-sulfatase, and N-sulfamidase. The enzyme was heat-inactivated after each step, and 20-μl aliquots were removed prior to the addition of the next enzyme. The initial reaction conditions included 20 mM Tris, pH 7.2, and 60 nmol of tetrasccharide in a 120-μl reaction volume, and enzyme reactions were carried out at 30 °C. The following specific conditions were also used: 1) 2-O-sulfatase, 1 μM enzyme, 6 h; 2) δ4,5-glycuronidase, 1 μM enzyme, 6 h; 3) 6-O-sulfatase, 5 μM enzyme, 5 mM CaCl₂, 12–15 h; 4) N-sulfamidase, same conditions as for 6-O-sulfatase. MALDI-mass spectrometry is an established method that was used for determination of the molecular masses of enzyme products (30).

The APTS derivatization protocol was adapted from Chen and Evangelista (31). Briefly, 2 μl of 100 mM APTS in 25% acetic acid (v/v) was mixed with 10 μl of 1 M sodium cyanoborohydride in tetrahydrofuran and 1 μmol of saccharide. The reaction mixture was incubated at 75 °C for 2 h and was diluted 1:100 prior to CE analysis. Capillary electrophoresis/laser-
Heparin/Heparan Sulfate Glycosaminoglycan from F. heparinum

FIGURE 1. Structure-based multiple sequence alignment of the sulfatases using ClustaW. The flavobacterial N-sulfamidase enzyme is a member of a large sulfatase family. The putative active site is boxed, with the critically modified cysteine 80 noted by an asterisk. Invariant residues are shaded in dark gray; those with partial identity in light gray, and conservative substitutions in charcoal. Multiple sequence alignment was generated by ClustalW using only select sequences identified from a BLASTP search of the protein data bank. Enzymes with the following GenBank accession numbers are abbreviated as follows: F. heparinum N-sulfamidase (Nuns); F. heparinum 6-O-sulfatase (FH6S); F. heparinum 2-O-sulfatase (FH2S); P. aeruginosa arylsulfatase (PARS, GenBank code AAC51350); human N-acetylgalactosamine-6-sulfate sulfatase or arylsulfatase B (Protein Data Bank code 1AUK), human arylsulfatase B (Protein Data Bank code 1FSU), and P. aeruginosa arylsulfatase B (Protein Data Bank code 1HDH) was performed using ClustalW (Fig. 1). The crystal structures of these enzymes were obtained from the Protein Data Bank and structure-based multiple superposition was obtained using C-MC (32) and SuperPose (33). The Homology module of Insight II molecular simulations package (Accelrys, San Diego) was used to obtain a structural model of N-sulfamidase with PDB 1AUK as template. The loop regions of the structure were modeled using the homology module of InsightII. The deletions were closed using the Discover module of InsightII, wherein most of the structure was held rigid except for the regions in the proximity of the deletion site, followed by 200 iterations of steepest descent and 300 iterations of conjugate gradient minimization. The loop regions and induced fluorescence was performed on a Beckman Coulter Proteomelab PA 800 with a 488-nm argon laser-induced fluorescence module. Samples were loaded onto a N-CHo capillary (50-μm inner diameter × 65-cm total length) using 0.5 p.s.i. of pressure at the anode for 20 s. Electrophoretic separations were performed using a 20-kV potential in a 100 mM sodium borate, pH 10.2 buffer for 15 min at 25°C. Fluorescence emission spectra were collected using a 520-nm narrow band filter.

Electrospray Ionization-Mass Spectrometry of Sulfated Glucosamine Monosaccharides—Electrospray ionization-mass spectrometry was performed generally as described for the 6-O-sulfatase (19) but with a few modifications. Reactions were carried out with 100 μM substrate, 2 mM CaCl2, 5 μM enzyme, and 50 mM sodium acetate buffer, pH 6.0, at 37°C overnight.
Heparin/Heparan Sulfate Glycosaminoglycan from F. heparinum

FIGURE 2. Obligatory substrate-product relationship of N-sulfamidase. Desulfation of the disulfated monosaccharide HNS,6S by the enzyme was followed by electrospray mass spectrometry. Panel i, substrate only shown here as the sodium adduct of a single ion species (M–1) with molecular mass of 350.8 Da; panel ii, treatment of GlcA,6S substrate with 6-O-sulfatase desulfates the 6-O position; panel iii, inability of N-sulfamidase to hydrolyze the original disulfated monosaccharide (compare with panel i); panel iv, co-treatment of the disulfated substrate with both 6-O-sulfatase and N-sulfamidase enzymes showing the disappearance of all sulfated monosaccharides and demonstrating a prerequisite 6-O-desulfation by the 6-O-sulfatase prior to sulfatide hydrolysis at the 2-amino position by the N-sulfamidase. Internal standard (458.1 Da) used to monitor ionization efficiency and mass calibration is noted by an asterisk.

RESULTS

Molecular Cloning and Recombinant Expression of F. heparinum N-Sulfamidase Gene—Screening of a genomic library with hybridization probes directed at the flavobacterial 2-O-sulfatase (26) revealed an overlapping phagemid clone that was then expanded by chromosomal walking and restriction mapping. An open reading frame of 1524 bp (orf) was identified by sequence analysis of this genomic region, and this was found to encode a protein of 500 amino acids in length. An examination of the putative orf gene product revealed that it possesses an NH2-terminal hydrophobic signal peptide and corresponding cleavage site sequence predicted by the von Heijne method for Gram-negative bacteria (25). The gene product also possesses a canonical sulfatase domain as described by the Protein Families Database identifier PF000884 (see under “Discussion”). The orf gene product was robustly expressed in E. coli as a soluble enzyme by removal of the amino-terminal signal sequence, as detailed earlier for the 6-O-sulfatase (19). The apparent molecular mass of 53,193 Da for the protein based on SDS-PAGE was consistent with the theoretical molecular weight calculated from putative amino acid composition based on translation of orf.

Biochemical Characterization of N-Sulfamidase and Optimization of in Vitro Reaction Conditions—We first considered the possibility of the orf gene product functioning as a generic arylsulfatase by testing it against 4-catechol sulfate and 4-MU sulfate, which are two different aromatic sulfates esters commonly used as substrates to assess generic sulfatase activity. It was found that the orf gene product exhibited only negligible activity relative to a known arylsulfatase from A. aerogenes that served as a positive control. To further test whether more robust activity could be determined against heparin and heparan sulfate, we used capillary electrophoresis wherein the two commercially available “heparin” monosaccharides 4-MU-GlcNAc,6S and 4-MU-GlcNS and the 6-O-sulfated galactose sugars 4-MU-Galα4S and 4-MU-Galβ3S (corresponding to the monosaccharide constituents of keratan sulfate and chondroitin/dermatan sulfate, respectively) were tested. This analysis showed that orf gene product could only hydrolyze the glucosamine sulfated at the 2-amino position. As such, we hereafter refer to the orf gene product as the N-sulfamidase (data not shown).

We further investigated the substrate specificities of N-sulfamidase by examining the influence of various substitutions at the 2-amino, 3-OH, and 6-OH positions of the glucosamine, using electrospray ionization-mass spectrometry. In this analysis, we found the N-sulfamidase activity at the 2-amino position was absolutely abolished when a 6-O-sulfate was also present (Fig. 2). Reciprocally related to this observation, in the accompanying paper (19) we determined that the 6-O-sulfatase side chains of all residues were then allowed to move freely by performing 500 iterations of steepest descent minimization. This was followed by further refinement of the structure by subjecting it to 500 iterations of steepest descent minimization without including charges and 500 iterations of conjugate gradient minimization, including charges to obtain the final predicted model of N-sulfamidase.

The coordinates of a heparin-derived hexasaccharide were obtained from co-crystal structure of FGF-1-heparin complex (Protein Data Bank code 2AXM). The hexasaccharide coordinates were imported into the Builder module of Insight II to generate the monosaccharide (GlcNS, GlcNS,6S, and GlcNS,6S) and disaccharide (ΔU,2s HNS,6S) substrates for N-sulfamidase. As with our previous studies on other sulfatases, the initial docking of the substrates in the enzyme groove was facilitated by the highly conserved position of the cleavable NH$SO_4$ group relative to the O-$\gamma$ of the germain diol. The enzyme-substrate complexes were then subjected to optimization with 300 steps of steepest descent followed by 400 steps of Newton-Raphson minimization, including charges, using the Discover module of Insight II. In this step, a force constant of 7000 kcal/mol was utilized to constrain the ring torsion angles to maintain the ring conformation of the substrates. The Amber force field was used to assign potentials to enzyme and sugar substrates for the energy minimization as described in the accompanying paper (19).
Heparin/Heparan Sulfate Glycosaminoglycan from F. heparinum

enzyme required a substituted amine (acetate or sulfate) at the 2-amino position. These findings imply that the 6-O-sulfatase and N-sulfamidase process substrates in a strictly sequential fashion with the former preceding the latter. Additionally, N-sulfamidase was completely inhibited by the presence of a 3-O-sulfate.

The CE-based assay was also used to determine the pH optimum for the enzyme. N-Sulfamidase exhibited a slightly acidic pH optima (between 5.5 and 6.5) but was active over a broad range, especially above pH 7.0. The enzyme showed higher activity in acetate buffer when compared with sulfonate buffers such as MES and MOPS when examined over this same pH range.

Having identified a suitable chromogenic substrate for measuring N-sulfamidase activity, we likewise used this substrate to develop a fluorescence-based, coupled enzyme assay as the means to define the optimal in vitro reaction conditions and enzyme kinetic parameters as generally described for the 6-O-sulfatase (19), with the exception that the secondary enzyme that was used was an α-glucosidase. These experiments demonstrated only modest hydrolysis of the 1→4-MU glycosidic linkage of the N-sulfated glucosamine by the α-glucosidase in the absence of an N-sulfamidase preincubation. In this assay, the N-sulfamidase exhibited a sensitivity to increasing ionic strength with 50% inhibition observed at ~200 mM NaCl and less than 20% activity remaining at 1 M NaCl relative to the zero NaCl control. Addition of sulfate or phosphate anions did not inhibit the action of N-sulfamidase.

We also investigated the action of the N-sulfamidase toward HSGAG disaccharides that possess a uronic acid at the nonreducing end. N-Sulfamidase was initially tested against a panel of unsaturated heparin disaccharides represented by ΔUs25HNS2568. For these experiments, standard reaction conditions were chosen as defined in the monosaccharide studies. None of the unsaturated disaccharides were desulfated by the enzyme (data not shown). The inability of N-sulfamidase to do so was confirmed in a related experiment in which all possible heparin disaccharides were first generated toward the nonreducing end. N-Sulfamidase was initially tested toward the absence of an N-sulfamidase preincubation. In this assay, that was used was an α-glucosidase. These experiments demonstrated only modest hydrolysis of the 1→4-MU glycosidic linkage of the N-sulfated glucosamine by the α-glucosidase in the absence of an N-sulfamidase preincubation. In this assay, the N-sulfamidase exhibited a sensitivity to increasing ionic strength with 50% inhibition observed at ~200 mM NaCl and less than 20% activity remaining at 1 M NaCl relative to the zero NaCl control. Addition of sulfate or phosphate anions did not inhibit the action of N-sulfamidase.

We also investigated the action of the N-sulfamidase toward HSGAG disaccharides that possess a uronic acid at the nonreducing end. N-Sulfamidase was initially tested against a panel of unsaturated heparin disaccharides represented by ΔUs25HNS2568. For these experiments, standard reaction conditions were chosen as defined in the monosaccharide studies. None of the unsaturated disaccharides were desulfated by the enzyme (data not shown). The inability of N-sulfamidase to do so was confirmed in a related experiment in which all possible heparin disaccharides were first generated toward the nonreducing end. N-Sulfamidase was initially tested toward the absence of an N-sulfamidase preincubation. In this assay, the N-sulfamidase exhibited a sensitivity to increasing ionic strength with 50% inhibition observed at ~200 mM NaCl and less than 20% activity remaining at 1 M NaCl relative to the zero NaCl control. Addition of sulfate or phosphate anions did not inhibit the action of N-sulfamidase.

We also investigated the action of the N-sulfamidase toward HSGAG disaccharides that possess a uronic acid at the nonreducing end. N-Sulfamidase was initially tested against a panel of unsaturated heparin disaccharides represented by ΔUs25HNS2568. For these experiments, standard reaction conditions were chosen as defined in the monosaccharide studies. None of the unsaturated disaccharides were desulfated by the enzyme (data not shown). The inability of N-sulfamidase to do so was confirmed in a related experiment in which all possible heparin disaccharides were first generated toward the nonreducing end. N-Sulfamidase was initially tested toward the absence of an N-sulfamidase preincubation. In this assay, the N-sulfamidase exhibited a sensitivity to increasing ionic strength with 50% inhibition observed at ~200 mM NaCl and less than 20% activity remaining at 1 M NaCl relative to the zero NaCl control. Addition of sulfate or phosphate anions did not inhibit the action of N-sulfamidase.

A structural basis for substrate specificity and enzymatic activity of N-sulfamidase—The molecular surface of the N-sulfamidase structural model revealed a groove region that was much wider than the 2-O-sulfatase that we have studied earlier (20), as well as the 6-O-sulfatase described in the accompanying paper (19). The theoretical model of the N-sulfamidase complexed with GlcNS substrate was used to identify the critical residues potentially involved in positioning of the substrate and/or catalysis (Fig. 3). Based on this theoretical model, the proposed roles of the putative active site residues are summarized in Table 1. The residues Arg-84 and Asp-40 that stabilize the FGly in the resting state in the active site are highly conserved among the different sulfatases. Furthermore, there is a pocket of oxygen atoms formed by Asp-40, Asn-61, Asp-246, Asn-247, the geminal diol of FGly, and N-sulfate group that can potentially coordinate a divalent metal ion such as Ca2+. The Ca2+ coordination formed by these residues resembles a similar arrangement in the other sulfatases (such as 6-O-sulfatase in the accompanying manuscript (19)) whose activity is critically influenced by Ca2+ ions. Surprisingly, two critical His residues that are highly conserved among the other sulfatases are absent in the putative active site of the N-sulfamidase. In the case of the 6-O-sulfatase homology model, His-130 is proposed to stabilize the FGly, whereas His-507 is proposed to be involved in the catalytic desulfation of the O-sulfate group. The absence of these critical His residues could be central to conforming the specificity of the N-sulfamidase enzyme to solely cleave the N-S bond (as against the O-S bond) in GlcNAc containing HSGAG substrates. Other O-sulfated hexosamine sugars such as GlcNAc,6GlcNAc,6GalNAc,4S,6S, and GalNAc,6S would therefore not be processed by N-sulfamidase, although these substrates might be accommodated in the active site. The modeled structural complex of N-sulfamidase with a 6-O-sulfate containing GlcNS,6S substrate showed that when the N-sulfate group is oriented toward the FGly residue, the 6-O-sulfate would make unfavorable steric contacts with Trp-249 in the active site (data not shown). This observation points to the obligatory sequential activity of the 6-O-sulfatase acting first to desulfate the 6-O-sulfate group, which is then followed by the action of N-sulfamidase. Similarly the complex of N-sulfamidase with GlcNS,6S showed that the 3-O-sulfate group would also have unfavorable steric contacts with Ile-188 and Asp-189 residues in the active site.

Using the homology model, we also investigated the structural rationale for the predominant exolytic action of N-sulfamidase. The unsaturated ΔUs25HNS2568 disaccharides were docked into the active site in an attempt to optimally position the N-sulfate group for catalysis. In so doing, the ΔU sugar and its sulfate group now made unfavorable steric contacts with several residues in the active site, including Trp-249, Leu-255, Ile-188, Trp-131, Trp-384, and Pro-363. This structural constraint imposed by the model is consistent with our observation that the N-sulfamidase will only process a GlcNS sugar at the nonreducing end of an oligosaccharide substrate (data not shown).

Specificity of N-Sulfamidase toward Longer Oligosaccharide Substrates—To experimentally validate the proposed exolytic activity of N-sulfamidase, we studied its activity on two structurally related sulfated trisaccharides (HNS65HNS65 and HNS65HNS65). Each of these was generated from the corresponding tetrasaccharide ΔUs25HNS65L2SHNS,6S by the tandem use of the 2-O-sulfatase and the Δ4-S-glycuronidase prior to the addition of 6-O-sulfatase and N-sulfamidase. Desulfation of the resultant trisaccharide was followed by MALDI-mass spectrometry (Fig. 4, data not shown for ΔUs25HNS65L2SHNS,6S). In this experiment, we observed that the 6-O-sulfatase was able to sin-
Heparin/Heparan Sulfate Glycosaminoglycan from *F. heparinum*

**FIGURE 1.** Theoretical structural model of *N*-sulfamidase-GlcNAc substrate. Shown at the top is the stereo view of the active site of *N*-sulfamidase with the docked GlcNAc substrate. The side chains of the key residues are shown and colored as follows: Arg and Lys, blue; Asp and Glu, red; Trp, Leu, and Be, brown; Asn and Gin, light blue; and FGly, purple. The GlcNAc substrate is colored by atom as follows: C, cyan; O, red; N, blue; and S, yellow. Shown at the bottom is the schematic of the active site with the key amino acids labeled for clarity. Also shown in the schematic is the location of the divalent Ca$^{2+}$ ion and its interactions with the active site and the substrate.

For this reason, we used an orthogonal CE-based method in an attempt to confirm the ability of the *N*-sulfamidase to hydrolyze odd-numbered oligosaccharides (lacking a 6-O-sulfate at their respective nonreducing ends) as strongly inferred in the previous experiment. Toward this end, the trisulfated pentasaccharide $H_{2}O_{2}H_{2}O_{2}H_{2}O_{2}$ was generated by $\Delta 4,5$-glycuronidase treatment of the purified hexasaccharide $H_{2}O_{2}H_{2}O_{2}H_{2}O_{2}$ and followed by incubation with the sulfamidase. All of the saccharides (untreated, $\Delta 4,5$ alone, and $\Delta 4,5$ followed by *N*-sulfamidase) were subsequently fluorescently labeled at their reducing end through reductive amination. End labeling of the sugars permitted their detection by laser-induced fluorescence following resolution of the products by capillary electrophoresis (Fig. 5). At each step in the experiment, saccharide peak assignment was inferred by observing discrete electrophoretic shifts in peak migration times as a function of exo-enzyme treatment and in accordance with their expected elution times based on differential charge densities. In this experiment, a unique peak (peak Z) eluting at 7.6 min appears following the sequential treatment of the starting material with the unsaturated glycuronidase and the *N*-sulfamidase. The late elution time of this peak is consistent with its possessing a lower sulfate density than either the trisulfated hexasaccharide starting material (peak X) or the pentasulfated pentasaccharide (peak Y, intermediate) resultant from digestion with the unsaturated glycuronidase. Such a dramatically shifted elution time is also suggestive of the presence of a positively charged amine that would reduce in the electrophoresis conditions described. From this analysis, it appears that the *N*-sulfamidase does desulfate oligosaccharides devoid of the 6-O-sulfate group on the terminal hexosamine, and it does so in a predictably exolytic fashion.

_Role of Calcium Metal Ions in Catalysis by *N*-Sulfamidase:_ *N*-Sulfamidase enzyme was found to be activated by the presence of calcium in a concentration-dependent manner (Table 2). Interestingly, the enzyme was found to be almost inactive in...
Heparin/Heparan Sulfate Glycosaminoglycan from F. heparinum

TABLE 1

<table>
<thead>
<tr>
<th>Active site residues</th>
<th>Proposed functional role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cys-80</td>
<td>The active site cysteine is modified into the hydrated form of the FGly-O-yl (20) and participates in nucleophilic attack on the sulfate group</td>
</tr>
<tr>
<td>Arg-84</td>
<td>This primarily stabilizes the hydrated FGly by interaction with O-yl. This is also well positioned for proton abstraction from O-yl after the catalytic process for elimination of sulfate and regeneration of diol</td>
</tr>
<tr>
<td>Asp-159</td>
<td>The negatively charged carbohydrate groups of these residues are likely to abstract the proton from the NH_T group of Lys-160</td>
</tr>
<tr>
<td>Lys-160</td>
<td>Upon loss of the proton to Asp-159, the nitrogen of this lysine residue is likely to be sulfated by the NHSO_3^- group, thereby breaking the existing nitrogen-sulfur bond. This process is mediated by the calcium ion. After this step, the sulfated lysine will be desulfated by the FGly residue. This lysine is also likely to increase electrophilicity of sulfur center by coordinating with the oxygen atoms of the sulfate group</td>
</tr>
<tr>
<td>Asp-246, Asn-247, Asp-40, Asn-41</td>
<td>Asp-246, Asp-247, Asp-40, and Asn-41 are well positioned to coordinate with the divalent Ca^2+ metal ion. The calcium ion is likely to play the crucial role of coordinating with the nitrogen of NHSO_3^- group, hence helping Lys-160 to break the N-S bond successfully. Hence, this tetrad is important for optimal N-sulfamidase function. Furthermore, Asp-40 is also understood to donate a proton and enhance nucleophilicity of O-yl</td>
</tr>
</tbody>
</table>

The absence of calcium, unlike previously analyzed sulfatases (Fig. 6A). For example, the 6-O-sulfatase described in the accompanying paper (19) was activated 2–3-fold by the presence of calcium but was somewhat active even in the presence of 1 mM EDTA. Furthermore, the divalent metal activation was found to be specific to calcium; inclusion of Mg^{2+} or Mn^{2+} had only negligible effects. To further examine this metal selectivity, we measured the potential for enzyme inhibition in the presence of the calcium-specific chelator EGTA. As expected, EGTA inhibited calcium-dependent N-sulfamidase activity (at 5 mM Ca^{2+}) in a concentration-dependent manner, with 50% inhibition occurring at ~3 mM EGTA (data not shown). In an attempt to determine the mechanism by which calcium exerts its effect on N-sulfamidase, we followed up these metal ion experiments by next measuring the effect of calcium on enzyme steady-state kinetics (Fig. 6B). Consistent with our previous results, the initial rate of the enzyme was significantly affected by calcium in a concentration-dependent fashion with both kinetic parameters being affected proportionally (Table 2).

Proposed Mechanism for Nitrogen-Sulfur Bond Cleavage by N-Sulfamidase Enzyme—The combination of our biochemical studies and the structural model of the N-sulfamidase active site and its interactions with the substrate led us to propose a mechanism for the N-S bond cleavage. This mechanism is novel given the uniqueness of N-sulfamidase active site and that it is the only enzyme from P. heparinum that cleaves N-S bond in HSGAGs. The proposed mechanism of N-sulfamidase action has been depicted as a step-by-step process pictorially (Fig. 7). In the resting state, the active site cysteine that is understood to be modified into the hydrated form of the FGly-O-yl (20) is likely to be stabilized by interactions with the carboxylate anion group of Asp-40 and the amino group of Arg-84 as shown. Furthermore, the divalent calcium ion has been proposed to coordinate with the diol form of FGly-O-yl and also with the negatively charged oxygen atom of the NHSO_3^- group upon addition of a suitable substrate N-NH-SO_3^- . Subsequently, the carboxylate anionic group of Asp-159 is well poised to abstract the proton from the NH_T group of Lys-160 residues. The nitrogen-sulfur bond cleavage is mediated by Ca^2+ ion and Lys-160 resulting in the transfer of the sulfate group to the formyl glycine. Alternatively, it is possible for the amine group of the deprotonated Lys-160 to form a bond with the sulfate group (potential transition state) that would in turn facilitate the transfer of this group to the formyl glycine (Fig. 7, dotted arrows). This crucial step that is unique to N-sulfamidase is likely to be motivated by the interaction of the Ca^{2+} ion with the nitrogen atom of the NHSO_3^- group and simultaneous protonation of this nitrogen atom by the NH_T group of Lys-158 as shown. This form of nitrogen-sulfur bond cleavage in the presence of metallic cations has been proposed for sulfonamide reactions (34). In our proposed mechanism, the divalent calcium cation and the basic Lys-160 residue provide the three interactions of this process, whereas in the sulfonamide mechanism proposed, a trivalent metal cation was used. The carboxylate group of Asp-40 is positioned to abstract a proton from the geminal diol leading to acquisition of the SO_3^- group from the sulfated form of Lys-160 as depicted. Following these distinguishing steps, the rest of the mechanism is likely to be similar to the other sulfatases (35). The basic Arg-84 residue has been indicated to interact with FGly-80 leading to the efflux of SO_3^- and formation of the aldehyde. This in turn is hydrolyzed back into the diol state, and the enzyme is ready to catalyze another desulfation process as and when a new substrate is encountered. The functions assigned to the key amino acids that participate in the proposed mechanism are summarized (Table 1).

An insight into the proposed mechanism for N-sulfamidase action based upon our structural model provides possible rationale for the observed behavior. First, calcium ion is absolutely required for coordinating the nitrogen-sulfur bond cleavage as outlined above. Without the presence of the divalent cation and the coordination of nitrogen of the N-SO_3^- group, the enzyme cannot act. Therefore, it is clear that the activity of N-sulfamidase is very much dependent on calcium ion concentration and that the enzyme becomes inactive in the absence of the cation. Moreover, it is well known that magnesium ions (Mg^{2+}) almost always exhibit octahedral coordination geometry with six being the preferred coordination number. Also, manganese ions (Mn^{2+}) prefer coordinating with four ligands thereby exhibiting predominantly tetrahedral coordination geometry. These are fewer ligands than calcium (Ca^{2+}) that is most commonly found coordinated by seven or eight ligands. The optimal coordination number in metal ions is a function of the ion size typically, because close packing of maximum pos-
Heparin/Heparan Sulfate Glycosaminoglycan from F. heparinum

A. Add 2-O sulfatase

B. Add A4,5 glycuronidase

C. Add 6-O sulfatase

D. Add N-sulfamidase

E. Add 6-O sulfatase → N-sulfamidase

TABLE 2
Steady-state kinetic parameters using 4-MU monosaccharide substrates

<table>
<thead>
<tr>
<th>Addition</th>
<th>$k_{cat}$</th>
<th>$K_m$</th>
<th>$k_{cat}/K_m$ ($\times 10^9$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>0.5 mM Ca$^{2+}$</td>
<td>5.1</td>
<td>45</td>
<td>11.3</td>
</tr>
<tr>
<td>5 mM Ca$^{2+}$</td>
<td>21.9</td>
<td>178</td>
<td>12.3</td>
</tr>
<tr>
<td>1 mM EDTA</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
</tbody>
</table>

* ND means not determined due to lack of activity.

DISCUSSION

In this paper, we have described the cloning, characterization, molecular modeling, and structure-function analysis of the flavobacterial N-sulfamidase enzyme. The N-sulfamidase (EC 3.10.1.1), commonly known as N-sulfoglucosamine sulfohydrolase, described here is one of only two nitrogen-sulfur cleaving enzymes currently identified by the Nomenclature Committee of the International Union of Biochemistry and the possibility of ligands around the ions is generally attempted. This analysis of the enzyme-substrate interactions explains the absolute necessity for calcium ions for N-sulfamidase action and also helps to rationalize the inability of other divalent cations such as manganese and magnesium to promote enzyme activity.

Each panel represent either peptide alone (~3216 Da) or oligosaccharide-peptide complex. The net mass of the oligosaccharide is listed in parentheses. A, addition of the 2-O-sulfatase; B, subsequent addition of the A4,5-glycuronidase; C, subsequent addition of the 6-O-sulfatase (note loss of a sulfate represented by a shift in net molecular mass from ~915 to ~835 Da); D, addition of N-sulfamidase directly after the A4,5-glycuronidase step (note the lack of any desulfation); E, addition of the N-sulfamidase subsequent to the 6-O-sulfatase (6-O then NS). The results in E are equivocal inasmuch as a double desulfated species with a net molecular mass of ~755 Da was not clearly detected in this experiment.
Molecular Biology (NC-IUBMB), the other enzyme being cyc-
lamate sulfatohydrolase (EC 3.10.1.2). Although significant pro-
gress has been made in understanding enzymatic O-sulfate
clavage mechanisms (35) involving oxygen-sulfur bond cleavage,
surprisingly little is known about the mechanisms of enzy-
matic action by the corresponding N-sulfate cleaving hydro-
lases that specifically break the normally stable nitrogen-sul-
fur bonds that are unique to HSGAGs. As such, the investiga-
tion presented here provides valuable insight into the unique prop-
erties of this enzyme.

One of the important observations that motivated our struc-
ture-function studies of N-sulfamidase was the notable absence of
key histidines that have been implicated as integral to the
function of O-sulfatases (35). The absence of these key histi-
dines could potentially govern the specificity of N-sulfamidase
to cleave nitrogen-sulfur as against oxygen-sulfur bond. This
observation prompted us to further investigate the mechanism of
nitrogen-sulfur bond cleavage by this unique active site of
N-sulfamidase. Using the theoretical structural model of the
enzyme-substrate complex, we propose a new mechanism for
the N-S bond cleavage by N-sulfamidase that was different
from oxygen-sulfur cleavage of O-sulfatases. To our knowledge,
this study is the first to propose such a mechanism of enzymatic
nitrogen-sulfur bond cleavage. We also experimentally vali-
dated the role of key active site residues by C80A or D40A as
described previously (20). These mutants showed no activity
(Table 3) in comparison with the wild type enzyme and hence
validate the critical role of these amino acids in the catalytic
activity of the enzyme.

Using this and previous studies, we have been able to recon-
struct the complete F. heparinum HSGAG degradation path-
way in vitro through a biochemical description of the respective
substrate specificities for each of the cloned enzymes. As such,
we are also able to place the activity of N-sulfamidase in a
sequential context related to the F. heparinum HSGAG degra-
dation pathway as it presumably exists in vivo, i.e. a degradation
pathway that begins with the heparin lyases (heparinas
es), which leads to small oligosaccharides with AU uronic acid at
the nonreducing end. These oligosaccharides are then acted

upon by the exolytic Δ4,5-glycuroni-
dase and the sulfatases. Previously, we
established that 2-O-sulfatase action
must precede the Δ4,5-glycuronidase
cleavage because the presence of a
2-O-sulfate group inhibits the glycuroni-
dase enzyme (18, 29). In the
accompanying paper (19) and this
study, we further establish that the
6-O-sulfatase enzyme should act
prior to N-sulfamidase because the
presence of the 6-O-sulfate group
inhibits the sulfamidase enzyme.
Interestingly, we still have to assign
within this sequence the functional
position of the 3-O-sulfatase activity
reported in the literature (36, 37).
The observations from this study
and the accompanying paper (19)
point to 3-O-sulfatase preceding the 6-O-sulfatase and N sul-
facidase. Therefore, the complete HSGAG degradation path-
way (after action by the depolymerizing heparinases) is likely
to proceed in the following sequence of enzymatic action: 2-O-
sulfatase, Δ4,5-glycuronidase, 3-O-sulfatase (putative), 6-
O-sulfatase, and N-sulfamidase.

Other questions related to the concerted activity of the
N-sulfamidase enzyme in vivo also remain. Chief among them
is the question of what precise form the substrates for these
end-of-the-line sulfatases actually take. Is it reasonable to
assume, for example, the "natural" substrates for the N-sulfa-
midase are actually monosaccharides? This assumption is at
least consistent with the sequentially exolytic nature of the
flavobacterial HSGAG-degrading pathway that we have
described. It is also in line with the HSGAG structure-activity
relationships and the active site architecture implied from these
relationships. Ultimately, the ability of these enzymes to act on
longer oligosaccharides in a manner predicted by their sub-
strate specificities is of great practical value toward the use of
these enzymes as discrete analytical tools for elucidating
HSGAG core structure.

N-Sulfated glucosamines are unique to heparin and heparan
sulfate, and it is not surprising that the mechanism proposed in
this paper is somewhat unique to this sulfate ester. The distinct
chemistry and atomic interactions obtained from an analysis of
our molecular model explained the empirically observed prop-
erties such as the substrate specificity, the novel mechanism of
nitrogen-sulfur cleavage, the activity of the enzyme only in the
presence of calcium ions but not other divalent metal cations,
the order of enzyme processivity, and the inhibition of the
enzyme by the presence of secondary sulfates within the gluco-
samine. As form follows function, these distinctions naturally
play out at the level of enzyme structure and, in the case of the
lysosomal N-sulfamidase, these are perceivable even at the pri-
mary sequence level where there is only about 10–25% identity
to O-sulfatases. We also point out that even when one compares
the heparan sulfamidase between divergent organisms such as
flavobacterium and mammals, discrete structural differences
are likely given the reversed order within the degradation

Heparin/Heparan Sulfate Glycosaminoglycan from F. heparinum

FIGURE 6. Effect of calcium metals on N-sulfamidase activity. A, calcium-specific requirement for N-sul-
famidase activity and inhibition by EDTA wherein the divalent metal effect was not observed when calcium was
replaced by either Mg2+ or Mn2+. (at either 1 or 5 mM concentrations). The following is the key: open bars (no
divalent metals added); black bars (1 mM EDTA added); light gray bars (1 mM divalent metal); stippled gray bars
(5 mM divalent metal); effect of calcium on steady-state kinetics of N-sulfamidase observed while varying
concentrations of Ca2+ or in the presence of 1 mM EDTA. Substrate saturation plots were fitted to pseudo
first-order Michaelis-Menten kinetics by nonlinear regression analyses of 0.5 mM Ca2+ (4), 1 mM Ca2+ (C), 5 mM
Ca2+ (©). For clarification, the EDTA result (showing a lack of activity) is omitted.
sequence in which the sulfatase enzymes act. In the lysosomal
pathway, the N-sulfamidase is a relatively early enzyme that precedes the 6-O-sulfatase, although our results indicate a reverse order for the flavobacterial enzymes. As such, the lysosomal heparin N-sulfamidase naturally possesses broader substrate specificity relative to the functional homologue from *F. heparinum*. It follows that the relative active site topologies should also differ, especially as it pertains to additional residues for the lysosomal enzyme that must accommodate secondary sulfate interactions.

**REFERENCES**

Heparin/Heparan Sulfate Glycosaminoglycan from F. heparinum

4. Influenza Virus Surface Proteins Interactions with Host Sialylated Branched Glycans

Summary

This chapter focuses on decoding the structure-function relationships mediating the interactions of Influenza virus surface proteins -- hemagglutinin (HA) and neuraminidase (NA) -- with sialylated branched glycans on the host. The findings are reported specifically from the context of the pandemic 2009 H1N1 swine-origin Influenza virus ("Swine Flu") that emerged during the course of this Thesis research, although they are broadly applicable to influenza-host interactions. Specifically, the structure-function relationships of the following Influenza proteins are analyzed for the Swine Flu influenza virus:

(1.) Hemagglutinin (HA)

Since HA is key for both the sialylated glycan receptor binding and viral-host cell membrane fusion steps, thus being a critical determinant of Influenza zoonosis, human-human transmission rates, extent of infection within human hosts, antigenic response and viral clearance kinetics, the structure-function relationship mediating pandemic 2009 H1N1 HA interaction with host sialylated glycans was investigated.

(2.) Neuraminidase (NA)

Since NA is critical for catalyzing the hydrolysis of terminal sialic acid residues from the newly formed Influenza virions and from the host cell receptors, in addition to being the primary target of popular antivirals Tamiflu and Relenza for which drug resistance was an emergent challenge, the structure-function relationship of 2009 H1N1 NA was investigated in detail.

(3.) Other key proteins: RNA polymerase (PB2) and Matrix protein (M2)

Since PB2 and M2 proteins have emerged as critical determinants of virulence and anti-viral therapy (with reported rampant resistance) respectively, the structure-function relationship of these proteins on 2009 H1N1 were also investigated.

These studies resulted in a peer reviewed publication in *Nature Biotechnology* in June 2009.
4.1 Introduction

4.1.1 Overview of Influenza A Viruses

Influenza A virus is a hypermutative negative strain RNA virus possessing eight gene segments [317-320]. Of these segments, three of the genes — hemagglutinin (HA), neuraminidase (NA), and the polymerase (PB2) — have emerged as critical for viral infection, antigenic response, human-to-human transmission rates, and pandemic potential (Figure 4.1) [321-325]. Within influenza A, five of the genome segments encoding the nucleocapsid protein (NP), the matrix proteins (M1 and M2), the nonstructural proteins (NS1 and NS2), and polymerase proteins (PB1, PB2 and PA) have maintained a relatively consistent evolutionary history in humans owing to antigenic pressure [326-330]. In contrast, the two genes encoding the major cell surface proteins (HA and NA) have been subjected to much more substantial evolutionary pressure, including mutation (antigenic drift) and wholesale reassortment (antigenic shift) being commonplace [321-324]. Due to such diversity, strains of influenza virus are identified based on their serotype of HA and NA [325-327]. There are currently 16 known serotypes of HA (H1-H16) and nine of NA (N1-N9) with various influenza strain subtypes being referred as a combination of HA and NA serotypes (Figure 4.2) [328-331].

Figure 4.1 Influenza Virus Lifecycle
The influenza viral life cycle may be briefly summarized as follows [332-337]. Hemagglutinin (HA) on the surface of the virus binds glycans terminated by sialic acid with a specific linkage (highlighted in green and red in Figure 4.1), initiating fusion of the virus with the host cell. This interaction is highly specific and is governed by the type of sialic acid linkage, the underlying sugars and branching pattern of the glycan receptors. The other viral surface proteins are the neuraminidase (NA) and the ion channel protein M2. Once the virus is internalized in the cell, fusion between viral and nuclear membranes occurs, and complexes of RNA and proteins termed viral ribonucleoprotein complexes (vRNPs) are transported into the nucleus of the host cell. The transcription to mRNA takes place in the host cell nucleus followed by export and protein synthesis. Also within the nucleus there is transcription of the RNA genome. Assembly of progeny vRNPs then occurs, with export, assembly of the virus progeny, and finally budding of the newly formed virus particles.

<table>
<thead>
<tr>
<th>HA</th>
<th>NA</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>H1</td>
</tr>
<tr>
<td>H2</td>
<td>H2</td>
</tr>
<tr>
<td>H3</td>
<td>H3</td>
</tr>
<tr>
<td>H4</td>
<td>H4</td>
</tr>
<tr>
<td>H5</td>
<td>H5</td>
</tr>
<tr>
<td>H6</td>
<td>H6</td>
</tr>
<tr>
<td>H7</td>
<td>H7</td>
</tr>
<tr>
<td>H8</td>
<td>H8</td>
</tr>
<tr>
<td>H9</td>
<td>H9</td>
</tr>
<tr>
<td>H10</td>
<td>H10</td>
</tr>
<tr>
<td>H11</td>
<td>H11</td>
</tr>
<tr>
<td>H12</td>
<td>H12</td>
</tr>
<tr>
<td>H13</td>
<td>H13</td>
</tr>
<tr>
<td>H14</td>
<td>H14</td>
</tr>
<tr>
<td>H15</td>
<td>H15</td>
</tr>
<tr>
<td>H16</td>
<td>H16</td>
</tr>
</tbody>
</table>

Figure 4.2 Influenza Hemagglutinin (HA) and Neuraminidase (NA) serotypes listed according to known species they have been reported in the past (gray shading).
4.1.2 Influenza Hemagglutinin: Importance of Host Sialylated Glycan Topologies

As highlighted in the influenza virus lifecycle (Figure 4.1), the key first step in the infection, transmission and virulence of influenza is viral surface HA binding to sialylated glycans on the epithelial cell surface [338-343]. The chemical structure distribution of Human Bronchial Epithelial (HBE) cell glycans has been studied with MALDI-MS previously and shown to consist of a heterogeneous mixture of complex, branched, predominantly sialic acid capped glycans with a variety of linkages and monosaccharide compositions (Figure 4.3).

Recently, our group defined glycan topology as a key determinant of influenza virus hemagglutinin (HA) binding to host sialylated glycan receptors, showing that influenza viruses adapted for infecting and transmitting between humans have predominantly umbrella topologies with α2,6 linkage, whereas those viruses that have adapted for predominantly avian infection have cone topologies with α2,3 linkage (Figure 4.4) [344, 345].

Figure 4.3 MALDI-MS glycan profile of Human Bronchial Epithelial (HBE) cells
Thus, it was established that transmission from birds to humans is due to the ability of HA to switch its preference from α2-3 sialylated glycans (cone topology) to α2-6 sialylated glycans (umbrella topology) that are extensively expressed in the human upper respiratory epithelial cells [344, 345].

Towards providing such molecular insight into the structure-function relationship of influenza HA, several crystal structures of influenza proteins have been determined in recent years (Table 4.1) [378]. These include crystal structures of HAs from H1 (human, avian, and swine), H2 (human), H3 (avian and human), H5 (avian), H7 (avian), H9 (swine) and their complexes with α2-3 and/or α2-6 oligosaccharides. These HA-glycan co-complex crystal structures provide atomic resolution information into the key residues and interactions governing the virus-host interactions for specific influenza strains. These structures have also greatly enhanced our understanding and appreciation for both the receptor binding and membrane fusion functions of influenza hemagglutinin (Figure 4.5).

Knowledge of the glycan-binding specificities obtained from analysis of these HA-receptor co-complex crystal structures was extended to homology-based molecular models of HA sequences from the 2009 H1N1 influenza virus in the enclosed manuscript, so as to understand the structural basis for human adaptation of the ongoing flu pandemic.
Table 4.1 Available Crystal Structures of Influenza Proteins Compiled from the PDB, with the HA proteins highlighted in *pink* and NA proteins in *blue*. Influenza subtypes that have resulted in 20th century pandemics are underlined.

<table>
<thead>
<tr>
<th>Influenza-A Subtype</th>
<th>Protein</th>
<th>PDB ID</th>
<th>Influenza-A Subtype</th>
<th>Protein</th>
<th>PDB ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1N1</td>
<td><strong>HA</strong></td>
<td>2WRG (3.0Å)</td>
<td>H5N1</td>
<td><strong>HA</strong></td>
<td>1JSM (1.9Å)</td>
</tr>
<tr>
<td></td>
<td><strong>NA</strong></td>
<td>3BEQ (1.6Å)</td>
<td></td>
<td><strong>NP</strong></td>
<td>2Q06:A (3.3Å)</td>
</tr>
<tr>
<td><strong>PA</strong> (partial)</td>
<td></td>
<td>2ZNL:A (2.3Å)</td>
<td></td>
<td><strong>NS</strong></td>
<td>3F5T (2.7Å)</td>
</tr>
<tr>
<td></td>
<td><strong>NP</strong></td>
<td>2IQH:A (3.2Å)</td>
<td></td>
<td><strong>M1</strong></td>
<td>1J16 (2.0Å)</td>
</tr>
<tr>
<td></td>
<td><strong>NS</strong></td>
<td>3L4Q (2.3Å)</td>
<td></td>
<td>(partial)</td>
<td>3CM8 (2.9Å)</td>
</tr>
<tr>
<td><strong>M1</strong> (partial)</td>
<td></td>
<td>1EA3 (2.3Å)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H2N2</td>
<td><strong>HA</strong></td>
<td>2WRE (3Å)</td>
<td>H9N2</td>
<td><strong>HA</strong></td>
<td>1JSD (1.8Å)</td>
</tr>
<tr>
<td></td>
<td><strong>NA</strong></td>
<td>1IVC (2.4Å)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H3N2</td>
<td><strong>NS</strong></td>
<td>1AIL (1.9Å)</td>
<td>H11N9</td>
<td><strong>NA</strong></td>
<td>1F88B:A (1.8Å)</td>
</tr>
<tr>
<td>RNA-binding Domain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>NS</strong></td>
<td>3EE8 (2.6Å)</td>
<td>H12N5</td>
<td><strong>NS</strong></td>
<td>3D6R (2.0Å)</td>
</tr>
<tr>
<td>Effector Domain</td>
<td></td>
<td></td>
<td></td>
<td>Effector Domain</td>
<td></td>
</tr>
<tr>
<td></td>
<td><strong>HA</strong></td>
<td>1MQN (3.2Å)</td>
<td>H13N9</td>
<td><strong>NA</strong></td>
<td>1NMB:N (2.2Å)</td>
</tr>
<tr>
<td></td>
<td><strong>NA</strong></td>
<td>2AEQ:A (3.0Å)</td>
<td>H7N3</td>
<td><strong>HA</strong></td>
<td>1TI8 (3.00Å)</td>
</tr>
<tr>
<td>PB2 Cap-binding</td>
<td></td>
<td></td>
<td>H11N6</td>
<td><strong>NA</strong></td>
<td>1V0Z (1.84Å)</td>
</tr>
<tr>
<td>Domain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PB2 C-terminal</td>
<td></td>
<td></td>
<td>H10N4</td>
<td><strong>NA</strong></td>
<td>2HTV (2.8Å)</td>
</tr>
<tr>
<td>Domain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.1.3 Influenza Neuraminidase: Prominent anti-viral target and emergent drug resistance

Neuraminidase (NA) is the enzyme on the surface of influenza viruses that cleaves the sialic acid groups on glycoproteins [378]. Specifically, influenza NA has two functions within the host system. First, while influenza HA is required for infection (as described in the last section), its presence on budding virions emerging from infected cells inhibits their release. Influenza NA cleaves the terminal neuraminic acid (sialic acid) residues from glycan structures on the surface of the infected cell, promoting release of progeny viruses and hence continued spread of influenza to uninfected cells. Second, influenza NA also cleaves sialic acid residues from viral proteins on other virions in the vicinity, thus preventing self-aggregation of viruses and auto-elimination thereof. It may be noted that in some other viruses (such as the measles virus), an integrated hemagglutinin-neuraminidase protein combines both their functions onto a
single motif, unlike in influenza wherein HA and NA are independently expressed on the viral surface.

Neuraminidase inhibitors (Figure 4.6) are a class of antiviral drugs targeted at influenza NA, which work by blocking NA's function of promoting budding from infected host cells. Oseltamivir (Tamiflu) a prodrug, Zanamivir (Relenza), and Peramivir are FDA-approved NA inhibitors. Unlike the M2 inhibitors (e.g. the adamantane derivatives) that are only effective against influenza A, these NA inhibitors are effective against both influenza A and influenza B.

Recent data from the Center for Disease Control (CDC) shows that there is rampant resistance to Oseltamivir (Tamiflu) amongst the seasonal influenza strains, with over 99.6% of all 2008 seasonal H1N1 strains having acquired resistance to Tamiflu treatment, up from a mere 12% in 2007-2008 flu season. This data suggests that drug resistance to NA inhibitors is an emerging issue of grave concern, since these drugs are the first line of defense to promote recovery upon infection and curtail spread of virus from humans-to-humans for both seasonal and pandemic influenza.

---

![Chemical structures of sialic acid (boxed) and popular influenza NA inhibitors](image)

**Figure 4.6** Chemical structures of sialic acid (boxed) and popular influenza NA inhibitors
Hence, one of the objectives of NA structure-function analysis in the context of the 2009 H1N1 pandemic virus was to analyze the potential for NA of this virus to acquire mutations that would enable the emergence of Tamiflu-resistant pandemic influenza. On the other hand, Relenza-resistant NA mutations have only rarely been reported till date in the literature and this aspect was also investigated. For this purpose, the crystal structures of influenza H1N1 NA from the PDB complexed to sialic acid (Figure 4.7) as well as the NA inhibitor molecules were utilized as templates to model the homology-based structure of the 2009 H1N1 influenza NA and analyze its interactions with the ligand and inhibitor molecules. The results of such analysis of 2009 pandemic H1N1 NA structure-function relationship as it pertains to predicting emergence of anti-viral drug resistance are described in detail in the enclosed manuscript.

4.1.4 Other Key Influenza Proteins: Polymerase (PB2) and Matrix (M2) Proteins

Influenza polymerase is an important protein responsible for replication of the virus within infected host cells. Mutation in the polymerase gene is believed to be responsible for migration of the virus from avian to human hosts. Understanding the structure-function relationship of influenza polymerase is hence key to devising new strategies in fighting seasonal and pandemic influenza. The influenza polymerase functions as a heterotrimeric assembly of
PB2, PB1 and PA proteins, with a small added open reading frame (ORF) that codes for the PB1-F2 peptide which plays a role in virus-induced cell death. While it remains to be seen how the polymerase complex must mutate to adapt to new hosts, single amino acid changes in PB2 (especially E627K) have been shown to be important determinants of human adaptation, pathogenicity, and virulence. In the enclosed manuscript,

The matrix (M) protein of influenza is the major internal protein of the influenza virion, constituting approximately 40% of the viral proteins. The M protein surrounds the inner core of the virion, has a molecular weight of approximately 27,000, and occupies the inner surface of the lipid bilayer of the envelope that contains the HA and NA subunits. The M protein is constituted of the M1 and M2 protein domains. The M2 protein is a fairly cross-type well-conserved (broad spectrum conserved) proton-selective ion channel. The channel itself is a homotetramer (consists of four identical M2 units) where the units are helices stabilized by two disulfide bonds and is activated by low pH.

Influenza M2 inhibitors based on adamantane backbones were first approved by the FDA as antivirals against influenza during the Asian Flu pandemic. Amantadine is the organic compound that is formally referred as 1-aminoadamantane and consists of an adamantane backbone substituted at one of the four methyne positions with an amino group (Figure 4.8). Amantadine is FDA-approved and sold commercially under the brand name "Symmetrel" for use as an antiviral against Influenza-A. Rimantadine is a closely related chemical derivative of adamantane (Figure 4.8) with similar biological properties and is also FDA-approved as an anti-influenza antiviral. Rimantadine is sold commercially under the brand name of Flumadine.
Figure 4.8 Chemical structures of Adamantane derivatives inhibiting influenza M2 protein

According to the CDC, 100 percent of seasonal H3N2 and seasonal H1N1 influenza samples tested in 2008 were resistant to adamantane-derivatives, and amantadine and rimantadine have since not been recommended for treatment of seasonal flu. With the emergence of the 2009 H1N1 pandemic flu, we analyzed the structure-function relationship of the M2 protein of this virus as it pertains to amantadine and rimantadine binding and potential resistance. A homology model of the 2009 swine flu M2 protein was derived to analyze potential resistance to adamantanes, based on template structure with PDB ID: 3C9J as shown (Figure 4.9).
Specifically, it was noted that the adamantane-derivative Amantadine (highlighted in green/white in Figure 3.8) binds to a deep hydrophobic pocket in the template MP2 protein of influenza viruses (colored gray/pink). However, from the homology model of the 2009 swine flu M2 protein (Figure 3.9), it was seen upon superposition onto the amantadine-bound template MP2 protein that the swine flu MP2 is polar/charged in the groove, thus presenting a very different molecular surface internal to the transmembrane region (as compared to the more hydrophobic patches in the wildtype MP2). Based on this analysis, we predicted almost no binding affinity of adamantanes to the 2009 California Swine Influenza MP2. In agreement with this prediction, recent experimental studies by the CDC have confirmed that the 2009 H1N1 virus is resistant to adamantane derivatives.
Figure 4.10 Amantadine (green/white) docked to homology model of 2009 H1N1 virus MP2 protein (yellow/cyan/gray) developed using template with PDB ID 2RLF (82% identity)

Similarly, superposition of the template M2 protein (yellow/orange) with bound rimantadine (cyan) and the modeled 2009 H1N1 M2 protein (pink/gray) was performed (Figure 4.11). Rimantadine binds to pockets on the outer surface of the M2 domain, causing a closure of this critical transmembrane protein in terms of pump/conformation change activities. However, the S31N mutation (boxed) is seen in the 2009 California Swine Strain. This mutation is known to confer drug resistance towards Rimantadine due to destabilization of the extreme end of the transmembrane domain (with the Asn-31 residue making unfavorable contacts with Leu and Ala hydrophobic residues), as seen from the figure. Such destabilization was predicted to result in poorer binding affinity of Rimantadine towards the MP2 protein, and was later confirmed by experimental binding assays.
4.1.5 Pandemic potential of Influenza Viruses: Emergence of the 2009 Swine Flu Virus

Given the past history of deadly influenza pandemics, there is substantial public health concern surrounding the prospect of influenza A pandemics and its associated potential global implications. The past 20th century itself has seen four influenza pandemics (1918, 1957, 1968, and 1977). The first and most deadly pandemic of the 20th century occurred in 1918 ("Spanish Flu") and involved an H1N1 subtype of influenza virus. The Spanish flu pandemic led to the death of over 40 million people worldwide.

Viruses containing the 16 HA and nine NA serotypes are naturally present in wild aquatic bird populations where they exist commensally without causing disease, allowing birds and pigs to become a reservoir for influenza strains, some with possible zoonoses potential (Figure 4.12). This is of specific concern because of the influenza pandemics of the last century, those that arose from H2N2 (1957) and H3N2 (1968), were avian-human reassortments that resulted in the humanization of an avian-adapted virus and efficient human-to-human transmission. Those genetic reassortments that led to an avian-to-human switch yield a number of important scientific and medical questions, not the least of which is what changes lead away from infectivity and propagation in avian species and towards human transmissibility?
In light of a particular influenza strain, H5N1 ("bird flu") with grave virulence potential reported, addressing such questions becomes extremely critical. Transmission of avian H5N1 influenza viruses to humans has been observed thus far only upon direct contact with infected poultry; the virus has not yet demonstrated efficient human-to-human transmission ability. Given that human infectivity has occurred and appreciating that this virus strain of H5N1 is extremely lethal (estimates ~ 60% of infected individuals succumb to the viral and immune-response onslaughts), the importance of comprehending the mechanism and specificity of viral entry and infection as well as identifying additional therapeutic and prophylactic strategies is evident.

Given the extensive damage to life and economies caused by multiple instances of zoonotic and pandemic influenza emergence within the 20th century, the present emergence of a novel H1N1 virus, viz., 2009 H1N1 or “swine” flu is of great concern; and has rightfully being declared the first Influenza pandemic of the 21st century by the World Health Organization (Figure 4.13). One of the primary cause for concerns regarding the 2009 swine flu virus has been the antigenic dissimilarity of this virus from seasonal, circulating H1N1s, thus implying little pre-existing immune protection against this virus in the human population of today. The 2009 H1N1 virus has been labeled a 'triple reassortant virus' owing to its acquisition of genes commonplace in human, swine, and avian influenza strains.
Gene Segments, Hosts, and Years of Introduction

~1918: HA, NP, NS

~1918: PB1

~1968: PB1

~1998: PB2, PA

~1998: PB2, PA

Triple Reassortant

Classical Swine

Eurasian Swine

2009 A(H1N1)

Figure 4.13 Host and lineage origins for the proteins of the 2009 H1N1 "swine flu" virus

Taken together, the recent incidence and spread in humans of the 'swine flu' influenza A virus hence raised global concerns regarding its virulence and pandemic potential. The first cases of human infection were reported in April 2009 in the Mexican town of La Gloria in Veracruz; soon after, reported infections occurred in areas of southern California and Texas of the United States of America (USA). Using the several recent studies that have focused on the necessary determinants for human adaptation and efficient human-to-human transmission of the H1N1 influenza A viruses (described in the last few sections), in the enclosed manuscript, a representative 2009 H1N1 strain was studied using homology modeling tools to predict the likely human adaptation and transmissibility of the 2009 H1N1 pandemic influenza viruses.

4.1.6 Published manuscript

Specific contributions of this Thesis research to the publication: Developed the homology-based structural models for the 2009 pandemic H1N1 'Swine Flu' Hemagglutinin (HA), Neuraminidase (NA), Polymerase (PB2), and Matrix (M2) proteins, in addition to proteins corresponding to the other genes segments of the virus. Each of the developed structural models was analyzed to predict biologically and medically salient structure-function relationships underlying emergence and spread of the pandemic, such as, human adaptation, human-to-human transmission, evolution of drug resistance, seasonal flu vaccine efficacy, antigenic drift, virulence, and overall pandemic potential.
Extrapolating from sequence—the 2009 H1N1 ‘swine’ influenza virus

To the Editor:

The recent incidence and spread in humans of the ‘swine flu’ influenza A virus has raised global concerns regarding its virulence and pandemic potential. The main cause of the so-called swine flu has been identified as human infection by influenza A viruses of a new H1N1 (hemagglutinin 1, neuraminidase 1) subtype, or ‘2009 H1N1 strain’. The first cases of human infection were reported in April in the Mexican town of La Gloria in Veracruz; soon after, reported infections occurred in areas of southern California and Texas. Several recent studies have focused on the necessary determinants for human adaptation and efficient human-to-human transmission of the H1N1 influenza A viruses. Here, using a representative 2009 H1N1 strain as our starting point, we offer a perspective on the likely human adaptation and transmissibility of 2009 H1N1 viruses.

At the time when this sequence analysis was performed, partial or complete sequences were available from 38 different human isolates of the 2009 H1N1 virus. These sequences were obtained from GISAID (Global Initiative on Sharing Avian Influenza Data; http://platform.gisaid.org/) and the NCBI Influenza Virus Resource (National Center for Biotechnology Information; http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html). Comparison of the amino acid sequences between the 38 isolates showed some intragenic differences: seven amino acid positions in HA (hemagglutinin), one in M1 (matrix 1), two in M2 (matrix 2), four in NA (neuraminidase), three in NP (nucleoprotein), two in PA and two in PB2 (both of which encode subunits of viral RNA polymerase). Given the few intragenic variations among the 38 isolates available at the time of this study, we use /California/04/2009 (Cal0409) as a representative 2009 H1N1 virus strain for further analysis. The top ranking hits of the BLAST search using the individual Cal0409 genes are shown in Supplementary Table 1 online.

Comparison of Cal0409 HA with the HA consensus sequences for human-adapted H1N1, avian-adapted H1N1 and swine-adapted H1N1 reveals important substitutions in positions 100–300, where the glycan receptor-binding sites and antigenic loops are located (Supplementary Table 2 online). Notably, the Cal0409 HA possesses the signature amino acids Asp190 and Asp225 that have been shown to play a key role in conferring specificity to the human α2-6 sialylated glycan receptors. We also observe amino acid substitutions that are unique to Cal0409 HA and have not been observed in previous human H1N1 HA. These include substitutions at sequence positions 74, 131, 145, 208, 219, 261, 263, 264, 365, 317, 368, 377 and 530. Among these residue positions, 131 and 145 are proximal to the glycan-binding site.

To determine the possible effect of these mutations on the glycan-binding properties of HA, we constructed homology-based structural complexes of Cal0409 with representative α2-3 and α2-6 sialylated oligosaccharides, as described previously (Fig. 1). The construction of theoretical HA-glycan structural complexes previously allowed us to provide a structural rationale for how specific amino acid mutations within the 1918 H1N1 HA can dramatically alter its relative α2-3/α2-6 binding affinity. Referring these previous efforts, we determined the potential glycan binding properties of Cal0409 HA by analyzing its contacts with the α2-3 and α2-6 sialylated glycans. On the basis of the observed contacts in the HA glycan complexes, we summarize in Table 1 the proposed roles of the residues in Cal0409 HA that provide binding specificity to α2-3 and α2-6 oligosaccharides, respectively. The main differences between the glycan-binding pockets of reference HAs and Cal0409 HA lie in the 140-loop region and the loop region preceding the 190-helix, Lys145, which is unique to Cal0409, along with Lys133 and Lys222, forming a positively charged ‘lysin fence’ at the base of the binding site that potentially are located to anchor the N-acetylneuraminic acid (Neu5Ac) and galactose (Gal) sugars of both α2-3 and α2-6 glycans. In the case of the Cal0409 α2-6 oligosaccharide structural complex, the lysine fence also includes Lys156, which is positioned to provide additional contact with the galactose. The orientation of Asp190 is typically stabilized by a network of interactions involving residues at 186, 187 and 189 that precede the 190-helix. In Cal0409 HA, the residues at these positions are Ser186, Thr187 and Ala189; this set of residues is unique to the 2009 H1N1 strains. These residues appear to retain the ability to stabilize the orientation of Asp190 such that it is positioned to make optimal contacts with the third N-acetylgalactosamine (GKNAc) sugar (starting from Neu5Ac toward the reducing end) of α2-6 glycans, defined previously.

Our observations of the Cal0409 HA-glycan interactions suggest that this HA has the necessary residues to provide optimal contacts for
high affinity binding to α2-6 glycans present in the human upper airways. Typically, the Glu190Asp substitution between avian and human-adapted H1 HA results in the loss of a critical contact with the Neu5Acα2-3Gal motif and a gain in contact with α2-6 glycans. In the Cal0409 HA, however, this loss in contact to α2-3 glycans appears to be compensated for by Lys145 of the lysine fence. In summary, our analysis suggests that Cal0409 HA possesses residues that can be positioned to make optimal contacts with α2-6 (a characteristic binding feature shared by human H1N1 HA) as well as α2-3 sialylated glycans. In future studies, it will be important to experimentally determine the relative α2-6 and α2-3 binding affinities of the 2009 H1N1 HAs using appropriate methods.

Comparison of the antigenic regions of Cal0409 NA with the consensus sequences of avian, human and swine-adapted N1 NAs shows that four positions—188, 331, 372 and 402—are novel in the 2009 H1N1 NA (Supplementary Table 3 online). NA is presently the primary target of therapeutic intervention for influenza infection, and oseltamivir (Tamiflu) and zanamivir (Relenza) are widely used NA-inhibiting drug molecules. Recently, it has been reported that there is an alarming increase in the oseltamivir resistance of the recent H1N1 viruses 13 , our analysis suggests that the recent emergence of Tamiflu-resistant viruses. Given the alarming proportion of circulating Tamiflu-resistant seasonal H1N1 viruses, 13, our analysis support for both augmentation of oseltamivir stockpiles with additional drugs (including zanamivir) and the prudent administration of antivirals in general.

To probe the resistance of the 2009 H1N1 viruses to adamantane-derivative drugs, the prototypic mutation associated with adamantane resistance is Ser31Asn (14,15) and the presence of this mutation in all the 2009 H1N1 strains is consistent with their observed resistance to adamantane-derivative drugs.

To provide additional functional context to the above analysis, the amino acid sequences of

**Figure 1** Glycan-binding properties of Cal0409 HA. A homology-based structural model of Cal0409 HA was constructed using the prototypic 1918 H1N1 HA (PDB ID: 1RUV) as a template. (a) Theoretical structural model of Cal0409 HA (gray) bound to an α2-3 oligosaccharide (carbon; green) and the key amino acids on HA (carbon; purple) that are positioned to make optimal contacts with the glycan. This model was constructed using the PR8 (A/Puerto Rico/8/34) HA-α2-3 oligosaccharide co-crystal structure (PDB ID: 1RVT; coordinates of trisaccharide Neu5Acα2-3Galβ1-4GlcNAcβ1-3Galβ1-4Glc are ordered). The analogous amino acids in PR8 that are different from Cal0409 HA (labeled red with PR8 residues in parenthesis) are also shown (carbon; yellow). (b) Theoretical structural model of Cal0409 HA bound to an α2-6 oligosaccharide (carbon; cyan) showing the key amino acids on HA (carbon; purple) that are positioned to make optimal contacts with the glycan. This model was constructed using A/Swine/Iowa/0120/2009 (AS130) HA-α2-6 oligosaccharide co-crystal structure (PDB ID: 1RVT; coordinates of tetrasaccharide Neu5Acα2-6Galβ1-4GlcNAcβ1-3Galβ1-4Glc are ordered). The analogous amino acids in AS130 HA that are different from Cal0409 HA (labeled red with AS130 residues in parenthesis) are also shown (carbon; orange). The oxygen and nitrogen atoms are colored red and blue respectively. (c) Molecular surface of the HA-α2-6 oligosaccharide complex highlighting the lysine fence (circled).
the HA and NA of the 2009 H1N1 strain were compared to those of the current vaccine H1N1 strain, A/Brisbane/59/07. A/Brisbane/59/07 has been associated with severe infectivity due, at least in part, to the fact that its HA and NA have low antigenic cross-reactivity to pre-existing humoral immunity. We then compared our results to the same analysis using A/Brisbane/59/07 and the previous H1N1 vaccine strain A/Solomon Islands/3/2006. This exercise provides an important reference point for a virus that is characterized as antigenically dissimilar to most circulating H1N1s. For the comparison between the NA of the 2009 H1N1 strain and the A/Brisbane/59/07 strain, we find that the overall percent sequence identity is 80.6%; however, this sequence identity drops to 38.0% within the antigenic regions (Supplementary Table 4 online). A similar trend is observed in HA (Supplementary Table 5 online), where the identity is reduced from 79.2% (overall) to 56.3% (antigenic). In contrast, when comparing A/Brisbane/59/07 with A/Solomon Islands/3/2006, the percent identity in NA shows only a marginal drop from 94.4% (overall) to 92.0% (antigenic). Similarly, with regards to HA alone, the percent identity...
drops from 98.6% (overall) to 97.2% (antigenic) (Supplementary Tables 6 and 7 online). Taken together, the results indicate that the substantial variability observed in the antigenic regions of the 2009 H1N1 viruses would most likely result in the presentation of new epitopes that may not cross-react with the antibodies generated using the current vaccine strains, thereby potentially having important implications toward the protective effect afforded by existing seasonal influenza vaccines.

In addition to the vital role of the viral coat proteins, most prominently HA, in governing transmission, virulence and human adaptation, recent studies have demonstrated the critical role of the viral RNA polymerase PB2 in the efficient respiratory droplet (or airborne) transmission of wild-type human H1N1 viruses and avian-human reassorted influenza viruses in the ferret model9. In this study, we demonstrated that a specific residue, Lys627, in human-adapted PB2 was critical for conferring efficient transmissibility. Conversely, mutation of this Lys627 to glutamic acid (which is typically found in avian and swine-adapted PB2) in PB2 of the 1918 pandemic strain (SC18) severely reduced its ability to transmit. Analysis of PB2 in all the 2009 H1N1 strains indicates that it has glutamic acid at position 627. On the basis of these earlier studies, we expect that the 2009 H1N1 viruses may be capable of transmission between humans, but the efficiency of transmission might be hampered by the absence of Lys627 in PB2. Although the HA of the 2009 H1N1 viruses is human-adapted, our analysis suggests that PB2 still requires an additional mutation to become fully human-adapted for efficient transmission.

A recent study9,10 evaluated the pandemic potential of the 2009 H1N1 viruses using the available epidemiological data. This study concluded that despite the substantial uncertainty in the data, the clinical severity of the 2009 H1N1 viruses is more comparable to the 1957 H2N2 pandemic outbreak than the 1918 H1N1 pandemic. Given the fact that the evolution of this virus is uncertain at best, we must remain vigilant for additional mutations that can render this strain more virulent.

In conclusion, we set out using sequence information to evaluate the HA, NA PB2 and M2 genes in the new H1N1 viral strain based on their known critical roles in the human adaptation, human-to-human transmission and resistance to currently used antiviral drugs (Table 2). As experimental data for 2009 H1N1 viruses become available, it will be possible to correlate those results with the analyses presented here.

**Table 2 Summary of predicted properties of HA, NA, PB2 and M2 for current 2009 H1N1 viruses**

<table>
<thead>
<tr>
<th>Protein</th>
<th>Predicted properties</th>
<th>Perspectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemagglutinin (HA)</td>
<td>* Glycan-binding site (including Asp190 and Asp225) is similar to that of human adapted H1N1 HA &lt;br&gt; * Binding site contains residues that are positioned to make optimal contacts with both a2-6 and a2-3 glycans &lt;br&gt; * Lys145, Ser186, Thr187 and Ala189 are novel substitutions that have not been observed in other human H1N1 HAs</td>
<td>* HA is human adapted and is expected to bind with high affinity to a2-6 and better affinity to a2-3 in comparison with other human H1N1 HAs &lt;br&gt; * Novel substitutions in the antigenic regions of HA might present new epitopes</td>
</tr>
<tr>
<td>Neuraminidase (NA)</td>
<td>* Novel substitutions in four positions 189, 331, 369 and 395 in the putative antigenic site &lt;br&gt; * Active site has not yet acquired the characteristic mutations such as His274Tyr that provides resistance to oseltamivir &lt;br&gt; * Zanamivir can potentially make optimal contacts with the oseltamivir-resistant NA mutants</td>
<td>* Novel substitutions in antigenic regions of NA might present new epitopes &lt;br&gt; * NA can acquire mutations that offer resistance to Tamiflu &lt;br&gt; * Zanamivir might be preferred to oseltamivir</td>
</tr>
<tr>
<td>PB2 and M2</td>
<td>* Lys627Glu mutation in PB2 reduced transmission efficiency of a prototypic human-adapted H1N1 virus &lt;br&gt; * M2 protein has acquired Ser314Asn mutation that provides resistance to adamantane-derived drugs</td>
<td>* Acquisition of the Glu627lys mutation could potentially improve transmission efficiency of the virus</td>
</tr>
</tbody>
</table>

**Note:** Supplementary information is available on the Nature Biotechnology website.

**ACKNOWLEDGMENTS**

The authors would like to acknowledge support from the Singapore-Massachusetts Institute of Technology Alliance for Research and Technology; and the National Institutes of Health (GM 57073 and U54 GM21160) to R.S.


2. *These authors contributed equally to this work. e-mail: rams@mit.edu*

**Table S1. BLAST hits for the gene sequences of Cal0409.** Results of BLAST nucleotide analysis are arranged into four columns: (1) gene name; (2) accession # of the best BLAST hit; (3) Name of the virus carrying the best hit from column 2; (4) percent sequence identity.

<table>
<thead>
<tr>
<th>Gene</th>
<th>Accession # of best BLAST hit</th>
<th>Virus name</th>
<th>Percent Identity</th>
</tr>
</thead>
<tbody>
<tr>
<td>HA</td>
<td>AF455680</td>
<td>Influenza A virus (A/Swine/Indiana/P12439/00 (H1N2))</td>
<td>95%</td>
</tr>
<tr>
<td>NA</td>
<td>AF250366</td>
<td>Influenza A virus (A/Swine/England/195852/92 (H1N1))</td>
<td>94%</td>
</tr>
<tr>
<td>PB2</td>
<td>EU301177</td>
<td>Influenza A virus (A/Swine/Korea/JNS06/2004(H3N2))</td>
<td>96%</td>
</tr>
<tr>
<td>PB1</td>
<td>AF342823</td>
<td>Influenza A virus (A/Wisconsin/10/98 (H1N1))</td>
<td>96%</td>
</tr>
<tr>
<td>PA</td>
<td>AF455722</td>
<td>Influenza A virus (A/Swine/Illinois/100084/01 (H1N2))</td>
<td>96%</td>
</tr>
<tr>
<td>NP</td>
<td>AF251415</td>
<td>Influenza A virus (A/Swine/Korea/JNS06/2004(H3N2))</td>
<td>96%</td>
</tr>
<tr>
<td>NS</td>
<td>AF153262</td>
<td>Influenza A virus (A/Swine/Iowa/533/99 (H3N2))</td>
<td>96%</td>
</tr>
<tr>
<td>MP</td>
<td>AY363575</td>
<td>Influenza A virus (A/Swine/Hong Kong/5212/99(H3N2))</td>
<td>97%</td>
</tr>
</tbody>
</table>

**Table S2. Sequence comparison of Cal0409 HA with consensus amino acids from avian, swine and human H1N1 HA.** The first row corresponds to the sequence of Cal0409 HA. The 2nd-4th rows represent sequence of positions where each position contains the consensus residue (the most predominantly occurring amino acid in all the sequences for that species in that position) in human, swine and avian H1N1 HAS respectively. The putative antigenic sites are highlighted in blue and glycan-binding sites are highlighted in yellow. The amino acids that are unique to Cal0409 HA and not found in any of the human H1N1 HAS are colored red.
Table S3. Sequence comparison of Cal0409 NA with consensus amino acids from avian, swine and human H1N1 NA. The first row corresponds to the sequence of Cal0409 NA. The 2nd-4th rows represent sequence of positions where each position contains the consensus residue (the most predominantly occurring amino acid in all the sequences for that species in that position) in human, swine and avian H1N1 NAs respectively. The putative antigenic sites are highlighted in blue and the active-site residues are highlighted in yellow. The amino acids in the antigenic region that are unique to Cal0409 NA and not found in any of the human H1N1 NAs are colored red.

Table S4. Sequence alignment between NAs of A/California/04/2009 (H1N1) (row 1) and A/ Brisbane/59/2007 (H1N1) (row 2). Antigenic sites are colored red. The antigenic similarity is measured as the percentage fraction of antigenic sites having the same amino acids (=19/50 = 38%).
### Table S5. Sequence alignment between HAs of A/California/04/2009 (H1N1) (row 1) and A/Brisbane/59/2007(H1N1) (row 2). Antigenic sites are colored red. The antigenic similarity is measured as the percentage fraction of antigenic sites having the same amino acids (=40/71 = 56.33%).

<table>
<thead>
<tr>
<th></th>
<th>Cal0409</th>
<th>Brisb5907</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MKAILVVLVLYTFATANADTLCIGYHANI NSTDTVTVEKNNTVTHSNVC LLEKHNGKLCK</td>
<td>MKIKLLVVLVCTTATYADTLCIGYHANI NSTDTVTVEKNNTVTHSNVC LLEKHNGKLCL</td>
</tr>
<tr>
<td></td>
<td>LRGQAPIPLHLGKNCIAIWGILGHP ECESLSLSSWYVEIFPS DNGTCYPGDFID YEELRE</td>
<td>LRGQAPIPLHLGKNCIAIWGILGHP ECESLSLSSWYVEIFPS DNGTCYPGDFID YEELRE</td>
</tr>
<tr>
<td></td>
<td>QLSVSSFVERFEIPF KSSNPNHSNKGVTAA CPHAGAKSFYNNLNLVVKGSFPLLSK</td>
<td>QLSVSSFVERFEIPF KSSNPNHSNKGVTAA CPHAGAKSFYNNLNLVVKGSFPLLSK</td>
</tr>
<tr>
<td></td>
<td>SYINDGKEVLVLVGIHPHSAD QSGLYQNADTVFVGSNRSYKFKPEIAIRPKFVRDQ</td>
<td>SYANNEKEVLVLVGIHPHSAD QSGLYQNADTVFVGSNRSYKFKPEIAIRPKFVRDQ</td>
</tr>
<tr>
<td></td>
<td>EGRMNYYWIFVEPGDKITEATNLVPRYAFAMERNAGSGGIISDTPVHCDNCTCQTPQ</td>
<td>EGRINYYWILFVEPGDKITEATNLVPRYAFAMERNAGSGGIISDTPVHCDNCTCQTPQ</td>
</tr>
<tr>
<td></td>
<td>GAINSLDFQNPQHIPITIGCKPVKYSTKLRATGLNPIQGFFGLFAIAFGIEGGWTQ</td>
<td>GAINSLDFQNPQHIPITIGCKPVKYSTKLRATGLNPIQGFFGLFAIAFGIEGGWTQ</td>
</tr>
<tr>
<td></td>
<td>MVDGMYQYHHTQNEQGSGYUADLKQSTQAIDENKVNVSIEKMNQTPTAVGKEFNLHKLK</td>
<td>MVDGMYQYHHTQNEQGSGYUADLKQSTQAIDENKVNVSIEKMNQTPTAVGKEFNLHKLK</td>
</tr>
<tr>
<td></td>
<td>IENLNKVDGDFLDDTNYAELULVLENERTLDYHDNSVKNYKEVQSLKQNNAEIGNG</td>
<td>IENLNKVDGDFLDDTNYAELULVLENERTLDYHDNSVKNYKEVQSLKQNNAEIGNG</td>
</tr>
<tr>
<td></td>
<td>CFQFYHSDKDNCMSVNYNTVDYPFYSTEAEKLQNEEIDGQKLSTRIYQILAYSTVASS</td>
<td>CFQFYHSDKDNCMSVNYNTVDYPFYSTEAEKLQNEEIDGQKLSTRIYQILAYSTVASS</td>
</tr>
<tr>
<td></td>
<td>LVLVSLGIAISFWMCNSGSLQCRIC</td>
<td>LVLVSLGIAISFWMCNSGSLQCRIC</td>
</tr>
</tbody>
</table>

### Table S6. Sequence alignment between NAs of A/Solomon Islands/3/2006 (H1N1) (row 1) and A/Brisbane/59/2007(H1N1) (row 2). Antigenic sites are colored red. The antigenic similarity is measured as the percentage fraction of antigenic sites having the same amino acids (=46/50 = 92%).

<table>
<thead>
<tr>
<th></th>
<th>SolIs306</th>
<th>Brisb5907</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MNPNQKIQITIGISIAIGISILMLILQIGNIISIMASHSGIQTSQGQHTGIACNQRIITYENST</td>
<td>MNPNQKIQITIGISIAIGISILMLILQIGNIISIMASHSGIQTSQGQHTGIACNQRIITYENST</td>
</tr>
<tr>
<td></td>
<td>WVANTYVNINNTNVVAKEDTSTVTAGNSSLCSISGWAITYKDMSIRISGKGDVFVIREP</td>
<td>WVANTYVNINNTNVVAKEDTSTVTAGNSSLCSISGWAITYKDMSIRISGKGDVFVIREP</td>
</tr>
<tr>
<td></td>
<td>FISCHLRECTFFLTRQALNLKHSNGTVKDRSPTYLMSCLPGEAFPYSNNRSFESAVS</td>
<td>FISCHLRECTFFLTRQALNLKHSNGTVKDRSPTYLMSCLPGEAFPYSNNRSFESAVS</td>
</tr>
<tr>
<td></td>
<td>ASACHDGQMIGILGSGPAVAVLYKNYIDITETIKNQKIRLRTCESECVMNSSCFT</td>
<td>ASACHDGQMIGILGSGPAVAVLYKNYIDITETIKNQKIRLRTCESECVMNSSCFT</td>
</tr>
<tr>
<td></td>
<td>IMTDGPNGAASYKFKIEKGGKVTHTKLNAPNHYECYCYDPTGTMMVCCRDNWHGSN</td>
<td>IMTDGPNGAASYKFKIEKGGKVTHTKLNAPNHYECYCYDPTGTMMVCCRDNWHGSN</td>
</tr>
<tr>
<td></td>
<td>RPWSFNQNLQYQIYIYCSGVSFGDNPRPKQGECSNPQTVDGAIVYVKGFSYKYNGNWIG</td>
<td>RPWSFNQNLQYQIYIYCSGVSFGDNPRPKQGECSNPQTVDGAIVYVKGFSYKYNGNWIG</td>
</tr>
<tr>
<td></td>
<td>RTESNRLEKGFEMIDPNNTNTO SDFSVQDVQVAYTDWGSYSGFVQHPELTGLDLCIRP</td>
<td>RTESNRLEKGFEMIDPNNTNTO SDFSVQDVQVAYTDWGSYSGFVQHPELTGLDLCIRP</td>
</tr>
<tr>
<td></td>
<td>CFWELVRLPRENNTIWTSGGSSIFSFCGVNSGTANNS</td>
<td>CFWELVRLPRENNTIWTSGGSSIFSFCGVNSGTANNS</td>
</tr>
</tbody>
</table>

---
**Table S7.** Sequence alignment between HAs of A/Solomon Islands/3/2006 (H1N1) (row 1) and A/Brisbane/59/2007 (H1N1) (row 2). Antigenic sites are colored red. The antigenic similarity is measured as the percentage fraction of antigenic sites having the same amino acids (=69/71 = 97.2%).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MKVKKLLVLLCTFTATAYDICTIGYHAN</td>
<td>NSTDTVDVTLKNTVTTHSVNLLEDSHNGKLC</td>
<td>MKVKKLLVLLCTFTATAYDICTIGYHAN</td>
<td>NSTDTVDVTLKNTVTTHSVNLLEDSHNGKLC</td>
<td>LKGIAPLQLGNCSVAGWILGNPCCELLIS</td>
<td>RESWSYIVKPNPENTCYPGHFADYEEELRE</td>
<td>LKGIAPLQLGNCSVAGWILGNPCCELLIS</td>
<td>RESWSYIVKPNPENTCYPGHFADYEEELRE</td>
<td>QLSSVSSPERFEIPKESSWPNNHTTTGVSASCSHNGESSFYRNNLLTGLNGLYPNLSKS</td>
<td>QLSSVSSPERFEIPKESSWPNNHTTTGVSASCSHNGESSFYRNNLLTGLNGLYPNLSKS</td>
<td>YANNKEKEVLVLWGVHHPNIGDQRALYHtenayYVSVSSHYSRKFPTPIAKRPKVDR</td>
<td>YANNKEKEVLVLWGVHHPNIGDQRALYHtenayYVSVSSHYSRKFPTPIAKRPKVDR</td>
<td>GRINYYWTLLEPGDIPEANGNIAPRAYAFLSRGFGSGIINSNAPMDECDAKCQTFQG</td>
<td>GRINYYWTLLEPGDIPEANGNIAPRAYAFLSRGFGSGIINSNAPMDECDAKCQTFQG</td>
</tr>
</tbody>
</table>
5. De novo Protein Structure Prediction and Lectin Sequence-to-Function Assignment

Summary

This chapter focuses on the development of a new tool for homology-based structural modeling based on our finding that protein core atomic interaction networks (PCAINs) in the solvent-unexposed core of protein domains are 'signature' of protein fold families. Specifically:

1. **Quantitative analysis of evolutionary tinkering and sequence divergence of protein folds**
   It is shown that evolutionary tinkering and sequence divergence are rampant across the protein universe, with a majority of protein folds being rich in 'twilight' and 'midnight' homology zones.

2. **PCAINs are universal signatures of the native folded structure of protein domains**
   It is deduced that PCAINs are well-conserved between domains of the same fold family, while significantly different from PCAINs for domains of other fold families, consistently across families the entire universe of proteins, emerging as "signature" of proteins native folded state.

3. **PCAIN is an effective tool for fold identification and structure prediction in the twilight zone**
   Building on the fold-specific nature of PCAINs, it is demonstrated that the use of PCAIN-based scoring schemes permits effective classification of protein sequences into their native folds and high-throughput, accurate homology-based (comparative) protein structure prediction. These results are further consistent for even proteins from the twilight zone and CASP experiments.

4. **The PCAIN methodology has tremendous applications for GBP structure prediction and furthering the field of Glycobiology**
   The developed de novo fold identification and structure prediction tools are shown to be particularly attractive for overcoming the challenges facing GBP structure elucidation, thus emerging as a useful tool for progressing towards de novo protein-glycan interaction analysis.

5. **Other significant biological applications of the PCAIN methodology are explored**
   The highlighted applications are in Immunology, specifically, modeling the structure of YopM effector protein from Bubonic Plague causative bacterium Yersinia Pestis and understanding its host immune modulatory properties, as proof-of-principle of PCAIN method applications.

These studies resulted in a peer reviewed publication in *PLoS ONE* in February 2010.
5.1 Pursuit of the Elusive Protein Fold Code: Implications of *Ab Initio* Mapping of Gene Sequences to Folded Protein Structures

The biopolymers DNA, RNA and Proteins are linear polymers, with each monomer connected at most to two other monomers. The sequence of their monomers effectively encodes biological information. The *genetic code* is the set of rules by which information encoded in the tri-nucleotide sequences of genetic material such as DNA or mRNA, is translated into the amino acid sequence of proteins, by all living cells (Figure 5.1). This information flow is described by the general transfers (normal flow) of the *Central Dogma of Molecular Biology* as follows: DNA information can be copied to DNA (*replication*), DNA information can be copied into mRNA, (*transcription*), and protein sequences can be synthesized using the information in mRNA as a template (*translation*). The transfers of information described by the central dogma are faithful, deterministic transfers, wherein one biopolymer's sequence is used as a template for the construction of another biopolymer with a sequence that is entirely dependent on the original biopolymer's sequence.

*Figure 5.1 Decoding Biological Information Flow in Nature's Lego Set: The Elusive Protein Fold Code*
The amino acid sequences of proteins that emerge from the translational machinery are folded into their native, functional structure, by a complex and as yet uncharacterized process termed protein folding. Protein folding is largely believed to be a templated process, going from the sequence of amino acids to secondary structural motifs (Chou-Fasman Code), to tertiary structural folds (Protein Fold Code), and ultimately quaternary structural assemblies (Symmetry Code) as depicted in Figure 5.1. The discovery of a protein fold code would hence significantly enhance universal understanding of biological information flow, providing for a direct, templated, 1-on-1 mapping of genetic nucleotide sequences to the folded native structure of proteins. However, the pursuit of the elusive protein fold code over the past five decades has met with limited progress owing to several challenges as outlined in the next few sections.

5.2 Anfinsen's Dogma in the Post-Genomics Era: Significance and Implications of the Twilight Zone of Homology

In the seminal research work published nearly four decades ago, Christian B. Anfinsen hypothesized that "information dictating the native fold of globular protein domains is fully encoded in their amino acid sequence" [388]. However, with the explosive amount of protein sequence, structure, and fold data generated since the time of Anfinsen during the Omics Era of the latter half of the 20th century, the emerging picture of the protein universe has challenged Anfinsen's dogma, for it has become evident that numerous protein fold families have incredible sequence diversity with a rampant "twilight zone" (less than 25% sequence identity between proteins of the same native fold) and no consistent "fold code" (Figure 5.2) [384]. In support of this observation, recent studies have shown that proteins with as low as 1-2% sequence identity may still adopt the same native fold [389-392], thus defying any tangible encoding of fold-dictating information into protein sequence. The pursuit of the elusive "fold code" has resulted in little more than patterns of amino acid sequence conservations that are specific to certain proteins [393-395], but no finding has been compelling enough to generalize universally or to utilize for biological applications. So as to quantitatively estimate the extent of evolutionary sequence divergence within the known universal set of protein fold families, we computed the Blosum62 score per amino acid pair on a fold family by fold family basis (Figure 5.2) to find that more than 60% of protein fold families are significantly divergent in their sequences, around 30% of protein fold families are moderately divergent in their sequences, and less than 10% of protein fold families are well conserved in their sequences. This result shows that sequence divergence is rampant in protein universe -- consequence of evolutionary tinkering, discussed in next section.
A snapshot of the protein folds in the universe exudes an extremely interesting and intricate set of architectures (Figure 5.3). The domains of the protein universe are classified in terms of secondary structural composition into architectures. The architectures are further classified into topologies/folds by incorporating the loops/connectivity information, and each fold is composed of several families on the basis of their evolutionary origin and function. As of 2010, there are around 1200 distinct protein folds that have been discovered, and it has been estimated that the total number of protein folds in the entire biological universe would be around 10,000. It is indeed extraordinarily fascinating that Nature employs merely a few thousand protein folds to generate the entire repertoire of her multimillion strong protein universe.
Figure 5.3 The Full-Spectrum of Known Protein Architectures in the Biological Universe
The observation that Nature has developed merely a few thousand folds to facilitate the multi-millions of vastly diverse functions in the biological universe had led scientists to remark that "Nature is a tinkerer rather than an innovator". Relentless evolutionary tinkering has resulted in several biologically, phylogenetically, and functionally promiscuous protein folds. Examples of such protein fold promiscuity has been widely appreciated in cases wherein function has demanded massive tinkering-driven sequence divergence. For instance, molecular recognition scaffolds are renowned to tolerate vast degrees of amino acid sequence divergence. A classical example of such a protein fold is the immunoglobulin-like (Ig-like) β-sandwich, the immune system's antibody and T-cell receptor (TCR) scaffold for universal molecular recognition, that exhibits unparalleled divergence in its complementarity determining regions (CDRs) (Figure 5.4). Another more recently characterized example is the C-type lectin fold, the evolutionary solution for massive sequence divergence necessitated by its coding for diversity-generating retroelements (DGRs) that permit its function as the dynamic host receptor-binding scaffold in temperate bacteriophages. Yet another interesting example of molecular-recognition driven diversity plays out in glycan-binding proteins (GBPs) that specifically recognize a multitude of structurally complex and chemically heterogeneous carbohydrates to moderate key biological processes including cell growth and development, cell-cell communication, host-pathogen interaction, and immune function, as discussed in the next section.

The uncompromised ability of protein domains to be foldable and stable despite their vast degrees of sequence divergence is a hallmark of Nature's intelligent design. With the exception of rare chameleon polypeptide fragments that have shown propensity to adopt multiple structures, a Lion's share of the known protein sequences still adhere to their one native structural fold, implying that information in the sequence necessarily and universally codes for its folded structure.

One of our long standing interests has been understanding principles of natural design to engineer protein scaffolds for specific recognition of cell surface glycan and glycoprotein motifs that regulate desired biological functions. As a first step, deducing the protein fold code and applying this for de novo identification of fold, and prediction of protein structure directly from gene sequence, emerged as the key goals to overcome the twilight-zone related challenges (from the protein perspective). Taken together with the earlier chapters that focused on tools to elucidate glycan-protein interactions (from the glycan perspective), this Thesis provides an integrated framework to enable design-based engineering of glycoproteins with desired function.
5.4 The Twilight Zone of Glycobiology: The Need for, and Challenges Posing, *Ab initio* GBP Structure Prediction

To accommodate the structural diversity of glycans arising from their non-template driven biosynthesis, there is substantial diversity in the sequence and three-dimensional structural folds of various GBPs (Table 5.1) [346-349]. The three-dimensional structure of the glycan binding domain of GBPs is therefore fundamental to understanding their specific molecular recognition of glycan motifs which in turn is critical to decode their structure-function relationships for a variety of biological and biotechnological applications [350-352]. Owing to the association of the glycan-binding domains for their multivalent presentation however, there are challenges in the recombinant expression and purification of GBPs [346-352], and therefore in obtaining their three-dimensional crystal structures.
Table 5.1  Diversity of the commonly occurring fold families in Glycan-binding proteins (GBPs) along with their occurrence frequency obtained from the cancer lectin database and sugar-binding proteins from the PDB (negligible occurrences are denoted by dashes).

<table>
<thead>
<tr>
<th>Fold family (α/β/αβ)</th>
<th>Glycan-binding Proteins (GBPs)</th>
<th>Frequency of fold occurrence in GBPs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mannose-binding proteins,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Snake coagglutinins, Surfactant proteins, Tetranectins, Selectins, DC-SIGNs, CD69, CD94</td>
<td>25.13%</td>
</tr>
<tr>
<td></td>
<td>CONCANAVALIN A-like (β)</td>
<td>18.24%</td>
</tr>
<tr>
<td></td>
<td>Calnexins, Galectins, Lectin leg-like, Legume lectins, Pentraxins</td>
<td>13.39%</td>
</tr>
<tr>
<td></td>
<td>β-trefoil (β)</td>
<td>20.33%</td>
</tr>
<tr>
<td></td>
<td>Fibroblast growth factors (FGF),</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ricin B-like lectins, Agglutinins,</td>
<td>9.22%</td>
</tr>
<tr>
<td></td>
<td>Glycosidases, Neurotoxins,</td>
<td>7.89%</td>
</tr>
<tr>
<td></td>
<td>Ribosome-inactivating proteins</td>
<td></td>
</tr>
<tr>
<td></td>
<td>β-prism (β)</td>
<td>13.39%</td>
</tr>
<tr>
<td></td>
<td>Mannose-binding lectins, Delta-endotoxins,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Vitelline membrane outer-layer (VMO) proteins</td>
<td>0.97%</td>
</tr>
<tr>
<td></td>
<td>β-sandwich (β)</td>
<td>3.86%</td>
</tr>
<tr>
<td></td>
<td>E set domains, Fibronectins,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Immunoglobulins, Tenascins</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Serpins (αβ)</td>
<td>1.89%</td>
</tr>
<tr>
<td></td>
<td>Antithrombins, Heparin cofactors,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Protein C inhibitors</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>TIM beta/alpha barrel (αβ)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Amylases, Chitinases, Glycerocanases</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>RNase A-like (αβ)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Angiogens</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Knottin (β)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hevein-like lectins</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>RAP domain-like (α)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Receptor associated proteins (RAPs)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Four-helical up-and-down bundle (α)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Apolipoproteins</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>β-propeller (β)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tachylectins, Fucose-specific lectins</td>
<td>-</td>
</tr>
</tbody>
</table>

Protein structure prediction is an extensive area of research aimed at bridging the gap between proteins with known primary amino acid sequences and those with available X-ray co-crystal structures. The methods for protein structure prediction can be classified broadly into *ab initio* energetics based approach to simulate folding of linear protein sequence and comparative methods that predict the structural fold and model the structure of a target protein sequence. Among the various methods for protein structure prediction, homology modeling is emerging as a valuable tool to bridge the gap between sequence and structure.
The fundamental principle behind the homology-based modeling approaches is the assignment of structure of a ‘target’ protein based on a ‘template’ (whose structure is known) using the sequence identity between the template and the target. Although there are plenty of outstanding automated/manual homology modeling servers and tools, their application to structure prediction of GBPs is limited due to several roadblocks. One major limitation is the generally poor pairwise sequence identity (PSI) that prevails in GBPs even within a given structural fold, as shown herein for the trefoil (top left), prism (top right), concanavalin (bottom left) and C-type lectin (bottom right) folds that are predominant in GBPs (Figure 5.5).

The extensive presence of the twilight zone in GBPs is believed to be a consequence of large evolutionary sequence drift that has occurred over several millions of years of binding by numerous chemically heterogeneous and structurally diverse glycan molecules to a handful of folds (Figure 5.6). For instance, a very diverse set of glycans including both GAGs and branched glycans (N-linked and O-linked) are found co-complexed to proteins with the β-trefoil fold. This example is an illustration of the structural and chemical diversity in glycan-binding to GBPs.
<table>
<thead>
<tr>
<th>Chemical Structure of Glycan</th>
<th>Glycan-binding Proteins (GBPs)</th>
<th>PDB IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Galactose (Gal)" /></td>
<td>Ricinus communis agglutinin, Viscum album lectin, Sambucus ebulus lectin, Earthworm lectin mutant, Tetanus toxin C-terminal domain</td>
<td>2AAI, 1PUU, 1YF8, 1TFM, 1HWP, 2D12, 1DLL</td>
</tr>
<tr>
<td><img src="image" alt="N-acetylgalactosamine (GalNAc)" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><img src="image" alt="Galactosamine (GalN)" /></td>
<td>Earthworm lectin mutant</td>
<td>2DS0</td>
</tr>
<tr>
<td><img src="image" alt="Glucose (Glc)" /></td>
<td>Clostridium botulinum neurotoxin-B</td>
<td>1F31</td>
</tr>
<tr>
<td><img src="image" alt="N-acetylgalucosamine (GlcNAc)" /></td>
<td>Tetanus toxin C-terminal domain</td>
<td>1FV3</td>
</tr>
<tr>
<td><img src="image" alt="Xylose (Xyl)" /></td>
<td>Amaranthus caudatus agglutinin</td>
<td>1JLX</td>
</tr>
<tr>
<td><img src="image" alt="N-Acetyneuraminic acid (Neu5Ac)" /></td>
<td>Marasmius Oreades Agglutinin</td>
<td>2IHO</td>
</tr>
<tr>
<td><img src="image" alt="N-Glycolyneuraminic acid (Neu5Gc)" /></td>
<td>Sambucus ebulus lectin</td>
<td>1HWN</td>
</tr>
<tr>
<td><img src="image" alt="2-Keto-3-deoxy-nononic acid (KDN)" /></td>
<td>Sambucus ebulus lectin</td>
<td>1HWO</td>
</tr>
<tr>
<td><img src="image" alt="Fucose (Fuc)" /></td>
<td>Sambucus ebulus lectin</td>
<td>1HWM</td>
</tr>
<tr>
<td><img src="image" alt="Glucuronic Acid (GlcA)" /></td>
<td>Viscum album lectin</td>
<td>1M2T</td>
</tr>
<tr>
<td><img src="image" alt="Iduronic Acid (IdoA)" /></td>
<td>Mouse mannose receptor</td>
<td>1FWU and 1FWV</td>
</tr>
<tr>
<td><img src="image" alt="Galacturonic Acid (GalA)" /></td>
<td>Aspergillus kawachii arabinofuranosidase</td>
<td>2D44</td>
</tr>
<tr>
<td><img src="image" alt="Mannuronic Acid (ManA)" /></td>
<td>Fibroblast Growth Factors</td>
<td>2AXM, 1BFC etc.</td>
</tr>
</tbody>
</table>

Figure 5.6 Chemical and Structural Diversity of Glycans Binding to β-trefoil Proteins
Furthermore, not all proteins that are constituted of folds prominent in GBPs -- such as beta-trefoil -- are necessarily glycan binding. In fact, a simple classification of proteins with the beta-trefoil fold into glycan-binding and non-glycan binding shows that both classes of proteins are quite prominent in this fold (Figure 5.7). Hence, in addition to the array of diverse chemical and structural glycan molecules that may have motivated the diversification of folds constituting GBPs, other non-glycan related factors may have also played a role in such evolutionary tinkering and widespread twilight zone in the folds majorly constituting GBPs.

<table>
<thead>
<tr>
<th>Non glycan-binding proteins with trefoil fold</th>
<th>Glycan-binding proteins with trefoil fold</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actin-binding structural protein (1dfc, 1hcd, 1hce)</td>
<td>Hydrolases: Abrin (1lab), Ebulin (1hwm, 1hwo, 1hwp), Xylanase (1sly, 1sav, 1lx, 1sy, 1nx, 1bx, 1kn, 1mm, 1mc9, 1v6u, 1v6v, 1v6w, 1v6x, 1v6y), GlycosyHydrolase (1ups), Glycosidase (2aai), Arabinofuranosidase (1wd3, 1wd4, 2d43, 2d44)</td>
</tr>
<tr>
<td>Hydrolase inhibitor (1ava, 1eyl, 1fzm, 1fno, 1r8n, 1r8o)</td>
<td>Serine protease inhibitor (1avu, 1avw, 1avx, 1ba7, 1bce, 1wbc, 2wbc, 4bc)</td>
</tr>
<tr>
<td>Unknown function protein (1j9f)</td>
<td>Unknown function protein (1f9j)</td>
</tr>
<tr>
<td>Interleukin-1beta (1hbb, 1ib, 1ir, 1ic, 1ebo, 1eq, 1erp, 1epb, 1jjs, 12h, 1md6, 1nx4, 1sx5, 1u4q, 1uf0, 1uwe, 1wpm, 1b1, 2b1a, 20a, 2rt, 2mib, 2mnh, 3bi, 4bi, 5bi, 5ib, 6ib, 7ib1, 8ib1, 9ib, 10j)</td>
<td>Mannose receptor (1dag, 1dqj, 1f6w)</td>
</tr>
<tr>
<td>DNA-binding protein (1ttu, 2fco)</td>
<td>Trichosanthes lectin-1 (1ggp), Moiote lectin (1ce7, 1cm2, 1onk, 1oq, 1oq, 1p8, 1pum, 1puu, 1sz6, 1fim, 1fyl, 2ml8), Hemolytic lectin (1vdv)</td>
</tr>
<tr>
<td>Storage protein (1wba)</td>
<td>Agglutinin (1jx, 1jy, 1jzo)</td>
</tr>
</tbody>
</table>

The extensive presence of the twilight and midnight zones in GBPs presents numerous challenges to the currently available homology based modeling methodologies. In the twilight or midnight zones, the homology model is very likely to have significant inaccuracies (high RMSDs from actual crystal structures), owing to errors in the selection of an optimal template, alignment of the target to the template amino acid sequence, and sometimes even in the very identification of the fold of a given target sequence (Figure 5.8). However, the divide between known protein sequences and known protein structures has been rapidly expanding (Figure 5.8). Taken together, there is an increasing need for new homology modeling tools that are efficient in the twilight zone, for successful application to de novo GBP structure prediction.
Figure 5.8. (Top) Inaccurate structure prediction by current homology modeling tools in the twilight zone; and (Bottom) Rapidly expanding known sequence - structure divide.

5.5 Evolutionary Tinkering as Noise Masking De Novo Deduction of Fold Codes: The Signal-to-Noise Ratio of Molecular Biology

We rationalized that, in order to perform the several orders of magnitude many functions as the number of folds at her disposal, Nature would need to carefully encode the sequence-to-structural fold mapping information in factors governed by just a handful of residues for each fold. Given that conventional sequence (1-D) and structural (3-D) analysis do not generically decode this "protein fold code", we hypothesized that such fold-dictating "signal" features are masked by evolutionary "noise", and that in order to "denoise" the tinkered information and improve the signal-to-noise ratio (SNR), one would need to exclusively examine the slowest evolving regions in each protein fold (Figure 5.9).
5.6 The Core of it All: Shielding from Water as a Factor Mitigating Evolutionary Tinkering in Protein Folds

We sought to decode conserved features within each fold family despite the vast degrees of sequence divergence and rampant twilight zone (Figure 5.2), so as to better understand the factors governing the protein fold code. Given that the residues constituting the "core" of proteins are generally amongst the slowest evolving regions of protein structures and are known to be central to folding and unfolding of protein chains, we focused on the core of proteins to elucidate fold-conserved features.
At the heart of a stable protein domain, are the solvent-unexposed residues in its core (Figure 5.10). The identity and packing of protein core residues are known to be key factors that mediate both the energetics of folding and the emergence of fold families. The quality of protein core packing has also proven useful to successfully refine and validate computationally generated structural models. Recent studies have further examined specific families of proteins from sequence and packing/volume perspectives to delineate factors governing protein stability. While different methods have been used to identify core residues of protein structures, in the enclosed study we used conserved solvent inaccessibility to automate the identification of residues constituting the core (termed *scaffold*) of domains from fold family alignments.

![Figure 5.10 Protein core/scaffold (gray circles; black sidechain lines), surface residues (red circles; blue sidechain lines), peptide bonds (brown lines); and water molecules (blue circles)](image)

5.7 Mining Protein Contact Maps from the Multi-dimensional Protein Universe

While information content of the protein universe has traditionally been analyzed from the three dimensions of amino acid sequence (1-D), protein contact maps (2-D), and structural coordinates (3-D), with the advent of databases such as pfam, SCOP, DALI, VAST, and CATH in the *Post-Genomic Era*, analysis of higher dimensions of the protein universe such as protein family (4-D), super-family (5-D), fold/topology (6-D), and architecture (7-D), are emerging as
attractive options for mining of the truly multi-dimensional protein universe. Owing to the fact that residue contacts are fundamental to defining protein structural folds, in this study, we considered the information content of protein contact maps (PCMs) — a function of the distance between atoms of all amino acids in a protein structure (Figure 5.11).

Figure 5.11 Protein Contact Maps from architectures in the protein universe
5.8 Introduction to the \((p,w)\) Landscape of Protein Structural Folds: A Valuable Tool for Homology Modeling in the Twilight Zone

The enclosed manuscript describes the denoising of PCMs to compute protein core atomic interaction networks (PCAINs) as signature of protein folds. The key steps of the methodology are highlighted herein; for a more detailed description of the methods, refer to the enclosed manuscript that follows this section.

![Structure-based multiple sequence alignment of \(\beta\)-trefoil protein fold family, highlighting the residue columns that are identified as consistently solvent unexposed (blue).](image)

**Figure 5.12** Structure-based multiple sequence alignment of \(\beta\)-trefoil protein fold family, highlighting the residue columns that are identified as consistently solvent unexposed (blue).
Get target sequence (e.g. 1QXM)

Identify secondary structure regions

STEP 1

Look up fold-specific scaffold position propensity database

Identify scaffold residues

STEP 2

Look up fold-specific scaffold residue interactions database

Estimate PCAIN of target sequence

Choose template based on scaffold and PCAIN correlations

Proceed to target-template alignment

Figure 5.13 The PCAIN-based methodology for protein fold and structure prediction
Input target sequence (e.g. 1QXM), chosen template sequence (e.g. 1ISW) and identified scaffold residue positions

Compute scaffold-anchored sequence alignment

STEP 3

QXM
vflepnsntnkv4a---isqse--kklwnklsganqknlyldntnkqmklkmvmdntslitn-----nap1
1ISW
ggq--kvgvsgvctclynaastdgtgqglydchsatnqpr---tda--nervygdkc--svefgtnn---

1QXM
ssnvktntngdqyvllqnyisrnvi3ynmnpmvlvgynidtt----lmvstqtnssnqfkknsn
1ISW
thqiy6cwgdnqk6l1----nsdgyv-qvqglc-davvgtantlqiy6cscnganqrn----

Build target model

STEP 4

1QXM
vflepnsntnkv4a---isqse--kklwnklsganqknlyldntnkqmklkmvmdntslitn-----nap1
1ISW
ggq--kvgvsgvctclynaastdgtgqglydchsatnqpr---tda--nervygdkc--svefgtnn---

1QXM
ssnvktntngdqyvllqnyisrnvi3ynmnpmvlvgynidtt----lmvstqtnssnqfkknsn
1ISW
thqiy6cwgdnqk6l1----nsdgyv-qvqglc-davvgtantlqiy6cscnganqrn----

For test cases

Evaluate output target model by superposition of actual and modeled structures

RMSD: 1.9Å

Figure 5.13 The PCAIN-based methodology for protein fold and structure prediction
Figure 5.14. Examples of structures obtained with PCAIN approach to homology modeling (cyan) superposed onto the actual crystal structures (green), demonstrating the excellent structure prediction capabilities of the developed tool. It may be noted that these structures are all of GBPs, constituted of β-trefoil, β-prism, Concanavalin A-like, and C-type lectin-like folds.
CASP-6 (TO203) - Model with least Ca RMSD of 1.29 angstroms was TO203TS393

PCAIN-based model (cyan) has Ca RMSD of 0.91 angstroms to 1vkpB00 structure using template 1xknA00 with target-template sequence identity of 29%

CASP-6 (TO197) - Model with least Ca RMSD of 1.37 angstroms was TO197AL291

PCAIN-based model (cyan) has Ca RMSD of 0.87 angstroms to 1yemA00 structure (pink) using template 2dc4A00 with target-template sequence identity of 60%

Figure 5.15. Structures predicted based on PCAIN methodology for CASP (Critical Assessment of Structure Prediction) target sequences (top) TO203 and (bottom) TO197 are provided (cyan) superposed on original crystal structures (pink), illustrating the universal potential of PCAIN-based structure prediction.
Published manuscript


*Specific contributions of this Thesis research to the publication: Conceived and developed the PCAIN method for de novo protein fold identification and structure prediction.*
Atomic Interaction Networks in the Core of Protein Domains and Their Native Folds

Venkataramanan Soundararajan, Rahul Raman, S. Raguram, V. Sasisekharan, Ram Sasisekharan*

Harvard-MIT Division of Health Sciences & Technology, Koch Institute for Integrative Cancer Research and Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

Abstract

Vastly divergent sequences populate a majority of protein folds. In the quest to identify features that are conserved within protein domains belonging to the same fold, we set out to examine the entire protein universe on a fold-by-fold basis. We report that the atomic interaction network in the solvent-unexposed core of protein domains are fold-conserved, extraordinary sequence divergence notwithstanding. Further, we find that this feature, termed protein core atomic interaction network (or PCAIN) is significantly distinguishable across different folds, thus appearing to be "signature" of a domain's native fold. As part of this study, we computed the PCAINs for 8698 representative protein domains from families across the 1018 known protein folds to construct our seed database and an automated framework was developed for PCAIN-based characterization of the protein fold universe. A test set of randomly selected domains that are not in the seed database was classified with over 97% accuracy, independent of sequence divergence. As an application of this novel fold signature, a PCAIN-based scoring scheme was developed for comparative (homology-based) structure prediction, with 1-2 angstroms (mean 1.61Å) C<sub>S</sub> RMSD generally observed between computed structures and reference crystal structures. Our results are consistent across the full spectrum of test domains including those from recent CASP experiments and most notably in the 'twilight' and 'midnight' zones wherein <30% and <10% target-template sequence identity prevails (mean twilight RMSD of 1.69Å). We further demonstrate the utility of the PCAIN protocol to derive biological insight into protein structure-function relationships, by modeling the structure of the YopM effector novel ES ligase (NEL) domain from plague-causing bacterium Yersinia Pestis and discussing its implications for host adaptive and innate immune modulation. Further, we find that PCAINs are well-conserved between domains of the same fold family, while significantly different from the PCAINs of different families. We suggest that the PCAIN is a fundamental fold feature that could be a valuable addition to the arsenal of protein modeling and analysis tools.

Introduction

Nature employs merely a few thousand protein folds to generate the entire repertoire of the multimillion strong protein universe [1]. Massively divergent amino acid sequences thus populate protein families of many folds (Figure S1), ostensibly challenging the notion that all information dictating fold mapping of proteins the protein fold code is programmed in the sequence [2,3]. We sought to decode conserved features within each fold family despite the vast degrees of sequence divergence, so as to better understand the factors governing the protein fold code. Given that the residues constituting the core are generally amongst the slowest evolving regions of protein structures [6] and are central to folding [5] and unfolding [6], we focused on the core of proteins to elucidate fold-conserved features.

At the heart of a stable protein domain, are the solvent-unexposed residues in its core [7,8]. The identity and packing of protein core residues are known to be key factors that mediate both the energetics of folding [9] and the emergence of fold families [10]. The quality of protein core packing has also proven fundamental to defining protein folds, in this study, we considered the information content of protein contact maps (PCMs) a function of the distance between atoms of all amino acids in a protein [11]. Further, in order to capture the information content in the solvent-unexposed core regions of protein structures, we defined the protein core atomic interaction network or PCAIN (Figure 1).

While different methods have been used to identify core residues of protein structures [7-14], we used conserved solvent inaccessibility as a metric to automate the identification of residues constituting the core of domains from protein family alignments. We then classified these residues to characterize each fold and compute a database of PCAINs. We find that PCAINs are well-conserved between domains of the same fold family, while significantly different from the PCAINs for domains of other fold families. The fold-specific structurated models [11]. Recent studies have further examined specific families of proteins from sequence and packing/volume perspectives to delineate factors governing protein stability [12,13]. Owing to the fact that atomic interactions are fundamental to defining protein folds, in this study, we considered the information content of protein contact maps (PCMs) a function of the distance between atoms of all amino acids in a protein [14]. Further, in order to capture the information content in the solvent-unexposed core regions of protein structures, we defined the protein core atomic interaction network or PCAIN (Figure 1). While different methods have been used to identify core residues of protein structures [7-14], we used conserved solvent inaccessibility as a metric to automate the identification of residues constituting the core of domains from protein family alignments. We then classified these residues to characterize each fold and compute a database of PCAINs. We find that PCAINs are well-conserved between domains of the same fold family, while significantly different from the PCAINs for domains of other fold families. The fold-specific
Figure 1. Computation of the protein core atomic interaction network (PCAIN) from the 2-D protein contact map (PCM). The PCM accounts for all atomic interactions in the 3-D protein structure while the PCAIN involves atomic interactions between just the conserved, solvent inaccessible residues in the 'core' of protein domains.

doi:10.1371/journal.pone.0009391.g001
nature of PCAINS is further found to be consistent across families from the entire universe of protein folds (numbering \(\sim 10^{18}\)), highlighting the PCAIN as "signature" of the native folded state of protein domains. Building on the fold-specific nature of PCAINS, we demonstrate the use of PCAIN-based scoring schemes for effective classification of protein sequences into their native folds and for high-throughput, accurate homology-based (comparative) protein structure prediction. We further highlight the potency of PCAINS for extending the current capabilities of homology modeling into the 'twilight' and 'midnight' zones \([15,16]\) of low target-template sequence identity (<30% and <10% respectively), including those from recent GASP experiments \([17]\). Having verified the utility of PCAINS, we proceed to estimate the sensitivity of PCAINS to threshold interaction distance (\(p\)) and conserved solvent accessibility (\(o\)) the two fundamental physical parameters that characterize the PCAIN thus defining a \((p, o)\) landscape for protein structures. From this analysis, we find that the PCAIN is most refined around specific windows of \((p, o)\) values and propose an adaptive approach for maximizing the fold signature "signal" to evolutionary sequence divergence "noise", thus enabling effective parameter-tuning of PCAINS for applications to derive biological insight into protein structure-function relationships. Finally, we showcase as an application of the developed protocols, PCAIN-based modeling of the hitherto unknown structure of the NEL domain from the YopM effector protein of plague-causative bacterium *Yersinia pestis*. We conclude with discussions on the biological implications of the modeled bacterial protein structure, especially from the perspective of adaptive and innate immune signaling modulation during host-pathogen interplay.

**Results and Discussion**

We used the CATH database \([18]\) as the source for our data on protein domains and their folds. At the time when this study was performed, the CATH database (Figure S2) had 112,450 protein domains classified into 1,018 folds. We chose 8,698 protein domains from across the 1,018 folds representing all the different homologous superfamilies in CATH to seed our database. The structure-based multiple sequence alignments for the seeded domains were obtained from DHS \([18]\) and conserved, solvent-unexposed core columns were identified for each alignment (Figure S3). Using the solvent accessibility parameters from DSSP/CATH-wolf \([19,21]\) for constructing the PCAIN database from the CATH database (Figure S4) as described in the methods section. As part of the PCAIN database, we developed a comprehensive framework to document key conserved interactions for each family of the protein universe was developed (Figure 2), permitting assignment of PCAIN scores to threaded structures.

In order to investigate the fold-specificity of PCAINS and contrast with that of PCMs, the averaged PCM and PCAIN scores for the seed domains from each of the 1018 folds were computed. The averaged PCM and PCAIN scores for all fold pairs were cross-correlated to obtain the correlation coefficients that provide for a quantitative estimate of variation in these scores for different folds (non-diagonal entries; Figure S5). The average degree of correlation in PCMs and PCAINS were also computed for each family, providing a quantitative estimate of the degree of fold-conservation for these scores (diagonal entries from top left to bottom right; Figure S5). From this data, it is clear that the PCM provides for no discernable fold-specificity owing to random correlations within (diagonal) and across (non-diagonal) folds. On the other hand, it is evident that the PCAIN is highly fold-specific with low inter-family correlation coefficient values (non-diagonal) and high intra-family correlation coefficient values (diagonal). In order to better illustrate this point, the PCM and PCAIN scores for several randomly selected fold families from architectures spanning a significant portion of the protein universe is also shown (Figure 3), from which the extremely high fold-specificity of PCAINS and low fold-specificity of PCMs is evident.

Given that the PCMs and PCAINS are functions of the threshold interaction distance (\(p\)) and conserved solvent accessibility (\(o\)) parameters, the entire analysis was repeated for various threshold interaction distances ranging from \(p = 3.5\) to \(p = 10\), and conserved solvent accessibility cutoffs ranging from \(o = 0\) to \(o = 10\), to observe consistently higher fold-specificity for PCAINS than PCMs (data not shown). This analysis suggests that despite the large degree of sequence divergence in a majority of fold families, atomic interactions between amino acids in the solvent-unexposed core of domains (PCAINS) are a highly fold-conserved feature. The poor fold-specificity of the PCM on the other hand, is tell-tale of high evolutionary tinkering noise \([22]\) drowning out the fold-conserved atomic interaction signals. Thus, it emerges that PCMs have high signal-to-noise (SNR) ratio and that the solvent accessibility parameter \((o)\) sifts out the function-driven evolutionary tinkering noise from PCMs. This implies that PCAINS are "de-noised filtrates" of PCMs - a result that corroborates the long-standing notion that exposure to solvent correlates with evolutionary-driven amino acid substitution \([23]\). Furthermore, from the perspective of 2-D and 3-D realms, this analysis suggests that solvent exposed atomic interactions are more liable to evolutionary tinkering than are solvent unexposed (buried) atomic interactions.

In order to examine the fold discriminating efficacy of PCAINS and PCMs with greater detail, a general screen of 50,000 randomly selected domains was considered from the universal set of 112,450 domains excluding the 8,698 representative domains from which the seed databases were constructed. While the PCAIN showed 97% accurate classification, the PCM showed only 14% accuracy in classification of domains into their respective folds (Figure 4A). Furthermore, the PCM's ability to classify folds was found to be heavily dependent on the target-template pairwise sequence identity (PSI), with an exponential decrease in classification accuracy with decrease in PSI (Figure 4B). It must be noted that in the higher PSI realm (>50%) wherein the PCM shows some marginal performance, sequence-based (1-D) methods are known to perform significantly well \([24]\) and the utility of the 2-D PCM based approach is defeated owing to the higher computational cost involved. On the other hand, the PCAIN is found to be largely uninfluenced by the drop in PSI and consistently shows over 95% fold-classification accuracy even in the twilight (<30% PSI) and midnight zones (<10% PSI) (Figure 4B). This analysis showcases the 2-D PCAIN as a useful tool to add to the existing methods for protein fold recognition such as profile pattern recognition and protein threading \([25-28]\).

While some existing methods are able to recognize folds accurately \([25-27]\), there is still an unmet need for methods that can proceed from fold recognition towards accurate homology-based structure prediction \([28]\) in the 'twilight' and 'midnight' zones wherein target-template sequence identity are <30% and <10% respectively \([15,16]\). Furthermore, this breakdown of homology modeling utility with low target-template identity challenges elucidation of structures for newly discovered proteins, several of which happen to fall into the twilight and midnight zones \([26,29,30]\). To address this issue, we systematically evaluated the potency of the PCAIN approach for homology-based structure prediction, motivated by the high fold-specificity of PCAINS. For this purpose, we developed a PCAIN-based scoring scheme (Figure S6) outlined in the methods section for template
Figure 2. Snapshots from the PCAIN database used for mining fold-distinguishing signatures. The solvent inaccessible core of domains (shaded brown) from all 1018 naturally occurring folds were identified and used to compute the PCAINs (as described in the methods section) as part of the PCAIN database. Shown herein are representative domains and PCAINs (with yellow arrow between) from the following fold families—(A.) Orthogonal α-bundle (DNA helicase RuvA subunit); (B.) Up-down α-bundle (coiled-coil); (C.) α-horseshoe (leucine-rich repeat variant); (D.) α-solenoid (peridinin-chlorophyll protein); (E.) α-barrel (glycosyltransferase); (F.) αβ-roll (HIV reverse transcriptase); (G.) αβ-complex (cytochrome); (H.) αβ-box (proliferating cell nuclear antigen); (I.) β-ribbon (seminal fluid protein PDC-109); (J.) β-sandwich (neurophysin); (K.) β-barrel (thrombin); (L.) β-propeller (pseudo β-propeller); (M.) β-clam (outer membrane lipoprotein receptor); (N.) β-trefoil (acidic fibroblast growth factor). Fold-distinguishing PCAIN patterns observed herein motivated systemic computation of intra-fold and inter-fold correlations on a family-by-family basis, as shown in supplementary figure S5. Fold-conserved interactions are evolutionary markers and are demarcated (red stars) on the corresponding sample set of the protein family alignments in supplementary figure S3.

doi:10.1371/journal.pone.0009391.g002
Figure 3. Contrasting the fold specificity of protein contact maps (PCMs) and protein core atomic interaction networks (PCAINs).

Averaged intra-family (diagonal) and inter-family (non-diagonal) correlation coefficients of (A.) PCMs and (B.) PCAINs were computed at 5 angstroms threshold distance \( p \) and normalized solvent accessibility/atom of \( \omega = 10 \) on a family-by-family basis for several prominent folds of the protein universe. The complete 1018 folds by 1018 folds correlations of PCMs and PCAINs for the entire fold universe is shown in supplementary figure S5.

From these figures it is clear that PCAIN is highly fold-specific but PCM shows no discernible fold specificity.

doi:10.1371/journal.pone.0009391.g003
Figure 4. Applications of PCAIN as a divergence-independent metric for protein classification, anchored sequence alignment, and structure prediction. (A) PCAINs were computed on a general screen of unselected protein domain sequences that were not part of the database and used to accurately classify these sequences as shown, confirming the fold-specific nature of PCAINs. PCMs of these domains are seen to be ineffective as classifiers in the general sequence space. (B) PCAIN is seen to be an effective classifier regardless of the sequence identity of the target domain towards members of its native fold and is observed to be effective even in the twilight (<0.30% PSI) and midnight (<10% PSI) zones. On the other hand, the PCM is observed to be highly dependent on this sequence identity and provides for some moderate classification accuracy only in the high sequence identity range. (C) The distribution of RMSD between PFCAIN-based predicted structures and the reference crystal structures for target sequences with mean RMSD of 1.61A highlights the structure prediction efficacy of the proposed method. (D) Pie chart of RMSD distribution for test sequences in the twilight and midnight zones is shown, indicating mean RMSD of 1.69A.

doi:10.1371/journal.pone.0009391.g004

selection, anchored sequence alignment, and homology-based structure prediction. This testing was performed with a general screen of randomly selected domains from the universal set of domains, excluding the representative domains of the seed database, and including those from recent CASP experiments.

The reference structure-based sequence alignments were seen to have extremely high correlations to the PCAIN-based anchored alignments with pearson's correlation coefficient of 0.91. It is interesting that atomic interactions are mined from 3-D structural coordinates and 2-D PCAINs are used to identify the fold-conserved set of atomic interactions that are finally mapped to thread 1-D amino acid sequences. This underscores the application of fold-conserved (including in twilight and midnight zones) higher dimensional data from structural (3D) and contact (2D) spaces for effective protein analysis. This also establishes that PCAIN-based anchored alignments closely mimic the actual structure-based sequence alignments, thus confirming the utility of PCAINs vis-a-vis sequence alignment. Furthermore, superposition of the modeled test structures onto the reference crystal structures demonstrated good structure prediction accuracy in the range of 1-2 angstroms, with mean RMSD of 1.61 angstroms (Figure 4C). In order to specifically estimate the efficacy of the PCAIN approach for structure prediction in the twilight and midnight zones of sequence identity, the RMSD range for the predicted
structures corresponding to the test domains in these zones was also computed (Figure 4D). The mean RMSD in the twilight and midnight zone was 1.69 angstroms with the overall RMSD distribution (Figure 4D) very similar to that obtained for the entire set of test domains (Figure 4C), thus proving that the PCAIN approach to structure prediction is sequence-identity-independent and hence notably potent in twilight-midnight zones. Successful prediction of structures for example targets from recent CASP (critical assessment of structure prediction) proceedings that are not part of the CATH database or the seed datasets further illustrate the generic, database-independent efficacy of the PCAIN approach (Figure S7). This analysis confirms the high-throughput accuracy of PCAIN-based structure prediction and showcases it as a valuable addition to the arsenal of structural modeling tools.

The significantly improved performance of PCAINs over PCMs [31] is due to three distinct advantages. Primarily, owing to denoising of "evolutionary tinkered" contacts from the PCM, the PCAIN enables exclusive retention of fold-specific signals. Next, the PCAIN scores for sequences generally best match with the representative domains from the same superfamily, rather than domains of other superfamilies even belonging to the same fold. Since protein folds are classified into superfamilies based on common functions and evolutionary relationships, it is likely that the PCAIN-based methodology enables handpicking of an optimal functionally-related model molecule for modeling the structure of the unknown protein, thus contributing significantly towards improving the accuracy of structure prediction. Finally, the PCAIN methodology provides for utilizing the fold-conserved residues as "anchors" in the target-template sequence alignment step, thus increasing efficacy of conventional alignment protocols. Taken together, these three factors contribute towards the potency of PCAINs for the discussed applications. With further improvements to the accuracy of secondary structure prediction methods and incorporation of additional fold-conserved features from solvent-exposed regions, it is conceivable that more accurate structures may be predicted as part of future advancements to the PCAIN methodology.

Given that the PCAIN is a function of two fundamental parameters, namely, threshold interaction distance (p) and conserved solvent accessibility (o), we investigated the effect of modulating these parameters (Figure 5). For this purpose, a parameter scan on (p, o) was performed and the effective operable landscape for PCAIN-based methods was mapped for the range p = 3.5 - 7.0 angstroms and o = 0 - 40%. Given that high intra-family PCAIN correlation scores and low inter-family PCAIN correlation scores are necessary for defining a refined fold signature with high SNR, the difference between these two scores provides a reliable measure of potency. We find that the PCAIN is sensitive to both the threshold interaction distance parameter (p) and the conserved solvent accessibility parameter (o), with higher sensitivity towards the former (Figure 5A). Specifically, the PCAIN is found to be most effective as a fold signature (high intra-family and low inter-family correlations) in the window p = 2 - 20% (Figure 5B) and similarly in the window p = 4.0 - 4.5 angstroms (Figure 5C).

The (p, o) landscape may be interpreted as follows. Protein structures are ensembles of backbone bonded dipeptide confirmations that are characterized by the (φ, ψ) plot [32, 34] and other side-chain interactions that are characterized by inter-residue distance [35]. Too much threshold interaction distance (p) implies accounting for non-influential residue pairs as interactions and such pseudo-interactions will add to the noise thus decreasing SNR and PCAIN potency. Too little threshold interaction distance (p), on the other hand, is not feasible, since it will be less than interatomic Van der Waals distances. The 'o' parameter accounts for the interplay between water molecules and the residues constituting the protein structure and from this perspective the PCAIN may be viewed as essentially the solvent unexposed network (SUN) of interacting residues. Specifically, a higher 'o' value implies accounting for partially solvent exposed (and hence possibly non-conserved) atomic interaction networks, thus adding to the noise factor and decreasing PCAIN effectiveness. A 'o' value close to zero, on the other hand, may be too stringent. Along the lines of this analysis, it is conceivable that fine-tuning of the PCAIN may be required for specific molecular biology applications. Having mapped the effective operable landscape for PCAIN-based methods with the goal of obtaining the maximal PCAIN effectiveness and highest possible SNR, we propose an adaptive framework (Figure 5D) for such fine-tuning of the (p, o) parameters as required by the application of interest.

Protein fold recognition and structure prediction have numerous biological applications [28, 30]. In addition to the previously demonstrated applications of sequence alignment, fold identification, template selection, and homology modeling, we demonstrate herein, application of the described PCAIN-based structure prediction methodology to derive biological insight into potential structure-function relationships of proteins with hitherto unresolved structure. As an example to highlight this application, we consider the effector protein YopM from the plague-causing bacterium Yersinia pestis [36]. While it is well-known that YopM is a critical virulence determining factor, structural insight into potential roles of YopM in F. pestis pathogenesis has been elusive, due to the unsolved structure of the YopM novel E3 ligase (NEL) domain [37].

We modeled the YopM NEL domain structure using the PCAIN methodology and investigated the putative ubiquitin ligase catalytic site (Figure 6A). From the modeled structure, we note remarkable correlation in molecular surface electrostatics including the highly-conserved patches (Figure 6B), in NEL domain structures from Salmonella TyphH [30], Salmonella SifB [39], Shigella IpaH [40], and Terrus pestis YopM, in addition to high correlation of the PCAINs for these domains (Figure 6C). Given that these patches constitute the NEL catalytic site [40] and the recently characterized Salmonella NEL domain interaction sites with human leukocyte antigen-DR (HLA-DR; a major histocompatibility complex (MHC) class II receptor) and thioredoxin (TRX) [38, 39], it is likely that the YopM NEL domain functions as an autoregulated E3 ubiquitin ligase and degrades human intracellular proteins, similar to NEL domains from Salmonella and Shigella. Such an ubiquitinase activity of YopM NEL has significant implications for modulation of host adaptive and innate immune response to plague (Figure 6D). The ubiquitination and subsequent degradation of HLA-DR by Salmonella effectors within antigen presenting cells like macrophages, B-cells, and dendritic cells, has been recently shown to diminish the surface expression of MHC class II antigens [41]. It is conceivable that a similar interaction of YopM NEL with HLA-DR could moderate the host adaptive immune response (Figure 6D). Confirmation of the proteolytic degradation of TRX by YopM will have important implications in the regulation of mitogen-activated protein kinase kinase kinase 5 (MAP3K5) signaling, for TRX interaction with MAP3K5 [43] provides YopM a plausible direct method to modulate innate immunity (Figure 6D). More specifically, future studies that biochemically characterize interactions of key host intracellular molecules to the YopM molecule modeled herein, will further our understanding of the specific mechanisms governing bacterial subversion of human adaptive and innate immune signaling pathways.

The Protein Fold Code
The modeling of YopM NEL domain demonstrated in this study amply highlights application of the PCAIN methodology to derive biological insight into protein structure-function relationships. Taken together with the previously described applications of the PCAIN methodology such as sequence alignment, fold identification, template selection, and structural modeling, our
Figure 6. Application of the PCAIN methodology to analyze potential structure-function relationships of the novel E3 ligase (NEL) domain from the YopM effector protein of the plague-causing bacterium *Yersinia Pestis*. (A) The YopM NEL domain structure was modeled using the PCAIN methodology and the putative ubiquitin ligase catalytic site was characterized, based on the recent experimental characterizations of *Salmonella* and *Shigella* NEL domains (38–41). The likely hydrogen bonds that stabilize the active site (black lines) and the key α-helices (H4, H7, and H9) are indicated. (B) Vacuum electrostatics of the molecular surfaces from superposed NEL domains of YopM, SspH2, and IPA9.1 were generated (see Methods) with negative, positive, and neutral patches colored red, blue, and white respectively. The finger-like extension (pink line), globular domain (orange arc), and active site location (black arrow) are indicated. (C) The solvent-unexposed residues that constitute the PCAIN of the modeled YopM NEL domain structure (gray) are shown as sticks (brown). The molecular surface of the YopM NEL domain is also shown alongside to highlight that the residues constituting the PCAIN (brown) are only very minimally solvent exposed. (D) This is a pictorial depiction of YopM in the intracellular context and the key structural implications for its modulation of human adaptive and innate immune signaling. Specifically, YopM is known to interact with protein kinase C-like 2 (PRK2) and ribosomal 56 protein kinase 1 (RSK1) resulting in increased activity and mobility of these kinases, in addition to potentiating natural killer (NK) cell depletion by suppressing expression of Interleukin-15 (IL-15) (37). YopM has also been shown to specifically interact with α1-antitrypsin (AAT) without affecting its anti-protease activity, due to which the biological significance of this interaction remains unknown (37). Also indicated by the question mark (?) symbols are hitherto unknown interactions for YopM, extrapolated based on the functions of the related proteins. Specifically highlighted in this regard are the degradation of human leukocyte antigen-DR (HLA-DR) and thioredoxin (TRX) that may cause suppression of adaptive immune response via modulation of antigen presentation and modulation of innate immune signaling via the MAPK cascade, respectively. It remains to be seen what precise intracellular molecules are targeted by YopM NEL for proteolytic degradation, considering the autoregulated ubiquitin ligase activity suggested by our PCAIN-based model and analysis.

doi:10.1371/journal.pone.0009391.g006
Study confirms the PCAIN as a fundamental fold feature that will be valuable addition to the arsenal of protein modeling and analysis tools. Additionally, the PCAINs computed as part of this work (such as those from the database shown in Figure 2) are likely to be useful resource for molecular engineering applications since they provide a rigorous starting framework or scaffold upon which rest of the protein design may be tailored based on the functions of interest. PCAIN computation and analysis may also be valuable for applications such as elucidating mechanisms of protein evolution, stability, folding, unfolding, and misfolding, given the central role of the protein core in governing these phenomena [43-46].

It has recently been shown that two specific amino acid sequences with overwhelming identity (~88%) adopt distinct folds, thus postulating that for the specific protein pair considered, only ~12% of the amino acid sequence codes for sequence-to-structure mapping [47]. PCAIN shreds light on a "fold code" that is consistently encoded into residues that constitute the networks of atomic interactions. The core region unexposed core regions of protein native structures. This suggests that the fold code is a network phenomenon along with sequence and structural phenomena, thus providing rationale as to why merely sequence-based or structure-based pattern analysis of proteins may not succeed in decoding fold signatures. The cores of the protein domains of the same fold as identified by our method can have low sequence identity and poor secondary structure motif matching, but high conservation of their PCAINS (Figure S8). Hence, defining protein cores based on treatment of protein structures as atomic networks characterized by the (p, w) plot and denoising of PCMS by recognition of signature network patterns, distinguishes our PCAIN methodology from the previously explored knowledge-based threading potentials. Our finding that the atomic interactions between just 15-20% of residues in native structures of each examined fold are conserved, further suggests that the PCAIN is a minimalistic fold code.

Finally, this study provides compelling evidence in support of Anfinsen's dogma [48] that information dictating the native structural fold of protein domains is encoded in its amino acid sequence. Herein we have shown that "a significant portion of the fold-dictating information is encoded by the atomic interaction network in the solvent-unexposed core of protein domains".

Materials and Methods

Automated Identification of 'Core' Residues and Construction of a Core Composition Database Characterizing All 1018 Folds of the Protein Universe

At the time when this study was performed, the CATH database [18] had 112,450 protein domains classified into 1,018 folds, from which 8,698 protein domains representing the different homologous super families were used to seed our database. CATH defines cores based on secondary structural element analysis, whereas in our method, the core can include non-secondary structural elements. Taken together with several other methodology distinctions, the cores identified by us are unique (as highlighted for the illustrative domain in Figure S8 for which more than 75% of CATH and PCAIN core residues are distinct). The structure-based multiple sequence alignments were obtained from DHS [18] (Figure S3) and the absolute solvent accessibility (ASA) factors from DSSP/CATH-wolf [21] were obtained for the amino acids of all 8,698 domains. The relative solvent accessibility (RSA) per atom was computed for each residue. The mean solvent accessibility (w) was then calculated for all columns of the seed alignments and a threshold was used to identify the consistently solvent-unexposed columns as shown (Figure S3). This set of consistently solvent inaccessible columns was mapped back onto the conserved residue positions thus defining the core for all the seeded protein domains from each alignment. This was compiled into a dataset of protein core residues, one corresponding to each protein family and each considered value of parameter w. The frequency of each amino acid at the core positions was also consolidated into a dataset of family-specific protein core residue propensities. The complete protein core characterization method, right from CATH mining until the construction of the datasets was automated with the implementation of a script in MATLAB 7.6.0 from The MathWorks, Inc. (Natick, MA).

Automated Construction of the PCM and PCAIN Databases for All 1018 Folds of the Protein Universe

A MATLAB script was written to automate the computation of protein contact maps (PCMs) for all seeded domains of the 1018 folds at various threshold interaction distance parameter (p) values (Figure S4). This was compiled into a database of PCMs on a fold-by-fold basis. The previously identified core residues for each domain of each fold at various w values was used to identify the rows and columns of interest from PCMs at various p values and these were concatenated into the corresponding PCAINs for each domain of each fold at various (p, w) values, as depicted pictorially (Figure 1). This step was automated with a MATLAB script, which was also ultimately used to compile the generated PCAMS into an integrated PCM-PCAIN database for various (p, w) values. A simple python script was written and executed in PyMol for visualization of all the protein cores and PCAINS shown in this study (Figure 2). The pearson's correlation coefficient was computed to quantitatively contrast PCMs and PCAINS both within and across all 1018 folds (Figure S5) and across 15 unselected folds for refined visualization purposes (Figure S3).

Automated Fold Classification of Randomly Selected Domains from the Protein Universe

A general screen of 50,000 randomly selected domains (obtained from the set of 112,450 domains excluding the 8,698 representative domains in the training set from which the PCM and PCAIN databases were constructed) were considered for testing the fold classification efficacy of PCAIN-based and PCM-based scoring schemes. The effectiveness of the classification approaches were then estimated (Figure 4A) using the actual folds of the test sequences as reference. Variations of the classification efficacies as a function of target-template sequence identity were also computed (Figure 4B).

Template Selection Based on Target PCAIN Estimation and Correlation with Protein Family PCAIN Signatures

An automated MATLAB script was written to compute the secondary structures of the target amino acid sequences based on secondary structure prediction consensus [49-51]. The type, quantity and distribution of secondary structures are partially characteristic of folds and offer a good first filter for the fold and template selection process. Potential amino acids that correlate with the propensity data for each core position of all the screened folds are then identified for the target sequences, providing an estimate of 'core fit' and serving as a second filter for fold and template selection. The algorithm for this step is also implemented in MATLAB 7.6.0 from The MathWorks, Inc (Natick, MA) and accepts three inputs, namely, target amino acid sequences, the corresponding secondary structural information, and the fold-specific core residue propensity dataset. The target sequences for
which all potential core residues are identifiable are deemed ‘core fit’ with respect to the screened folds and the target PCAIN scores for these are computed using the PCAIN database. For a majority of cases, the identical residue pairs are present in the database and hence their corresponding pairwise score is directly utilized. In other cases, an average of pairwise interactions between the two considered core positions from all other members of the screened fold family is used in this step. The target PCAIN scores are subsequently back-correlated with the averaged PCAIN score of each family and the resulting correlation coefficients provide an additional estimate of the degree of ‘core fit’. A simple threshold step is used at this stage as the third and final filter to determine the protein family, thus providing for selection of the optimal template molecule.

Automated Anchored Sequence Alignment and Comparative Structure Prediction for Randomly Selected Protein Domains

The steps of this algorithm are depicted as a flowchart in Figure S6. Briefly, a general screen of randomly selected domains were obtained from the set of 112,450 domains (excluding 6,698 representative domains for which PCM and PCAIN databases were constructed) and their PCAINs were estimated as detailed above. The computed target PCAIN scores were then correlated with the PCAIN scores (from the seed database) of every representative homologous superfamly member of the identified fold family in order to compute the optimal template, based on similarities at the level of evolutionary origin and function. The corresponding scaffold residues of the target and template sequences are then ‘anchored’ and pairwise sequence fragments between subsequent anchors are aligned using standard functions from the MATLAB bioinformatics toolbox with the BLOSUM62 scoring matrix and default gap penalties. The process involving fold identification, template selection and anchored alignment is maximally automated with the design of a MATLAB-based model. The structure-based sequence alignments are correlated with the PCAIN-based anchored alignments to estimate the efficacy of the PCAIN approach to sequence alignment (Figure 4C). Once the optimal anchored target-template alignments were computed, these were input to the automated homology modeling script of Discovery Studio from Accelrys, Inc. (San Diego, CA) that uses standard force fields to determine the energy minimized 3-D structural coordinates for the test sequences, including those from recent CASP experiments (as illustrated by examples in Figure S7). Each modeled 3D structure was then superposed onto the actual crystal structure obtained from the PDB using an automated MATLAB function and the root mean square deviations upon superposition were computed (Figure 4D).

Modeling NEL Domain Structures with the PCAIN Methodology and Analysis of Their Putative Structure-Function Relationships

The molecular structures of NEL domains from *Yersinia pestis* YopM (NCBI Reference Sequence: ZP_02316950.1) and *Salmonella typhimurium* Shp (GenBank: AAD39928.1) were modeled using the described PCAIN methodology with the identified optimal template structure of *Staphylococcus aureus* effector IpaH (PDB ID: 3CKD). All structure-function relationship analysis, including vacuum electrostatics generation for the modeled *Yersinia pestis* YopM NEL, modeled *Salmonella typhimurium* Shp NEL, crystal structures from *Staphylococcus aureus* IpaH NEL (PDB ID: 3CKD), and *Salmonella* SapH2 NEL (PDB ID: 3G06), were performed with PyMol.

Supporting Information

**Figure S1** Evolutionary sequence divergence of protein families. More than 60% of protein families from the Pfam database were found to be significantly divergent in their sequences (High range), around 30% of protein families were found to be moderately divergent in their sequences (Medium range) and less than 10% of protein families were found to be well conserved in their sequences (Low range). This shows that evolutionary tinkering and sequence divergence are rampant across the protein universe.

Found at: doi:10.1371/journal.pone.0009391.s001 (0.14 MB JPEG)

**Figure S2** The diversity of protein folds. Representative protein domains from CATH showcasing the fold diversity, classified according to their class (mainly α/β/αββ) and architecture. Found at: doi:10.1371/journal.pone.0009391.s002 (0.11 MB JPEG)

**Figure S3** Sample sets from fold family alignments highlighting the solvent-unexposed (core) conserved positions (blue columns). (A) Sample proteins from a family of the architecture - Orthogonal bundle. (B) Sample proteins from a family of the architecture - Up-down bundle. (C) Sample proteins from a family of the architecture - Alpha-horseshoe. (D) Sample proteins from a family of the architecture - Alpha-alpha Barrel. (E) Sample proteins from a family of the architecture - Beta-Barrel. (F) Sample proteins from a family of the architecture - Beta-Trefoil. (H) Sample proteins from a family of the architecture - Beta-Propeller. (K) Sample proteins from a family of architecture - αβ Roll. (L) Sample proteins from a family of architecture - αβ Box. (M) Sample proteins from a family of architecture - αβ Complex.

Found at: doi:10.1371/journal.pone.0009391.s003 (1.50 MB JPEG)

**Figure S4** A sample dataset from the protein contact maps (PCM) database. The inter-residue contact maps at 5 angstroms threshold distance are shown for representative domains from a diverse set of topologies/folds spanning all natural architectures in the protein universe.

Found at: doi:10.1371/journal.pone.0009391.s004 (0.19 MB JPEG)

**Figure S5** Protein contact maps (PCMs) versus protein core atomic interaction networks (PCAINs) intra- and inter- fold family correlations reveal striking specificity for PCAIN across the universe of folds. Averaged intra-fold (diagonal) and inter-fold (non-diagonal) correlation coefficients of (a.) PCMs and (b.) PCAINs at 5 angstroms threshold, shows cleas that the PCAIN is highly fold-specific whereas the PCM shows no discernible fold specificity.

Found at: doi:10.1371/journal.pone.0009391.s005 (0.43 MB JPEG)

**Figure S6** Flowchart governing PCAIN-based fold recognition of target sequence, template selection, anchored target-template alignment, and homology-based structure prediction. The detailed procedures associated with each step are described in the methods section. Briefly, a combination of secondary structure distribution and PCAIN scores from the key interaction positions was used to (i.) identify the fold of the target sequence, (ii.) compute the ideal template structure based on the closest functional homolog estimated from the superfamilies of the identified fold, (iii.) converge on the set of ‘anchor’ positions between the target and template sequences based on protein core amino acid frequencies to compute the optimal anchored target-template alignments, and (iv.) determine the target domain’s 3-D structural coordinates from...
the anchored alignments with an automated homology modeling script. 

Found at: doi:10.1371/journal.pone.0009391.s006 (0.05 MB JPG)

**Figure S7** Superposition of structures predicted based on PCAIN methodology for CASP (Critical Assessment of Structure Prediction) target sequences (a) TO203 and (b) TO197, illustrates PCAIN-based structure prediction. PCAIN-based structures predicted (cyan) are superposed onto reference crystal structures (pink) for (a) TOP203 and (b) TO197 from CASP6 with RMSDs of 0.51A (at 29% target-template sequence identity) and 0.87A (at 60% target-template sequence identity) respectively. The corresponding results of structure prediction accuracy from the CASP models shows as tables shows minimum RMSDs of 1.29A and 1.37A respectively.

Found at: doi:10.1371/journal.pone.0009391.s007 (0.26 MB JPG)

**Figure S8** Defining protein cores and extracting their information with the PCAIN methodology. (A.) Polar and charged residues (yellow) are also part of the core of protein domain as identified by our method, as shown with E.coli thioredoxin (cyan) as an example. (B.) Only 7% identity (shaded green) is present in the sequence of residues that constitute the core of glutaredoxin and thioredoxin that adopt the same fold, whereas 93% of the core residues are different in identity (shaded yellow). However, the PCAINs of these two proteins are seen to have 58% correlation, over the PCMs that have only 41% correlation. This example further illustrates that the identity or hydrophobicity of residues are poor tools for extracting information from protein cores, whereas the PCAIN is optimal for extracting conserved information from protein cores. Similarly, very poor overlap is seen between residues used for CATH alignments (underlined) and the residues that contribute to the PCAIN, thus illustrating the novelty in determination of PCAIN residues.

Found at: doi:10.1371/journal.pone.0009391.s008 (0.07 MB JPG)

**Acknowledgments**

The authors thank Dr. Kaman Thanarakarnan for helpful discussions.

**References**


Bibliography


109. Doxorubicin Hydrochloride for Injection, USP. Pfizer Pharmaceuticals


260. Dierks, T; Schmidt, B; Borissenko, Lv; Peng, J; Preusser, A; Mariappan, M; Von, Figura, K (May 2003). "Multiple sulfatase deficiency is caused by mutations in the gene encoding the human C(alpha)-formylglycine generating enzyme". Cell 113 (4): 435–44.


