

# Freezing Transition of Heteropolymers

by

Vijay Satyanand Pande

Submitted to the Department of Physics  
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Physics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 1995

© Vijay Satyanand Pande, MCMXCV. All rights reserved.

The author hereby grants to MIT permission to reproduce and  
distribute publicly paper and electronic copies of this thesis  
document in whole or in part, and to grant others the right to do so.

Author .....

Department of Physics

July 26, 1995

Certified by .....

Toyoichi Tanaka  
Professor of Physics  
Thesis Supervisor

Accepted by .....

George Koster  
Chairman, Departmental Committee on Graduate Students

MASSACHUSETTS INSTITUTE  
OF TECHNOLOGY

SEP 26 1995

ARCHIVES

LIBRARIES



# Freezing Transition of Heteropolymers

by

Vijay Satyanand Pande

Submitted to the Department of Physics  
on August 1, 1995, in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy in Physics

## Abstract

The ability to create synthetic heteropolymers with the protein-like capabilities of renaturation to a particular conformation capable of specific molecular recognition is of great technological importance. To create protein-like heteropolymers, we suggest "Imprinting," which dictates that monomers should be equilibrated at some low temperature prior to polymerization, then polymerized such that this monomer prearrangement is somewhat preserved. We argue that this optimization of monomers to themselves and to the target molecule prior to polymerization leads respectively to the protein-like properties of stability and functionality since the prearrangement of monomers is analogous to optimization performed by nature over evolutionary time. Thus, many of our results are applicable to proteins and may shed light on their biological and prebiological origins.

To study Imprinting, we employ a variety of theoretical techniques. Since we must prove that the single heteropolymer conformation capable of specific molecular recognition dominates equilibrium, to computationally study the thermodynamics of Imprinted heteropolymers, we must enumerate every possible globular conformation; using a massively parallel supercomputer, we found the conformations in which the chains were polymerized were indeed the lowest energy conformation. We also validated Imprinting by the Monte Carlo kinetics of lattice models, thereby showing that the chains we proved to have the lowest energy conformations were able to kinetically access this conformation. Finally, we used replica mean field theory to examine the relationship between the nature of interactions chosen and the success of Imprinting by calculating the phase behavior for Imprinted heteropolymers directly from the microscopic Hamiltonian of monomer-monomer interactions; we have theoretically shown that Imprinting is feasible and indeed robust with respect to variations in several aspects, including the chemistry employed, interactions of monomers in the soup versus on the polymer, and the target molecule intended for recognition.

Thesis Supervisor: Toyochi Tanaka

Title: Professor of Physics

## Acknowledgments

It is clear that modern science is not only deeply enriched by extensive collaborations but almost requires it. Certainly my thesis is no exception to this and I am therefore deeply indebted to my collaborators and those who kindly gave me valuable feedback and comments. First and foremost, my work as well as my character as a scientist has been invaluablely molded due to the fortunate circumstance of having two advisors. In this manner, Profs. Toyo Tanaka and Shura Grosberg have not just helped me perform my research but have taught me something much more fundamental: how to do research. I am very grateful for the guidance, encouragement, and friendship.

As for other collaborators, I am grateful for Chris Joerg's contributions in coding of the enumeration. I would also like to acknowledge helpful discussions with and encouragement from Profs. Eugene Shakhnovich, Sasha Gutin, Mehran Kardar, and Gene Stanley. The computational aspects of this work would not be possible without the use of the CM-5 time, provided by Dr. Tom Greene and Project SCOUT. I would also like to thank Profs. Chris Turner and Tadashi Tokuhiro for their help with NMR. I feel very lucky to have worked in a "family-like" group and would like to acknowledge the various members who have helped me throughout my stay: Dr. Massahiko Annaka, Rose Du, Tony English, Dr. Michal Orkiz, Bill Robertson, Bernhard Schnur, Changnan Wang, Kevin Wasserman, Hua Yang, and Dr. Xiao Hong Yu. Finally, I would like to thank my parents and sister Nalini for their loving encouragement.



# Contents

|           |  |           |
|-----------|--|-----------|
| <b>I</b>  | <b>Introduction</b>  | <b>15</b> |
| <b>1</b>  | <b>Background</b>  | <b>17</b> |
| 1.1       | Polymer Basics . . . . .   | 17        |
| 1.2       | Molecular Biology Basics . . . . .                                   | 18        |
| 1.3       | Protein Folding . . . . .  | 19        |
| <b>2</b>  | <b>Imprinting</b>  | <b>21</b> |
| <b>II</b> | <b>Computational</b>   | <b>33</b> |
| <b>3</b>  | <b>Enumeration</b>   | <b>35</b> |
| <b>4</b>  | <b>Computer Simulation</b>   | <b>45</b> |
| 4.1       | Introduction . . . . .   | 45        |
| 4.2       | Description of the Model . . . . .                                   | 48        |
| 4.3       | Thermodynamics . . . . .   | 53        |
| 4.3.1     | Potts Interactions (27-mers and 36-mers) . . . . .                   | 53        |
| 4.3.2     | Potts Interactions: Polymer with Target Molecule (26-mers) . . . . . | 60        |
| 4.3.3     | Different Interactions (36-mers) . . . . .                           | 62        |
| 4.3.4     | Comparison to other polymer ensembles (36 mers) . . . . .            | 63        |
| 4.4       | Kinetics . . . . .   | 67        |
| 4.4.1     | Potts Interactions (27-mers) . . . . .                               | 70        |
| 4.4.2     | Polymer and Target Molecule Kinetics (26-mers) . . . . .             | 74        |

|            |   |            |
|------------|---|------------|
| 4.4.3      | Different Ensembles (36-mers, SMJ Matrix Interactions) . . .      | 75         |
| 4.5        | Monomer-monomer correlations along the polymer sequence . . . . . | 77         |
| 4.6        | Discussion . . . . .  | 80         |
| 4.7        | Conclusions . . . . .   | 83         |
| <b>III</b> | <b>Analytic</b>   | <b>85</b>  |
| <b>5</b>   | <b>Two Letter Designed</b>  | <b>87</b>  |
| 5.1        | Introduction . . . . .  | 87         |
| 5.2        | Formulation of the Model . . . . .                                | 88         |
| 5.3        | Replica Theory Analysis . . . . .                                 | 91         |
| 5.4        | Replica Symmetry Breaking . . . . .                               | 94         |
| 5.5        | Phase Diagram . . . . .   | 98         |
| 5.6        | Discussion . . . . .  | 102        |
| 5.7        | Conclusions . . . . .   | 106        |
| <b>6</b>   | <b>Random Heteropolymers</b>                                      | <b>107</b> |
| 6.1        | Introduction . . . . .  | 107        |
| 6.2        | Development of the Formalism . . . . .                            | 109        |
| 6.2.1      | The Model and its Hamiltonian . . . . .                           | 109        |
| 6.2.2      | Replicas . . . . .  | 111        |
| 6.2.3      | Effective Energy in Replica Space . . . . .                       | 114        |
| 6.2.4      | Effective Entropy in Replica Space . . . . .                      | 116        |
| 6.2.5      | Freezing Transition . . . . .                                     | 116        |
| 6.3        | Discussion . . . . .  | 118        |
| 6.3.1      | What is $\widehat{\Delta}$ ? . . . . .                            | 118        |
| 6.3.2      | Two Exactly Solvable Models . . . . .                             | 119        |
| 6.3.3      | Reduction Theorems . . . . .                                      | 123        |
| 6.3.4      | Freezing Temperature: General Consideration . . . . .             | 124        |
| 6.3.5      | Independent Interaction Model . . . . .                           | 128        |
| 6.3.6      | Random Sequences of Real Amino Acids . . . . .                    | 128        |

|          |  |            |
|----------|--|------------|
| 6.4      | Conclusion . . . . .   | 130        |
| 6.5      | Appendix: Proof of equation (20) . . . . .   | 131        |
| <b>7</b> | <b>Designed Heteropolymers</b>   | <b>137</b> |
| 7.1      | Introduction . . . . .   | 137        |
| 7.2      | Development of the Model . . . . .   | 139        |
| 7.2.1    | Disordered Short-Range Two-Body Interactions . . . . .                                   | 139        |
| 7.2.2    | Self-Averaging over the Sequences . . . . .  | 142        |
| 7.2.3    | Self-Averaging over Preparation Conformation . . . . .                                   | 144        |
| 7.2.4    | Manipulations with Replicas . . . . .  | 145        |
| 7.2.5    | Free Energy of Replica System . . . . .  | 149        |
| 7.3      | Discussion . . . . .   | 152        |
| 7.3.1    | Phase Diagram . . . . .  | 152        |
| 7.3.2    | An Exactly Solvable Model: The Generalized Potts Model . .                               | 154        |
| 7.3.3    | Expansion around the triple point . . . . .  | 157        |
| 7.3.4    | Flexible chain limit . . . . .   | 157        |
| 7.3.5    | Stiff Chain Limit . . . . .  | 159        |
| 7.3.6    | Miyazawa-Jernigan Matrix . . . . .   | 160        |
| 7.4      | Conclusion . . . . .   | 162        |
| 7.5      | Appendix: Rotation of Replica Space . . . . .  | 164        |
| <b>8</b> | <b>Design and renaturation with different interactions</b>                               | <b>167</b> |
| 8.1      | Introduction . . . . .   | 168        |
| 8.1.1    | What is this work about? . . . . .   | 168        |
| 8.1.2    | Sequence design and folding are governed by <i>different</i> inter-<br>actions . . . . . | 168        |
| 8.2      | The Model . . . . .  | 170        |
| 8.3      | Free Energy of the Model . . . . .   | 172        |
| 8.4      | Analysis of the Free Energy and Phase Diagram . . . . .                                  | 175        |
| 8.5      | Discussion . . . . .   | 178        |
| 8.6      | Simplification of Equation (7) . . . . .   | 181        |

|           |   |            |
|-----------|---|------------|
| 8.7       | Relationship between the average number of species-species contacts<br>and the interaction matrix . . . . . | 187        |
| <b>9</b>  | <b>Designed Heteropolymer in an External Field</b>  | <b>189</b> |
| 9.1       | Introduction . . . . .  | 189        |
| 9.2       | The model . . . . .   | 192        |
| 9.3       | Discussion . . . . .  | 201        |
| <b>10</b> | <b>Quenched and Annealed Disorder in the REM</b>  | <b>207</b> |
| <b>IV</b> | <b>Experimental</b>   | <b>215</b> |
| <b>11</b> | <b>Protein Correlations</b>   | <b>217</b> |
| 11.1      | Introduction . . . . .  | 217        |
| 11.2      | Brownian Bridge Representation for Protein Sequences . . . . .  | 219        |
| 11.3      | Brownian Bridges for Some Particular Sets of Proteins . . . . .   | 221        |
| 11.4      | Discussion . . . . .  | 224        |
| <b>12</b> | <b>NMR Analysis</b>   | <b>241</b> |
| 12.1      | Introduction . . . . .  | 241        |
| 12.2      | Experimental . . . . .  | 243        |
| 12.3      | Results and Discussion . . . . .  | 243        |
| <b>V</b>  | <b>Conclusions</b>  | <b>249</b> |
| <b>13</b> | <b>Summary</b>  | <b>251</b> |
| <b>14</b> | <b>Future Work</b>  | <b>255</b> |
| 14.1      | Experimental Realization of Imprinting . . . . .  | 255        |
| 14.2      | Correlations in Protein Sequences . . . . .   | 256        |
| 14.3      | Solution of the Protein Folding Problem . . . . .   | 257        |

# List of Figures

|     |  |    |
|-----|--|----|
| 2-1 | Monomer soup and resulting polymer . . . . .   | 24 |
| 2-2 | Energy spectrum for random and Imprinted sequences . . . . .   | 26 |
| 2-3 | Phase diagram for Imprinted polymers . . . . .   | 28 |
| 2-4 | How to decode protein sequences in a physical manner in order to study correlations in their sequences . . . . . | 29 |
| 2-5 | Correlations in protein sequences reflect energy optimization in evolution . . . . .                             | 30 |
| 3-1 | Symmetry example . . . . .   | 37 |
| 3-2 | Logarithm of the number of walks vs the number of sites . . . . .  | 39 |
| 3-3 | Ways to break symmetries in enumeration . . . . .  | 43 |
| 4-1 | The Imprinting Model . . . . .   | 50 |
| 4-2 | The three elementary moves employed in the Monte Carlo kinetics of 3D lattice polymers performed. . . . .        | 51 |
| 4-3 | Probability for renaturation for 27-mers and 36-mers with Potts interactions. . . . .                            | 54 |
| 4-4 | Two-letter 27-mers: Variation in composition . . . . .   | 55 |
| 4-5 | Two-letter 27-mers: heat capacity and probability of renaturation . . . . .                                      | 56 |
| 4-6 | REM order parameter $\langle X(T) \rangle$ (averaged over the ensemble) vs the number of Potts species . . . . . | 57 |
| 4-7 | $T(X = 0.8)$ vs the number of Potts species for Imprinted and random ensembles. . . . .                          | 58 |

|      |  |     |
|------|--|-----|
| 4-8  | $X(T)$ for each sequence from the ensemble of $q = 7$ chains with unique ground states and an energy gap. . . . .                              | 59  |
| 4-9  | Examples of the energy of the active site of a 26-mer vs temperature for chains designed with and without a target molecule . . . . .          | 61  |
| 4-10 | Probability distribution for a chain with a given energy gap size for Imprinted design, SG design, and random chains . . . . .                 | 66  |
| 4-11 | Example energy spectrum for $q = 9$ Imprinted 36-mer with Potts interactions. . . . .  | 67  |
| 4-12 | Histogram of folding times for Potts 27-mers . . . . .   | 69  |
| 4-13 | Characteristic folding time vs Temperature. . . . .  | 71  |
| 4-14 | Characteristic free energy barrier height vs Temperature. . . . .  | 72  |
| 4-15 | Probability of the system falling into a trap (metastable energy state) vs Temperature. . . . .  | 73  |
| 4-16 | Brownian bridges for Imprinted, SG, and Random sequences . . . . .   | 78  |
| 4-17 | Brownian bridges for an ensemble average of Prokaryote catalysts. . . . .  | 79  |
| 5-1  | Phase diagram for designed copolymers. . . . .   | 101 |
| 5-2  | Sample energy spectra for sequences imprinted at different polymerization temperatures. . . . .  | 103 |
| 6-1  | Plot of the inverse reduced freezing temperature vs the effective flexibility . . . . .  | 121 |
| 6-2  | For the Miyazawa and Jernigan matrix of amino acid interactions, we plot the flexibility vs the reduced inverse freezing temperature . . . . . | 129 |
| 7-1  | Cartoon of the Imprinting process . . . . .  | 140 |
| 7-2  | Renaturation of an Imprinted heteropolymer . . . . .   | 141 |
| 7-3  | Phase diagram for designed heteropolymers . . . . .  | 161 |
| 8-1  | Phase diagram for different values of the matrix similarity factor $g$ . . . . .   | 179 |
| 9-1  | Cartoon contour plot of a random field . . . . .   | 193 |
| 9-2  | Plot of the $g$ factor . . . . .   | 202 |

|       |  |     |
|-------|--|-----|
| 11-1  | Brownian bridges for hydrophilic, hydrogen bonding, and coulomb mappings for catalysts and coils . . . . . | 222 |
| 11-2  | Brownian bridges for hydrophilic, hydrogen bonding, and coulomb mappings for prokaryotes . . . . .         | 223 |
| 11-3  | Brownian Bridges for a series of evolutionary groups: Coulomb mapping                                      | 226 |
| 11-4  | Brownian Bridges for a series of evolutionary groups: Hydrophilic mapping . . . . .                        | 227 |
| 11-5  | Bridges of different species for Coulomb mapping . . . . .   | 233 |
| 11-6  | Bridges of different species for Coulomb mapping . . . . .   | 234 |
| 11-7  | Bridges of different species for Coulomb mapping . . . . .   | 235 |
| 11-8  | Bridges of different species for Coulomb mapping . . . . .   | 236 |
| 11-9  | Bridges of different species for Hydrophobic/hydrophilic mapping .   | 237 |
| 11-10 | Bridges of different species for Hydrophobic/hydrophilic mapping .   | 238 |
| 11-11 | Bridges of different species for Hydrophobic/hydrophilic mapping .   | 239 |
| 11-12 | Bridges of different species for Hydrophobic/hydrophilic mapping .   | 240 |
| 12-1  | Pulse sequence and parameters used in the NMR analysis. . . . .  | 244 |
| 12-2  | NMR Spectra for AAc/MAPTAC heteropolymer gel. . . . .  | 245 |





# List of Tables

|      |   |     |
|------|---|-----|
| 3.1  | Summary of enumeration data . . . . .   | 39  |
| 3.2  | Number of Hamiltonian walks for $3 \times 3 \times 4$ cubic sublattice for each<br>different starting point unrelated by symmetry . . . . . | 41  |
| 4.1  | Number of conformations (Hamiltonian walks) not related by sym-<br>metry on the cubic sublattice . . . . .                                  | 52  |
| 4.2  | Renaturation properties for different types of interactions . . . . .   | 62  |
| 4.3  | Comparison of design methods in thermodynamics . . . . .  | 64  |
| 4.4  | Comparison of design methods in kinetics . . . . .  | 76  |
| 11.1 | Legend for plots of bridges for different species . . . . .   | 232 |

R

# **Part I**

## **Introduction**





# Chapter 1

## Background

Proteins play a fundamental role in the biochemistry of all life on earth. Apart from structural purposes such as fibers, proteins play a major role in the molecular biochemistry of life in that they act to catalyze reactions (enzymes) or recognize and render aid in rendering inert various foreign molecules (antibodies). Indeed, Linus Pauling had the insight to recognize these two operations as specific examples of a more general concept: *specific molecular recognition* [Pau65]. The ability to understand how proteins are capable of specific molecular recognition may also shed light on how one may synthetically create artificial heteropolymers with protein-like properties as well as potentially reveal some of the secrets of the origin of life on earth.

### 1.1 Polymer Basics

A *polymer* is a long chain molecule consisting of many *monomers*, much like beads on a necklace. One of the most fundamental statistical physical properties of polymers is the phase transition between ordered and disordered states. The disordered state of a polymer is a *coil*. In this phase, we must maximize the entropy; this is achieved when the polymer behaves much like a random walk. In this case, the mean variance of the polymer size scales just like a random walk; i.e. for  $R \sim N^\nu$ ,  $\nu$  is 1/2. If we include the fact that the polymer cannot go through itself (and therefore is

not exactly a random walk), we expect that the polymer should swell a bit and  $\nu$  increases slightly.

In the ordered phase in which there are overall attraction between monomers, the polymer collapses into a *globule*, which is completely dense, i.e.  $\nu = 1/d$ . The phase transition from the disordered coil to the ordered globule is actually much like more common order-disorder transitions such as the vapor-liquid transition. In both examples, a convenient order parameter is the density of particles  $\rho$ , and a phase transition in  $\rho$  in the polymer case has been predicted analytically [Lif78] and seen experimentally [Nis79].

For *homopolymers* (polymers consisting of only one type of monomer) the globular phase in this case is not comprised of any particular arrangement of monomers in space; indeed, since all of the monomers are the same, all arrangements with the same density have the same energy. On the other hand, *heteropolymers* (polymers consisting of different types of monomers) may have an overall net attraction between monomers in order to be in a globular phase, but different arrangement of monomers in space yields different energies. Thus, in the case of heteropolymers, the *conformation* (arrangement of monomers in space) of the polymer plays a role in the determination of the energy of the polymer.

## 1.2 Molecular Biology Basics

Molecular recognition occurs when there is a large interaction energy benefit when the *target* (molecule one wishes to recognize) and *substrate* (part of the molecule which performs the recognition) are bonded. In order to make this recognition *specific*, one needs this bond to be strong for the target molecule and weak (or even repulsive) for all other molecules.

Proteins are heteropolymers comprised of monomers of the twenty amino acids. Proteins are capable of specific molecular recognition of a particular target since the amino acids on the protein substrate form energetically and entropically (“how the target fits”) favorable contacts with the target. However, for this to occur, one

of course needs a very specific placement of particular amino acids in space, or in other words, a particular protein conformation. Thus, the protein must do two jobs: fold to a conformation which can perform specific molecular recognition and be able to be stable in that single conformation.

How can the protein do this? Since proteins are heteropolymers, the only elements which can play any role at all in determining the equilibrium protein conformation are the *sequence* of amino acids along the protein and the nature of *interactions* between the amino acids. Thus, one can consider the desired protein conformation to be *written* along the protein sequence in the *language* of amino acids. Indeed, amino acids very much form a language since if we replaced each type of amino acid in a protein with a different type of monomer, the sequence would very much be the same, but due to the different language, the original conformation could not be derived. Also, just like any language, there is some room for errors and thus some mutations in the protein sequence can still lead to correct “communication” of the desired target conformation.

### 1.3 Protein Folding

Thus, the “information” detailing to which conformation the protein should fold is encoded in the protein sequence, but how does one go from the sequence to the equilibrium state of the polymer. This very question is one of the fundamental problems of modern biophysics and is called the “protein folding problem.” A related and interesting problem is the inverse question: given a desired equilibrium conformation, what sequence leads to that conformation?

We have previously described how conformation and sequence enters into the determination of heteropolymer energy: different globular conformations have different energies due to the fact that the polymer consists of different types of monomers. Thus, the protein must somehow have a sequence that which when arranged in the desired conformation is much lower in energy than when arranged in all other conformations. Therefore, one may suspect that evolution has “se-

lected sequences” such that they optimize the energy of the desired conformation with respect to the energy of all other conformations. The means by which evolution has selected sequences is unknown (although there are several hypotheses [Sha93b,Lau89,Bry87]).

However if one’s goal is to create renaturable heteropolymers which can recognize a particular target molecule, perhaps we do not need to explicitly know the solution to the direct or inverse protein folding problem but merely need some way to select sequences, whether it mimics the method used by evolution or not, to yield the desired result. In the next chapter, we propose such a procedure.



# Chapter 2



## Imprinting

The synthesis of a man-made polymer capable of functioning in a protein-like fashion can be of tremendous technological importance, and may also shed light on the natural creation of the molecular basis of life. In case of proteins, the unique 3D fold, responsible for the particular functionality of the molecule, is determined by the particular sequence of monomer units. We suggest a procedure, which we call Imprinting, to control the monomer sequence of an artificial heteropolymer during its synthesis in order to obtain a heteropolymer with the protein-like properties of quick and reliable renaturability to some unique spatial fold capable of certain functional properties. To control the sequence formation, our procedure employs interactions between monomers. We will show that this leads to renaturable chains, because renaturation is governed exactly by the same interactions between monomer units. We present here both analytical and computational study of Imprinting, yielding the requirements on the set of monomers chosen and further more specific prescriptions for the experimental verification of this theory.

The Imprinting procedure is formulated as follows: Consider a dense solution of monomers prior to polymerization. Monomers are allowed to equilibrate their spatial arrangement and are then rapidly polymerized, such that the monomers have

not sufficiently moved from their equilibrium positions in the “monomer soup.” As the monomers have been equilibrated prior to polymerization, the resulting polymer, which interacts with itself using the same interactions present in the soup of its monomers, should also be in a low energy conformation. If this energy is lower than that of all other conformations, we expect that the polymer will thermodynamically tend to renature to this very conformation. Thus, we will examine (i) whether the polymerization conformation is indeed the non-degenerate ground state conformation, and (ii) whether the folding process leads indeed kinetically to this conformation.

The desirable aspects of the “native” conformation, such as an “active site,” can be controlled in Imprinting by appropriate external molecules or fields. For example, if the monomers are allowed to equilibrate in the presence of some target molecule, a hole with complementary monomeric contacts will remain in the polymerization conformation.

We stress several important advantages of Imprinting: (i) it does not employ any products of biological evolution, such as synthetic apparatus of the living cell; (ii) it is not restricted to the use of the amino acid chemistry of real proteins; (iii) from the theoretical perspective, neither the solution of the direct nor inverse protein folding problem is required. Indeed, we do not purport either to compute the  $3D$  structure for the given  $1D$  sequence (direct) or to compute the  $1D$  sequence to fold to the given  $3D$  structure (inverse); (iv) our approach also does not involve evolutionary time scales, and is supposed to work thermodynamically, as a sort of “evolution in a test tube.”

To study Imprinting, we employ both analytic and computational treatments in which we assume only very general properties of monomers and polymer chain. Specifically, we examine the monomer interaction energy of the form

$$\mathcal{H}(\text{seq}, \text{conf}) = \sum_{I,J}^N B(s_I, s_J) \Delta(\mathbf{r}_I, \mathbf{r}_J) \quad , \quad (2.1)$$

where  $B(i, j)$  gives the interaction energy between monomers of *species*  $i$  and  $j$ ,  $s_I$

is the species of monomer number  $I$ ,  $\mathbf{r}_I$  is the position of monomer  $I$  in space, and  $\Delta(x, y) = 1$  for  $x = y$  and zero otherwise. Eq (2.1) has the simple interpretation that monomers number  $I$  and  $J$  interact based upon their proximity and the interaction energy between their respective species ( $s_I$  and  $s_J$ ).

In the Imprinting procedure, the set of interactions between monomer species in the monomer soup prior to polymerization is the same as that for the polymerized monomers. Thus, in modeling Imprinting, we use  $\mathcal{H}$  for calculating the interaction energy of the monomer soup as well as the interaction energy of the polymer. We examine Imprinting for *general*  $B(i, j)$ . In fact, this does not restrict, but rather extends the applicability of our approach to real chemical systems: while it may not be possible to calculate the interaction matrix  $B(i, j)$  for a specific chemical system to sufficient accuracy, the examination of Imprinting for a general set of  $B(i, j)$  includes the interaction matrix for *any* real chemical system. Furthermore, we can show that Imprinted chains can be created for broad variety of  $B(i, j)$ , as well as which properties of  $B(i, j)$  are useful for optimizing Imprinted sequences.

To study Imprinting analytically, we employ the replica method (see, for example [Bry87, Gar88a, Sha89a, Sfa93]) and examine the ensemble of Imprinted chains in the thermodynamic limit. Computational simulations of Imprinting have been performed using lattice models [Sha90a, Pan94d]. As we would like to prove that Imprinted chains renature to the polymerization conformation, we expect that this particular conformation should dominate equilibrium. The benefit of lattice models is that for sufficiently short chains, we can enumerate all globular conformations and thereby confirm that the conformation with the lowest energy is non-degenerate and is indeed the polymerization conformation [Sha90a, Pan94a]. Using a massively parallel supercomputer, we were able to model compact polymers on a  $3 \times 3 \times 4$  size cubic sublattice (which has 84,731,192 conformations [Pan94a]). Also, Monte Carlo kinetics was used to examine whether a particular sequence would fold to the polymerization conformation in some finite time.

Imprinting consists of two stages. We examine the *prearrangement stage* by modeling an ensemble of Imprinted sequences, each with a randomly chosen poly-

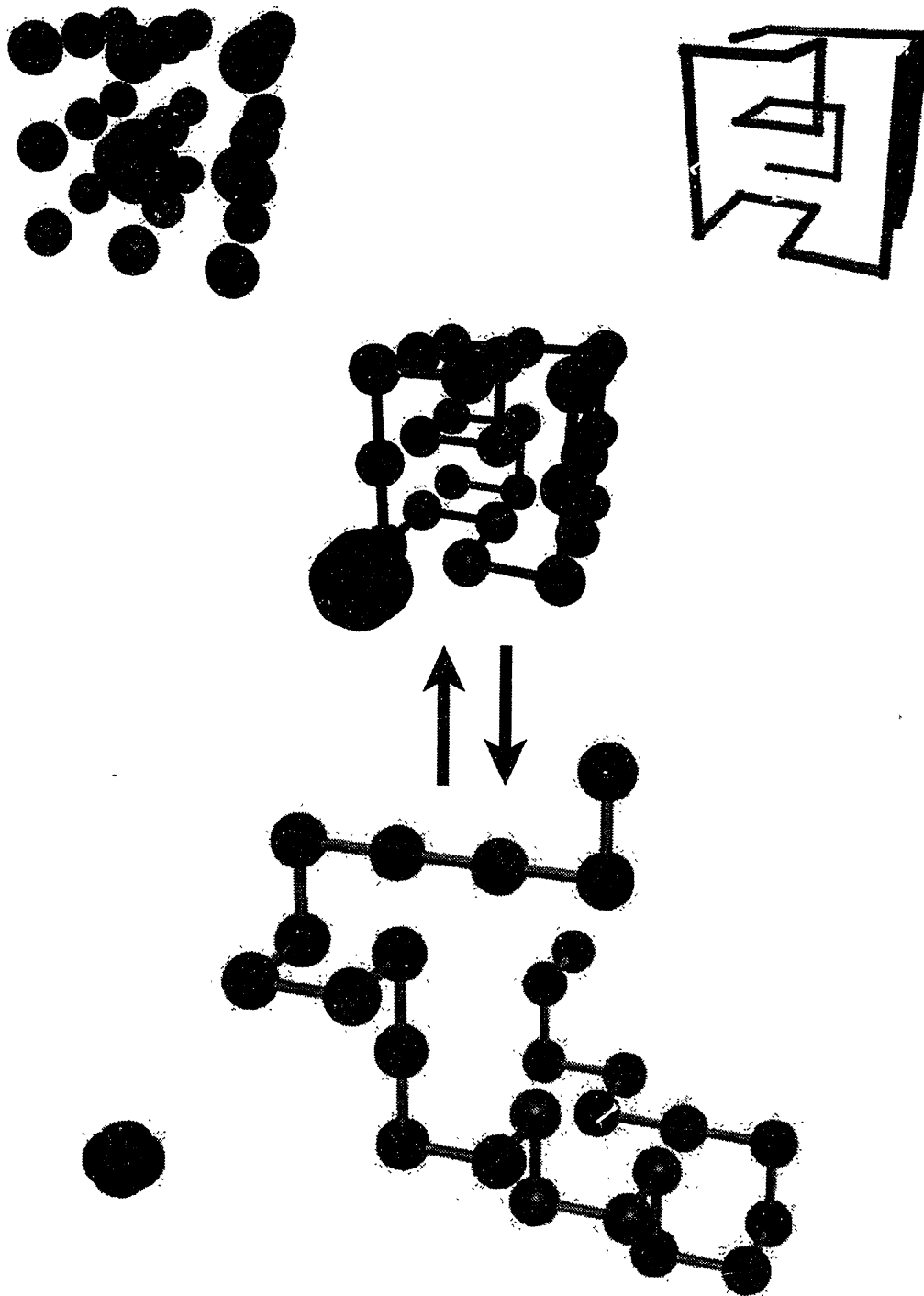


Figure 2-1: *Monomer soup and resulting polymer.* Monomers are allowed to equilibrate their spatial arrangement with respect to themselves and a particular target molecule. Upon polymerization, the resulting polymer conformation contains an active site capable of specific recognition of the target molecule introduced during polymerization.

merization conformation. The probability that a particular sequence is in the ensemble under investigation is given by the Boltzmann weight  $P \sim \exp(-\mathcal{H}/T_p)$ , where  $T_p$  is the temperature at which the chains are polymerized. Constructively, the modeling of the ensemble of Imprinted chains can be performed computationally by Monte Carlo sampling of sequence space or the weighted average over all sequences in an analytic model.

To model the *renaturation stage*, we examine whether the members of the ensemble of Imprinted chains each fold to their respective polymerization conformation; as the polymerization conformation contains an active site for specific molecular recognition, renaturation to the polymerization conformation is sufficient for specific molecular recognition. Computationally, we found that the ground state conformation of approximately 50% of the Imprinted chains was the polymerization conformation [Pan94d]. Upon examination of the Monte Carlo kinetics of successfully Imprinted chains, we found protein-like folding behavior: there is an optimal temperature at which the mean folding time is minimized and within a small temperature range around this temperature, folding to the native state was quick and reliable.

To understand these results, we employ the approach of Shakhnovich and Gutin [Sha93b] and consider the density of states as a function of energy (which is roughly the number of conformations with a particular energy). As the interaction energy is the sum of pairwise interactions, in the simpler case of a random polymer sequence the distribution of energies is known to be gaussian [Bry87,Sha89a]. This gaussian distribution of statistically independent states exactly corresponds to the so-called Random Energy Model (REM) [Mez84,Bry87,Gar88a,Sha89a]. Thus, the density of states for random sequences is similar to that shown in Fig 2-2a: we see that even random sequences have a large probability of a unique ground state, and thus “freeze” upon lowering the temperature beyond  $T_f$ . This freezing transition to a unique ground state is the result of competition between the reduced energy of the ground state and the large entropy of the numerous higher energy states. The description of this transition is simplified for sufficiently flexible chains. Indeed,

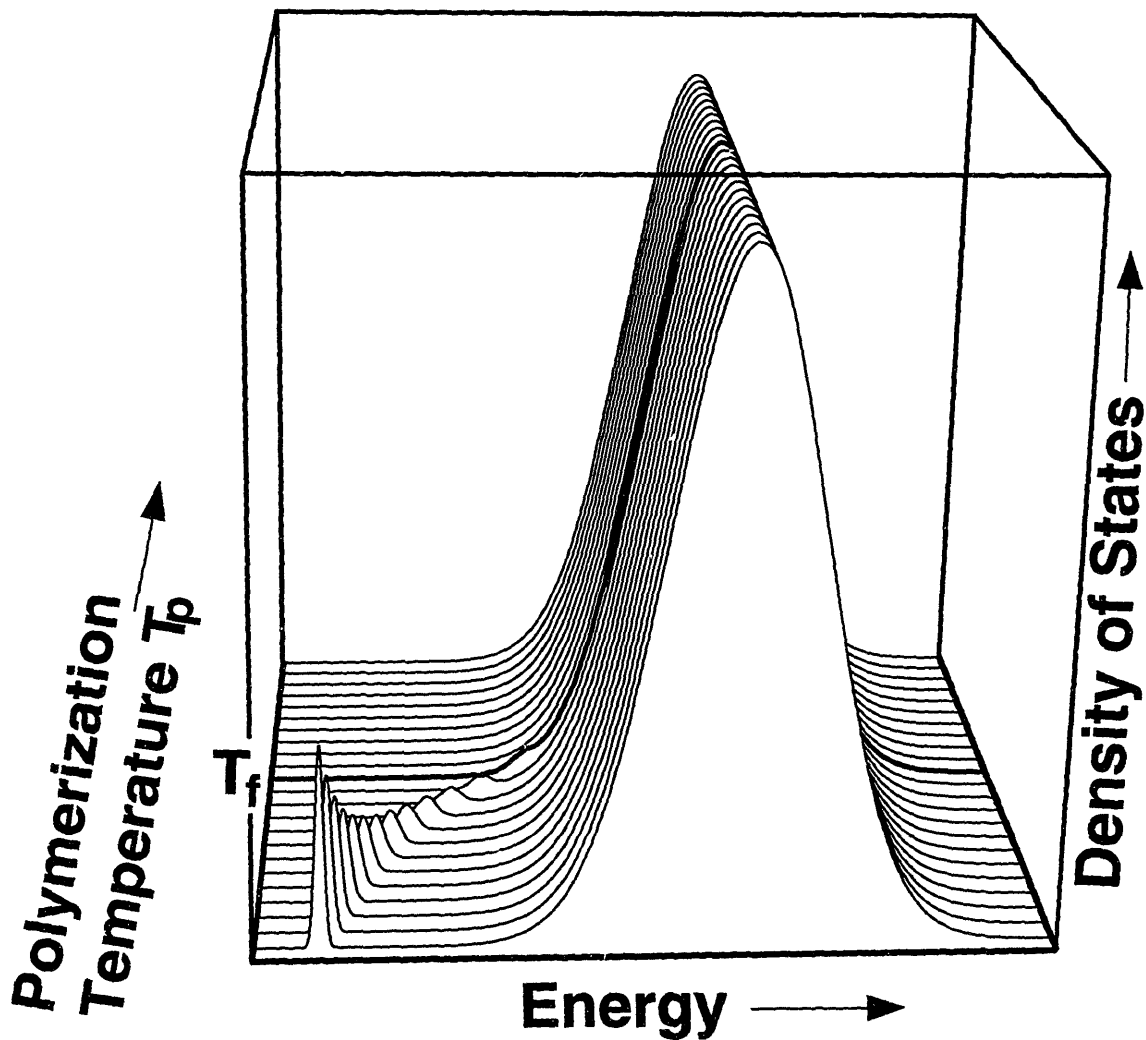


Figure 2-2: *Energy spectrum for random and Imprinted sequences.* Imprinting lowers the energy of the polymerization conformation below that of the REM ground state.

in this case, monomer interaction generally takes place on the level of a mixture of monomers, or quasi-monomers. As each quasi-monomer of the flexible chain involves a large number of monomers, the later are effectively averaged, and thus we find [Pan95b] that the freezing temperature for flexible chains is simply proportional to the standard deviation of the species interaction matrix:  $T_f \sim \langle B^2 \rangle_c^{1/2}$ . We see that for any heteropolymeric choice of monomer species, there is a non-zero freezing temperature. Both the REM nature of random sequences [Sfa93, Pan95b] and the general form for  $T_f$  [Pan95b] can be derived directly from eq. 2.1.

For Imprinted sequences, the prearrangement of the monomers lowers the energy of one particular conformation, and if the polymerization temperature is sufficiently low ( $T_p < T_p^c$ ), the energy of the polymerization conformation can be lowered below that of the REM ground state. Thus, the critical polymerization temperature, which distinguishes between Imprinted ( $T_p < T_p^c$ ) and random ( $T_p > T_p^c$ ) sequences, is the temperature at which the REM ground state is stable, i.e.  $T_p^c = T_f$ . Since the ground state energy for Imprinted sequences is much lower than that of random sequences, the Imprinted ground state is more stable and thus Imprinted chains freeze at a temperature  $T_{\text{tar}} > T_f$ ; also, we expect that  $T_{\text{tar}}$  increases as we lower the ground state energy, which is accomplished by lowering  $T_p$ . Using a more rigorous treatment, it was found that for flexible chains,  $T_{\text{tar}}$  can be expressed simply in terms of  $T_p$  and  $T_f$

$$T_{\text{tar}} = \frac{T_f^2}{T_p} + T_f \sqrt{\left(\frac{T_f}{T_p}\right)^2 - 1}, \quad (2.2)$$

Thus, to this order, the freezing transition to the polymerization conformation for Imprinted sequences does not depend on any specific nature of monomer species interactions. While higher order corrections do involve the interaction matrix more explicitly, these terms do not qualitatively change the behavior described by eq. 2.2.

Thus, as shown in Fig 2-3, we find three globular (i.e., compact) phases: *random*, in which equilibrium consists of many conformations; *frozen*, in which random sequences freeze to the REM ground state; and *target*, in which Imprinted sequences fold to the polymerization conformation. Within the target phase, we distinguish between two temperature ranges: for  $T < T_f$ , folding to the polymerization conformation is slowed since the low energy REM states are metastable and therefore act as traps; for  $T_f < T < T_{\text{tar}}$ , folding is quick and reliable. This behavior has also been seen by Monte Carlo kinetics [Sali94a,Pan94d,Soc94].

Thus, we have found that Imprinted sequences behave much like proteins. This is due to the optimization of the energy of the native state. A natural question is how does this optimization compare with that of proteins. Of course, any direct comparison between Imprinted sequences and proteins is difficult. One quick

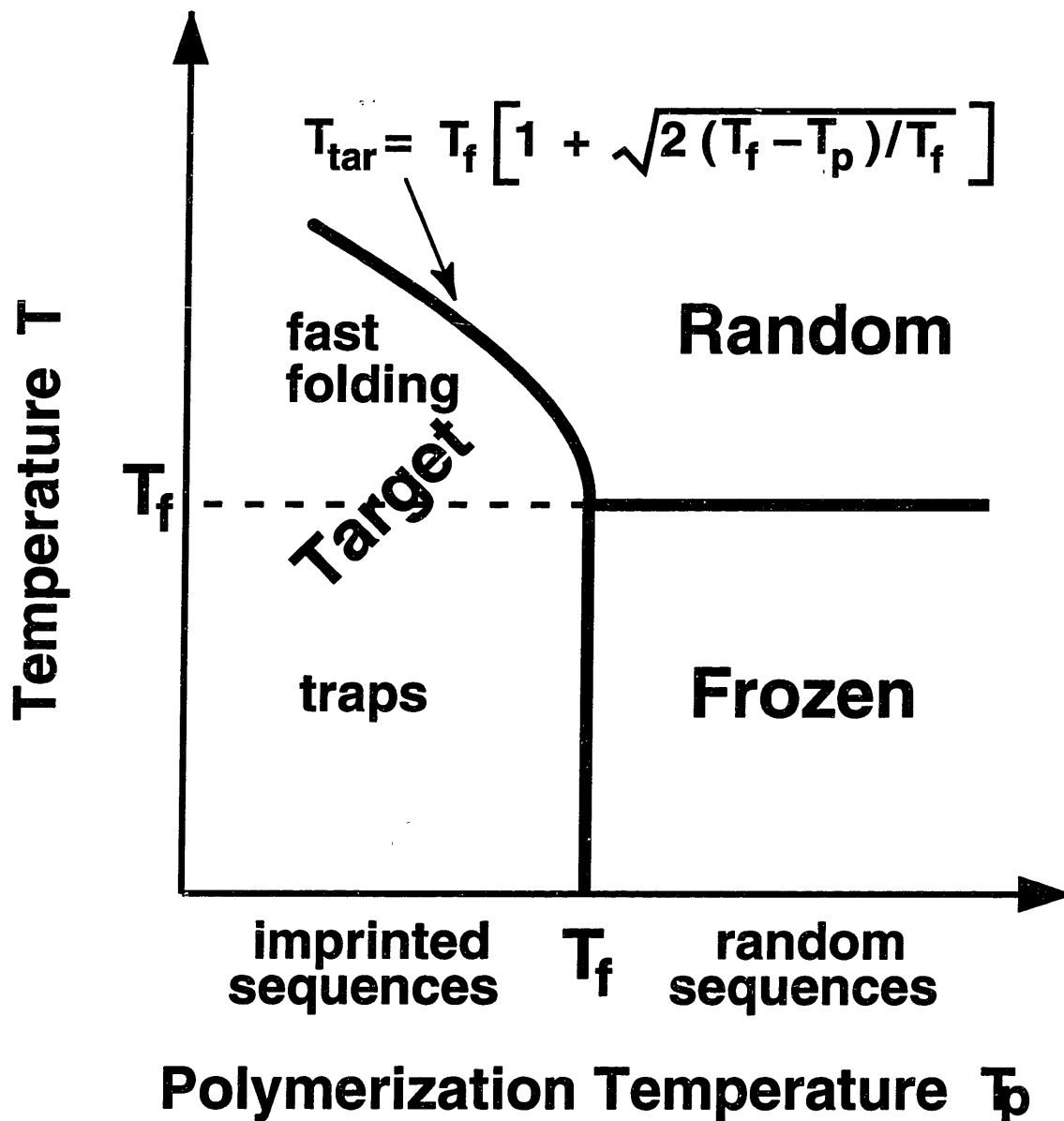


Figure 2-3: *Phase diagram for Imprinted polymers.* At sufficiently high polymerization temperature ( $T_p > T_f$ ), prearrangement leads to random sequences; for this case, there is a transition to a frozen phase at  $T_f$ . For  $T_p < T_f$ , chains are prearranged and there is a transition to the polymerization conformation at a temperature  $T_{\text{tar}}$  greater than  $T_f$ . In the temperature range  $T_{\text{tar}} > T > T_f$ , Imprinted chains fold quickly and reliably to the polymerization conformation. Folding Imprinted chains at temperatures below  $T_f$  is considerably slowed down as the polymeric frustrations which lead to a frozen state in random sequences lead to strong metastable states in Imprinted sequences.



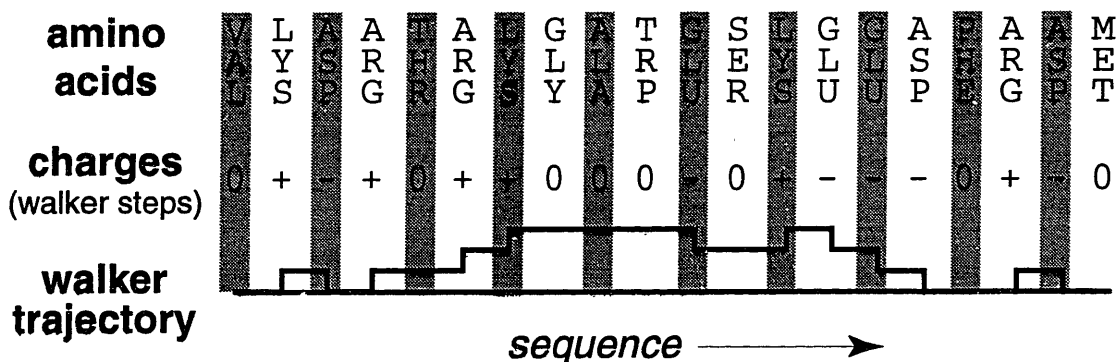


Figure 2-4: *How to decode protein sequences in a physical manner in order to study correlations in their sequences.* In the optimization of the interactions between all neighboring monomers in the monomer soup, Imprinting creates correlations along the sequences of Imprinted chains such that monomer species with attractive (repulsive) monomer volume interactions are likely (unlikely) to be neighbors along the chain. As this correlation involves only the linear sequence of monomers, and not volume interactions, it equally effects the energy of all conformations; therefore, there would be no evolutionary pressure to induce correlations other than an Imprinting-like process. To examine whether an ensemble of real protein sequences have been evolutionarily optimized in an Imprinting-like fashion, we “translate” each sequence in a given ensemble, using three decodings related to the fundamental interactions of amino acid chemistry: hydrophobic, hydrogen bonding, and Coulomb interactions.

comparison is to examine  $T_{\text{tar}}$  for amino acids. According to the interpretation given in [Fin93], energies of interaction between amino acids were determined by Miyazawa and Jernigan [Mia85] in  $T_f^a$  units, where  $T_f^a$  is freezing temperature for random polypeptide; in other words, the matrix elements of the MJ matrix are, in our designations,  $B(i, j)/T_f^a$ . For this MJ matrix, we obtain  $T_f/T_f^a \approx 1.1$ , which is remarkably close to the expected ratio of 1.

Also, recently, correlations reflecting energy optimization have been found in protein sequences [Pan94c]; these correlations are similar to those found in Imprinted sequences and indicate that perhaps protein and Imprinted sequences perhaps share a common past.

We summarize our discussion with a prescription to experimentally create Imprinted chains. First note that monomer species should be chosen for their heteropolymeric interactions and their ability to be polymerized in the regime  $T_p < T_f$ .

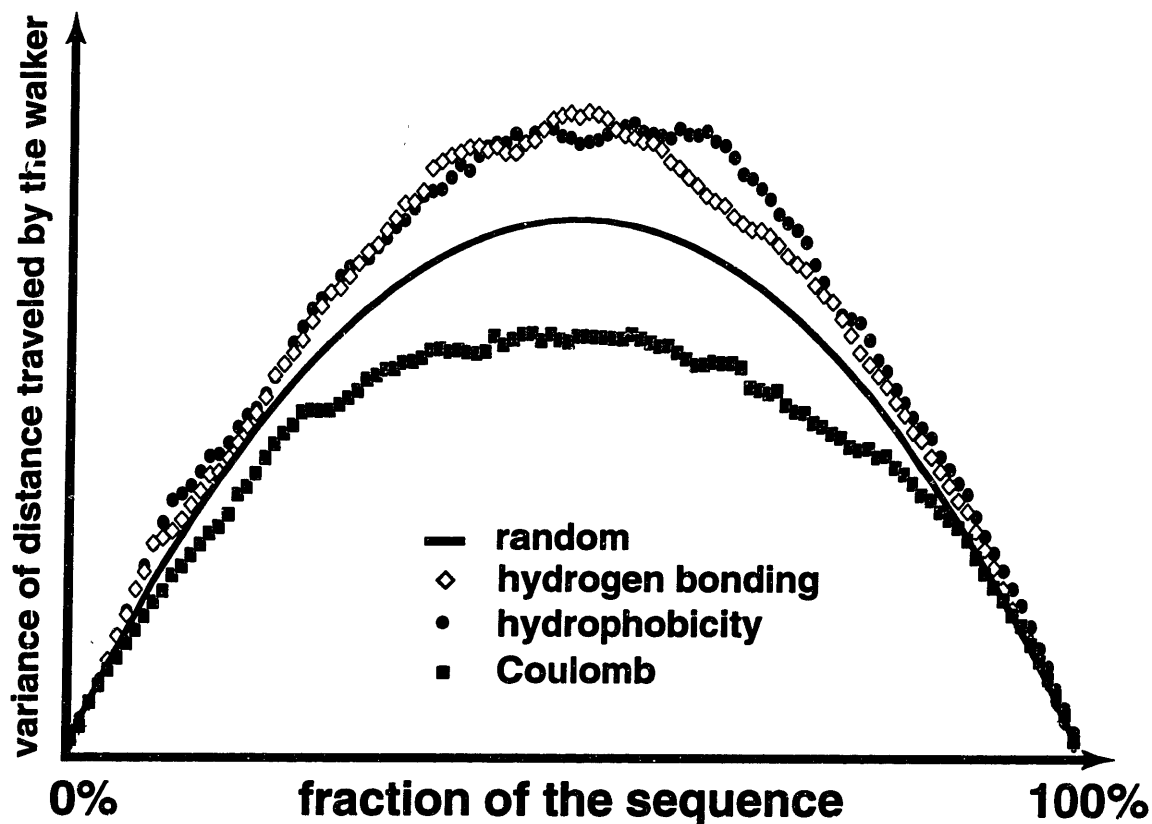


Figure 2-5: *Correlations in protein sequences reflect energy optimization in evolution.* For hydrophobic, hydrogen bonding, and Coulomb decodings, we examined the degree of correlations in the translated sequences by calculating the variance of the trajectory of a walker whose series of steps are dictated by the translated sequence, rescaling this trajectory to compensate for differences in protein length and composition, and then averaging the trajectories of the walkers over the ensemble of protein sequences. If there are no correlations in the translated sequences, then we would expect the averaged trajectories to match that of a random walker. Alternating correlations lead to a walker which generally switches directions, thus not traveling as far as the random walker, whereas persistent correlations lead to a walker that generally continues to travel in its current direction, therefore traveling farther than the random walker. Examining correlations in the sequences of an ensemble of globular proteins capable of molecular recognition (prokaryote catalysts), we find clear correlations in accordance with energy minimization.

Perhaps this may be accomplished through more exotic polymerization schemes, such as UV polymerization, reversible polymerization, microemulsion polymerization, or a combination of these and others. Furthermore, a dense solution of monomers should be used so molecules interact through volume interactions and not merely along the chain, as in a conventional Markovian polymerization scheme. Finally, as in the case of proteins, we expect quick and reliable folding to the polymerization conformation only in a relatively small region of acting temperature  $T_f < T < T_{\text{tar}}$ .

In conclusion, we find that, theoretically, chains produced by the Imprinting method will be renaturable and able to recognize a particular target molecule. Our method allows the monomers themselves to “design” the polymer, thus not requiring complicated computer simulations for design and allowing a great savings in time, as compared to evolutionary methods. Indeed, this great savings of time, relatively simple design scheme, and modest requirements on the nature of constituent monomer species also makes Imprinting a candidate for a scheme for prebiotic evolution *in vivo* as well as a model for molecular recognition *in vitro*.

The rest of the thesis explores these facets in much greater detail and the chapters are primarily derived from the analogous papers. In chapter 3, previously published in [Pan94a], we describe the computational methodology involved in the enumeration of the conformations on cubic sublattices and the surprising result of the applicability of Flory theories to these small systems. In chapter 4, we combine the results of [Pan94b] and [Pan94d] and detail the nature of the thermodynamics and kinetics of Imprinted sequences analyzed by computer simulation; we find that Imprinting optimization of the monomer soup leads to a large percentage of thermodynamically and kinetically renaturable sequences.

Next, we detail the analytical examinations of Imprinting. In chapter 5, we introduce the formalism, originally found in [Pan94e] for replica analysis of sequences with optimized ground state conformations using a simple black and white model. Chapter 6 consists of the work [Pan95a] to examine random heteropolymer freezing for all heteropolymers which interact through short range interactions. We find

that freezing is not heavily dependent on the nature of interactions. As detailed in [Pan95b], chapter 7 combines the formalisms of chapters 5 and 6 to examine the freezing transition of designed heteropolymers with short range interactions. Chapters 8 and 9 build upon this formalism, allowing for different interactions during design vs. folding [Pan95d] and interactions with an external field [Pan95c]. Finally, we comment in chapter 10 on the nature of approximations made in the Random Energy Model (REM) for heteropolymers used throughout this thesis as well as a means to derive the REM results without the replica trick [Pan95e].

In Part IV, we detail some experimental work. Chapter 11 discusses the nature of correlations in proteins first described in [Pan94c] and the relationship between these correlations and a possible Imprinting-like stage of protein evolution. Chapter 12 describes some NMR work performed on heteropolymer gels with multiple phases; while this work may be at present only vaguely related to physical questions of heteropolymer freezing at the moment, hopefully further analysis will provide more clearer links.

Finally, the work is summarized in chapter 13 and potential future aspects detailed in chapter 14.

**Part II**

**Computational**



# Chapter 3

## Enumeration

A massively parallel supercomputer was used to exhaustively enumerate all of the Hamiltonian walks for simple cubic sublattices of four different sizes (up to  $3 \times 4 \times 4$ ). The behavior of the logarithm of the number of walks was found to be linear in the number of vertices in the lattice. The linear fit is shown to agree also with the asymptotic limit of the Flory mean field theoretical estimate. Thus, we suggest that the fit obtained yields the number of walks for any size fragment of the cubic lattice to logarithmic accuracy. The significance of this result to the validity of polymer models is also discussed.

A Hamiltonian walk is defined to be a walk over some graph such that each vertex is visited once and only once. In general, Hamiltonian walks are known to be one of the most challenging and important issues in the graph theory. As for graphs of cubic sublattices, exhaustive enumeration of Hamiltonian walks is especially important in the physics of heteropolymers. Indeed, Hamiltonian walks on the sublattices are naturally identified with maximally compact conformations of polymer chains. In heteropolymers, such as proteins, there may be one single conformation, which is practically fully compact and which strongly dominates the partition function of the system. Thus, Monte Carlo sampling is not sufficient in this case, and exhaustive enumeration of conformations is required.

This was first performed by Shakhnovich and Gutin [Sha90a] when enumerating

the 103346 Hamiltonian walks on the  $3 \times 3 \times 3$  cubic sublattice in order to verify the phase transition of heteropolymers predicted analytically. The fact delicate effects of the analytic theory were reproduced shows that even a small sublattice can be an effective model. However, there are some properties not present in the  $3 \times 3 \times 3$  case, such as pseudo-knots. Thus, enumeration of even the  $3 \times 3 \times 4$  case (which includes pseudo-knots) can shed light on new physical properties.

The enumeration algorithm is formulated as follows. We can consider any lattice in terms of the graph connecting the lattice sites. Consider all of the (not necessarily self-avoiding) walks of length  $N$  on an infinite lattice of coordination number  $z$ . At each lattice point, we have  $z$  possible different directions to travel in order to reach a new site. These walks can be described as a tree of  $N$  levels with  $z$  branches at each node, each corresponding to a possible choice of direction to the next site. The enumeration of the possible walks is merely the counting of the number of branches of length  $N$  of this “ideal” tree. We now impose the condition that the lattice is finite, say  $l \times m \times n$ . We must now remove the branches of the ideal tree which correspond to walks that are not contained in the new boundaries (for example, the walk consisting of  $N$  steps in a single direction is no longer in the set of possible walks when  $l, m, n$  are all less than  $N$ ). The addition of the constraint of self-avoidance further removes branches from the tree. We study the case of Hamiltonian walks, i.e. in the above notation  $N = l \cdot m \cdot n$ . Thus, the enumeration of all of the Hamiltonian walks is the counting of the number of branches of length  $N$  of this new “restricted” tree.

In order to ascertain which sub-branches of the original ideal tree are removed, we must follow down the sub-branches of the ideal tree until we reach the end of the branch. A branch ends either when the walk is of length  $N$  or when there are no other possible sub-branches (for example, when a self-avoiding walk blocks itself off). Now, we back up one level of the tree and continue the procedure on a new sub-branch. In this way, all of the branches of the tree are exhaustively traversed in a very systematic manner.

Using the prescription above, there will be some walks related by symmetry (eg



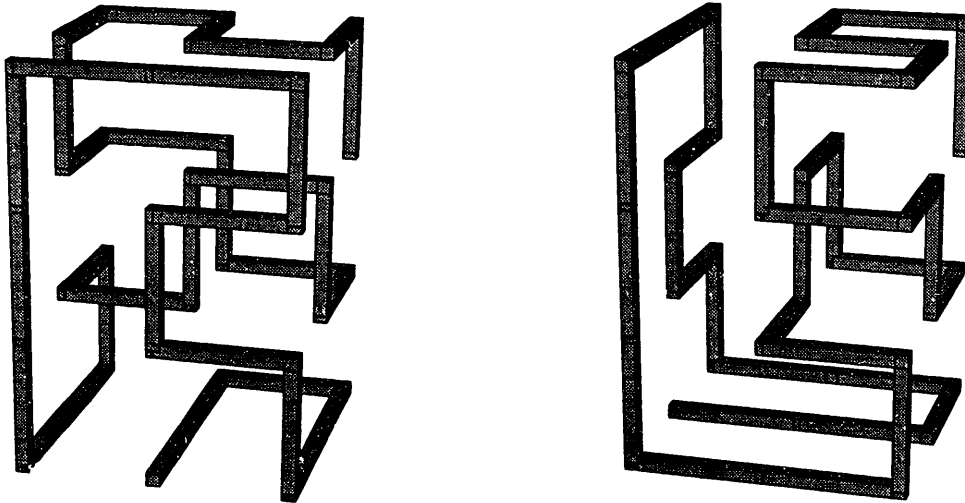


Figure 3-1: These two 36 site walks are related by mirror symmetry. Thus, only one is included in the enumeration procedure.

rotations and reflections). For example, consider the two walks shown in Figure 3-1. They are related by symmetry, in this case a reflection. We do not wish to include both of these walks, so we used “starting paths” which break all possible symmetries. We start enumeration, i.e. the traversal of the ideal tree, only for those sub-branches of the last node in each starting path. In this way, we remove branches related by symmetry. There are in fact many starting paths necessary for several reasons: 1) there are several different points (unrelated by symmetry) which one can start the walk; 2) there are many symmetries to break. Therefore, we have devised an algorithm to generate these starting paths. This algorithm will be discussed in the Appendix.

Note that we have neglected one transformation: the reversal of the start and end of the walk. For heteropolymers, we want to include walks related by this symmetry, as the polymer sequences are not invariant with respect to sequence reversal. However, this may not be appropriate for other applications of Hamiltonian walks and should therefore be addressed accordingly. We also note that the arguments presented here and in the Appendix can be easily modified to handle unusual lattices, such as unvisitable sites (used to model a “target site” in polymer models), lattice dislocations, and other lattice aberrations, since unusual lattice topologies can be

easily described in terms of the graph connecting the sites and the symmetries relating orientations of this graph.

The number of Hamiltonian walks increases exponentially with the number of vertices, so in order to gain the necessary computational speed to calculate the number of walks on larger sublattices, we employed two techniques. The most significant technique utilized was the use of a massively parallel computer (128 node Thinking Machines CM-5) and a parallel version of the tree enumeration algorithm. This parallel version used the method of “Continuation-Passing Threads” [Hal94], i.e. a random work stealing scheduler able to assign subtrees to different processors and dynamically pass work (i.e. sub-branches to enumerate) to inactive processors as necessary. The throughput of the parallel algorithm was found to scale linearly with the number of processors.

The second technique used was the addition of simple checks to see if we can end the search down a branch early. Each time a node is added to the walk, we check each neighbor of that node to see if it is surrounded by nodes which have already been visited. If so, then the node can never be visited, and if that node has not yet been visited, then the partial path produced so far can never lead to a valid walk; thus, we do not need to search down this path any further. Also, we keep track of how many unvisited nodes have only one unvisited neighbor. Clearly, in a successful walk, such a node must be the last node of the walk. So if we ever find two such nodes, we can safely stop the search down this partial path. These “blocked neighbor” checks provided one to two orders of magnitude speed improvement over prior algorithms.

These two improvements yielded sufficient computational power to enumerate the Hamiltonian walks on the  $3 \times 3 \times 4$  and  $3 \times 4 \times 4$  sublattices. The results are summarized in Table 3.1.

With four lattice sizes ( $N = 18, 27, 36, 48$ ), it may be possible to see some trend in the number of walks ( $M$ ) as a function of the number of lattice sites ( $N$ ). In Figure 3-2, the natural logarithm of the number of walks is plotted versus  $N$ . We

| $N$ | $M$             | CPU time†      | starting paths |
|-----|-----------------|----------------|----------------|
| 18  | 1,711           | $\ll$ 1 second | 27             |
| 27  | 103,346         | 0.2 second     | 35             |
| 36  | 84,731,192      | 5 minutes      | 816            |
| 48  | 134,131,827,475 | 64 hours       | 3579           |

Table 3.1: Summary of enumeration data, where  $N$  is the number of sites and  $M$  is the number of walks unrelated by symmetry. †CPU time given for 128 node CM-5

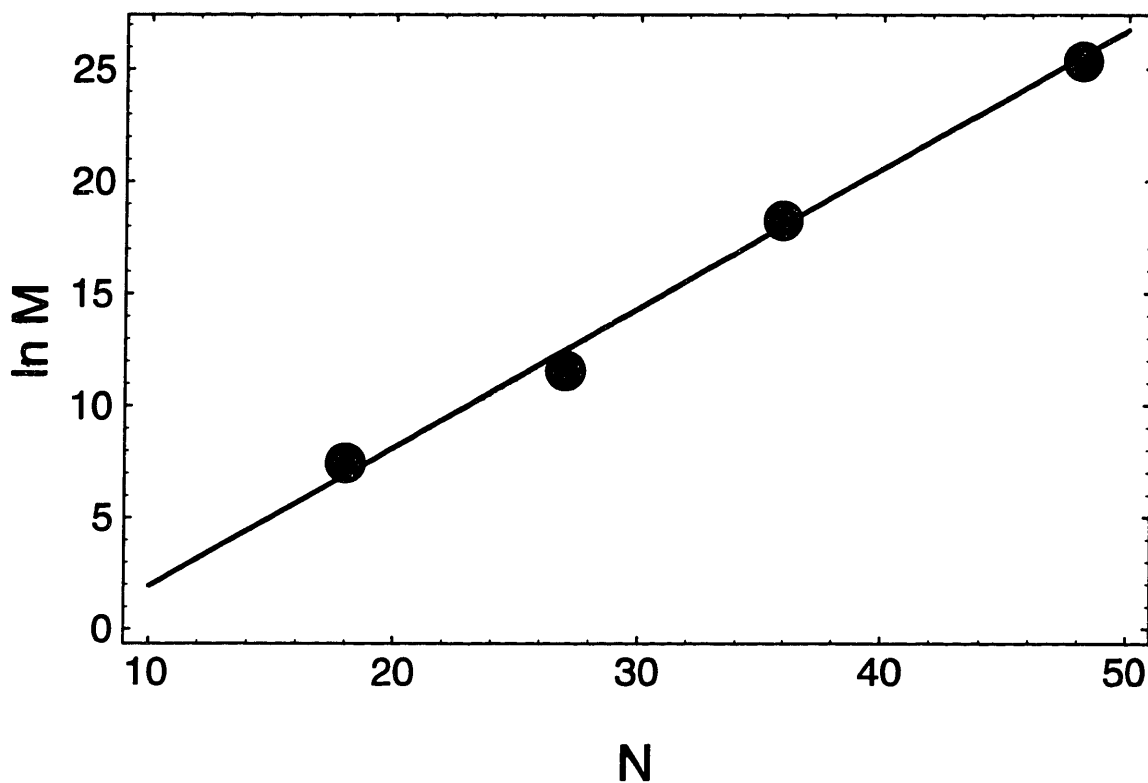


Figure 3-2: Logarithm of the number of walks ( $M$ ) vs the number of sites ( $N$ ), for  $N = 18, 27, 36, 48$ . We see that the curve is essentially linear.

fitted a linear relation of the form

$$\ln M = \alpha + \beta N \tag{3.1}$$

with  $\alpha = -4.3 \pm 1.2$  and  $\beta = 0.62 \pm 0.04$  ( $R^2$  of the fit: 0.99). Note that while this is trivial to calculate the number of walks for  $N < 18$  (i.e.  $N=8$  and  $12$ ), the inclusion of these points does not alter (within error) the linear fit or the arguments to follow; however, as discreteness effects should become great in these cases, we exclude them. Thus, we find that this fit works well for the region of small  $N \leq 48$ .

On the other hand, the Flory [Flo53] mean field calculation of the entropy of polymer melt is known to be applicable to the estimation of the number of compact globular conformations in the  $N \rightarrow \infty$  limit. Indeed, the conceptual foundation of the Flory treatment is the restriction imposed on the addition of new monomers within the constraints of the avoidance of occupied sites and chain connectivity. This kind of argument is equally applicable to both a macroscopic melt of different long chains, and a large globule of one single chain, as two systems differ only in the contributions of independent chains mixing entropy, which is negligible in the long-chains melt, and of surface effects, which are negligible in thermodynamic limit. Therefore, in the  $N \rightarrow \infty$  limit we have the estimate

$$M \approx \left(\frac{z-1}{e}\right)^N \tag{3.2}$$

where  $z$  is the coordination number of the lattice. The question is, however, how large  $N$  should be to validate this approximation. This problem is similar in spirit to the nature of the convergence of other mean field theories, or even the central limit theorem.

It turns out that in fact eqs (3.1) and (3.2) agree very well, thus validating the extrapolation of (3.1) for the entire region of  $N \rightarrow \infty$ . We can formally transform (3.2) into (3.1) by saying that

$$z = 1 + \exp[1 + \beta + \alpha/N] \tag{3.3}$$

| Site type    | Starting site | number of walks   |
|--------------|---------------|-------------------|
| corner       | 0             | 28,186,048        |
| short edge   | 1             | 13,648,609        |
| long edge    | 9             | 16,166,505        |
| small face   | 4             | 5,298,397         |
| large face   | 10            | 18,287,284        |
| inside       | 13            | 3,144,349         |
| <b>total</b> |               | <b>84,731,192</b> |

Table 3.2: Number of Hamiltonian walks for  $3 \times 3 \times 4$  cubic sublattice for each different starting point unrelated by symmetry. We use the following convention for numbering sites on a  $l \times m \times n$  sublattice:  $p(x, y, z) = x + ly + lmz$ .

In the  $N \rightarrow \infty$  limit, we have  $z = 1 + \exp[1 + \beta]$ . Using our fit for  $\beta$ , we calculate  $z = 1 + \exp[\alpha] = 6.1 \pm 0.2$ , which compares well with the exact value of 6 for the simple cubic lattice. As eq (3.1) agrees with the results of exact enumeration in the regime  $N \approx \mathcal{O}(10^2)$  as well as the Flory theory in the  $N \rightarrow \infty$  limit, we suggest that eq (3.1) may be used to derive the number of walks for arbitrary  $N$  to logarithmic accuracy.

It is worthwhile to note that the point for  $N = 27$  in the Fig. 3 is definitely below the interpolation straight line. This might be related to the fact that this is the case of maximally symmetric cubic shape. We are indebted to Dr. A. Gutin for the comment on similar effect on  $2D$  lattice [Lau89].

Thus, in terms of models of polymers, the polymeric entropy of small cubic lattice polymer models seems to be valid at least to the mean field approximation, and therefore the results which rely heavily on the nature of the conformations, such as heteropolymer theory, obtained from even small lattice models have some physical meaning. As one examines longer chains, the system starts to exhibit other physical properties, such as the presence of pseudo-trefoils in 36-mers [Diao94] and more complicated topologies in larger sublattices. However, in these cases the effect of the lattice model in modeling of polymer topology, for example, is unclear.

In conclusion, as eq (3.1) yields  $M \approx 2 \times 10^{15}$  for  $N = 64$ , it seems that

the enumeration of the  $4 \times 4 \times 4$  sublattice is several orders of magnitude out of reach using our current algorithm and supercomputer power. However, perhaps this estimate is slightly pessimistic, as sublattices with a cubic shape are expected to have less conformations than predicted by our fit. Also, the case  $N = 48$ , while possible to enumerate, is still extremely CPU time consuming and therefore cannot be used routinely in any current polymer modeling scheme. However, enumeration of  $N = 36$  is not very CPU time consuming. Furthermore, there are fundamental differences between the previously enumerated case of  $N = 27$  and  $N = 36$ , such as the presence of pseudo-knots. Thus, the use of the case  $N = 36$  will allow much richer modeling of the thermodynamics of lattice polymers [Pan94d]. Finally, while the cases  $N = 64$  and greater cannot even be enumerated at present, hopefully the estimate on the number of conformations given will be useful, for example in the analysis of Monte Carlo kinetics on cubic lattices [Pan94d,Sali94b].

## Appendix: Enumeration of starting paths

We wish to enumerate the different paths which completely break all of the symmetries. First, we must enumerate all of the symmetries. Consider all of the vertices of the graph to be numbered consecutively. Any transformation (eg. rotation, mirror inversion, etc) can be expressed as a permutation of these indices. The number of transformations, and therefore permutations, is calculated as follows in terms of the number of ways a  $d$ -dimensional hyper-cube can be re-oriented: i) we first have the symmetry by the number of corners of the cube ( $2^d$ ); ii) next, once we choose a corner to fix, we have  $d!$  ways to choose how we arrange the edges (for example, for  $d = 3$ , we have 3 ways to place the first edge, leaving 2 ways to place the second ). Thus, there are  $d!2^d$  transformations for a simple hyper-cubic lattice in  $d$  dimensions.

To generate the starting paths, we traverse the tree and compare sub-branches

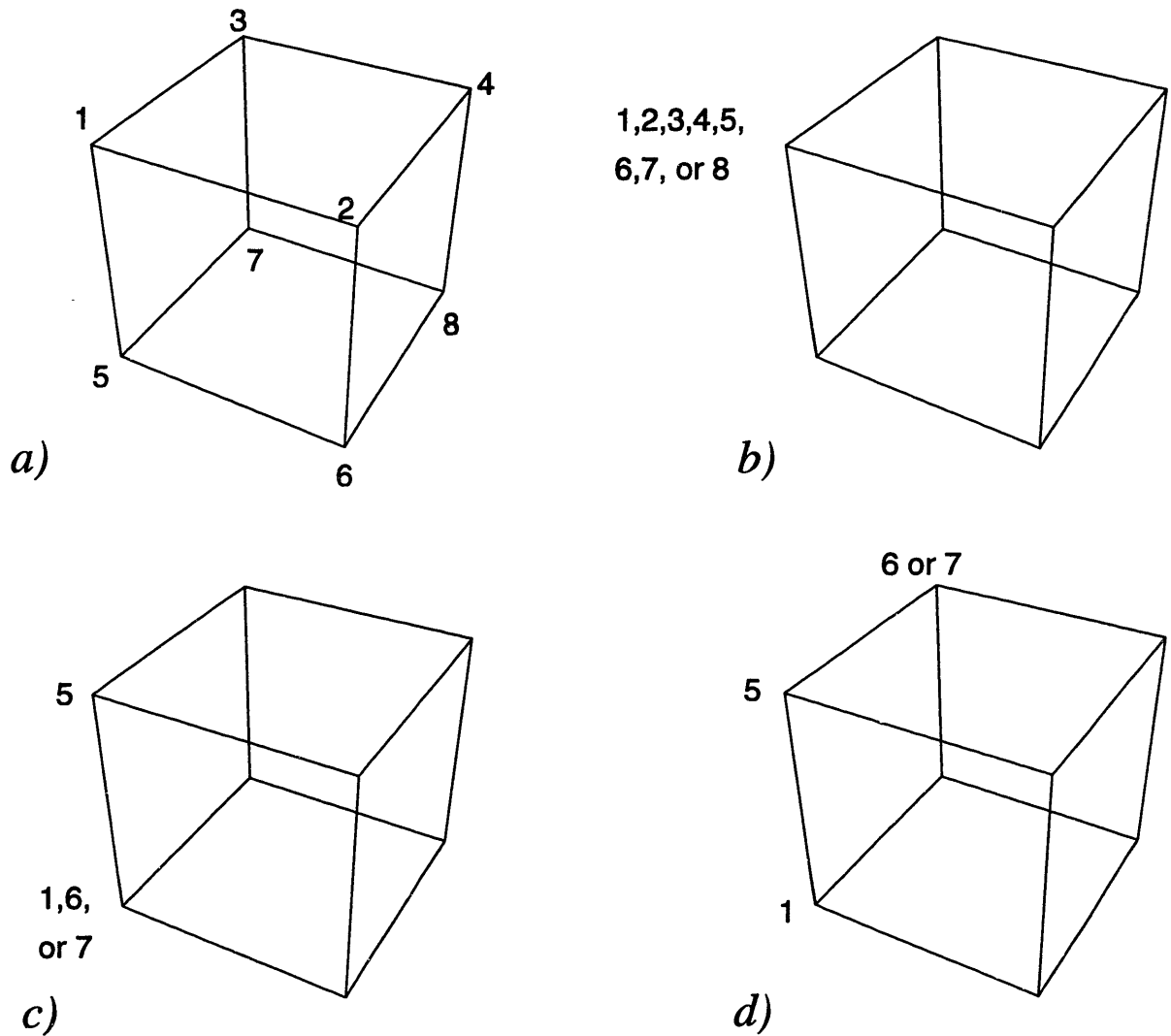


Figure 3-3: For three dimensional space, there are 48 ways to orient a cube: a) we label the corners of the cube from 1 to 8; b) first, we fix one corner of the cube: there are eight corners from which to choose; c) now we fix another corner: there are 3 edges to choose from; d) for the final corner, there are only 2 edges left from which to choose. The choice of a corner and two edges completely describes the orientation of the cube.

for symmetries. At each node, we transform the trajectories formed by each sub-branch using all of the enumerated transformations (i.e. applying all of the permutations). If any transformation can map one sub-branch into another, then the sub-branches are related by symmetry, and we can discard one of them. The remaining sub-branches themselves will now be enumerated using the same procedure. If none of the sub-branches are related by symmetry, and if all of the  $d!2^d$  symmetries have been broken by the current path, then the current path is a starting path, and we can back track and continue the enumeration with the unexplored branches.

For example, consider a walker starting from the corner of a cube. It is at the top of the tree of Hamiltonian walks. It now has three possible paths, but each path can be transformed into the other by a mirror symmetry. Thus, we can discard two of the sub-branches, choose the third, and continue the process. When none of the sub-branches of a given node are related by symmetry, then each sub-branch is a starting path. Then the walker backs up one level of the tree in order to traverse through the sub-branches left behind.

The enumeration of all of the walks and the enumeration of the starting paths are deeply related. Each traverse the ideal tree, only differing in when the walk has completed and when sub-branches are to be discarded.



# Chapter 4

## Computer Simulation

In Chapter 1, a procedure was suggested to synthesize polymers with characteristics similar to those observed in globular proteins: renaturability and the existence of an “active site” capable of specifically recognizing a given target molecule. This procedure is studied using a computer simulation of the thermodynamics of lattice 27-mers and 36-mers for different types of short range interactions. We found, in the best conditions, a 50% success rate of creating renaturable heteropolymers, thus confirming the original results. The folding kinetics as examined by Monte Carlo simulation shows that the imprinted sequences can reach the ground state reliably and quickly. Finally, we compare the correlations found in the imprinted sequences with those found in natural proteins. We interpret these results as the confirmation of the efficacy of the polymerization procedure.

### 4.1 Introduction

The inverse protein folding problem is a challenging problem of biophysics. It is also related to theoretical descriptions of prebiotic evolution and the origin of life. The entire question stems from the fact that proteins are capable of having a unique

space conformation which is thermodynamically stable and accessible kinetically. This particular “native” conformation is encoded in the sequence of chain links, or in other words, “written” in form of the monomer sequence in the “language” of volume interactions between monomers. For a protein to function, it must be in its native conformation, which may be capable of highly specific molecular recognition. Accordingly, there are several formulations of the inverse protein folding problem. Specifically, one may wish to design a sequence which will

- 1a. be stable in a *given* conformation;
- 1b. kinetically fold to this given conformation;
- 2a. have some stable unique conformation, no matter which one;
- 2b. kinetically fold to this conformation.

If the desired conformation is in a sense close to the native conformation of one of the known proteins, then problem (1), both (a) and (b), can be approached with biotechnology, i.e. using the synthetic apparatus of the living cell. This is of course very important and fruitful for numerous applications, such the improvement of some enzymes, etc. As for the physics involved, the solution to problem (1a) in the framework of lattice toy models has been suggested recently [Sha93b] and the capability of this approach to the solution of question (1b) is now under investigation [Sali94b]. On the other hand, it was recently shown theoretically [Sha89a,Sfa93] that problem (2a), but not (2b), can be solved by simply taking random sequences. And this solution of the problem (2a) is in a sense the best (or it is among the best) possible, as the fraction of the chains with unique ground state conformation is of order one in the ensemble of random chains.

There are, however, important pitfalls aspects which are not addressed either by the formulation (1) or (2). Indeed, let us think of some distant goal, such as artificial antibody, or let us think of primary polymerization in the primordial soup. As opposed to the formulation (1), we are not interested in reproduction of any particular conformation, especially the one close to the conformation of

any existing protein. Instead, the desirable conformation, or some part of the 3D structure, such as active site, is dictated in each particular case by the goal. If we are speaking about artificial antibody, then its conformation must be only capable in having specific active site to the given antigen. If we are speaking about the appearance of organization in the primordial soup, the question is the conformation which matches to some other molecules presented, etc. On the other hand, if there is some need to provide a conformation with some particular properties, then there is no chance to obtain the desirable properties using random chains.

Following similar arguments, we have suggested a new formulation of the inverse protein folding problem [Pan94b]: we wish to design a sequence which will

- 3a.** be stable in a conformation chosen at random *prior* to polymerization;
- 3b.** kinetically fold to this conformation.

The underlying idea of our approach is to employ as the driving force in the synthesis process the same molecular interactions which may be responsible for recognition, self-recognition or renaturation of an already prepared chain. To use interactions from the very beginning, we suggest polymerization in a dense mixture of different monomers, possibly in the presence of the given target molecule, where necessary correlations have been already created due to monomer-to-monomer volume interactions. Thus, the monomer solution, possibly in the presence of a target molecule (which plays the role of an antigen, ligand, etc), is energetically minimized prior to polymerization.

Polymerization leads, of course, to some random conformation, but, given that there are strong enough interactions between monomers and the target molecule, this *polymerization conformation* has an active site which matches the target molecule in shape and complementary interactions. In order for the polymer to be capable of molecular recognition, the polymerization conformation must be the reproducible unique ground state. The main purpose of this work is a more detailed investigation of the imprinting model in both (a) thermodynamic and (b) kinetic aspects. The polymerization conformation may, or may not, include some active

site for target molecule.

In what follows, we investigate the Imprinting model, a polymerization procedure suggested to be capable of solving the third formulation of the inverse folding problem. The great advantage of our approach is that polymer synthesis, including possibly the design of active site, is carried out by the polymerization procedure *thermodynamically*. This means the usual laboratory, and not evolutionary, time scale. Also, this procedure employs neither the machinery of the living cell, nor the chemical compounds of real biochemistry. Note that these arguments seem also applicable to the case of prebiotic synthesis of chains. In the prebiotic scenario, there is no biochemical machinery available, but all of the elements necessary for the Imprinting model are believed to be present in the primordial soup. It is especially compelling to consider the creation of polymers capable of molecular recognition starting from only monomers and the necessary target molecules, and to have this process be capable on relatively short time scales.

## 4.2 Description of the Model

To model polymerization in the presence of volume interactions and to explore the conformational properties of the emerging polymer chains, a  $l \times m \times n$  fragment of the simple cubic lattice with  $N = l \cdot m \cdot n$  vertices is considered. Each Hamiltonian walk, i.e. a walk which covers every point on the lattice once and only once, is identified with a possible globular (completely dense) conformation of the polymer chain with  $N$  monomers. The great advantage of this model is all the conformations can be exhaustively enumerated, so that the partition function and all the thermodynamic properties can be found exactly (which is important for the system in which one particular microstate is expected to give the overwhelming contribution to the partition function).

To construct a chain of  $N$  monomers, we first place  $N$  particles of  $q$  different species on the lattice vertices (one and only one particle in each vertex). We then swap monomers and let them equilibrate at some given temperature  $T_p$  using

the standard Metropolis algorithm [Met53]. The energy of interactions between monomers is assumed to be short range. Thus, some symmetric interaction matrix  $J_{ij}$ , where  $i, j = 1 \dots q$ , can be used to define the interaction energy between monomers of species  $i$  and  $j$ .

This is our model of a condensed mixture of monomers. When this mixture reaches equilibrium, we instantly break its movement at the current microstate, and polymerize the monomers by applying a globular conformation randomly chosen from our list of enumerated conformations. This is the “polymerization conformation.” This procedure is illustrated in Figure 4-1.

To determine whether the prepared chain is renaturable, we have to explore its conformational space. Using our list of all compact conformations, we calculate the total energies of volume interactions for each of the possible conformations of the prepared chain, and ask (i) is the polymerization conformation the ground state (minimal energy among other conformations), and (ii) is this ground state non-degenerate. If these conditions are both met, the chain is said to be thermodynamically renaturable.

With chains that have been shown to be thermodynamically renaturable, we have also addressed the question of kinetic renaturability, i.e. do the chains fold quickly and reliably to the polymerization conformation. Kinetic renaturability was addressed by using a Monte Carlo procedure which starts with an imprinted heteropolymeric chain at high temperature (therefore, the chain is a coil) and then the temperature is sharply lowered. The kinetics algorithm employs only three fundamental movements of the monomers in the chain; these are described in Figure 4-2. At each Monte Carlo iteration, a monomer along the chain is chosen at random and the partial partition function corresponding to all of the possible elemental moves is calculated. Using the Metropolis criterion, a move is selected and the chain accordingly changed. We have included the null move, i.e. no change, in the partition function. Hopefully, this will improve the correspondence between Monte Carlo time and the real time of the folding kinetics of real chains.

In order to show that this polymerization procedure in general is capable of

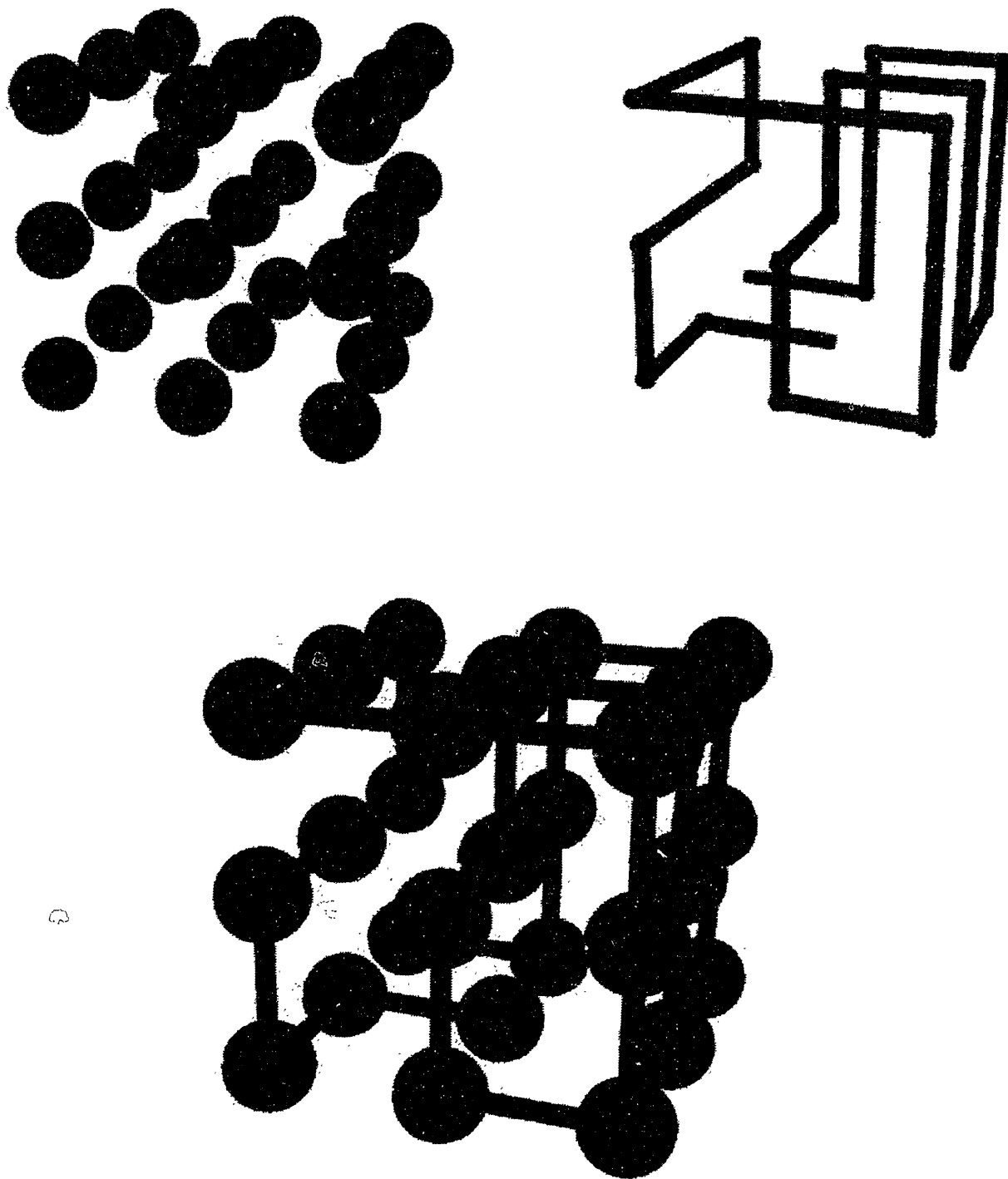


Figure 4-1: The Imprinting Model. Clockwise from top left: monomer solution, polymerization conformation, and prepared polymer.

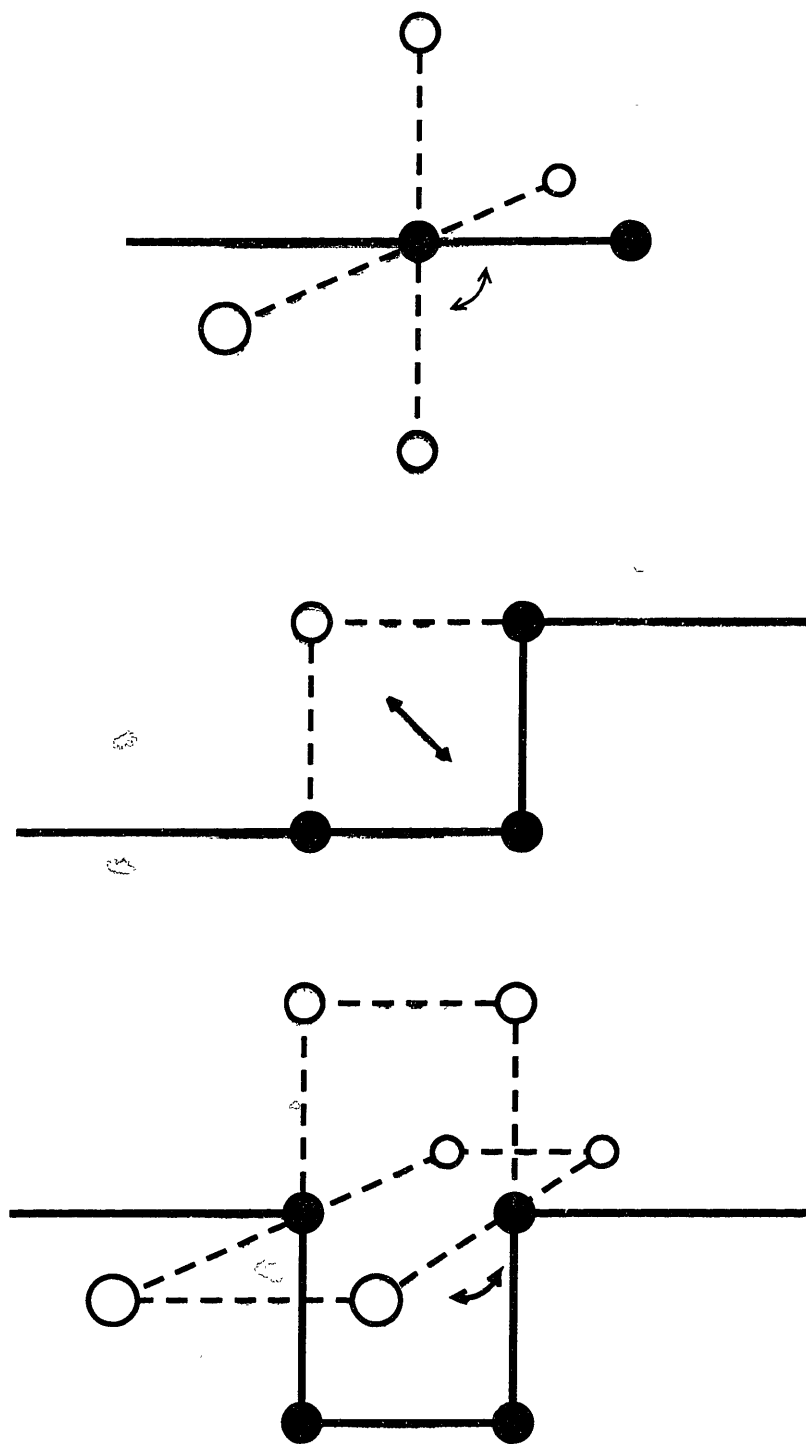


Figure 4-2: The three elementary moves employed in the Monte Carlo kinetics of 3D lattice polymers performed. From top to bottom: movement of the end monomer of the chain, L flip, and crankshaft. These 3 elementary moves can be combined to create more complex moves, and are believed to yield ergodic folding kinetics.

| $N$ | Conformations |
|-----|---------------|
| 26† | 174,056       |
| 26‡ | 564,368       |
| 27  | 103,346       |
| 36  | 97,720,079    |

Table 4.1: Number of conformations (Hamiltonian walks) not related by symmetry on the cubic sublattice. †The empty site is only in a face. ‡The empty site can be either in a face or in a corner.

creating renaturable polymers, i.e. independent on details such as the interactions chosen, etc., we have implemented the general scheme described above for different size polymers and different interaction matrices. Specifically, we have investigated the cases  $N = 27$  and  $N = 36$  (without target molecule) and  $N = 26$  (with one lattice site for a target molecule). We have enumerated these cases [Pan94a] and the results are summarized in Table 4.1.

We used three types of interactions: i) *Potts Interactions*: We made the simplest supposition on the character of nearest-neighbors interactions between those particles, namely, we attribute attraction energy  $-J < 0$  for each pair of identical interacting particles (i.e., occupying the neighboring lattice sites) and repulsion energy  $J > 0$  of the same absolute value to each pair of neighboring particles of different types. This exactly corresponds to the standard  $q$ -state Potts model. ii) *Random Matrix*: We constructed symmetric random matrices  $J_{ij}$  where each element  $(i, j)$ , with  $j \geq i$ , of the matrix was a random number chosen from a Gaussian distribution. iii) *Miyazawa and Jernigan*: We used a version of the matrix of interaction energies derived from amino acids by Miyazawa and Jernigan (MJ) [Mia85]. Specifically, we simplified the MJ potentials (denoted SMJ) by truncating the strongest interactions ( $J_{ij} < -0.2 \rightarrow -1$ ) and setting the rest to zero [Sali94b].



## 4.3 Thermodynamics

### 4.3.1 Potts Interactions (27-mers and 36-mers)

We have performed an examination of the thermodynamics of 27- and 36-mers whose monomers interact via Potts interactions. We have addressed several cases with different number of species  $q$  ( $q = 2, 3, 4, 5, 7, 8, 9, 14, 20$  for 27-mers and  $q = 4, 6, 7, 9, 12, 18$  for 36-mers). For each  $q$ , many polymer sequences (5000 for 27-mers, 300 for 36-mers) were created using the above procedure. For each sequence, the energy spectrum, i.e. the degeneracy at each energy level, was calculated.

As  $q = 1$  (all monomers are attracted to each other) and  $q = N$  (all monomer repel each other) are both homopolymers for Potts interactions, we expect to find some maximum in the heteropolymeric properties at intermediate  $q$ . This maximum is seen perhaps most dramatically in the probability of creating a renaturable chain  $P_{\text{renat}}$ . As shown in Figure 4-3, we find a peak in heteropolymeric properties at  $q \approx 7$  for  $N = 27$  and  $q \approx 9$  for  $N = 36$ . Note that the value of  $P_{\text{renat}}$  is surprisingly high, i.e. more than half the chains were renaturable. Also, as the  $P_{\text{renat}}(q/N)$  curve seems independent of  $N$ , we conclude that there are most likely no “dangerous” terms in  $P_{\text{renat}}$  such as  $\exp(-N)$ .

It has been argued [Sali94b] that the favorability of the kinetics of a given sequence is strongly linked with the energy gap between the ground and the first excited energy states. We will address this later in Section 4. As we have calculated the energy for all of the possible conformations of the polymers, we can directly obtain statistics of this energy gap. We find similar behavior as that found for  $P_{\text{renat}}$  i.e., we see again a peak at intermediate  $q$ , and the values of  $P_{\text{gap}}$  are also reasonably large. Thus, we would expect that a large fraction of the chains produced are also kinetically renaturable. This will be addressed in the Kinetics section of this chapter.

We can describe the thermodynamics of the transition from high temperature, where many conformations contribute more or less equally, to low temperature,

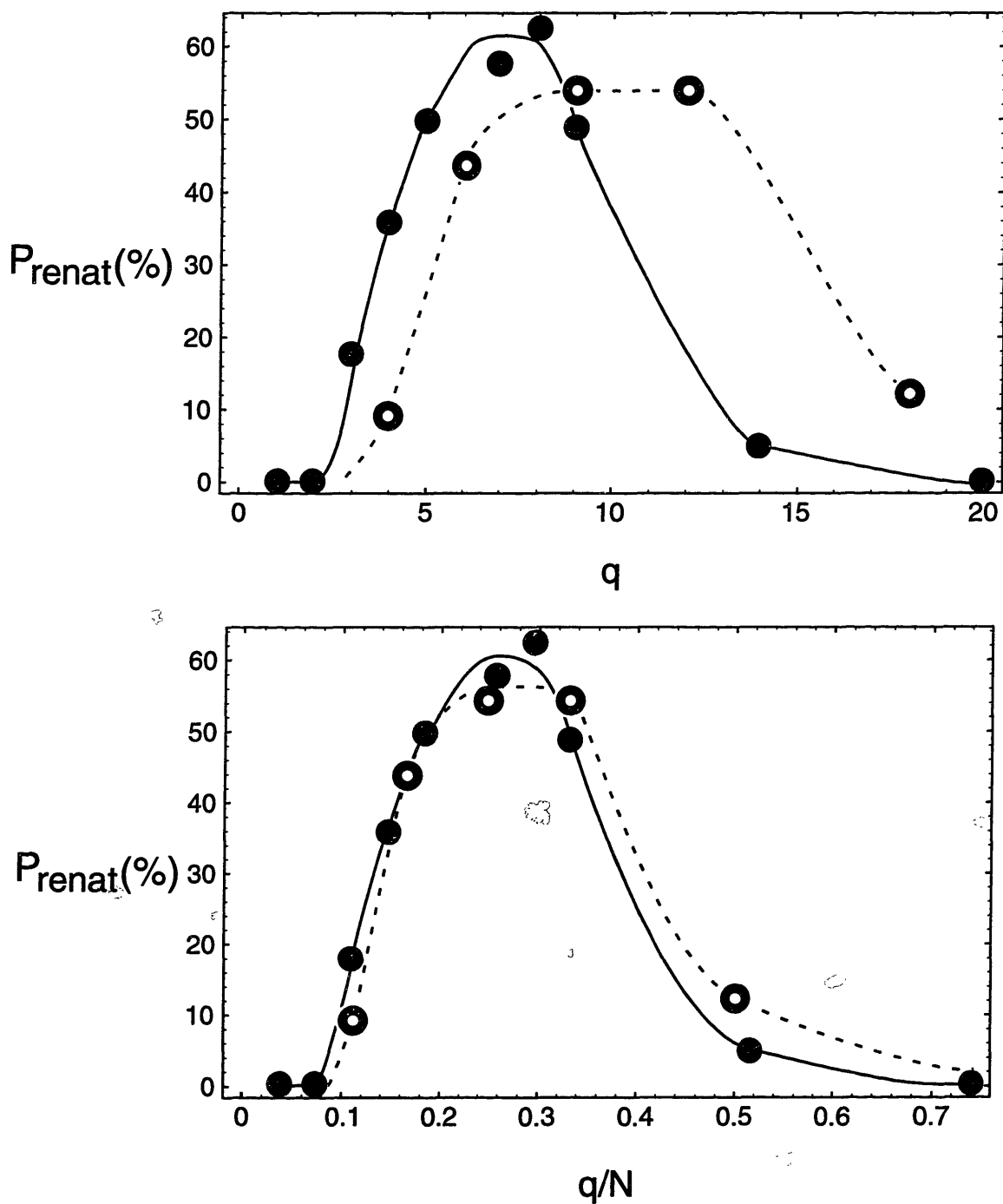


Figure 4-3: Probability for renaturation ( $P_{\text{renat}}$ ) for 27-mers ( $\bullet$ ) and 36-mers ( $\circ$ ) with Potts interactions; the lines are meant solely to guide the eye. a) As the limiting cases  $q = 1$  and  $q = N$  are both homopolymers, we expect to find a peak in heteropolymeric properties at intermediate  $q$ . b) Note that the values of  $P_{\text{renat}}$  for different  $N$  coincide when  $P_{\text{renat}}$  is plotted vs  $q/N$ .

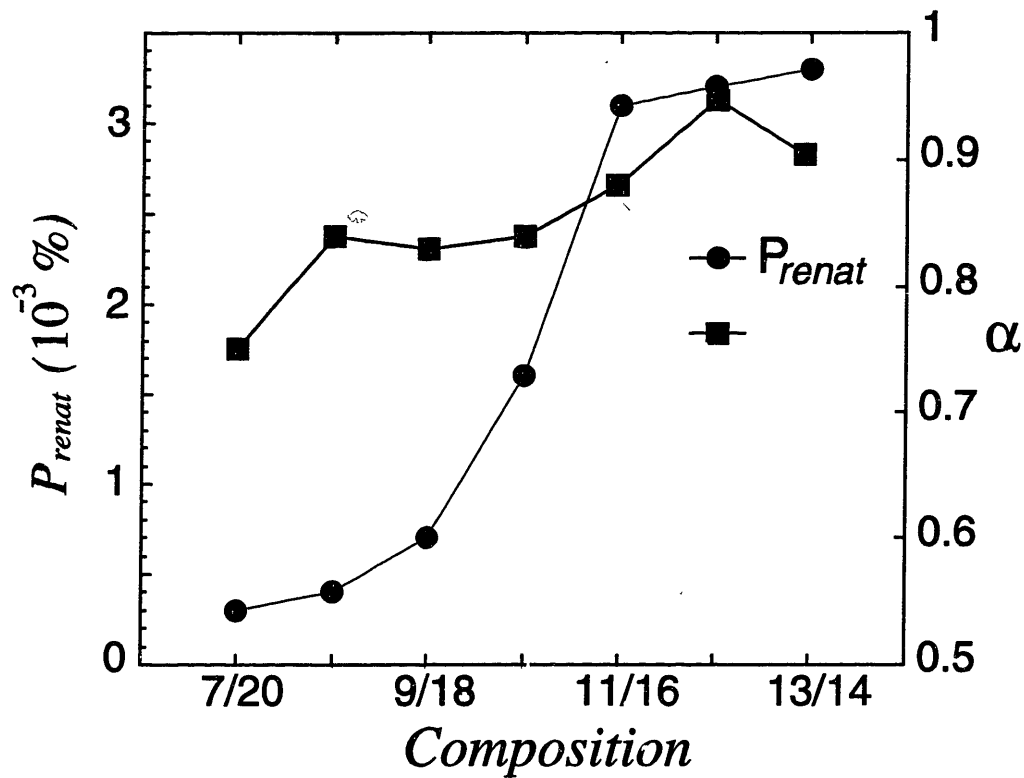


Figure 4-4: 27-mers: For the two letter model, we vary the asymmetry in composition (i.e. how many black/white). We see that as we go farther from an even composition 13/14, the probability of renaturation  $P_{renat}$  decreases quickly. Physically, this is due to the fact that we are approaching the homopolymeric limit.

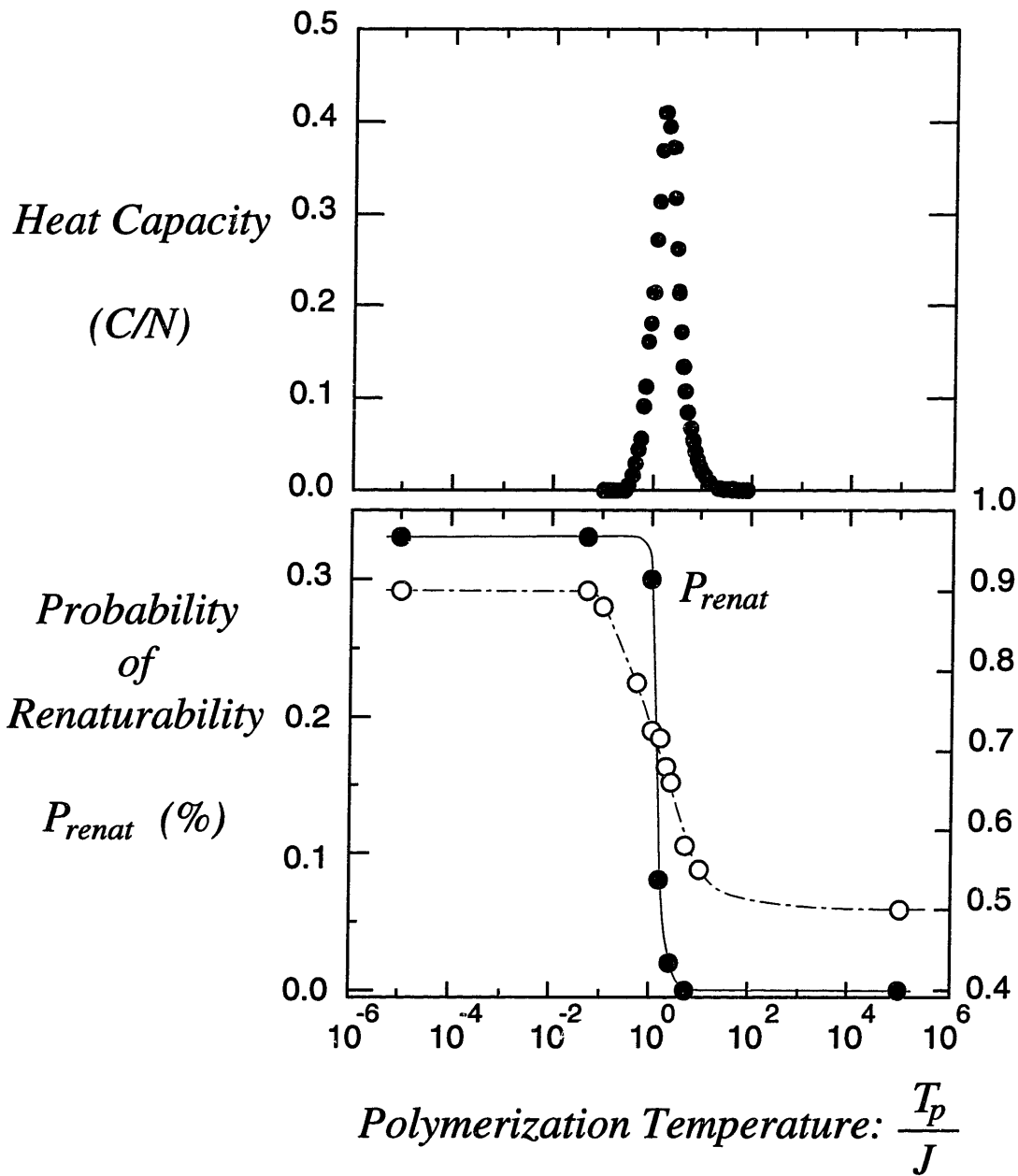


Figure 4-5: 27-mers: If we plot the heat capacity with  $P_{\text{renat}}$ , we see that there is indeed a sharp phase transition. Interestingly, the value of the correlation exponent  $\alpha$  is more sensitive to temperature change than  $P_{\text{renat}}$ .

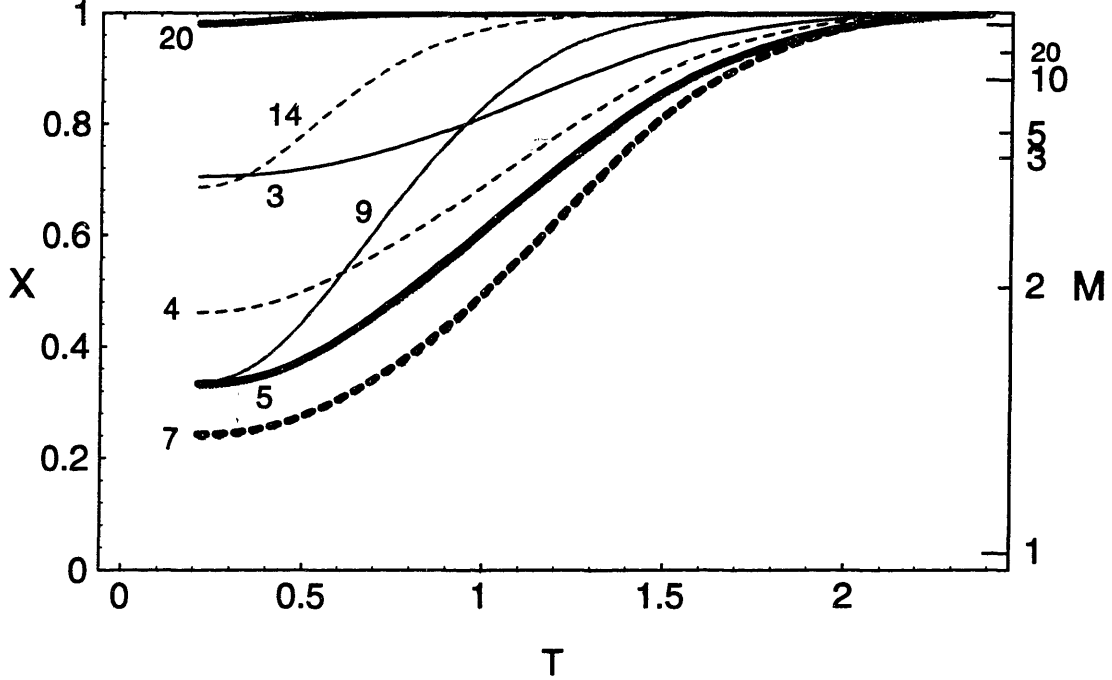


Figure 4-6:  $\langle X(T) \rangle$  (averaged over the ensemble) vs  $q$  for  $q = 3, 4, 6, 7, 9, 14, 20$ . The number near each curve denotes  $q$ . There are two important effects to observe. As the number of thermodynamically relevant states is related to  $X$  as  $\mathcal{M} = 1/(1-X)$ , we see that there is a minimum in the average number of relevant states at low temperature for intermediate  $q$ . The sharpness of the curves varies with  $q$  as well.

where one or a few state(s) dominate, by the order parameter

$$X(T) = 1 - \sum_i^{N_{\text{confs}}} p_i^2 \quad (4.1)$$

where

$$p_i = \frac{\exp(-\epsilon_i/T)}{Z}, Z = \sum_i^{N_{\text{confs}}} \exp(-\epsilon_i/T) \quad (4.2)$$

$p_i$  is the Boltzman probability of finding the system in the state  $i$  with energy  $\epsilon_i$  at temperature  $T$  and  $N_{\text{confs}}$  is the total number of conformations (microstates).  $X$  can be related to the total number of thermodynamically relevant states  $\mathcal{M}$  by  $\mathcal{M} = 1/(1-X)$ . Indeed, for the case where only one state is thermodynamically relevant ( $p_1 = 1$  and all other  $p_i = 0, i > 1$ ) then  $X = 0$  and  $\mathcal{M} = 1$ . For the case where all states have equal probability ( $p_i = 1/N_{\text{confs}}$ ), then  $X = 1$  and  $\mathcal{M} = \infty$ .

Figure 4-6 shows  $X(T)$  for different  $q$ . The average degeneracy of the ground

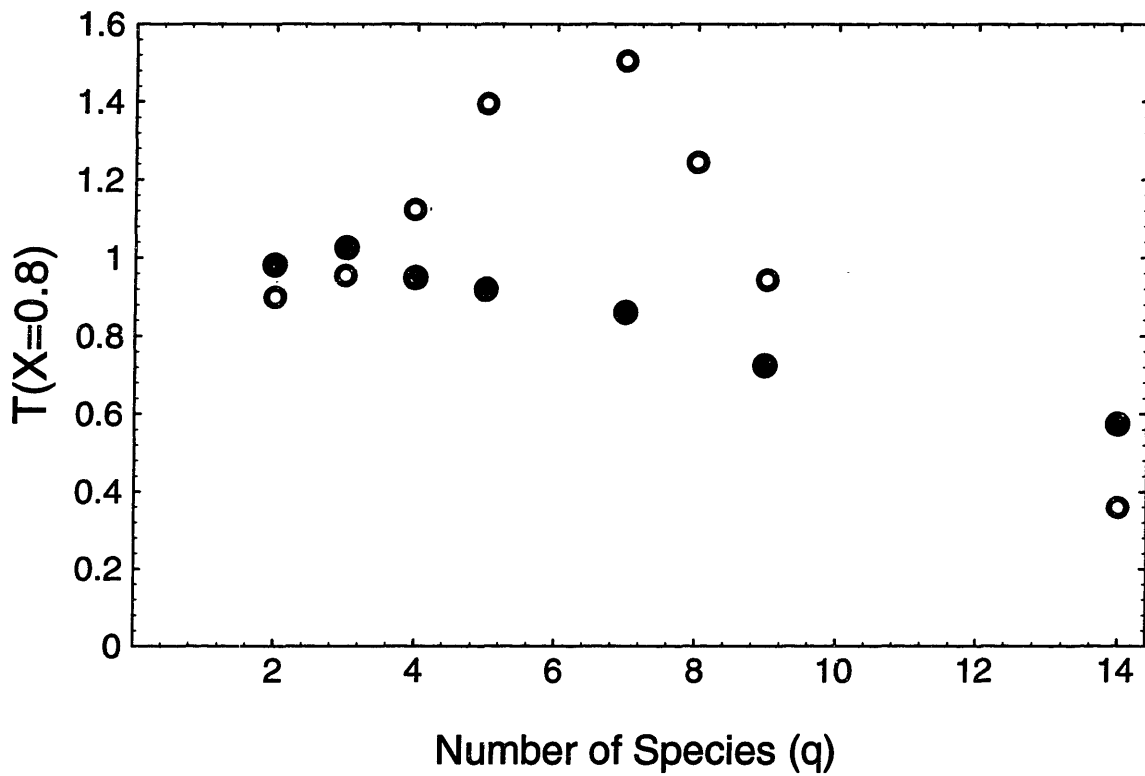


Figure 4-7:  $T(X = 0.8)$ , which is related to the freezing temperature  $T_f$ , vs  $q$  for Potts model Imprinted ( $\bullet$ ) and random ( $\circ$ ) ensembles. We see a peak in heteropolymeric properties for Imprinted chains, but not for random chains.

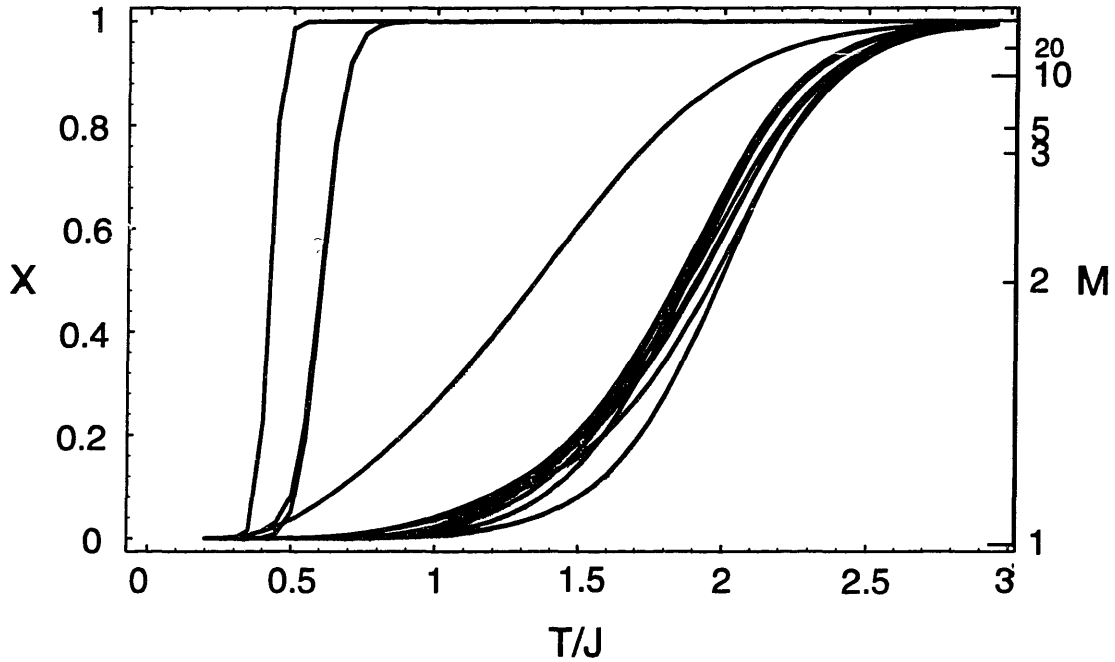


Figure 4-8:  $X(T)$  for each sequence from the ensemble of  $q = 7$  chains with unique ground states and an energy gap. There is a clear bimodality between sequences with a sharp  $X(T)$  and a smooth  $X(T)$ .

state is given by  $1 - X(T \approx 0)$ . Apart from variations in this vertical offset, we see that the curves differ in the temperature at which  $X(T) \approx 1$ . If we plot  $T(X = 0.8)$ , which is related to the freezing temperature of the polymer [Sali94b], vs  $q$ , as shown in Figure 4-7, we see that there is again a maximum at  $q = 7$  for 27-mers and  $q = 9$  for 36-mers. This maximum means that freezing occurs most easily at the optimal heterogeneity of the chain. Indeed, at small  $q$ , freezing is prevented by the segregation of the monomer mixture which leaves many possibilities for the chain to rearrange conformation within the homogenous domain. On the other hand, large  $q$  means that there are very few monomers of each kind, which again means that inside certain domains, the chain can be rearranged without a change in energy. This optimum of  $T_f$  at intermediate  $q$  is exactly the manifestation of the principle of minimal frustration [Bry87].

Finally, we consider  $X(T)$  for  $q = 7$  (chosen as it has the greatest heteropolymer effects and therefore is the most interesting) for sequences which have a unique

ground state and an energy gap between the ground state and first excited state. As shown in Figure 4-8, there is a clear bimodality seen in the  $X(T)$  for these sequences. This fact is very interesting, as it is unclear what causes the differences in  $X(T)$  between the two possible groups. One conjecture was that there should be some differences in the kinetics of the two groups. This was not found. Therefore the physical meaning of this bimodality remains unclear.

In summary, for Potts model one would expect that heteropolymer properties, such as the presence of unique structure, would be most visible when the nature of the interactions are “most heteropolymer.” As  $q = 1$  and  $q = N$  are both heteropolymers, this translates into maximums in heteropolymer behavior at intermediate  $q$ . It is interesting that we did not find any significant differences between the cases  $N = 27$  and  $N = 36$ . In fact, when we examine heteropolymeric aspects, such as the probability of renaturability ( $P_{\text{renat}}$ ) vs  $q/N$ , we found no particular dependence on  $N$ . While this cannot prove that this polymerization procedure should work for arbitrary  $N$ , it strongly indicates that extrapolation to realistic polymer lengths is reasonable.

### 4.3.2 Potts Interactions: Polymer with Target Molecule (26-mers)

On the  $3 \times 3 \times 3$  sub-lattice, we have left the center site on a face purposely empty during enumeration. This yields all of the 26-mers on the  $3 \times 3 \times 3$  sub-lattice. The intentional hole is used to model a target site. As the maximum for 27-mers was found at  $q \approx 7$ , we used the 7 letter Potts model to imprint sequences for a particular 26-mer conformation in the presence of a “target molecule.” The target molecule was modeled as a cube which physically fits in the empty site in the 26-mer globule, but has a different species on each face. We calculated the energy spectrum for these sequences *without* the target molecule and found that a large percentage of them (64 %) were thermodynamically renaturable. Thus, thermodynamically, one expects that the chains should fold back to the polymerization conformation



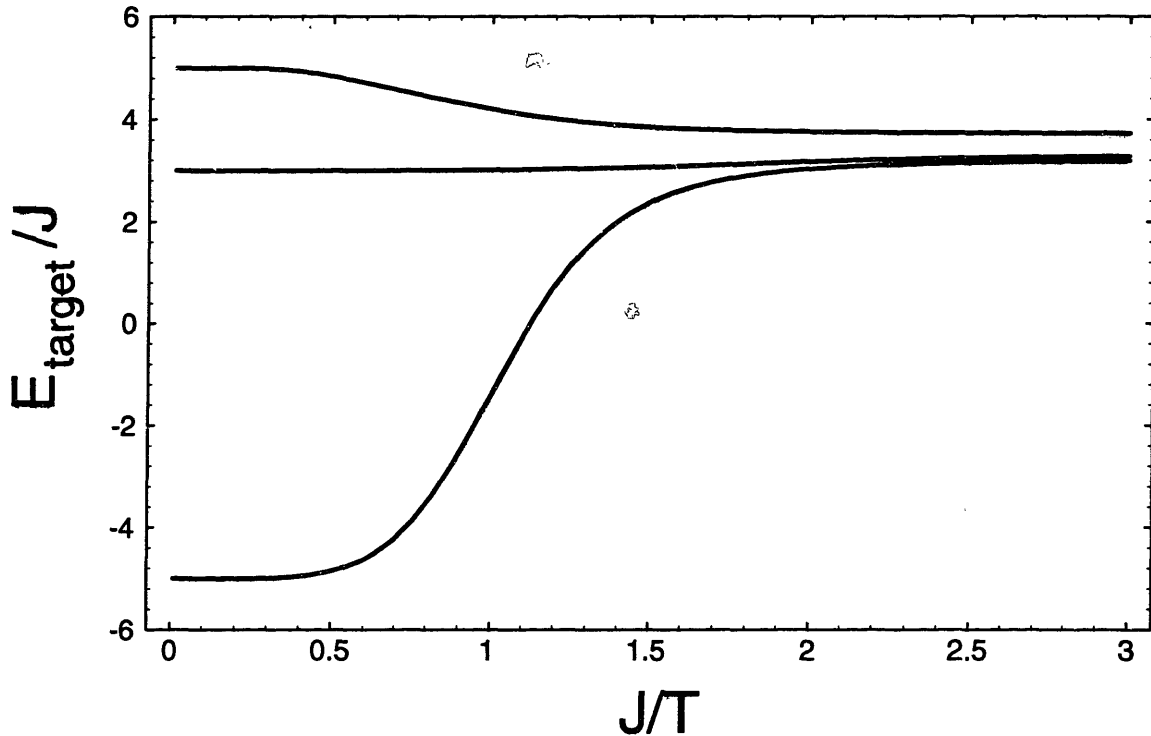


Figure 4-9: Examples of the energy of the active site of a 26-mer vs temperature for chains designed with (bottom) and without a target molecule (top two). At infinite acting temperature, there is little difference in the affinity with the target molecule. At low acting temperatures, there is a large difference between sequences imprinted with the target molecule and without.

without the target molecule, and would therefore create a site ready to accept this target molecule with high specificity.

We also calculated the energy of the active site vs temperature for 26-mers of three ensembles: 1) designed in the presence of the target molecule, 2) designed without a target molecule 3) random sequences. All of the renaturable 26-mers designed (i.e. low  $T_p$ ) in the presence of the target molecule had a completely complementary active site ( $E_{\text{target}} = 5J$ ) in the ground state conformation. For the complete ensemble of chains designed with the target molecule (i.e. including even the sequences with degenerate ground states), we found  $\langle F_{\text{target}} \rangle = -3.9J$ . As the acting temperature is increased,  $\langle F_{\text{target}} \rangle$  undergoes a phase transition around  $T \approx J$  and reaches a high temperature value of  $\langle F_{\text{target}} \rangle = 3.3J$ . On the other hand, chains designed without the target molecule had  $\langle F_{\text{target}} \rangle = 3.8J$  at low acting temperature

|                      | 9-Potts | 9-letter Random | SMJ |
|----------------------|---------|-----------------|-----|
| $P_{\text{renat}}$   | 54%     | 26%             | 73% |
| $P_{\text{des}}$     | 53%     | 15%             | 72% |
| $P_{\text{gap}}$     | 28%     | 12%             | 12% |
| $P_{\text{des gap}}$ | 28%     | 8%              | 12% |

Table 4.2: The probability of renaturation ( $P_{\text{renat}}$ ), probability of renaturation to the polymerization conformation ( $P_{\text{des}}$ ), the probability of the presence of a gap between a unique ground state and the first excited state ( $P_{\text{gap}}$ ), and the probability that the ground state is unique, the polymerization conformation, and has an energy gap ( $P_{\text{des gap}}$ ) for 36-mer Imprinted chains for different types of interactions: 9-letter Potts, 9-letter random matrix, and the SMJ matrix.

and undergoes a phase transition to  $\langle F_{\text{target}} \rangle = 3.5J$ . This is shown in Figure 4-9. The differences in the high temperature limit of  $\langle F_{\text{target}} \rangle$  are not significant. For comparison, a randomly arranged monomer mixture has  $\langle F_{\text{target}} \rangle = 3.5J$ . Thus, it is clear that 1) imprinting allows the formation of a renaturable chain with the active site present in the monomer mixture prior to polymerization and 2) the specificity of this active site, even in this simple 26-mer model, is significant.

### 4.3.3 Different Interactions (36-mers)

In order to examine whether this polymerization procedure is valid for a variety of different types of monomer-monomer interactions, we repeated the procedure above for different types of interactions. For the three types of interactions examined (Potts, Random matrix, Miyazawa and Jernigan), we generated 300 chains as described above and calculated the energy spectrum of each chain, i.e. the energy distribution of the particular sequence over all possible conformations. Table 4.2 shows the results for the three types of interactions. We see that the yields vary less than an order of magnitude. Thus, we find that the “efficiency” of production of our method is not dependent on the specific nature of interactions, but merely on the heteropolymeric character of the interactions. This was suggested earlier in the chapter for  $q$ -state Potts model interactions.

We have found that a significant fraction of the chains produced have an energy gap, again independent of the type of interactions used. However, while there is no great qualitative difference in these interactions as far as the thermodynamic quantities are concerned, they do have some effect on kinetics, but only due to the differences in the average interaction energy. When this is compensated for, there are no significant differences.

In summary, we have established that the polymerization scheme described earlier and quantitatively studied above can indeed produce a significantly large yield of polymers which are capable of renaturing to their polymerization conformation for lattice model chains of length  $N = 36$ .

#### **4.3.4 Comparison to other polymer ensembles (36 mers)**

In order to understand the meaning of these results, we compare our method of preparing polymer sequences from two other ensembles: i) polymers prepared using the “sequence annealing” method of Shakhnovich and Gutin (SG) [Sha93b] and ii) random sequences. The SG method minimizes the energy of a polymer by swapping monomers. Formally, the SG method is similar to ours as both minimize energy, except we include all of the monomer interactions (the energy of the monomer soup) whereas the SG method excludes the polymer bonds (the energy of the polymer). In spirit, however, these methods are very different as our method models an experimentally realizable procedure whereas the SG method is perhaps best a model of biological evolution. Random sequences are a useful control group with which to compare.

In order to quantitatively compare these three methods, we generated 300 chains using the SMJ interaction matrix for all three design methods and calculated the energy spectrum for each sequence. We now examine several criteria. First, consider the “efficiency” of the three methods at producing renaturable polymers ( $P_{\text{renat}}$ ), as shown in Table 4.3. We must emphasize that this criterion of renaturability only requires that the ground state be non-degenerate, not that the ground state be the target conformation in the SG method or the polymerization conformation in our

| Characteristic                                     | Imprinted | SG    | Random |
|--|-----------|-------|--------|
| $P_{\text{renat}}^{\text{one}}$                    | 50 %      | 94.5% | 29.0%  |
| $P_{\text{renat}}^{\text{many}}$                   | 73%       |       | 29.0%  |
| $P_{\text{des}}^{\text{one}}$                      | 48 %      | 95%   | 0      |
| $P_{\text{des}}^{\text{many}}$                     | 72%       |       | 0      |
| $P_{\text{gap}}^{\text{one}}$                      | 9.5%      | 63%   | 0%     |
| $P_{\text{gap}}^{\text{many}}$                     | 12%       |       | 0%     |
| $\langle E_{\text{gnd}} \rangle$                   | -30       | -36   | -21    |
| $\langle E_{\text{gnd all}} \rangle$               | -54       | -42   | -28    |
| $\langle E_{\text{chain}} \rangle$                 | -25       | -6    | -7     |
| $\langle E_{\text{prob}} - E_{\text{gnd}} \rangle$ | 20        | 25    | 14     |

Table 4.3: Comparison of design methods in thermodynamics. Averaging can be performed over many runs with a particular polymerization (target) conformation ( $P_{\dots}^{\text{one}}$ ) or over an ensemble of polymerization conformations ( $P_{\dots}^{\text{many}}$ ). The probability of the existence of a unique ground state ( $P_{\text{renat}}^{\dots}$ ), successful design and renaturability ( $P_{\text{des}}^{\dots}$ ), and successful design and renaturability with an energy gap ( $P_{\text{gap}}^{\dots}$ ) are examined. Also, characteristics of the ground state such as the average ground state polymer energy ( $\langle E_{\text{gnd}} \rangle$ ), i.e. not including bonds along the linear sequence of the polymer, monomer energy ( $\langle E_{\text{gnd all}} \rangle$ ), i.e. including all nearest neighbor bonds, chain energy ( $\langle E_{\text{chain}} \rangle$ ), i.e. including only the bonds along the chain, and energy difference between the most degenerate energy state and the ground state ( $\langle E_{\text{prob}} - E_{\text{gnd}} \rangle$ ) are useful in comparing design methods.

method. This major difference is seen in the probability of successful design ( $P_{\text{des}}$ ). While random chains do have a reasonably high chance of having a unique ground state (i.e. large  $P_{\text{renat}}$ ), the probability that this ground state is the polymerization conformation is given by  $P_{\text{des}} \approx 1/M$ , where  $M$  is total number of possible conformations.

Interestingly, there is a difference when we calculate  $P_{\text{renat}}$  and  $P_{\text{des}}$  using one design conformation to generate many different sequences or using a new conformation for each new sequence. The former method is more natural to the SG design procedure, as one has a given desired target conformation, and the latter is used in the Imprinting method, where a conformation is chosen randomly corresponding to some random polymerization. The discrepancies between using one or many design conformations reflects biases due to the geometry of a particular conformation. These biases are averaged out when many conformations are used.

The percentage of chains which have an energy gap is shown in Table 4.3. It is also interesting to examine the probability distribution of gap sizes in the chains produced using the three methods, as shown in Figure 4-10. We see that it is extremely unlikely that random chains have an energy gap. This is a well known property of random heteropolymer chains [Sali94b]. It is interesting to note that while the SG method produces a greater yield of chains with an energy gap overall, and a larger average gap, the Imprinting method does indeed produce chains with an energy gap and for small gap sizes, produces more of them.

In addition to the size of the energy gap between the ground and first excited states, it is also interesting to consider the depth of the “energy well” of the ground state. From Figure 4-11 we see clearly that there exists a continuous Gaussian-like part of the energy spectrum. The energy difference between the peak of this spectrum, i.e. the most likely energy, and the ground state also is very telling in terms of folding kinetics and ground state stability. Again, we find that the random chains have the least difference in energies and the SG method the most. This effect is also reflected in the average ground state energy itself.

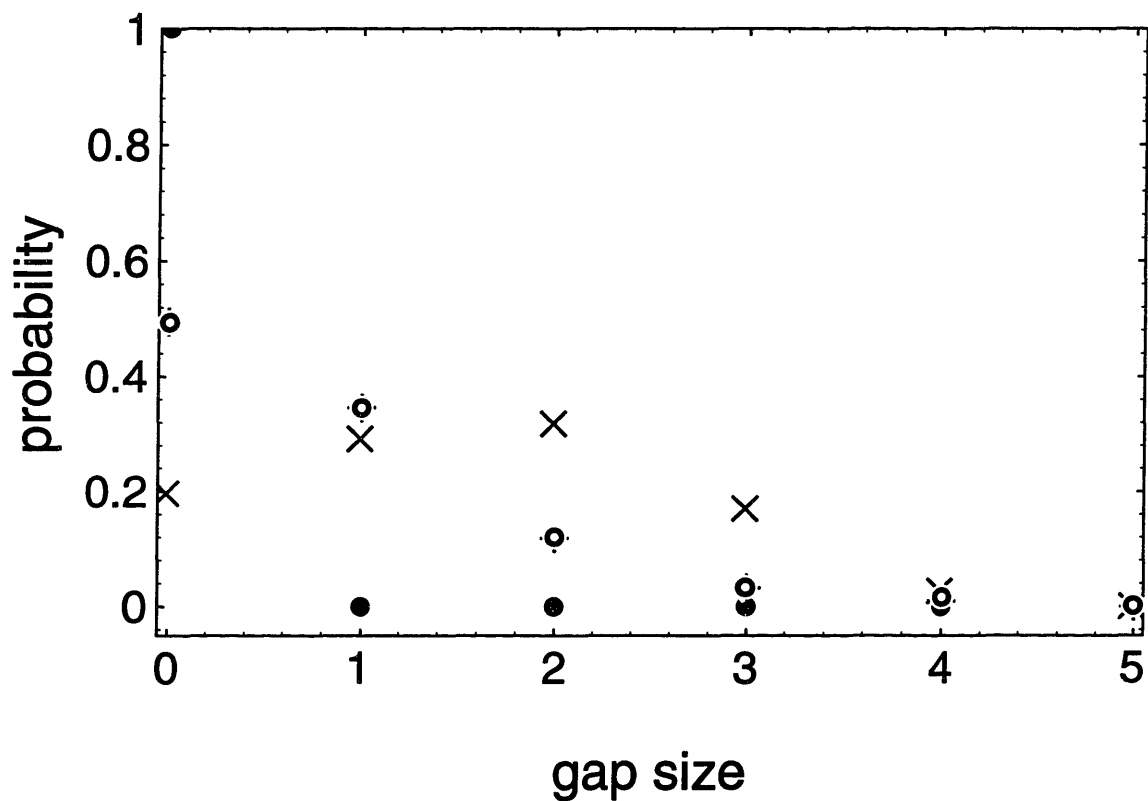


Figure 4-10: Probability distribution for a chain with a given energy gap size for Imprinted design (○), SG design (×), and random chains (●). For small gap sizes (gap=1), Imprinting produces a slightly higher yield. Both design procedures do significantly better than random chains.

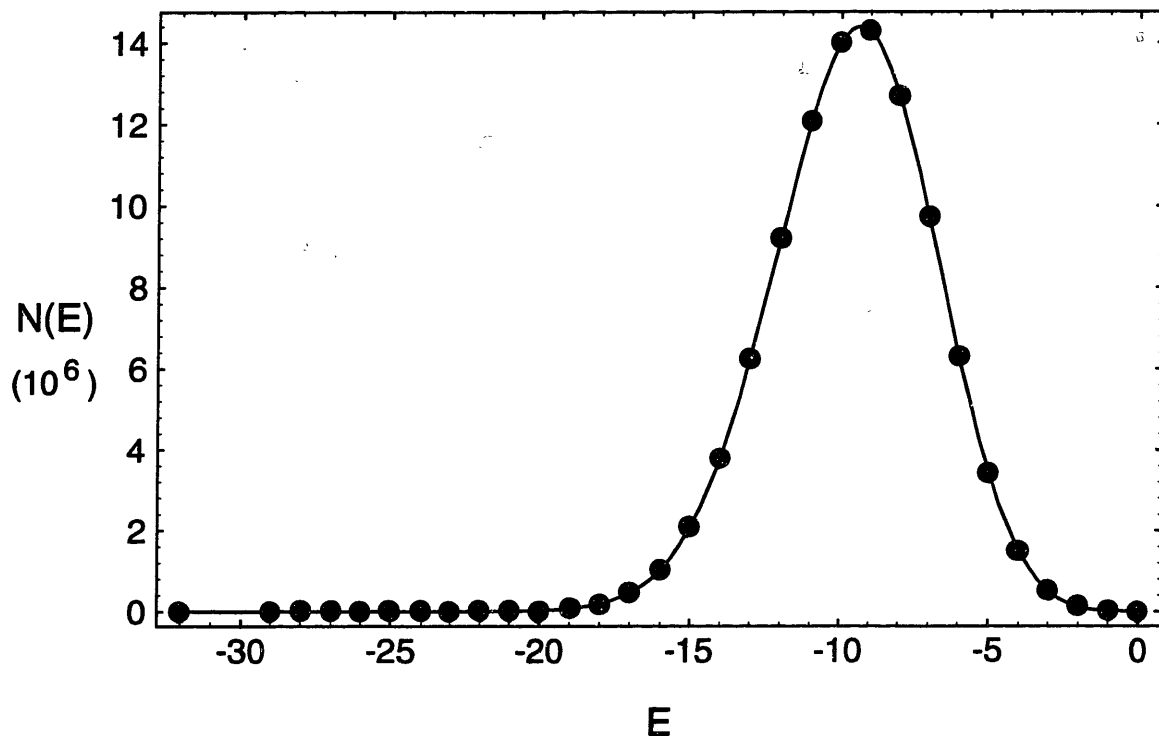


Figure 4-11: Example energy spectrum for  $q = 9$  Imprinted 36-mer with Potts interactions. Note the gap between the ground state and the first excited state.

## 4.4 Kinetics

We have stated that we must enumerate all of the conformations in the thermodynamic approach to examining the imprinting model as we expect that only one microstate should dominate the Hamiltonian. In fact, this is exactly what we are trying to prove. We furthermore stated that it is difficult to examine the conformation space by some Monte Carlo approach. The most effective and physically meaningful Monte Carlo search in the conformation space of a heteropolymer sequence is the process of Monte Carlo folding of a given sequence. As we have designed the sequences with a specific set of interactions, we can use this exact set in the Monte Carlo simulations as well. In this sense, like the calculation of the energy spectrums above, we can examine the method of imprinting self-consistently, using the same potentials involved in the design procedure as well as the Monte Carlo folding.

The major pitfall in the examination of renaturability via kinetics is that in

order to guarantee that a target conformation is in fact the ground state, one must perform many folding runs and hope that, if the target conformation is not the ground state, that the true ground state will be reached. However, as non-compact states have fewer polymer-polymer contacts and more polymer-solution contacts, we would expect that the ground state of a given sequence will be compact. Therefore, we have enumerated the compact states of both 27 and 36-mers. When performing kinetics studies, we can thus be certain that a given ground state is in fact unique. Also, we can now examine the relationship between the energy spectrum and folding kinetics.

In laboratory experiments of heteropolymer kinetics, one can begin with an ensemble of denatured chains and observe the time dependence of renaturation. In our computer experiments, we essentially do the same procedure. We start each Monte Carlo kinetics run with the chain at some random coil configuration, by setting the temperature to effectively infinity and allowing the chain to relax. Next, we reduce the temperature sharply to some acting temperature  $T$ , and record how many Monte Carlo iterations were needed for the chain to fold to the polymerization conformation. We then repeat this process for many runs. Thus, we have folded an ensemble of chains from the denatured state and have observed the time dependence on renaturation.

There are some general results which seem to be independent of the nature of interactions or chain length. First, the distribution of folding times of the ensemble of runs for a given sequence has been observed to be exponential. Thus, this distribution is quite broad and one must perform several kinetics runs in order to get a reasonable measure of the characteristic folding time  $\tau$ , where  $P(\tau) \approx e^{-t/\tau}$  is the probability of folding at time  $t$ . A sample distribution is shown in Figure 4-12. An exponential distribution is characteristic of the kinetics of overcoming a *single* free energetic barrier of height  $F_{\text{barr}}$  at temperature  $T$ . In this case, one would expect the characteristic time for crossing the barrier to be

$$\tau_{\text{barr}} = \exp\left(\frac{F_{\text{barr}}}{T}\right) \quad (4.3)$$



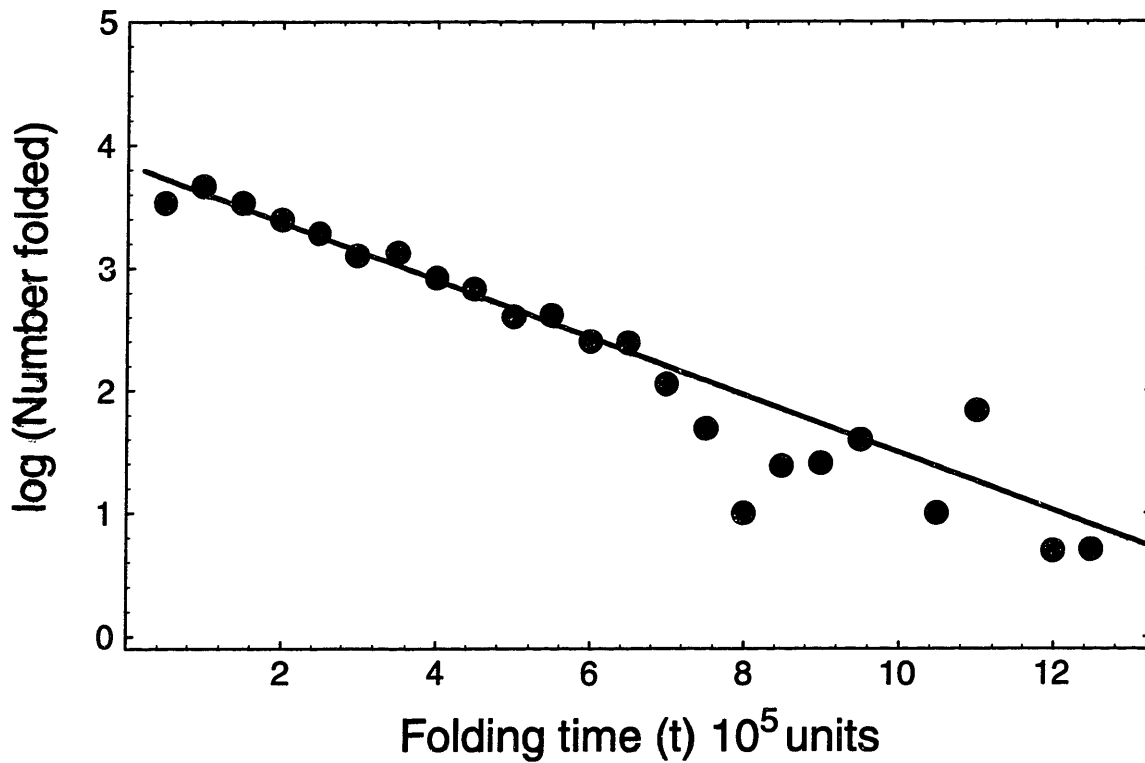


Figure 4-12: Histogram of folding times for Potts 27-mers ( $T = 0.8$ ):  $\log_{10}$  of the number of chains folded vs folding time ( $t$ ). The points are fit well by an exponential curve for short folding times, thus allowing the definition of a characteristic folding time,  $\tau$ . As for the long folding time regime, there are much fewer counts and therefore fluctuations dominate the histogram.

Thus, in our lattice model, folding to the native state is a process involving only a single barrier height, in comparison to, for example, many barrier heights with different energy heights. This free energy barrier can be calculated from our measurements of  $\tau$ .

Due to the CPU intensive nature of these calculations, the runs were performed on a massively parallel supercomputer (CM-5). Also, as we do not want to exclude extremely long runs which may be ignored due to finite length of computer time in a given computer job run, etc., we associate with each run a number used as the random number seed in the beginning of the run. Thus, the results were reproducible and any runs which were terminated prior to completion could be rerun.

Second, we have reproduced the results of Shakhnovich [Sali94b] that the folding time is decreased and chain stability is increased when the energy difference (“gap”) between the ground state and the first excited state is increased and the energy of the ground state is decreased. Finally, we have found that random chains are generally slower than designed chains in kinetics. This statement is of course related to the previous one, as the effect of design is to increase the gap and lower the ground state energy [Sha93b].

#### 4.4.1 Potts Interactions (27-mers)

It has been previously asserted [Sali94b] that all 27-mers, either random or designed, fold quickly and easily. In general, we have also found the same result. However, 27-mers still are a sufficiently robust model such that interesting temperature dependencies can be observed. We start with a specific sequence whose energy spectrum has been enumerated and it was found that the ground state of this sequence is relatively low in energy and non-degenerate. We now perform Monte Carlo kinetics on this sequence and perform many runs at various temperatures. At very high temperatures, the chain is coiled, as expected. As we lower the temperature, the coil to globule phase transformation occurs, but this globule is a “random globule” i.e. with no unique structure. In other words, this is essentially a homopolymer phase transition. As we gradually lower temperature further (below  $T_{\text{renat}}$ ), we find

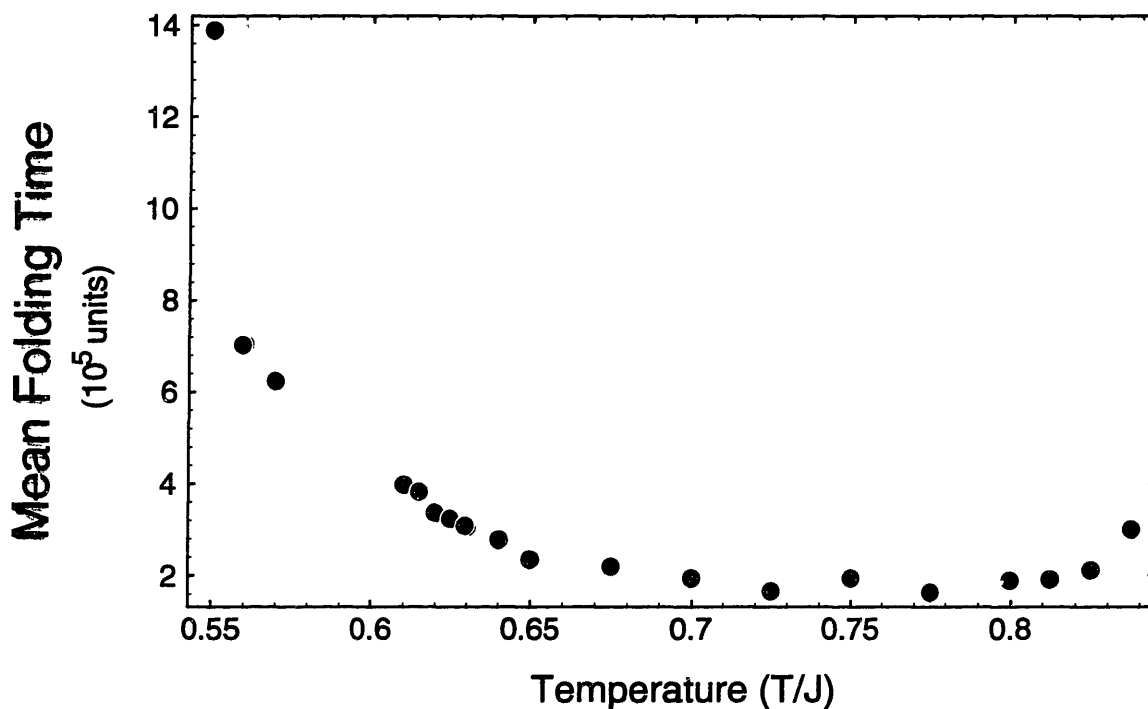


Figure 4-13: Characteristic folding time ( $\tau$ ) vs Temperature ( $T$ ). At high temperatures, entropy, i.e. the large number of relevant states, slows down folding where as at low temperatures, kinetics are slow as it takes longer to overcome energetic barriers. Thus, there is a minimum in  $\tau$  vs  $T$ .

that the system folds into the polymerization conformation. Upon even further temperature decrease (below  $T_g$ ), the system can easily get stuck in traps, i.e. the low energy (but not ground) states become metastable and thus further slow down the kinetics of folding to the ground state.

As it was previously stated, the probability distribution of folding times is exponential. Thus, many runs ( $\approx 1000$ ) were performed at each temperature. The temperature dependence of the characteristic folding time is shown in Figure 4-13. We see that there is an optimal temperature. If we exclude runs which have been slowed by traps (where the chain defined to have fallen in a trap is here defined as remaining in the same state for more than  $10^5$  Monte Carlo iterations), we get a slightly different temperature dependence of the characteristic folding time. From the above, we can calculate the height of this (free) energy barrier vs temperature, as shown in Figure 4-14. We see that there is some characteristic energy and the curve

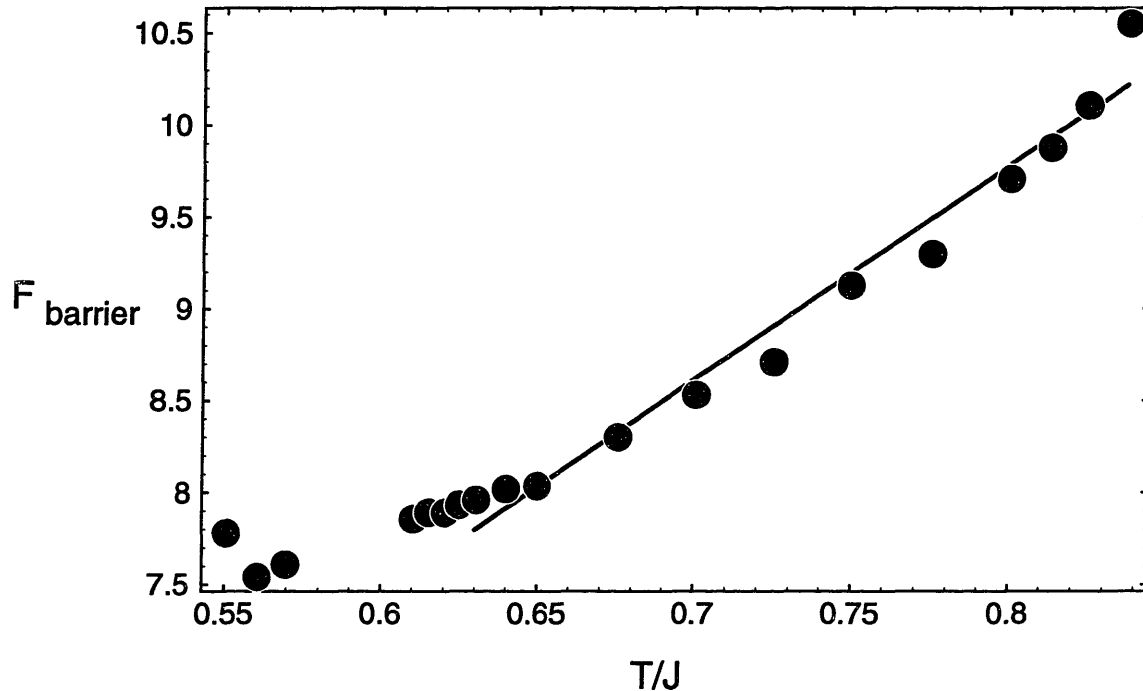


Figure 4-14: Characteristic free energy barrier height vs Temperature ( $T$ ). At low temperatures, the free energy is mostly constant, as the free energy is dominated by the energetic barrier.

is virtually constant at low temperatures. As temperature is increased, the entropy contribution becomes more pronounced and the free energy of the barrier increases. It is the competition between these two forces which creates a characteristic folding time minima.

It is also interesting to examine the probability of falling into a trap. This is shown in Figure 4-15. We see that there is a sharp transition between essentially never falling into a trap and always falling into a trap.

The results discussed above are very general and have been found to be qualitatively the same for different sequences and number of species. Thus, the very sensitive nature of temperature dependence of the folding kinetics seen in proteins, for example, is also seen in the Imprinted  $q$ -Potts 27-mers.

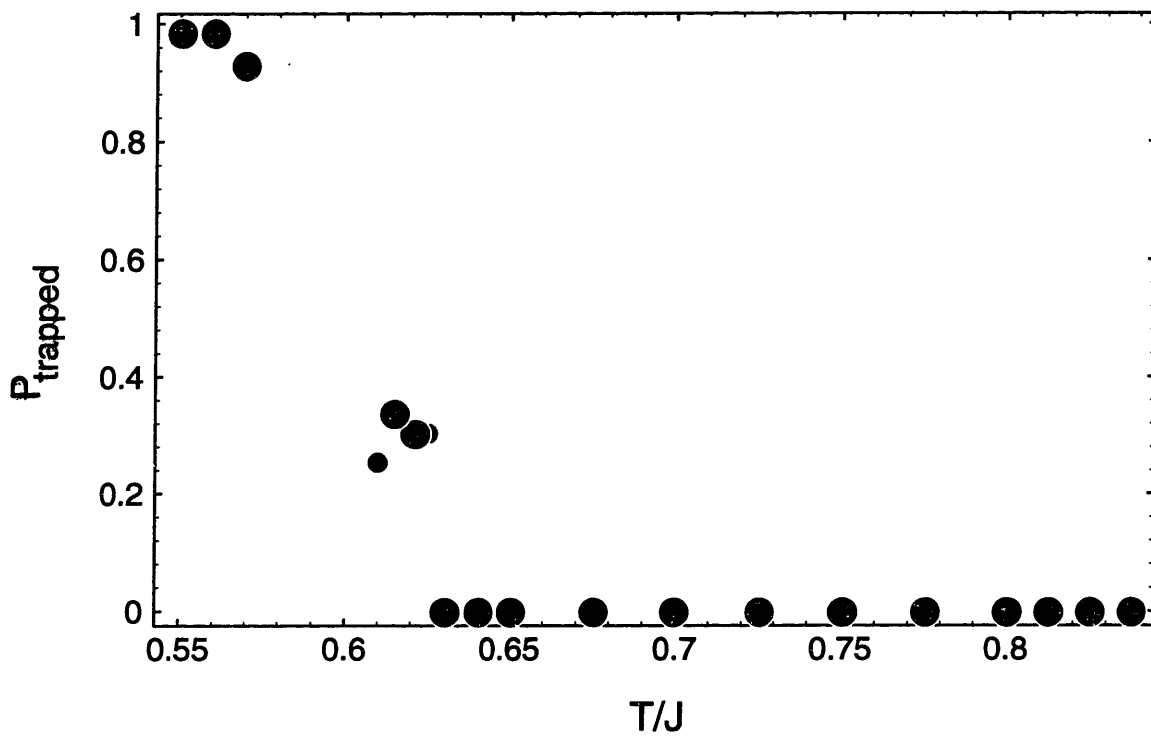


Figure 4-15: Probability of the system falling into a trap (metastable energy state) vs Temperature ( $T$ ). We see a transition between the temperature region where traps are rare to the low temperature case where almost all runs encounter traps.

#### 4.4.2 Polymer and Target Molecule Kinetics (26-mers)

For the 26-mers enumerated (cf Section 3.1) to be renaturable to the polymerization conformation (which has an absorption site for the target molecule present during polymerization), we ran Monte Carlo kinetics. First, we examined whether the chains would kinetically renature to the polymerization conformation without the target molecule. We found that folding behavior was in general similar to that of 27-mers, with the only difference due to different acting temperatures necessary. This is an important result, as with out the target molecules, bonds present during polymerization were absent during kinetic renaturation; however, the thermodynamically renaturable chains were able to successfully renature kinetically, and are thus ready to recognize the original target molecule.

We also examined the case where the target molecule is introduced during folding kinetics. The target molecule was allowed to rotate one turn and translate one lattice site in each possible direction. The Boltzman weights of these elementary moves were included in the partition function during Monte Carlo runs. We observed several different polymer-target molecule interactions:

1) *synergetic folding*: Often polymer folding was slowed due to the fact that the chain was stuck in a local minimum. Many of the polymer-polymer contacts were correct, so the polymer energy was relatively low, but often a part of the polymer in what should be the active site of the chain, or more simply the polymer was in a low energy, but not ground state. Upon introduction of the target molecule, the chain was able to quickly renature and the target molecule was recognized. This circumstance is reminiscent of the induced fit hypothesis in molecular biology.

2) *independent folding*: In this case, the non-specific attraction of the target to the polymer was not sufficient to keep the two molecules attached. Therefore the target molecule diffused around in the solution and the polymer folded independently. The probability of capture in this case has nothing to do with how well the target site is designed or how well the polymer folds but merely the probability of binary collision between these two particles. When the polymer has created an active site, there is enough non-specific attraction to keep the target molecule in

contact with the polymer such that the target molecule quickly diffuses to the active site, where it is caught in the energy well created by the specific active site.

3) *disruptive folding*: It was also observed, less frequently than synergetic folding, that the target molecule was capable of destroying a kinetics run which appeared to be near to the native state. Often this was due to the target molecules disruption of the critical nucleus or its capture into something other than the designed active site. As a stable polymer conformation cannot be obtained and the target is not strongly bound energetically to the very specific designed active site, the result of this circumstance was always the separation of target and polymer and the recommencement of refolding.

In general, independent folding was by far the most common because the non-specific attraction was insufficient to keep the target molecule and polymer together during polymer folding.

#### **4.4.3 Different Ensembles (36-mers, SMJ Matrix Interactions)**

For 36-mers, the amount of time necessary to fold is considerably more (in general around 2 orders of magnitude) than that of 27-mers. Thus, it is much more difficult to produce the folding time vs temperature relations discussed in the previous section. However, we have found qualitative agreement with the results found for 27-mers, i.e. as temperature is lowered, there are a series of transitions: coil to random globule, random globule to target globule, and the presence of traps. We have discussed these transitions using 27-mers as the quality of the statistics is greater and there are no differences found in 36-mers.

However, we now discuss a comparison of polymer ensembles from the aspect of the speed of folding kinetics. As previously discussed, we have three ensembles: SG designed, Imprinted, and random sequences. We previously designed 36-mer sequences using the SMJ interaction matrix and tested their thermodynamic renaturability. Here, we will discuss the results of the Monte Carlo kinetics simulations

| Characteristic                        | Imprinting      | SG              | Random               |
|---------------------------------------|-----------------|-----------------|----------------------|
| $\langle \tau \rangle_{\text{all}}$   | $1 \times 10^8$ | $5 \times 10^7$ | $\geq 2 \times 10^9$ |
| $\langle \tau \rangle_{\text{best}}$  | $2 \times 10^7$ | $1 \times 10^7$ | $1 \times 10^9$      |
| $\langle \tau \rangle_{\text{worst}}$ | $8 \times 10^8$ | $3 \times 10^8$ | $\geq 4 \times 10^9$ |

Table 4.4: Comparison of design methods in kinetics. We have performed 250 runs for sequences in each of the three ensembles: SG, Imprinted, and random.  $\langle \tau \rangle_{\text{all}}$  gives the characteristic time averaged over all sequences. “Best” and “worst” refer to the characteristic time of the *sequence* with the best and worst energy spectrum, from the point of view of kinetics (eg size of energy gap, etc.). Due to the long folding times of random chains, lower bounds for folding times are shown where necessary.

on these sequences to compare folding performance.

As one might suspect, the best performance was obtained for the highly optimized SG chains, then the imprinted chains, and the random chains were the worst. A comparison of average folding times for the three methods is shown in Table 4.4. Note that in order to compare methods, this average is over the ensemble of sequences and runs. There are two aspects that strongly separated the designed sequences from the random sequences in the thermodynamics analysis: reduced energy of the ground state and the presence of an energy gap between the ground state and the first excited state. These differences, in our simulations, do in fact make a significant difference in the folding kinetics. However, the *degree* to which these aspects are present, eg. how large is the gap, does not seem to be as important as the existence of these features. Thus, the best SG chains perform  $100\times$  better than the best random sequences, but only  $2\times$  better than the best imprinted sequences.

Thus, imprinted sequences, while not as optimized as SG sequences, are not significantly worse in folding kinetics than SG chains, and are significantly better than random chains. In a sense, the “incomplete” optimization (as compared to SG chains) seen is enough to make the sequences significantly better than random sequences.



## 4.5 Monomer-monomer correlations along the polymer sequence

Monomer-monomer correlations were recently discovered in ensembles of real proteins [Pan94c]. This was accomplished by “decoding” the protein’s amino acid sequence using a physically relevant mapping. For example, a “Coulomb” mapping was used to translate a sequence of amino acids into a sequence consisting of 1, 0,  $-1$  based upon the charge of the corresponding amino acid. The monomer-monomer correlations found in these “physically translated” sequences were such that they reflected some energy minimization procedure. For example, the Coulomb map lead to alternating correlations in proteins.

It is interesting to use the same procedure to detect correlations in polymers from the three ensembles considered in the last section. It was shown [Peng92,Pan94c] that a quantitative estimate of the degree of correlations could be obtained by the definition of a “critical exponent”  $\alpha$ . Following [Peng92,Pan94c] we map the sequence of monomers onto the trajectory of a random walk

$$x(t, s) = \sum_{i=s}^{s+t} \xi_i \quad (4.4)$$

where  $\xi_i = +1$  if the monomer  $i$  is A (white) and  $\xi_i = -1$  if it is B (black). Standard way to define  $\alpha$  is to calculate power law behavior of

$$\langle [x(t, s) - \langle x(t, s) \rangle_s]^2 \rangle \sim t^{2\alpha}, \quad (4.5)$$

where  $\langle \dots \rangle_s$  is the average along the chain. In order to include the case of relatively short chains, we have modified the definition. First, we introduce the “Brownian bridge”  $y(t)$  for the given sequence as

$$y(t) = x(t, 0) - \frac{t}{N} x(N, 0) \quad (4.6)$$

and calculate the value of  $r(t) = \langle y(t)^2 \rangle$ , where the average is performed over the

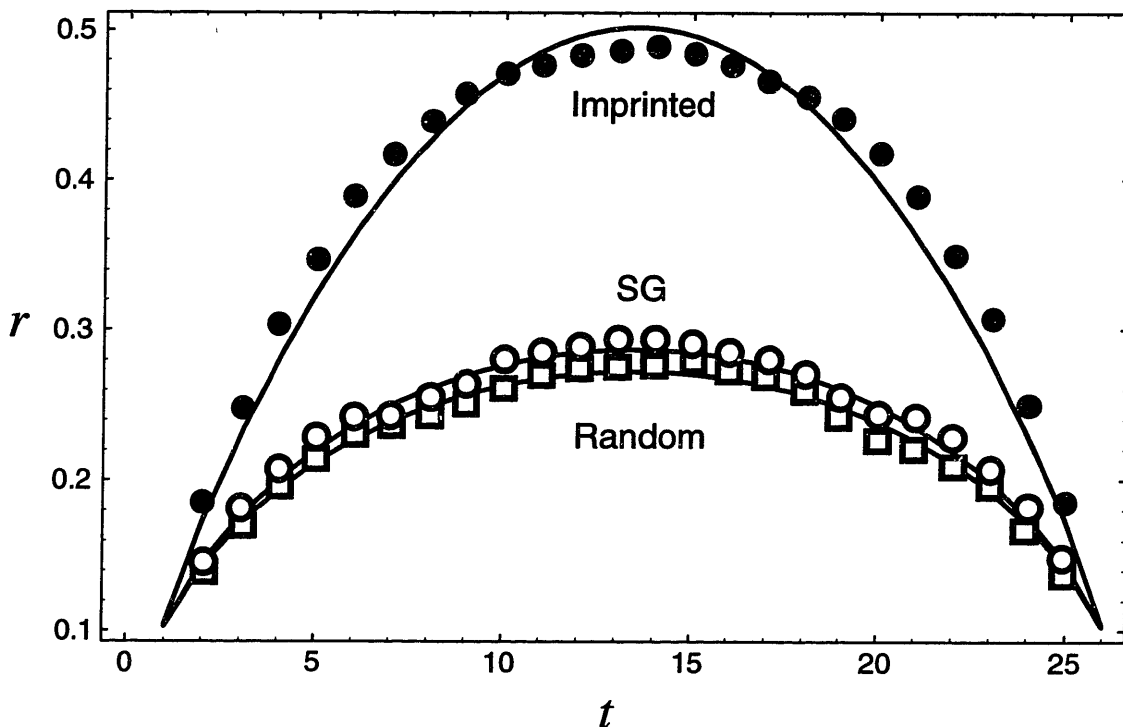


Figure 4-16: Brownian bridges for three polymer sequence ensembles (from top to bottom): Imprinted, SG, and Random sequences. Note that bridges for Random and SG sequences are almost indistinguishable, whereas the bridge for Imprinted sequence shows strong correlations. Imprinted and SG sequences were created at low polymerization (selection) temperatures  $T_p$ ; as  $T_p$  is increased, these bridges approach the Random sequence bridge.

ensemble of chains prepared under the same condition, i.e. at the same  $T_p$ . The obtained  $r(t)$  functions at different  $T_p$  are shown to obey accurately the interpolation relation of the type

$$r(t) = [(t^{-2\alpha} + (N - t)^{-2\alpha}]^{-1/2}, \quad (4.7)$$

thus allowing the determination of  $\alpha$ .  $\alpha = \frac{1}{2}$  is predicted by central limit theorem for the non-correlated sequence, thus, sequences with random, persistent, and alternative correlations lead to  $\alpha = \frac{1}{2}$ ,  $\alpha > \frac{1}{2}$  and  $\alpha < \frac{1}{2}$  respectively.

The ensemble of random chains lead to  $\alpha = \frac{1}{2}$  as expected. For 2 letter Potts interactions, where similar types of monomers attract each other, we would expect persistent types of correlations. Figure 4-16 shows Brownian Bridges for Imprinted, SG, and Random chains. We see that there is a significant degree of correlations

## *Prokaryote Catalysts*

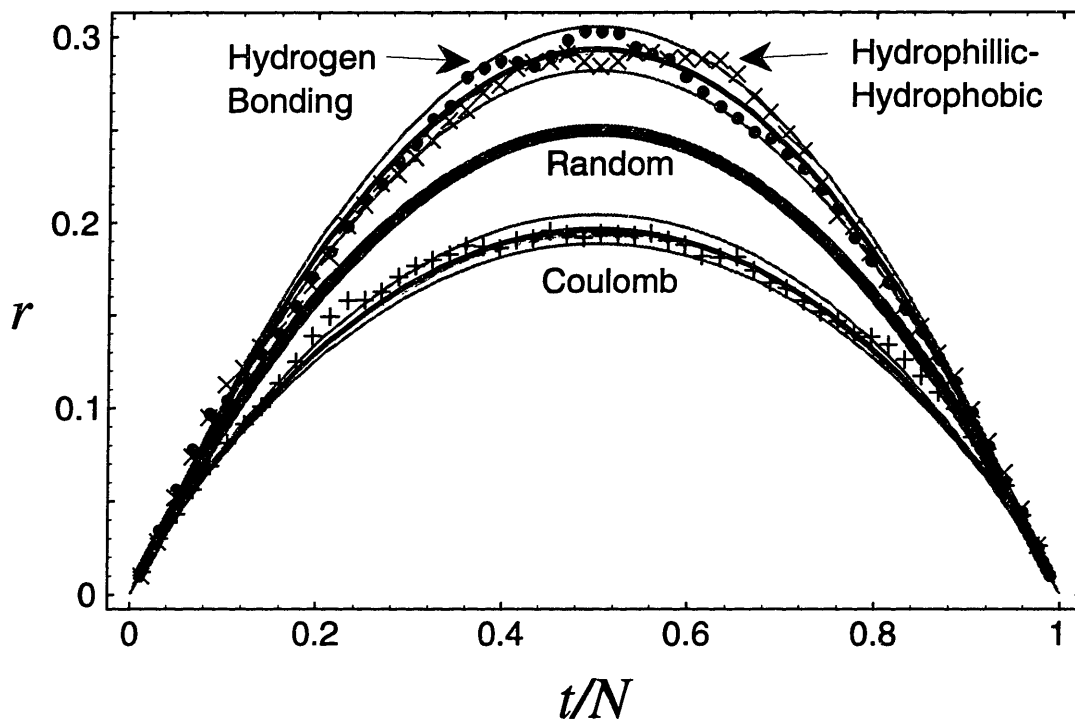


Figure 4-17: Brownian bridges for an ensemble average of Prokaryote catalysts. The three different physical mappings employed (Coulomb, Hydrogen bonding, and Hydrophilic-Hydrophobic interaction) yield three different bridges. For the Coulomb mapping, energy is minimized for alternating charges, and anti-correlations (i.e. below the bridge for random sequences) are found. For the other two mappings, energy is minimized for persistent correlations.

in Imprinted chains, whereas SG chains are just marginally more correlated than random chains. There is a simple physical argument to explain this: in the design Hamiltonian, the Imprinted chain Hamiltonian includes interactions along the chain (it includes interactions between all monomers) and the SG Hamiltonian excludes interactions along the chain. Thus, correlations are a first order effect (neighbor-neighbor) for Imprinted chains, but only a higher order effect (nearest-neighbor coupling) for SG chains.

Furthermore, higher order correlations are themselves strong only in simple interactions such as Potts interactions. For more complicated interactions, such as the SMJ matrix, higher order correlations are suppressed. It is difficult to analyze

monomer-monomer correlations for the SMJ matrix as there is no simple physical mapping. However, we do know the interaction matrix *exactly*, so we can examine the energy of interactions along the linear sequence of the polymer,  $E_{\text{chain}}$ . Normally the energy is ignored in polymer models such as the examination of the energy through enumeration or kinetics as  $E_{\text{chain}}$  is independent of conformation and depends only on the sequence. However, this energy does reflect the nature of the preparation of the sequence. We found that the energy of the linear sequence for SG chains is essentially the same as those of random chains. This is reasonable, as the polymer bonds are removed from the SG Hamiltonian, the free energy of the linear sequence will be entropy dominated, thus yielding similar results to random sequences. This is seen quantitatively in Table 2, by examining the average energy between the interactions along the linear sequence of the chain. However, the chain energy for Imprinted chains is much lower than random chains, which is reasonable as Imprinting minimizes the total energy of monomers, which includes the chain energy.

The fact that proteins also have analogous correlations in their sequences is extremely interesting in the light of the results of this model. Further comments will be left to the discussion.

## 4.6 Discussion

The main result of this work is that for polymers with strong heteropolymeric behavior, imprinted chains have a probability of renaturability of essentially order 1. By this, we mean that there are not dangerous terms such as  $\exp(-N)$  that would reduce the renaturability. What defines “strong heteropolymeric behavior” depends on the system in question. For example, for Potts Polymers (i.e. Potts interaction matrix), small  $q$  (i.e.  $q \sim 1$ ) and large  $q$  (i.e.  $q \sim N$ ) are both homopolymers. Thus, we would expect, and in fact we found, a large probability of renaturability for Potts Polymers with intermediate  $q$ .

We have said that the “sequence annealing” sequence design method of Shakhnovich

and Gutin (SG) is very different in spirit, but similar formally. We now discuss the differences between the methods and the consequences of these differences. First, the SG method excludes polymer bonds in the Hamiltonian, whereas the Imprinting Hamiltonian includes them. This is of course the difference between the energy of the polymer and the energy of the monomer solution. Here is the essence of the difference between the two methods. As the Imprinted polymers were optimized for interactions in the monomer solution, they will not necessarily be as optimized in terms of monomer-monomer interactions after polymerization. However, the S-G method minimizes the energy of monomer-monomer interactions for an already polymerized chain. Thus, the Imprinting method cannot give a larger yield of renaturable or design chains than the SG method. The obvious question, of course, is to what degree is the Imprinting method less optimized?

This immediately brings us to the question of the quasi-monomer renormalization of the polymer. Specifically, the degree to which these models differ depends on the difference due to the presence of the interactions along the polymer chain during design. However, these interactions become irrelevant upon renormalization, and thus they should not play a major role in the nature of renaturability.

On the other hand, it may be possible to detect this difference, i.e. the difference in the energy of interactions along the chain, in existing biopolymers, which are assumed to be a product of another design procedure: evolution. As mentioned in section 5, one way to examine the energy along the polymer chain is by looking at the correlations in monomer sequences. This was done for proteins [Pan94c], and correlations corresponding to energy minimization between monomers were found. Correlations of this type was also found for imprinted chains, but are only faintly seen in SG chains, where interactions along the chain are ignored and all correlations must therefore be derived from higher order (i.e. next-nearest neighbor at best) interactions.

However, it was asserted that the correlations in proteins diminished, i.e. the sequences became more random, when more evolutionary advanced species were examined. This has the following possible interpretation in terms of the two design

procedures under examination here. We speculate that proteins were created in some prebiotic synthesis similar to Imprinting. This would cause correlations in protein sequences, but more importantly, imprinting is a reasonable model for the prebiotic evolution of biopolymers. Polymerization after monomer equilibration does not involve any pre-existing biochemical systems, and, has a large probability of creating renaturable chains which can, for example, accept some target molecule whose site has been imprinted.

The SG method has always purported to be a model of biological evolution. In this way, monomer substitution is performed by some sort of genetic algorithm performed by some biological system. Thus, we speculate that proteins have evolved with their organisms in order to minimize the protein's ground state energy [Sha93b]. If the mechanism for this was "sequence annealing," then one would expect that the correlations along the chain would diminish and eventually become random. This is exactly what has been found in proteins.

Finally, we speculate on what is the physical foundation of the ability of design procedures to yield chains 1) which are renaturable and 2) which are renaturable to the target conformation. The existence of a unique state, and therefore thermodynamic renaturability, for random chains was first shown analytically by Shakhnovich and Gutin [Sha89a]. In general, one can view a design procedure as some factor which chooses as subset of all of the possible sequences. Each chain in this subset is a local minima in sequence space of the energy functional minimized in the design procedure. Thus, in the SG design procedure, it is clear that the polymer energy in the target conformation will be minimized. In any other conformation, this designed sequence is essentially random and therefore should have a higher energy. In the imprinting model, there is not an exact correspondence between the energy of the monomer solution and the energy of the polymer. However, this brings us merely back to the question of the nature of the monomer to quasi-monomer renormalization, and thus we would expect that this difference should be minimal for sufficiently long chains. Even in the model 27-mers and 36-mers we see that this difference is noticeable, but not significant. In a sense, one can draw an anal-

ogy between the design procedure and the learning algorithm of a neural network [Sha89b,Pan94e]. In both cases, the interactions have been previously minimized, leading to the reproduction of the “learned” configuration.

## 4.7 Conclusions

In conclusion, we find that the polymers created using our method fall in between the highly optimized sequences of the SG method and the kinetically unfavorable random sequences. The fact that the probability of creating a chain with an energy gap suggests that the folding kinetics should be much closer to that of the SG method than that of random chains. This was indeed found by running Monte Carlo kinetics simulations.

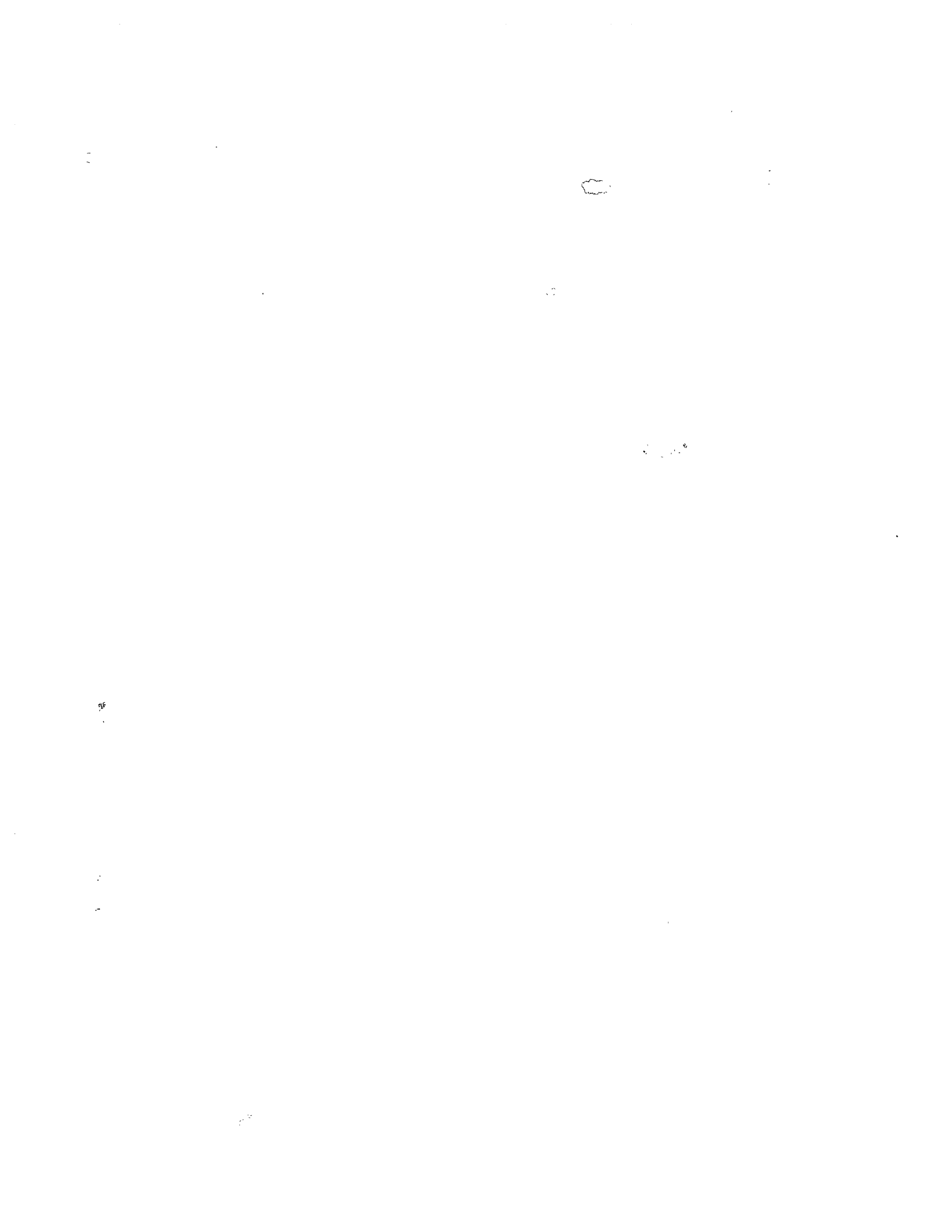
We have stated that the Imprinted chains are not as optimized as those created by SG design. This is directly the effect that SG chains minimize polymer energy, leading to random monomer-monomer correlations along the chain, and Imprinted chains minimize total monomer energy, thus creating correlations. Thus, the presence of these correlations reflects imperfect optimization. Thus, while our method does not create as highly optimized proteins as the SG method, perhaps such high optimization is not necessary, as shown in the example of natural proteins. Furthermore, perhaps Imprinted chains also share a common “history” with natural chains, as the polymerization procedure described here may describe *in vivo* synthesis of prebiotic proteins in addition to the *in vitro* synthesis of protein-like polymers.





## **Part III**

## **Analytic**



# Chapter 5

## Two Letter Designed

In the previous chapter, Imprinting was examined computationally. Using mean field replica theory, we find, in addition to the freezing transition of random chains, a transition to the target “native” state. The stability of this state is shown to be greater than that of the ground state of random chains. The results derived here should at least qualitatively be applicable to known biopolymers, which are conjectured to be *in vivo* “designed” by evolution. Furthermore, we present a crude prescription for a laboratory procedure in which chains can be synthesized *in vitro*.

### 5.1 Introduction

Due to its biological importance and physical complexity, the problem of the folding transition of copolymers has attracted considerable attention [Wol91,Sfa93,Gar88a]. Biologically, proteins represent a sort of “designed” heteropolymer, in this case the result of evolution. It is also known that proteins have a unique structure. It is intriguing to consider that the existence of a *specific* (i.e. designed) unique ground state could be the result of evolution. This has been studied computationally: a Monte Carlo procedure dubbed “sequence annealing” swapped monomers via the Metropolis criterion such that the polymer energy is minimized [Sha93b]. However, in the mean field approximation, sequence annealing and Imprinting are formally

indistinguishable, as both choose sequences with a fixed monomer composition such that the energy of interaction is minimized [Pan94b].

In this chapter, we employ the replica approach to describe the effect of design on the freezing transition previously predicted for random chains [Sfa93,Sha89a]. The mean field replica approach is believed to be applicable to disordered polymers, as the polymer problem is similar to the *long range* SK spin-glass [She75]. Indeed, due to polymer flexibility, all monomers can come in contact with each other in real space, and, therefore, interact with each other, no matter how far they are along the chain. In this sense, the heteropolymer is perhaps the best physical realization of the SK system, with a truly infinite radius of interaction [Sha89a].

The freezing transition in random copolymers is due to the competition between the entropic favorability of a large number of conformations and the energetic tendency toward one or a few conformations with distinctively low energies. Qualitatively, we expect that the design procedure should lead to sequences whose unique ground state conformation is the target conformation, as we have, in a sense, exerted a “field” which chooses an ensemble of sequences which have been optimized for the particular target conformation.

## 5.2 Formulation of the Model

Consider a heteropolymer chain with a frozen sequence of monomers  $s_I$ , where  $I$  is the number of monomer along the chain ( $1 \leq I \leq N$ ) and  $s_I$  is the type of monomer in the given sequence. In the present model, we consider only two values for  $s$ ,  $s = \pm 1$ , and have the interaction Hamiltonian of the form

$$\mathcal{H} = -\frac{1}{2}B \sum_{I,J}^N s_I s_J \delta(\mathbf{r}_I - \mathbf{r}_J) \quad (5.1)$$

where  $\delta(\mathbf{r})$  is the Dirac delta function. As we wish to concentrate on heteropolymeric effects, we do not explicitly write, but implicitly assume, an overall attractive second virial coefficient as well as a repulsive third virial coefficient. Specifically, we

assume that the complete Hamiltonian is given by the sum of heteropolymeric and homopolymeric terms:

$$\mathcal{H}' = -\frac{1}{2}B \sum_{I,J}^N s_I s_J \delta(\mathbf{r}_I - \mathbf{r}_J) + B_0 \rho + C \rho^2 \quad (5.2)$$

where  $B_0$  and  $C$  are the mean second and third virial coefficients. As we assume that  $|B_0| \gg B$ , we can optimize the free energy with respect to  $\rho$  independently of any heteropolymeric properties. Thus, these homopolymeric terms lead to a compact globular state with constant density  $\rho = -B_0/2C$ . Furthermore,  $B$  is due to heteropolymeric effects; it is the “preferential” energy: for two types of species labeled 1 and 2, the preferential energy is the energy difference  $E_{12} - \frac{1}{2}(E_{11} + E_{22})$ . The meaning and value of the preferential energy for any real system depends on the nature of the actual interactions involved (eg. hydrogen bonding, hydrophobic forces, etc.). Essentially, some conformations with a given density (fixed due to the homopolymeric terms) might be more thermodynamically favorable than others, due to heteropolymeric effects. This will be the main subject of our analysis.

The partition function is expressed as

$$Z(\text{seq}) = \sum_{\text{conformations}} \exp \left[ -\frac{1}{T} \mathcal{H}(\text{conf}, \text{seq}) \right]. \quad (5.3)$$

Note that the Hamiltonian depends on both conformation and sequence. The standard way to approach the partition function of a system with frozen disorder is to employ, first, the principle of self-averaging of free energy and, second, the replica trick:

$$F(\text{seq}) \simeq F = \langle F(\text{seq}) \rangle_{\text{seq}} = -T \langle \ln Z(\text{seq}) \rangle_{\text{seq}} = -T \lim_{n \rightarrow 0} \frac{\langle Z^n(\text{seq}) \rangle_{\text{seq}} - 1}{n}, \quad (5.4)$$

where  $\langle \dots \rangle_{\text{seq}}$  means average over the set of sequences.

In the works [Sfa93,Sha89a], while averaging, the sequences were considered to be random. The main purpose of this work is to incorporate the fact that sequences

are somehow selected. This means

$$\langle \dots \rangle_{\text{seq}} = \sum_{\text{seq}} \dots P_{\text{seq}} , \quad (5.5)$$

where  $P_{\text{seq}}$  is the probability distribution for different sequences which appear in the process of design or synthesis of chains.

Both of the recently suggested models of sequence preparation [Pan94b,Sha93b] (see above) employ in the selection process the same volume interactions with which the links of chains interact. In both cases,  $P_{\text{seq}}$  is governed by the Boltzmann factor related to the same Hamiltonian (5.1) taken for the “target” conformation  $\star$ . Since we are not interested in any particular  $\star$  conformation and, besides, this conformation seems to be out of control in any real (not computer) experiment, we average over the conformation  $\star$ :

$$P_{\text{seq}} = \frac{1}{z} \sum_{\star} \exp \left[ -\frac{1}{T_p} \mathcal{H}(\star, \text{seq}) \right] , \quad (5.6)$$

where  $T_p$  is “polymerization” temperature at which design procedure is performed and

$$z = \sum_{\text{seq}} \sum_{\star} \exp \left[ -\frac{1}{T_p} \mathcal{H}(\star, \text{seq}) \right] \quad (5.7)$$

is the normalization constant. The probability  $P_{\text{seq}}$  includes all possible sequences, not only the ones with any given composition.

Collecting the above equations, we can write the  $n$ -replica partition function, up to the constant factors, as

$$\begin{aligned} \langle Z^n(\text{seq}) \rangle_{\text{seq}} &= \frac{1}{z} \sum_{\text{seq}} \sum_{\star} \exp \left[ -\frac{1}{T_p} \mathcal{H}(\star, \text{seq}) \right] \left\{ \sum_{\text{conformations}} \exp \left[ -\frac{1}{T} \mathcal{H}(\text{conf}, \text{seq}) \right] \right\}^n \\ &= \frac{1}{z} \sum_{\text{seq}} \sum_{C_0, C_1, \dots, C_n} \exp \left[ -\frac{1}{T_p} \mathcal{H}(C_{\alpha=0}, \text{seq}) - \frac{1}{T} \sum_{\alpha=1}^n \mathcal{H}(C_{\alpha}, \text{seq}) \right] , \quad (5.8) \end{aligned}$$

where  $C_{\alpha} = C_0, C_1, \dots, C_n$  stand for conformations of replica number  $\alpha$ , and index  $\alpha = 0$  is attributed to the target conformation  $\star$ . As expected, we return to the usual case of completely random sequences at  $T_p \rightarrow \infty$ . The new physics which

appears at finite  $T_p$  is the main subject of further analysis.

### 5.3 Replica Theory Analysis

We write the  $n$ -replica partition function

$$\begin{aligned} \langle Z^n(\text{seq}) \rangle_{\text{seq}} &= \frac{1}{z} \sum_{\text{seq}} \sum_{C_0, C_1, \dots, C_n} & (5.9) \\ &\times \exp \left\{ \sum_{\alpha=0}^n \frac{B}{2T_\alpha} \sum_{I, J=1}^N \int d\mathbf{R}_1 d\mathbf{R}_2 s_I \delta(\mathbf{r}_I^\alpha - \mathbf{R}_1) \right. \\ &\left. \times s_J \delta(\mathbf{r}_J^\alpha - \mathbf{R}_2) \delta(\mathbf{R}_1 - \mathbf{R}_2) \right\} , \end{aligned}$$

where  $T^\alpha = T_p$  for  $\alpha = 0$  and  $T^\alpha = T$  for  $\alpha > 0$ , and we perform Hubbard-Stratonovich transformation

$$\begin{aligned} \langle Z^n(\text{seq}) \rangle_{\text{seq}} &= \frac{1}{z} \int \mathcal{D}\phi \sum_{C_0, C_1, \dots, C_n} \exp \left\{ -\frac{1}{2} \sum_{\alpha=0}^n \int d\mathbf{R}_1 d\mathbf{R}_2 \right. & (5.10) \\ &\left. \times \frac{T^\alpha}{B} \phi^\alpha(\mathbf{R}_1) \phi^\alpha(\mathbf{R}_2) \delta(\mathbf{R}_1 - \mathbf{R}_2) \right\} \end{aligned}$$

$$\times \sum_{\text{seq}} \exp \left\{ \sum_{\alpha=0}^n \int d\mathbf{R} \sum_{I=1}^N \phi^\alpha(\mathbf{R}) s_I \delta(\mathbf{r}_I^\alpha - \mathbf{R}) \right\} \quad (5.11)$$

Here  $\phi^\alpha(\mathbf{R})$  are the fields conjugated to the corresponding densities  $\sum_{I=1}^N s_I \delta(\mathbf{r}_I^\alpha - \mathbf{R})$ ,  $\int \mathcal{D}\phi \dots$  means functional integration over all the fields  $\{\phi^\alpha(\mathbf{R})\}$ , and we have dropped all irrelevant multiplicative constants from the partition function. Note that the sum over sequences enters only in the last "source" term of (5.11). The summation over sequences can be easily performed to yield

$$\begin{aligned} \exp \{\text{source term}\} &= \sum_{s_1, s_2, \dots, s_N = \pm 1} \prod_{I=1}^N \exp \left\{ s_I \sum_{\alpha=0}^n \int d\mathbf{R} \phi^\alpha(\mathbf{R}) \delta(\mathbf{r}_I^\alpha - \mathbf{R}) \right\} \\ &= \prod_{I=1}^N \sum_{s = \pm 1} \exp \left\{ s_I \sum_{\alpha=0}^n \int d\mathbf{R} \phi^\alpha(\mathbf{R}) \delta(\mathbf{r}_I^\alpha - \mathbf{R}) \right\} \\ &= \prod_{I=1}^N 2 \cosh \left\{ \sum_{\alpha=0}^n \int d\mathbf{R} \phi^\alpha(\mathbf{R}) \delta(\mathbf{r}_I^\alpha - \mathbf{R}) \right\} & (5.12) \end{aligned}$$

We perform now the expansion over  $\phi$ . It is the most important approximation of this work. The corresponding conditions of applicability will be given later. Keeping the terms up to  $\mathcal{O}(\phi^2)$ , we get the  $n$ -replica partition function in the form:

$$\langle Z^n(\text{seq}) \rangle_{\text{seq}} = \frac{1}{z} \sum_{C_0, C_1, \dots, C_n} \int \mathcal{D}\phi \exp[-\mathcal{E}\{Q_{\alpha\beta}\}] , \quad (5.13)$$

where the effective energy of  $n$ -replica system is given by

$$\begin{aligned} \mathcal{E}\{Q_{\alpha\beta}\} &= \frac{1}{2} \int d\mathbf{R}_1 d\mathbf{R}_2 \\ &\times \left[ \sum_{\alpha=0}^n \frac{T^\alpha}{B} \phi^\alpha(\mathbf{R}_1) \phi^\alpha(\mathbf{R}_2) \delta(\mathbf{R}_1 - \mathbf{R}_2) \right. \\ &\quad \left. - \sum_{\alpha, \beta=0}^n \phi^\alpha(\mathbf{R}_1) \phi^\beta(\mathbf{R}_2) Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) \right] , \end{aligned} \quad (5.14)$$

and

$$Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) = \sum_{I=1}^N \delta(\mathbf{r}_I^\alpha - \mathbf{R}_1) \delta(\mathbf{r}_I^\beta - \mathbf{R}_2) \quad (5.15)$$

is the standard two replica overlap order parameter [Sha89a, Gar88a]. Recall that the value of  $Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2)$  has the very simple physical meaning: it is proportional to probability of finding simultaneously one monomer of the replica  $\alpha$  at the point  $\mathbf{R}_1$  and one monomer of the replica  $\beta$  at the point  $\mathbf{R}_2$ . Also note that the normalization conditions

$$\int d\mathbf{R}_1 Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) = \rho_\beta(\mathbf{R}_2) \quad \text{and} \quad \int d\mathbf{R}_1 d\mathbf{R}_2 Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) = N \quad (5.16)$$

are obvious from the definition of  $Q_{\alpha\beta}$ , eq. (5.15), where  $\rho_\beta(\mathbf{R}_2)$  is the density. As we are concerned here with a large globule, density is assumed constant throughout the globule, such that

$$\rho_\beta(\mathbf{R}) = \rho \quad (\text{constant in space, same for all replicas}) ; \quad (5.17)$$



and therefore

$$Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) = Q_{\alpha\beta}(\mathbf{R}_1 - \mathbf{R}_2) . \quad (5.18)$$

We also mention that the diagonal element is given by

$$Q_{\alpha\alpha}(\mathbf{R}_1, \mathbf{R}_2) = \rho \delta(\mathbf{R}_1 - \mathbf{R}_2) . \quad (5.19)$$

We can therefore rewrite the effective  $n$ -replica energy in the form

$$\begin{aligned} \mathcal{E}\{Q_{\alpha\beta}\} &= \frac{1}{2} \sum_{\alpha,\beta=0}^n \int d\mathbf{R}_1 d\mathbf{R}_2 \phi^\alpha(\mathbf{R}_1) \phi^\beta(\mathbf{R}_2) \\ &\times \left[ \left( \frac{T^\alpha}{B} - \rho \right) \delta_{\alpha\beta} \delta(\mathbf{R}_1 - \mathbf{R}_2) - Q_{\alpha\neq\beta}(\mathbf{R}_1 - \mathbf{R}_2) \right] . \end{aligned} \quad (5.20)$$

Now we pass from summation over conformations (microstates) to functional integration over  $Q_{\alpha\beta}$  (macrostates).  $Q_{\alpha\beta}$  is the only relevant order parameter. The corresponding entropy is given by [Sha89a]

$$e^{S\{Q_{\alpha\beta}\}} = \sum_{C_0, C_1, \dots, C_n} \delta \left( Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) - \sum_{I=1}^N \delta(\mathbf{r}_I^\alpha - \mathbf{R}_1) \delta(\mathbf{r}_I^\beta - \mathbf{R}_2) \right) , \quad (5.21)$$

and therefore

$$\langle Z^n(\text{seq}) \rangle_{\text{seq}} = \int \mathcal{D}Q \int \mathcal{D}\phi \exp \{ -\mathcal{F} + \mathcal{F}_0 \} \quad (5.22)$$

$$\mathcal{F} = \mathcal{E}\{Q_{\alpha\beta}\} - S\{Q_{\alpha\beta}\} \quad (5.23)$$

$$\mathcal{F}_0 = -\ln z \quad (5.24)$$

The mean field evaluation of this partition function implies a saddle point approximation for the integral over  $Q$ , eq. (5.24). Normally, this means taking the maximal value of the integrand. It is commonly believed, however, that in order to find the correct analytic continuation in the  $n \rightarrow 0$  limit, one has to take the *maximal* rather than the minimal value of the relevant free energy  $\mathcal{F}$ , because there are  $n(n-1)/2$  off-diagonal elements in the  $Q_{\alpha\beta}$  matrix, and therefore for the  $0 < n < 1$  case, the integral over  $Q_{\alpha\beta}$  represents summation over a *negative* number

of variables. Following this principle, we write

$$\langle Z^n(\text{seq}) \rangle_{\text{seq}} = \int \mathcal{D}\phi \exp[-\text{Max}_{\{Q\}} \mathcal{F}\{Q\}] = \text{Min}_{\{Q\}} \int \mathcal{D}\phi \exp[-\mathcal{F}\{Q\}] , \quad (5.25)$$

i.e., we have to *maximize* the effective free energy functional (5.24).

## 5.4 Replica Symmetry Breaking

We choose some standard function  $\varphi(\mathbf{x})$ , say gaussian, with the normalization condition  $\int d\mathbf{x} \varphi(\mathbf{x}) = 1$ , and say that

$$Q_{\alpha\beta}(\mathbf{R}_1 - \mathbf{R}_2) = \frac{\rho}{(R_t^{\alpha\beta})^d} \varphi\left(\frac{\mathbf{R}_1 - \mathbf{R}_2}{R_t^{\alpha\beta}}\right) , \quad (5.26)$$

where  $d$  is the dimensionality of space,  $R_t^{\alpha\beta}$  can be interpreted as the diameter of the tube in which replicas  $\alpha$  and  $\beta$  coincide, and the normalization condition defines the coefficient. We now repeat the arguments of [Sha89a]: as the entropy scales like  $-(R_t^{\alpha\beta})^{-2}$  at  $n < 1$ , we get each  $\alpha \neq \beta$  term of the free energy functional of the form

$$\mathcal{F}_{\alpha\beta} = -\frac{A_1}{(R_t^{\alpha\beta})^2} + \frac{A_2}{(R_t^{\alpha\beta})^d} \quad (5.27)$$

where  $A_1$  and  $A_2$  are positive numbers. For  $d > 2$ , which is the main concern of this work, we find two maxima, namely  $R_t^{\alpha\beta} = \infty$  and  $R_t^{\alpha\beta} = 0$  (in the later case, see the discussion in [Sha89a] concerning the short distance cut-off  $R_t^{\alpha\beta} = v^{1/3}$ , where  $v$  is the excluded volume). The first corresponds to two replicas,  $\alpha$  and  $\beta$  which are independent and do not overlap at all ( $Q_{\alpha\beta} = 0$ ), while the second corresponds to replicas which coincide at the microscopic level ( $Q_{\alpha\beta} = \rho \delta(\mathbf{R}_1 - \mathbf{R}_2)$ ). Thus, from these scaling arguments in  $R_t^{\alpha\beta}$ , we conclude that  $Q_{\alpha\beta}$  is of the form

$$Q_{\alpha\beta}(\mathbf{R}_1 - \mathbf{R}_2) = \rho q_{\alpha\beta} \delta(\mathbf{R}_1 - \mathbf{R}_2) \quad (\alpha \neq \beta) , \quad (5.28)$$

where off-diagonal matrix elements of the new matrix  $q_{\alpha\beta}$  are either 0 or 1. If we additionally define diagonal matrix elements  $q_{\alpha\alpha}$  as

$$q_{\alpha\alpha} \equiv 1 - \frac{T^\alpha}{B\rho}, \quad (5.29)$$

we can write

$$\langle Z^n(\text{seq}) \rangle_{\text{seq}} = \frac{1}{z} \text{Min}_{\{q_{\alpha\beta}\}} e^{S\{q_{\alpha\beta}\}} \left\{ \int \mathcal{D}\phi \exp \left[ \frac{1}{2} \sum_{\alpha,\beta=0}^n q_{\alpha\beta} \phi^\alpha \phi^\beta \right] \right\}^N, \quad (5.30)$$

where integration over  $\mathbf{R}$  disappears leaving the product of  $N = \rho \int d\mathbf{R}$  integrals. Moreover, we can perform Gaussian integration over  $\phi$  yielding <sup>1</sup>

$$\langle Z^n(\text{seq}) \rangle_{\text{seq}} = \exp \left[ -\text{Max}_{\{q_{\alpha\beta}\}} \mathcal{F}\{q_{\alpha\beta}\} \right] \quad (5.33)$$

$$\mathcal{F}\{q_{\alpha\beta}\} = \frac{N}{2} \ln [\det(-q_{\alpha\beta})] - S\{q_{\alpha\beta}\} - \ln z. \quad (5.34)$$

where we have dropped all irrelevant additive constants.

To maximize the  $n$ -replica free energy over  $q_{\alpha\beta}$  means in fact finding the optimal grouping of replicas. There is the following obvious transitivity rule: if, say,  $R_t^{\alpha\beta} = 0$  and  $R_t^{\beta\gamma} = 0$ , meaning that conformations of replicas  $\alpha$ ,  $\beta$  and  $\gamma$  are all the same, then  $R_t^{\alpha\gamma} = 0$  as well. In other words, if  $q_{\alpha\beta} = 1$  and  $q_{\beta\gamma} = 1$ , then  $q_{\alpha\gamma} = 1$  as well. Using matrix row and column operations, we can organize any such matrix into block diagonal form. This means gathering replicas that overlap in the groups and placing replicas of the same group into the same diagonal block in the matrix. One

---

<sup>1</sup>It is instructive to perform first the Gaussian functional integrals over  $\phi^0$  yielding

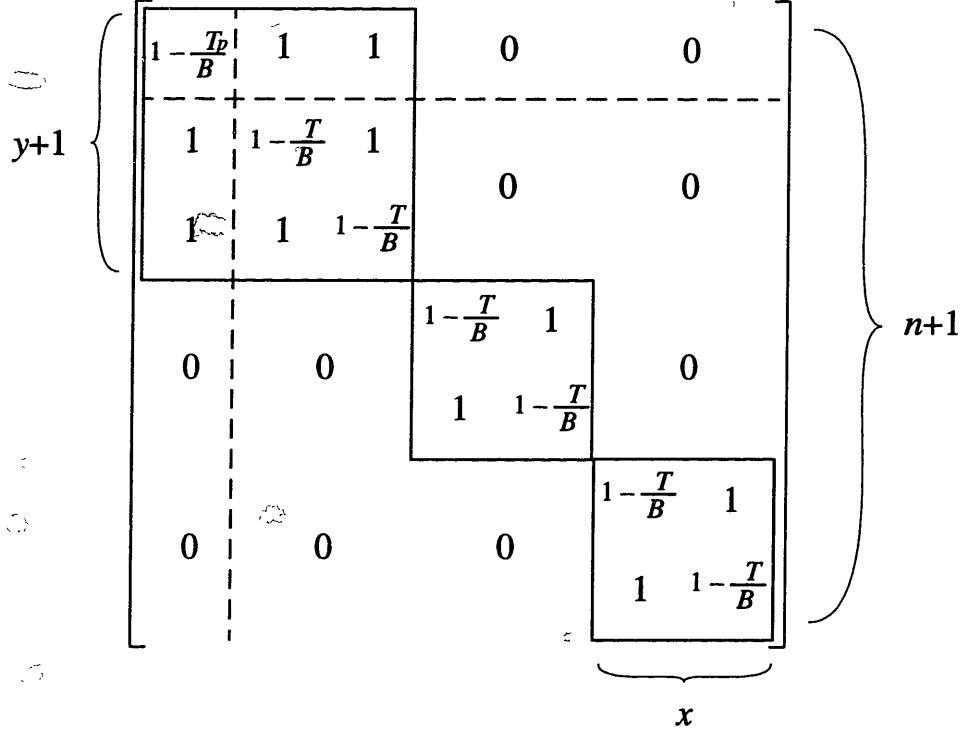
$$\langle Z^n(\text{seq}) \rangle_{\text{seq}} = \frac{1}{z} \text{Min}_{\{q_{\alpha\beta}\}} e^{S\{q_{\alpha\beta}\}} \left\{ \int \mathcal{D}\phi \exp \left[ \frac{1}{2} \sum_{\alpha,\beta=1}^n \hat{q}_{\alpha\beta} \phi^\alpha \phi^\beta \right] \right\}^N, \quad (5.31)$$

where

$$\hat{q}_{\alpha\beta} = q_{\alpha\beta} + \frac{q_{\alpha 0} q_{0\beta}}{q_{00}} \quad (5.32)$$

In this form, we reduce the problem to a form similar to that of Sfatos *et al* [Sfa93], with a new effective order parameter  $\hat{q}_{\alpha\beta}$ . This form explains that replica 0 plays the role of external field in replica space, adsorbing other replicas. In other words, two replicas  $\alpha, \beta > 0$  attract each other directly, as in the random polymer, and, additionally, because both are attracted to target replica.

of the blocks is comprised of some  $y + 1$  replicas which do overlap (i.e., practically coincide) with the “target” replica 0. Other  $(n + 1) - (y + 1) = n - y$  replicas belong to  $n/x$  groups, some  $x$  replicas in each:



One can say that  $y$  replicas here are “adsorbed” on the target conformation, which plays the role of external field for  $n$  other replicas. A similar situation exists in neural networks [Amit87], where the memorized image plays a similar role to the target conformation. On the other hand, the grouping of other replicas is due to spontaneous replica permutation symmetry breaking.

The determinant of the  $q_{\alpha\beta}$  matrix can be directly calculated. First, since the matrix is block-diagonal, its determinant is the product of block determinants. Each  $x \times x$  block has a  $(x - 1)$ -fold degenerate eigenvalue  $[\tilde{q} - q]$  and one distinct eigenvalue  $[\tilde{q} - q + xq]$ , where  $\tilde{q} = 1 - T/B\rho$  and  $q = 1$  are diagonal and off-diagonal matrix elements, respectively. As to the  $(y + 1) \times (y + 1)$  block, it has one distinct diagonal element  $q_{00} = 1 - T_p/B\rho = \tilde{q}_p$  and for this reason the eigenvalue  $[\tilde{q} - q]$  is only (matrix size - 2) =  $(y - 1)$ -fold degenerate, while the two others are  $(1/2) \left[ [\tilde{q} + \tilde{q}_p + q(y - 1)] \pm \sqrt{[\tilde{q} - \tilde{q}_p + q(y - 1)]^2 + 4yq^2} \right]$ . Taking the product of all eigenvalues throughout all blocks, and noting that  $\det(-q_{\alpha\beta}) = (-1)^{n+1} \det(q_{\alpha\beta})$ ,

we obtain

$$\ln [\det (-q_{\alpha\beta})] = \frac{n-y}{x} \ln \left[ 1 - \frac{B\rho}{T} x \right] + \ln \left[ 1 - \frac{B\rho}{T} \left( y + \frac{T}{T_p} \right) \right] + n \ln \left[ \frac{T}{B\rho} \right] + \ln \left[ \frac{T_p}{B\rho} \right], \quad (5.35)$$

To estimate the entropy  $S\{q_{\alpha\beta}\}$  related to the grouping of replicas, we follow Ref. [Sfa93] to argue that due to the polymeric bonds connecting monomers along the chain, once one monomer is fixed in space, the next must be placed within a volume  $a^3$ . Since replicas that belong in the same group coincide within a tube of radius  $R_t \sim v^{1/3}$ , there are  $a^3/v$  ways to place the next monomer and thus the entropy per monomer is  $\ln(a^3/v)$ . But since all replica conformations coincide within the group, we must restrict the position of the next monomer to a single place. Thus, the entropy loss for each group is  $s(x-1)$ , where  $s = \ln(a^3/v)$  is related to the flexibility of the chain, and therefore

$$S = Ns \left[ \frac{n-y}{x} (x-1) + y \right]. \quad (5.36)$$

As to the last term in (5.34),  $-\ln z$ , it is formally related to the normalization condition for the probability  $P_{seq}$ , but physically it is the free energy of single replica 0 taken at the polymerization temperature  $T_p$ . It can be therefore easily found by taking  $n = 0$ ,  $y = 0$  in the preceding formulae:

$$-\ln z = N \cdot \ln \left[ \frac{T_p}{B\rho} - 1 \right]. \quad (5.37)$$

Collecting equations (5.34), (5.35), and (5.36), we obtain

$$\frac{1}{N} \mathcal{F} = \frac{n-y}{2x} \ln \left[ 1 - \frac{B\rho}{T} x \right] + \frac{1}{2} \ln \left[ 1 - \frac{B\rho/T}{1 - B\rho/T_p} y \right] + s \left[ n - \frac{n-y}{x} \right]. \quad (5.38)$$

where we have employed the fact that the last two terms in (5.35) cancel with normalization constants from gaussian integration not explicitly written. We are left, therefore, only with maximization over  $x$  and  $y$  in the  $n \rightarrow 0$  limit, yielding the opportunity to comment on the physics of the possible phases.

## 5.5 Phase Diagram

Let us discuss the possible values of  $x$  and  $y$  in the  $n \rightarrow 0$  limit. For the replica system, when  $n$  is positive integer, we have  $1 \leq x \leq n$  and  $0 \leq y \leq n$ . Clearly,  $x = 1$  means there is no grouping, i.e. no replica symmetry breaking. On the other hand,  $x = n$  means all the replicas belong to the same group, or replica symmetry is broken. When  $n$  becomes less than 1 and goes to 0, inequalities flip. Nevertheless,  $x$  must remain in between of  $n$  and 1, and approaching  $x$  to 1 means disappearance of replica permutation symmetry breaking. In other words,  $x = 1$  corresponds to the freezing transition. This transition has been investigated in [Sfa93]. Maximizing  $x$  in equation (5.38), we recover the result of [Sfa93] for random chains <sup>2</sup>

$$2s = \ln \left[ 1 - \frac{B\rho}{T}x \right] + \frac{(B\rho/T)x}{1 - (B\rho/T)x} . \quad (5.39)$$

The solution of this equation is of the form  $x = T\xi(s)/B\rho$ , where  $\xi(s)$  is the function defined by the equation  $2s = \ln(1 - \xi) + \xi/(1 - \xi)$ . According to our discussion, this solution is valid when it gives  $x \leq 1$ , i.e. at  $T \leq T_f$ , where  $T_f$  is given by

$$T_f = \frac{B\rho}{\xi(s)} \quad \text{or} \quad 2s = \ln \left[ 1 - \frac{B\rho}{T_f} \right] + \frac{B\rho/T_f}{1 - B\rho/T_f} . \quad (5.40)$$

$T_f$  is the temperature of freezing transition for the chain with random sequence [Sfa93], and we can write

$$x = \begin{cases} T/T_f & \text{when } T \leq T_f \\ 1 & \text{otherwise} \end{cases} \quad (5.41)$$

---

<sup>2</sup>In fact, this result corresponds exactly to the so-called Parisi *ansatz* with one-step replica symmetry breaking. In our model, however, it can be obtained in a more sophisticated manner, without any *ansatz*. Indeed, we can easily consider the general case of some  $g$  groups of replicas, with different numbers  $x_i$  of replicas in each group. We have then  $\sum_{i=1}^g \ln \left[ 1 - \frac{B\rho}{T}x_i \right]$  instead of  $\frac{n-y}{x} \ln \left[ 1 - \frac{B\rho}{T}x \right]$  in the  $\ln(\det(-q_{\alpha\beta}))$  term and  $s \sum_{i=1}^g (x_i - 1)$  instead of  $s \frac{n-y}{x} (x - 1)$  in the entropy term. Maximization with respect to  $x_i$  within the constraint  $\sum_{i=1}^g x_i = n - y$  gives that all  $x_i = x$  are the same, leaving us with the simplified version considered above.

We find that  $x_0$  and  $T_f$  are independent of any design parameters such as  $y$  and  $T_p$ . This has a clear physical meaning: if one considers the chain prepared by our procedure in some particular conformation  $\star$ , then for almost all of the conformations except  $\star$ , this chain behaves as if it had a completely random sequence. This is why freezing into a random conformation is not at all affected by the procedure of sequence selection.

Consider now maximization with respect to  $y$ . The condition  $0 \leq y \leq n$  is obvious for positive integer  $n$ :  $y = 0$  means no replicas in the target group,<sup>3</sup> while  $y = n$  means all replicas are in the target conformation. When  $n$  becomes less than 1 and goes to 0,  $y$  remains in between 0 and  $n$ , which is also 0. Since  $y$  is always small, the second logarithmic term in (5.38) can be linearized to obtain

$$\frac{1}{N}\mathcal{F} = \frac{n}{2x} \left\{ \ln \left[ 1 - \frac{B\rho}{T}x \right] - 2s \right\} + sn - \frac{y}{2x} \left\{ \ln \left[ 1 - \frac{B\rho}{T}x \right] + \frac{(B\rho/T)x}{1 - (B\rho/T_p)} - 2s \right\}. \quad (5.42)$$

Thus, the effective free energy (5.42) is linear in  $y$ . The maximal value is therefore reached always at the boundary of the interval, i.e. either at  $y = 0$  (no replicas in the target group) or at  $y = n$  (all the replicas are in the target group). The corresponding phase transition occurs when the  $y$  dependence of the free energy flips sign, and the transition point  $T_p^{cr}$  can be easily found, since linear in  $y$  term of the free energy (5.42) vanishes at the transition point. We substitute  $s$  from the condition (5.40) and find

$$T_p^{cr} = \begin{cases} B\rho \left[ 1 - \frac{B\rho/T}{\ln \left[ \frac{1-B\rho/T_f}{1-B\rho/T} \right] + \frac{B\rho/T_f}{1-B\rho/T_f}} \right]^{-1} & \text{when } T > T_f \\ T_f & \text{otherwise} \end{cases} \quad (5.43)$$

Clearly, this is a first order transition.

Summarizing this discussion, we conclude that there are three different globular phases for heteropolymers prepared by our procedure: (i) *random globule*, essential-

---

<sup>3</sup>In thermodynamic limit, the probability to obtain given target conformation out of random choice is negligible.

ly similar to homopolymeric one, where energetical preferences between monomers are not sufficient to stabilize any particular conformation, so that the thermodynamic equilibrium is realized as the mixture of astronomically large number of conformations; (ii) *frozen globule*, where each chain chooses some small number of the minimally frustrated [Bry87] conformations, but the choice is essentially unpredictable and remains out of control; (iii) *target globule*, where chain chooses exactly the conformation prescribed in the preparation procedure. This is shown in phase diagram, Fig. 1.

Now we are prepared to finally perform the  $n \rightarrow 0$  limit. Indeed, for both  $y = 0$  and  $y = n$  cases, the effective free energy (5.42) is linear in  $n$ . According to the original expression of the replica approach (5.4), the real free energy of the heteropolymer chain equals to

$$F = -T \lim_{n \rightarrow 0} \frac{\langle Z^n(\text{seq}) \rangle_{\text{seq}} - 1}{n} = -T \lim_{n \rightarrow 0} \frac{\exp(-\mathcal{F}) - 1}{n} \simeq \frac{T\mathcal{F}}{n}. \quad (5.44)$$

From this, we write the free energies of all three globular phases: random ( $x = 1$ ;  $y = 0$ ), frozen ( $x = T/T_f$ ;  $y = 0$ ), and target ( $x = T/T_f$  for  $T < T_f$ ,  $x = 1$  for  $T \geq T_f$ ;  $y = n$ )

$$\frac{1}{N} F_{\text{random}} = T \ln \left[ 1 - \frac{T_f \xi(s)}{T} \right] \quad (5.45)$$

$$\frac{1}{N} F_{\text{frozen}} = T \ln [1 - \xi(s)] + \frac{T_f \xi(s)}{1 - \xi(s)} \left[ \frac{T}{T_f} - 1 \right] \quad (5.46)$$

$$\frac{1}{N} F_{\text{target}} = T \ln [1 - \xi(s)] + \frac{T_f \xi(s)}{1 - \xi(s)} \left[ \frac{T}{T_f} - \frac{1 - \xi(s)}{1 - \xi(s)T_f/T_p} \right]. \quad (5.47)$$

Note that these are already *real* free energies, so that a lower free energy corresponds to a more stable phase, according to usual physical logic. By looking at the free energies above, one can easily reproduce phase diagram, Fig. 1: in each region of the diagram the corresponding free energy is minimal.



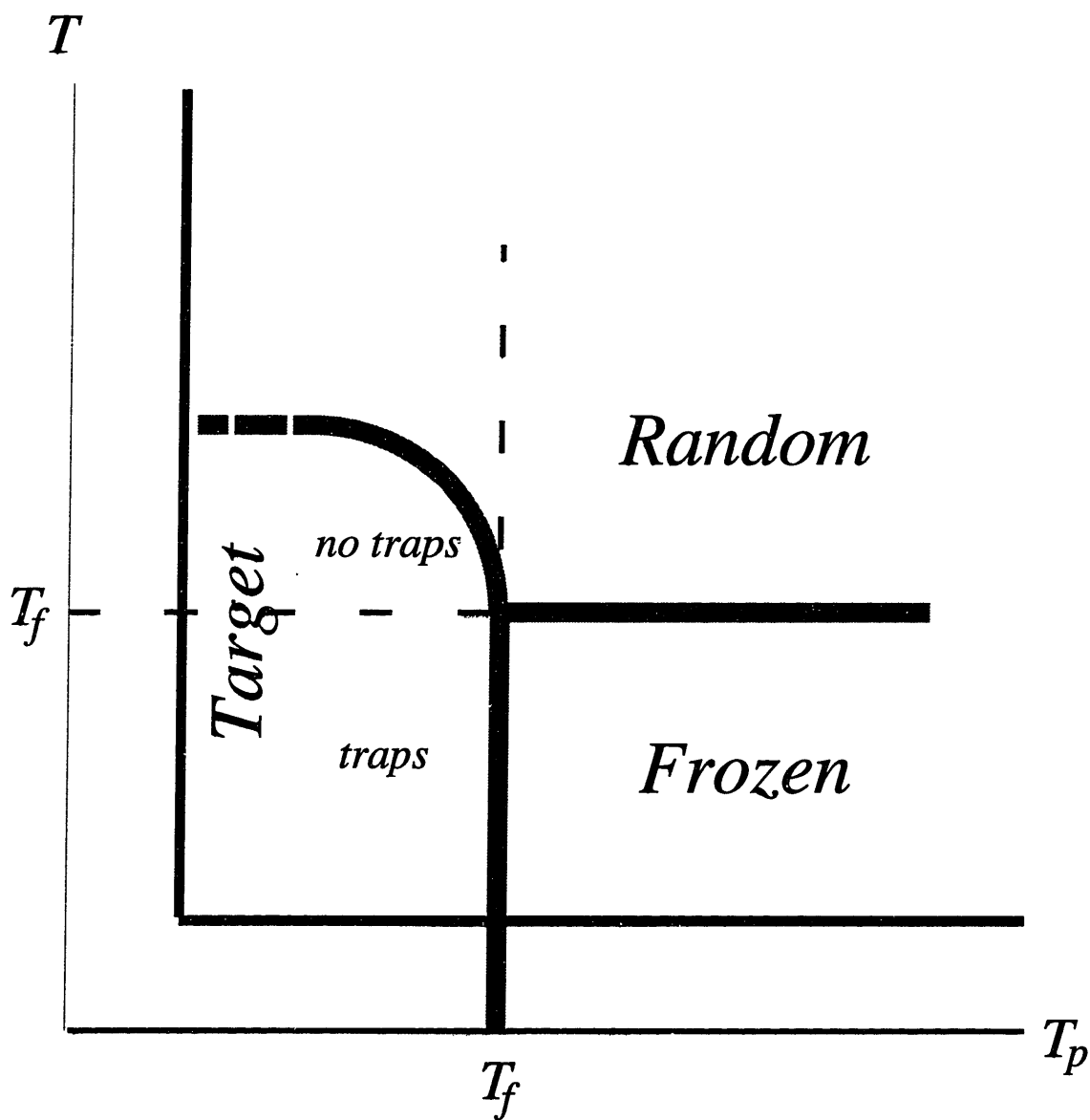


Figure 5-1: Phase diagram for designed copolymers. There are three phases: 1) *random globule*, in which a vast number of conformations (folds) are allowed for the chain in the equilibrium; 2) *frozen globule*, in which only a few conformations or even one conformation are allowed; 3) *target globule*, in which the designed conformation ( $\star$ ) is the only allowed one. Note that the target globule phase region of phase diagram can be divided in two parts: the target conformation is the most stable state in both, but a few of the other random conformations may be thermodynamically either metastable, thus serving as *traps* in kinetics, or unstable *without traps*. Lines at low  $T$  and  $T_p$  represent the areas of inapplicability of the theory.

## 5.6 Discussion

The free energies of both frozen and target phases do not depend on temperature in the low  $T$  limit:

$$\frac{1}{N}F_{\text{frozen}}(T \rightarrow 0) \equiv E_{\text{frozen}}^{\text{gnd}} = -\frac{\xi(s)T_f}{1 - \xi(s)} \quad (5.48)$$

$$\frac{1}{N}F_{\text{target}}(T \rightarrow 0) \equiv E_{\text{target}} = -\frac{\xi(s)T_f}{1 - \xi(s)T_f/T_p} \quad (5.49)$$

These limits are naturally interpreted as the energies of ground state conformations for random chain and for the chain with selected sequence, respectively. The ground state energy for a random sequence is independent of  $T_p$ , while the energy of the target conformation increases with  $T_p$ . We see that the selection of sequences, or preparation of heteropolymers by our synthetic procedure, reduces the energy of the ground state. This implies a very peculiar character of the density of states of the selected chains (Fig. 2). Indeed, since the selected sequence looks random for all conformations except for the target one, its energy spectrum includes the target conformation as the ground state and the typical ground state of random chain as the first excited state.

As was recently understood [Sha93a], this kind of spectrum is very important from the point of view of the kinetic accessibility of the ground state. Of course, one cannot analyze kinetics purely by thermodynamic considerations. In general, self-organization of the correct globular structure includes coil-to-globule compaction and some search for the correct globular conformation. We are not in the position to estimate the time scales involved in those processes. However, we can qualitatively compare the kinetics of the target phase self-organization for the two cases  $T < T_f$  and  $T > T_f$ .

It is instructive to see a realistic representation of the very bottom part of the energy spectrum, as is shown here in the magnified section. As was shown in [Sha90b], conformations of low energy are absolutely different structurally, and therefore, different pairs of monomers are in contact and are contributing to the

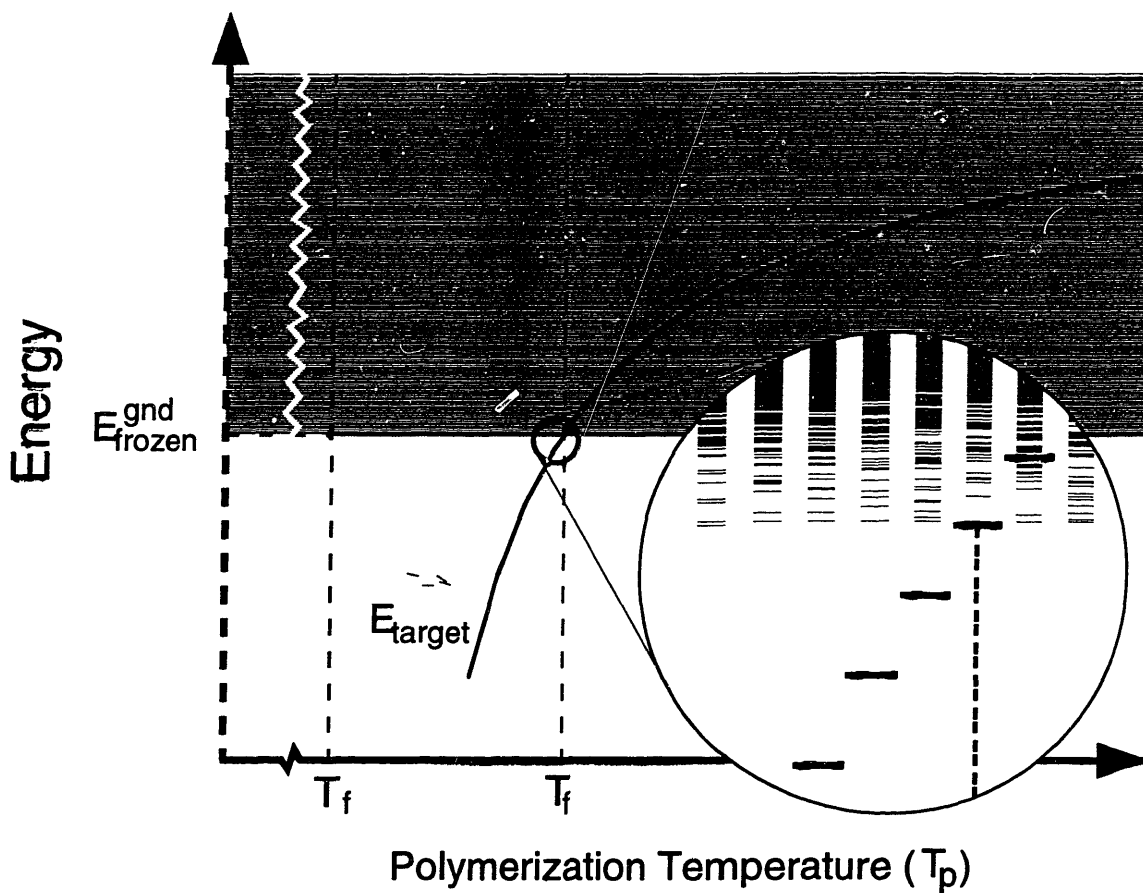


Figure 5-2: Sample energy spectra for sequences imprinted at different polymerization temperatures ( $T_p$ ). The energy of the target conformation ( $E_{\text{target}}$ ) vs polymerization temperature ( $T_p$ ) is plotted. As  $T_p$  is increased to  $T_f$ ,  $E_{\text{target}}$  increases. In the region  $T_p \approx T_f$  (magnified section), we see that  $E_{\text{target}}$  is equal to  $E_{\text{random}}^{\text{gnd}}$ , the average ground state energy of a random chain. This is related to the phase transition between target and frozen phases (see phase diagram, Fig. 1).

energy in those conformations. For this reason, the bottom part of the spectrum obeys the random energy model (REM) [Der80]. With the change of  $T_p$ , the energy of the target conformation changes in a regular fashion, as plotted. Other states represent different independent realizations of the REM system. Eight examples are shown in the inset. For  $T_p > T_f$ , the average energy for the target conformation state is larger than  $E_{\text{random}}^{\text{gnd}}$ . Note that  $E_{\text{target}}$  is the average energy, and that for  $T_p \gg T_f$ , the probability distribution of the energy of the target conformation becomes (up to normalization) equal to the density of all other states.

Indeed, consider the target phase on the phase diagram and examine first the  $T < T_f$  case. In this case, the frozen phase is, from a thermodynamic point of view, metastable. Even though it is less stable than the target state, metastability means that a macroscopic free energy barrier must be overcome to leave this state. It is, therefore, a very strong trap along the way of chain self-organization into the target conformation. We conclude, that at  $T < T_f$ , the target conformation may not be kinetically accessible, even though thermodynamically it is the most stable. On the other hand, at  $T > T_f$ , the randomly frozen conformation is not stable at all; thus, there are no effective long-living traps on the pathway of self-organization, and, therefore self-organization is expected to be considerably faster and more reliable.

We now analyze the conditions of applicability of our approach. In fact, besides the fact that we were doing mean field theory, there is only one delicate approximation which comes in eq. (5.15), where we neglect higher order terms in the expansion over  $\phi$ . It is easy to show, that all the subsequent terms in  $\phi$  are positive (and therefore do not cause the divergence of the integrals over  $\phi$  like eq. (5.13)). In particular, the next term in  $\phi$  looks like

$$\int d\mathbf{R}_1 d\mathbf{R}_2 d\mathbf{R}_3 d\mathbf{R}_4 \sum_{\alpha, \beta, \gamma, \delta=0}^n \phi^\alpha(\mathbf{R}_1) \phi^\beta(\mathbf{R}_2) \phi^\gamma(\mathbf{R}_3) \phi^\delta(\mathbf{R}_4) Q_{\alpha\beta\gamma\delta}(\mathbf{R}_1, \mathbf{R}_2, \mathbf{R}_3, \mathbf{R}_4) . \quad (5.50)$$

This should be negligible compared to the  $Q_{\alpha\beta}$ -term in eq. (5.15) throughout the region of  $\phi$  contributing to the integration over  $\phi$ . In other words, if we treat eq. (5.15) in terms of effective  $\phi$ -dependent Landau free energy and write it schematically in

the form  $f = Q_2\phi^2 + Q_4\phi^4$ , then the fourth power term should be negligible up to where quadratic term is of order one. From eq. (5.30), it is clear that the quadratic term can be estimated as  $N\lambda\phi^2$ , where  $\lambda$  is the smallest, and therefore most dangerous, eigenvalue of the  $-q_{\alpha\beta}$  matrix. On the other hand, the normalization condition for  $Q_{\alpha\beta\gamma\delta}$  implies that  $Q_4 \sim N$ . Therefore, the condition of applicability is  $N\lambda\phi_0^2 \gg N\phi_0^4$ , where  $\phi_0$  is given by  $N\lambda\phi_0^2 \sim 1$ , yielding  $\lambda \gg N^{-1/2}$ . Note, that this has a clear physical meaning:  $\lambda$  goes to 0 means that the  $\phi$ -dependent Landau free energy approaches a phase transition, which is known as microphase segregation [Sfa93]. Thus, our theory becomes inapplicable close to the microphase segregation regime. As to  $\lambda$ , we know all of the eigenvalues: they are  $T/B\rho$ ,  $T/B\rho - x$ , and (at  $y = 0$ )  $T_p/B\rho - 1$ . From (5.40), we have  $B\rho = T_f\xi(s)$ ; therefore, the condition of applicability of the approach can be written in the form

$$\frac{T}{T_f\xi(s)} - 1 \gg \frac{1}{\sqrt{N}} \quad \text{and} \quad \frac{T_p}{T_f\xi(s)} - 1 \gg \frac{1}{\sqrt{N}}. \quad (5.51)$$

On the other hand, the mean-field approach for the globule is valid at  $a^3/v \gg 1$ , or  $s \gg 1$ . In this case,  $\xi(s) \simeq 1 - 1/2s$ . If we define  $\tau$  and  $\tau_p$  according to  $T = T_f(1 - \tau)$  and  $T_p = T_f(1 - \tau_p)$ , then the conditions of applicability (5.51) take the form

$$\tau \ll \frac{1}{2s} - \frac{1}{\sqrt{N}} \quad \text{and} \quad \tau_p \ll \frac{1}{2s} - \frac{1}{\sqrt{N}}. \quad (5.52)$$

Thus, our results are valid only in a rather narrow region below the phase transition. This is understandable physically: at low polymerization temperature phase segregation occurs in preparation system of two types of monomers, giving rise to very long homopolymeric parts of the prepared sequence. This of course prevents effective freezing of the chain to either random or target conformation. For this reason we expect, that not only our theory breaks closely below the freezing temperature, but also the very phenomenon of freezing and imprinting exists only in rather narrow region of parameters for the two-letter heteropolymer. To improve the situation, one has to pass to a richer set of monomer species, as it is indicated in computer simulations [Pan94b]. The corresponding analytic theory is therefore

a challenging problem.

## 5.7 Conclusions

In conclusion, we comment on the relevance of our results. First, in the mean field approximation, there is no difference between the sequence design model of biological evolution [Sha93b] and the imprinting model [Pan94b]. Thus, the above results should be valid for both. As for general heteropolymers, including proteins, we expect that the qualitative results found here should also be valid, as the physical origins of the transition to the target state is not deeply connected to the nature of the polymer investigated, but the existence of designed sequences.

The goal of this chapter was to introduce design aspects into the replica formalism for the black/white copolymer model. In the next chapter, we introduce a model of heteropolymers with *arbitrary* short range interactions, parameterized by a matrix of monomer species interaction energies (a generalization of the black/white model) and solve it for random sequences. In chapter 7, we combine the design and the matrix interaction formalisms.

# Chapter 6

## Random Heteropolymers

Mean field replica theory is employed to analyze the freezing transition of random heteropolymers comprised of an arbitrary number ( $q$ ) of types of monomers. Our formalism assumes that interactions are short range and heterogeneity comes only from pairwise interactions, which are defined by an arbitrary  $q \times q$  matrix. We show that, in general, there exists a freezing transition from a random globule, in which the thermodynamic equilibrium is comprised of an essentially infinite number polymer conformations, to a frozen globule, in which equilibrium ensemble is dominated by one or very few conformations. We also examine some special cases of interaction matrices to analyze the relationship between the freezing transition and the nature of interactions involved.

### 6.1 Introduction

The relationship between the sequence and conformation of a heteropolymer is one of the most challenging unsolved problems in biophysics. In the case of proteins, it is widely believed that the native functional conformation is, in a sense, “written” in the sequence of the heteropolymer in the “language” of the interactions between monomer species. This conformation is also believed to be both the ground state

from thermodynamic point of view (better to say, it is structurally very close to the ground state, up to some short scale thermal and/or frozen fluctuations) and reliably accessible from the kinetic point of view.

The fact that even chains with random sequences can have a unique frozen ground state was first discussed in terms of phenomenological models [Bry87], where the freezing transition was shown to be similar to that of the Random Energy Model (REM) [Der80]. The REM-like freezing transition was also derived starting from a microscopic Hamiltonian in which the interactions between pairs of monomers were assumed to be random, independently taken from a Gaussian distribution [Sha89a]. In this model, the nature of interactions between species was parameterized in terms of the mean and width of the monomer-monomer interaction distribution. Thus, in this sense, polymer sequence was not explicitly included in this model, since it is absent from the Hamiltonian. As for models with polymer sequences explicitly present, two have been considered so far: 2 letter Ising-type model [Sfa93] and the so-called  $p$ -charge model [Gar88b,Sfa94]. These models were shown to also exhibit a freezing phase transition for random chains.

Therefore, it is natural to conjecture that any sort of random heteropolymer will have this kind of transition, and the question is whether we are able to understand the properties and characteristic temperature of this transition for realistic models of heteropolymers. Indeed, proteins, for example, are comprised of 20 kinds of different monomers, which interact to each other in a complicated manner. There are several relevant types of interactions between different monomers, such as van-der-Waals interactions, dipole-dipole interactions, hydrogen bonds, and hydrophobic interactions.

As long as we are speaking about short-range interactions, interactions can be described in terms of a matrix: if there are  $q$  types of monomers, we have a  $q \times q$  matrix, where each  $(i, j)$  matrix element represents the energy of interaction between monomers of the types  $i$  and  $j$ , given that they are in spatial contact. There were several attempts in the literature to derive this kind of “interaction” matrix for real amino acids (see, in particular, [Mia85]). It is rather difficult, however, to



derive this kind of matrix. Furthermore, the sensitivity of heteropolymer properties to deviations of the interaction matrix is unclear. For computer simulations, for example, it is important to know how precise one should be in choosing the interaction energies in order to reproduce the native state and to avoid the appearance of some other state, structurally completely different, which may appear as the ground state of a simulated system due to an imperfect interaction matrix. Of course, other non-protein heteropolymers might be also of interest.

In this chapter, we consider the freezing transition for a heteropolymer with an arbitrary interaction matrix. We derive a general formalism for the analysis of the freezing transition of random chains in which only short range interactions are assumed. In addition to the formal benefit that the general treatment establishes a formalism with which other short range species interaction models can be derived as special cases by using specific interaction matrices, this theory can be used to analyze what properties of a species-species interactions matrix effect the freezing transition and in what way.

## 6.2 Development of the Formalism

### 6.2.1 The Model and its Hamiltonian

Consider a heteropolymer chain with a frozen sequence of monomers  $s_I$ , where  $I$  is the number of monomer along the chain ( $1 \leq I \leq N$ ) and  $s_I$  is the sort of monomer  $I$  in the given sequence. Let  $q$  be the total number of different monomer species,  $1 \leq s(I) \leq q$ . In the condensed globular state, the spatial structure of the chain is governed by volume interactions between monomers. The disorder and heteropolymer effects of different monomer species comes mainly through pairwise monomer-to-monomer interactions. On the other hand, higher order interactions provide the non-specific excluded volume effect, while chain connectivity defines the set of available placements of monomers in space. This is clear when one considers the lattice model, where subsequent monomers are nearest neighbors on the lattice

(chain connectivity): a site on the lattice can be occupied by only one monomer (excluded volume effect), and the energy is given as a sum of pairwise interactions of the nearest neighbors on the lattice. The complicated set of monomer-monomer interactions, related to frozen-in sequence, appears then due to the restricted set of pairings of monomers in the space. The interaction part of Hamiltonian can be therefore written in a rather simple way:

$$\mathcal{H} = \sum_{i,j}^q \sum_{I,J}^N B_{ij} \delta(\mathbf{r}_I - \mathbf{r}_J) \delta(s_I, i) \delta(s_J, j) + \mathcal{H}' \quad (6.1)$$

where  $B_{ij} \delta(\mathbf{r}_I - \mathbf{r}_J)$  gives the Mayer function of short range interaction between monomers of *species*  $i$  and  $j$ , placed in space at the distance  $\mathbf{r}_I - \mathbf{r}_J$  apart from each other,  $s_I$  is the species of monomer number  $I$  (“spin” of monomer  $I$ ), and  $\delta$  is either Kronecker or Dirac delta. Eq (6.1) has the simple interpretation that monomers number  $I$  and  $J$  interact based upon their proximity,  $\delta(\mathbf{r}_I - \mathbf{r}_J)$ , and the second virial coefficient of interaction between the species of the two monomers,  $B_{s_I s_J}$ . The  $\mathcal{H}'$  contribution contains all higher order interactions of monomers. We assume that it is “homopolymeric” in form, i.e. it does not depend on the monomer species, but only on the overall density  $\rho$ . It can be written as  $\mathcal{H}' = C\rho^2 + D\rho^3 + \dots$ , where all virial coefficients  $C, D, \dots$  are assumed to be positive (repulsive).

Throughout the chapter, we will use the following notation: upper case Roman characters label monomer numbers, i.e. bead number along the chain ( $1 \leq I \leq N$ ), lower case Roman characters label monomer species numbers ( $1 \leq i \leq q$ ), and lower case Greek characters are for replica indices ( $1 \leq \alpha \leq n$ ), which will be defined later. We will be also using the notation for vectors and operators (matrices) with the clear indication of the dimensionality of the corresponding space, as we consider several different spaces simultaneously. For example, the interaction matrix with matrix elements  $B_{ij}$  will be denoted as  $\hat{B}^{(q)}$ . In this notation, vector  $\vec{\rho}^{(q\infty)}$  means the density distribution  $\rho_i(\mathbf{R})$  for all species ( $q$ ) over  $3D$   $\mathbf{R}$ -space ( $\infty$ ).

## 6.2.2 Replicas

The statistical mechanics of a heteropolymer chain is expressed through the partition function, which can be somewhat formally written as

$$Z(\text{seq}) = \sum_{\text{conformations}} \exp \left[ -\frac{1}{T} \mathcal{H}(\text{conf}, \text{seq}) \right], \quad (6.2)$$

where we have clearly indicated that our Hamiltonian depends on both conformation and sequence. The standard way to approach the partition function of a system with frozen disorder is to employ, first, the principle of self-averaging of free energy and, second, the replica trick:

$$F = \langle F(\text{seq}) \rangle_{\text{seq}} = -T \langle \ln Z(\text{seq}) \rangle_{\text{seq}} = \lim_{n \rightarrow 0} \frac{\langle Z^n(\text{seq}) \rangle_{\text{seq}} - 1}{n}, \quad (6.3)$$

where  $\langle \dots \rangle_{\text{seq}}$  means average over the set of all possible  $q^N$  sequences.

In this chapter, we consider random sequences, meaning that the species  $1, 2, \dots, q$  appear independently along the chain with the probabilities  $p_1, p_2, \dots, p_q$  ( $\{p_i\} = \bar{p}^{(q)}$ ), so that the probability of realization of the given sequence ( $\text{seq} = s_1, s_2, \dots, s_I, \dots, s_N$ ) is written as

$$P_{\text{seq}} = p_{s_1} p_{s_2} \dots p_{s_I} \dots p_{s_N} = \prod_{I=1}^N p_{s_I} \quad (6.4)$$

Collecting the above equations, we can write the key value of  $n$ -replica partition function as

$$\begin{aligned} \langle Z^n(\text{seq}) \rangle_{\text{seq}} &= \sum_{\text{seq}} P_{\text{seq}} \left\{ \sum_{\text{conformations}} \exp \left[ -\frac{1}{T} \mathcal{H}(\text{conf}, \text{seq}) \right] \right\}^n \\ &= \sum_{\text{seq}} P_{\text{seq}} \sum_{C_1, \dots, C_n} \exp \left[ -\frac{1}{T} \sum_{\alpha=1}^n \mathcal{H}(C_\alpha, \text{seq}) \right], \end{aligned} \quad (6.5)$$

where  $C_\alpha = C_1, \dots, C_n$  stand for conformations of replica number  $\alpha$ .

For each conformation and each replica, we introduce density distributions of

all species as

$$m_i^\alpha(\mathbf{R}) = \sum_{I=1}^N \delta(s_I, i) \delta(\mathbf{r}_I^\alpha - \mathbf{R}) ; \quad \{m_i^\alpha(\mathbf{R})\} \equiv \vec{m}^{(qn\infty)}. \quad (6.6)$$

For simplicity, we will not explicitly include the sequence independent terms  $\mathcal{H}$  from the original Hamiltonian (6.1). We then write in terms of the densities

$$\begin{aligned} \langle Z^n(\text{seq}) \rangle_{\text{seq}} &= \sum_{\text{seq}} P_{\text{seq}} \sum_{C_1, \dots, C_n} \exp \left\{ -\frac{1}{T} \sum_{\alpha=1}^n \sum_{i,j=1}^q \int d\mathbf{R}_1 d\mathbf{R}_2 m_i^\alpha(\mathbf{R}_1) B_{ij} \delta(\mathbf{R}_1 - \mathbf{R}_2) m_j^\alpha(\mathbf{R}_2) \right\} \\ &= \sum_{\text{seq}} P_{\text{seq}} \sum_{C_1, \dots, C_n} \exp \left\{ -\frac{1}{T} \langle \vec{m} | \hat{B} | \vec{m} \rangle^{(qn\infty)} \right\}, \end{aligned} \quad (6.7)$$

where  $\langle \dots \rangle^{(qn\infty)}$  means scalar product in which all vectors and operators are supposed to have dimensionality as indicated ( $q \times n \times \infty$  in this case). Operator  $\hat{B}^{(qn\infty)}$  is  $B_{ij}$  with respect to monomer species, and it is diagonal in both replica space and real coordinate space, meaning that it has matrix elements  $B_{ij} \delta_{\alpha\beta} \delta(\mathbf{R}_1 - \mathbf{R}_2)$ . The next step is to perform Hubbard-Stratonovich transformation of the form

$$\langle Z^n(\text{seq}) \rangle_{\text{seq}} = \mathcal{N} \sum_{C_1, \dots, C_n} \int \mathcal{D}\{\phi\} \exp \left\{ \frac{T}{4} \langle \vec{\phi} | \hat{B}^{-1} | \vec{\phi} \rangle^{(qn\infty)} \right\} \times \sum_{\text{seq}} P_{\text{seq}} \exp \left\{ \langle \vec{\phi} | \vec{m} \rangle^{(qn\infty)} \right\}. \quad (6.8)$$

Here  $\{\phi_i^\alpha(\mathbf{R})\} = \vec{\phi}^{(qn\infty)}$  are the fields conjugated to the corresponding densities and  $\mathcal{N}$  is normalization factor which comes from integration over  $\phi$ .

Note that the sum over sequences enters only in the last ‘‘source’’ term of this expression:

$$\exp \{\text{source term}\} = \sum_{\text{seq}} P_{\text{seq}} \exp \left\{ \langle \vec{\phi} | \vec{m} \rangle^{(qn\infty)} \right\}. \quad (6.9)$$

The summation, or average, over the sequences is easier to describe in non-vector notation:

$$\begin{aligned} \exp \{\text{source term}\} &= \sum_{s_1, s_2, \dots, s_N} \prod_{I=1}^N p_{s_I} \prod_{i=1}^q \exp \left\{ \delta(s_I, i) \sum_{\alpha=1}^n \int d\mathbf{R} \phi_i^\alpha(\mathbf{R}) \delta(\mathbf{r}_I^\alpha - \mathbf{R}) \right\} \\ &= \prod_{I=1}^N \sum_{s_I=1}^q p_{s_I} \prod_{i=1}^q \exp \left\{ \delta(s_I, i) \sum_{\alpha=1}^n \int d\mathbf{R} \phi_i^\alpha(\mathbf{R}) \delta(\mathbf{r}_I^\alpha - \mathbf{R}) \right\} \end{aligned}$$

$$= \prod_{I=1}^N \sum_{i=1}^q p_i \exp \left\{ \sum_{\alpha=1}^n \int d\mathbf{R} \phi_i^\alpha(\mathbf{R}) \delta(\mathbf{r}_I^\alpha - \mathbf{R}) \right\} \quad (6.10)$$

As in case of two-letter heteropolymer, to extract the relevant order parameters, we expand over the powers of the fields  $\phi$  (high temperature expansion) and keep terms up to  $\mathcal{O}(\phi^2)$ :

$$\begin{aligned} \text{source term} &= \sum_{i=1}^q \sum_{\alpha=1}^n \int d\mathbf{R} \rho^\alpha(\mathbf{R}) p_i \phi_i^\alpha(\mathbf{R}) \\ &+ \frac{1}{2} \sum_{i,j=1}^q [p_i \delta_{ij} - p_i p_j] \sum_{\alpha,\beta=1}^n \int d\mathbf{R}_1 \int d\mathbf{R}_2 \phi_i^\alpha(\mathbf{R}_1) Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) \phi_j^\beta(\mathbf{R}_2) \end{aligned}$$

where we use standard definitions [Gar88a,Sha89a,Pan94e]

$$Q_{\alpha_1, \dots, \alpha_k}(\mathbf{R}_1, \dots, \mathbf{R}_k) = \sum_{I=1}^N \prod_{\kappa=1}^k \delta(\mathbf{r}_I^{\alpha_\kappa} - \mathbf{R}_\kappa), \quad (6.12)$$

$$Q_\alpha(\mathbf{R}) \equiv \rho^\alpha(\mathbf{R}) = \sum_{I=1}^N \delta(\mathbf{r}_I^\alpha - \mathbf{R}) ; \quad Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) = \sum_{I=1}^N \delta(\mathbf{r}_I^\alpha - \mathbf{R}_1) \delta(\mathbf{r}_I^\beta - \mathbf{R}_2). \quad (6.13)$$

Note that the total density of the polymer chain  $\rho^\alpha(\mathbf{R})$  in equilibrium does not depend on replica number and, within a large globule, does not depend on  $\mathbf{R}$ . Replicas are interpreted as pure states of the polymer chain [Mez84,Sha89a,Pan94e], and the  $k$ -replica order parameter  $Q_{\alpha_1, \dots, \alpha_k}$  is interpreted as the overlap between replicas  $\alpha_1, \dots, \alpha_k$ .

The  $n$ -replica partition function is now written in the form:

$$\begin{aligned} \langle Z^n(\text{seq}) \rangle_{\text{seq}} &= \mathcal{N} \sum_{C_1, \dots, C_n} \int \mathcal{D}\{\phi\} \\ &\times \exp \left\{ \left\langle \vec{\phi} \left| \frac{T}{4} B_{ij}^{-1} \delta(\mathbf{R}_1 - \mathbf{R}_2) \delta_{\alpha\beta} + \frac{1}{2} Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) \Delta_{ij} \right| \vec{\phi} \right\rangle^{(qn\infty)} + \left\langle \vec{\rho} \left| \vec{\phi} \right\rangle^{(nq\infty)} \right\} \end{aligned} \quad (6.14)$$

where

$$\Delta_{ij} = p_i \delta_{ij} - p_i p_j \quad \text{and} \quad \vec{\rho}^{(qn\infty)} \equiv \rho_i^\alpha(\mathbf{R}) = p_i \sum_{I=1}^N \delta(\mathbf{r}_I^\alpha - \mathbf{R}). \quad (6.15)$$

We are left with a Gaussian integral (6.14) for the  $n$ -replica partition function, which is simplified by the argument given in [Sha89a,Sfa93,Pan94e], showing that the  $\mathbf{R}$ -dependence of  $Q_{\alpha\beta}$  is of  $\delta$ -type, so that

$$Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) = \rho q_{\alpha\beta} \delta(\mathbf{R}_1 - \mathbf{R}_2) \quad , \quad (6.16)$$

where diagonal matrix elements of new matrix  $\hat{q}^{(n)}$  are 1, while off-diagonal elements are either 0 or 1. This means physically that two replicas  $\alpha$  and  $\beta$  might be either uncorrelated (independent), so that  $Q_{\alpha\beta} = 0$ , or they may be correlated so that one repeats the 3D fold of the other down to the microscopic length scale, so that  $Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) = \rho \delta(\mathbf{R}_1 - \mathbf{R}_2)$ . We do not repeat this argument here, as it is explained elsewhere (see the argument presented in [Pan94e] which is slightly different from the original one [Sha89a]).

### 6.2.3 Effective Energy in Replica Space

With the simplified form of  $Q$  matrix, we evaluate the Gaussian integral over all  $\phi^\alpha$  variables. This yields

$$\langle Z^n(\text{seq}) \rangle_{\text{seq}} = \sum_{C_1, \dots, C_n} \exp[-NE\{Q\}] \quad (6.17)$$

with the energy of the form

$$E = \left\langle \rho^{(nq)} \left| \left[ T \left( \hat{B}^{(q)} \right)^{-1} \otimes \hat{I}^{(n)} + 2\rho \hat{q}^{(n)} \otimes \hat{\Delta}^{(q)} \right]^{-1} \right| \rho^{(nq)} \right\rangle + \frac{1}{2} \ln \det \left[ \frac{T}{4} \left( \hat{B}^{(q)} \right)^{-1} \otimes \hat{I}^{(n)} + \frac{1}{2} \rho \hat{q}^{(n)} \otimes \hat{\Delta}^{(q)} \right] + \frac{1}{2} \ln \det \left( 4\hat{B}^{(qn)} / T \right) \quad (6.18)$$

Here  $\otimes$  means the direct product, eg. for the block matrix  $\hat{B}^{(qn\infty)} = B_{ij} \delta_{\alpha\beta} \delta(\mathbf{R}_1 - \mathbf{R}_2) = \hat{B}^{(q)} \otimes \hat{I}^{(n)} \otimes \hat{I}^{(\infty)}$ . In general,  $\hat{A}^{(r)} \otimes \hat{B}^{(s)}$  produces block matrix of the total size  $rs$ , according to the rule: instead of each matrix element of  $\hat{A}^{(r)}$  matrix, say  $A_{uv}$ , we substitute the block equal to  $A_{uv} \hat{B}^{(s)}$ . The last term in (6.18) comes from normalization factor  $\mathcal{N}$  in eq (6.14); it is easy to check that the normalization factor

created by Gaussian integration  $\mathcal{N}$ , simply eliminates normalization factors first introduced by the Hubbard-Stratonovich transformation.<sup>1</sup> Noting that  $\det(\widehat{A}) \cdot \det(\widehat{B}) = \det(\widehat{A}\widehat{B})$ , we can simplify the last relationship as

$$E = \frac{1}{2} \ln \det \left[ \widehat{I}^{(qn)} + \frac{2\rho}{T} \widehat{q}^{(n)} \otimes \widehat{\Delta}^{(q)} \widehat{B}^{(q)} \right] + \left\langle \widehat{\rho}^{(nq)} \left| \frac{1}{T} \widehat{B}^{(q)} \otimes \widehat{I}^{(n)} \left[ \widehat{I}^{(qn)} + \frac{2\rho}{T} \widehat{q}^{(n)} \otimes \widehat{\Delta}^{(q)} \widehat{B}^{(q)} \right]^{-1} \right| \widehat{\rho}^{(nq)} \right\rangle, \quad (6.19)$$

We have assumed that the Gaussian integral converges and can be calculated. This is guaranteed only by the appropriate form of  $\widehat{q}^{(n)}$  matrix, i.e., by replica symmetry breaking. We make an *ansatz* that  $\widehat{q}^{(nq)}$  is of the form of a Parisi matrix with one-step replica symmetry breaking [Par80,Sfa93]. We say that replicas can be gathered into  $n/x$  groups each of which consists of  $x$  replicas. The conformations of all of the replicas in a given group coincide to the microscopic scale, i.e. for  $\alpha, \beta \in$  group A and  $\gamma \in$  group B, then  $q_{\alpha\beta} = 1$  and  $q_{\alpha\gamma} = q_{\beta\gamma} = 0$ . Thus  $\widehat{q}^{(nq)}$  can be written as a block matrix (in replica space) which is partitioned into  $n/x$  blocks of size  $x \times x$  along the diagonal. Inside each diagonal block,  $q_{\alpha\beta} = 1$ , and outside  $q_{\alpha\beta} = 0$ . In fact, it was recently shown that this form can be derived by energy minimization in the two letter case [Pan94e], and we can easily repeat this argument for the general  $q$ -letter case at hand. For the sake of simplicity, however, we omit the derivation, thus, formally employ the *ansatz*.

We can substantially simplify both terms in the energy (6.19), and convert them into the form

$$E = \frac{n}{2x} \ln \det \left( \widehat{I} + \frac{2\rho x}{T} \widehat{\Delta} \widehat{B} \right) + n \left\langle \widehat{p} \left| \rho \frac{\widehat{B}}{T} \left( \widehat{I} + \frac{2\rho x}{T} \widehat{\Delta} \widehat{B} \right)^{-1} \right| \widehat{p} \right\rangle \quad (6.20)$$

Here we have dropped the labels of the dimensionality of the vectors and operators, as all of them are of the same dimensionality ( $q$ ). This is because we have diagonal-

---

<sup>1</sup>The normalization constant appears first due to the Hubbard-Stratonovich transformation introducing functional integration over  $\phi$ . Therefore, formally we must also introduce in  $\mathcal{N}$  some cutoff volume which is essentially the underlying "lattice spacing" in our model which avoids an ultra-violet divergence. However, all of these constants are removed upon Gaussian functional integration, and therefore it is unnecessary to give any detailed description.

ized the energy in both  $\mathbf{R}$  ( $\infty$ ) and replica ( $n$ ) spaces, so only the species dimension ( $q$ ) remains.

The proof of the simplification leading to (6.20) is given in the Appendix. We now turn to its analysis.

### 6.2.4 Effective Entropy in Replica Space

In order to get the free energy, we must also consider the entropy change due to the constraint on  $q_{\alpha\beta}$ . Following Refs. [Sfa93,Gro94], we argue that due to the polymeric bonds connecting monomers along the chain, once one monomer is fixed in space, the next must be placed within a volume  $a^3$ , where  $a$  is the distance between monomers along the chain. Since replicas that belong in the same group coincide within a tube of radius  $R_t \sim v^{\frac{1}{3}}$ , where  $v$  is the excluded volume of a single monomer, there are  $a^3/v$  ways to place the next monomer and thus the entropy per monomer is just  $\ln(a^3/v)$ . But since all replica conformations coincide within the group, we must restrict the position of the next monomer to a single place. Following the Parisi ansatz for one-step RSB, for  $n$  replicas, there are  $n/x$  groups with  $x$  replicas per group. The entropy loss is therefore

$$S = Ns \frac{n}{x}(x - 1) \quad (6.21)$$

where  $s = \ln(a^3/v)$  is related to the flexibility of the chain.

### 6.2.5 Freezing Transition

Recall that, for notational convenience, we drop the indication of dimensionality, as all operators and vectors are now assumed to be in species space, i.e. dimensionality  $q$ . We optimize the free energy

$$F = \frac{n}{2x} \ln \det \left( \hat{I} + \frac{2\rho x}{T} \widehat{\Delta} \hat{B} \right) + n \left\langle \vec{p} \left| \frac{\rho}{T} \hat{B} \left( \hat{I} + \frac{2\rho x}{T} \widehat{\Delta} \hat{B} \right)^{-1} \right| \vec{p} \right\rangle + s \frac{n}{x}(x - 1) \quad (6.22)$$



with respect to  $x$ , yielding

$$2s = \ln \det \left( \hat{I} + \frac{2\rho x}{T} \widehat{\Delta} \widehat{B} \right) - \text{Tr} \left[ \frac{2\rho x}{T} \widehat{\Delta} \widehat{B} \left( \hat{I} + \frac{2\rho x}{T} \widehat{\Delta} \widehat{B} \right)^{-1} \right] + \left\langle \vec{p} \left| \frac{2\rho x}{T} \widehat{B} \frac{2\rho x}{T} \widehat{\Delta} \widehat{B} \left( \hat{I} + \frac{2\rho x}{T} \widehat{\Delta} \widehat{B} \right)^{-2} \right| \vec{p} \right\rangle. \quad (6.23)$$

As is clear from the very structure of this equation, its solution is of the form  $x = T\xi/2\rho$ , where  $\xi$  is given by

$$2s = \ln \det \left( \hat{I} + \xi \widehat{\Delta} \widehat{B} \right) - \text{Tr} \left[ \xi \widehat{\Delta} \widehat{B} \left( \hat{I} + \xi \widehat{\Delta} \widehat{B} \right)^{-1} \right] + \left\langle \vec{p} \left| \xi^2 \widehat{B} \widehat{\Delta} \widehat{B} \left( \hat{I} + \xi \widehat{\Delta} \widehat{B} \right)^{-2} \right| \vec{p} \right\rangle. \quad (6.24)$$

Recall that  $x$  is the number of replicas in one group, i.e., the number of replicas which have the same conformation down to microscopic fluctuations. This interpretation is clear when  $n$  is integer and  $n > 1$ . While taking the  $n \rightarrow 0$  limit, we have to consider  $x$  to be in between  $n$  and 1, so that  $x < 1$  means the existence of grouping of replicas, or broken replica permutation symmetry, while  $x$  approaching 1 means the restoration of replica symmetry. Therefore,  $x = 1$  defines the point of phase transition between the frozen globular phase with broken replica symmetry and the phase of a random “liquid-like” replica symmetric globule. The corresponding freezing temperature is given by  $T_f = 2\rho/\xi$ .

Thus, from the  $n$ -replica free energy, we obtain the real free energy

$$F = \begin{cases} \frac{T_f}{2} \ln \det \left( \hat{I} + \frac{2\rho}{T_f} \widehat{\Delta} \widehat{B} \right) + \left\langle \vec{p} \left| \rho \widehat{B} \left( \hat{I} + \frac{2\rho}{T_f} \widehat{\Delta} \widehat{B} \right)^{-1} \right| \vec{p} \right\rangle - s(T_f - T) & \text{for } T < T_f \\ \frac{T}{2} \ln \det \left( \hat{I} + \frac{2\rho}{T} \widehat{\Delta} \widehat{B} \right) + \left\langle \vec{p} \left| \rho \widehat{B} \left( \hat{I} + \frac{2\rho}{T} \widehat{\Delta} \widehat{B} \right)^{-1} \right| \vec{p} \right\rangle & \text{for } T > T_f \end{cases} \quad (6.25)$$

## 6.3 Discussion

### 6.3.1 What is $\widehat{\Delta}$ ?

We first examine the physical meaning of the operator  $\widehat{\Delta}$  and the term  $\widehat{\Delta}\widehat{B}$ . From the definition of  $\widehat{\Delta}$ , we have

$$(\widehat{\Delta}\widehat{B})_{ik} = \sum_j (p_i\delta_{ij} - p_i p_j) B_{jk} = p_i B_{ik} - \sum_j p_i p_j B_{ij} \quad (6.26)$$

We can always write  $B_{ij}$  in terms of a sum of a homopolymeric attraction ( $B_0$ ) and heteropolymeric deviations ( $b_{ij} = B_{ij} - \langle B \rangle$ ). From (6.26), we see that  $\widehat{\Delta}$  removes the mean interaction of species  $k$  from all matrix elements  $B_{kj}$ . In other words,  $\widehat{\Delta}$  removes all homopolymeric effects.

It is instructive to examine what happens to the energy (6.20) when one formally takes  $\widehat{\Delta}\widehat{B} = 0$ ; in this case

$$E = n(\rho/T) \langle \vec{p} | \widehat{B} | \vec{p} \rangle = n(\rho/T) \sum_{ij} p_i p_j B_{ij} = n(\rho/T) \langle B \rangle, \quad (6.27)$$

which is simply the averaged second virial term. Note that as this term is not coupled to  $x$ ,  $\langle B \rangle$  does not enter into the calculations of the freezing temperature. We note that the terms  $n\mathcal{H}' = nC\rho/T + \dots$  from the original Hamiltonian (6.1) are not explicitly written, but must be considered when optimizing the free energy. Thus, for  $|\langle B \rangle| \gg |b_{ij}|$ , we can optimize the free energy with respect to  $x$  and  $\rho$  independently. However, if this condition is not valid, the coupling between density and the replica overlap order parameter becomes significant; this should lead to other interesting physical phenomena, which are beyond the scope of this chapter. The ‘‘homopolymeric’’ attractive second virial term, in competition with the repulsive higher order terms in  $\mathcal{H}'$ , is responsible for the formation and maintenance of the globular conformation with a reasonably high density. Therefore,  $\langle B \rangle$  primarily enters into homopolymer effects, such as the coil to globule transition. Other effects, such as the freezing transition, are purely heteropolymeric, and are due to

$b_{ij}$ , or  $\widehat{\Delta\hat{B}}$  terms; they are related to the choice of some energetically preferential conformations out of the total vast number of globular conformations.

For the homopolymer case ( $q = 1$  or  $B_{ij} = B_0$ ) or the effective homopolymer case (a heteropolymeric interaction matrix is rendered homopolymeric due to the choice of composition  $\vec{p}$ ; say,  $p_1 = 1$ , while others  $p_i = 0$ ), we immediately see that  $\widehat{\Delta\hat{B}} = 0$ , so  $T_f = 0$  and thus there is no freezing transition. (This is of course just trivial check of consistency of our equations).

### 6.3.2 Two Exactly Solvable Models

There are some models which can be solved exactly from eq (6.24). We will see that the exact solution of simple models yields insight which will be important in the more general consideration of the next section.

#### Potts Model

Potts interactions are defined by the interaction matrix  $B_{ij} = b\delta_{ij} + B_0$ . The freezing temperature can be found exactly for this model for the case of even composition, i.e.  $p_i = 1/q$ . From (6.20), we see that the relevant matrix to address is  $\hat{I} + 2\rho x \widehat{\Delta\hat{B}}/T$ . As the diagonal elements of this matrix are  $1 + 2b\rho x(q - 1)/Tq^2$  and all the off diagonal elements are equal  $-2b\rho x/Tq^2$ , we find a  $(q - 1)$ -fold degenerate eigenvalue  $1 + 2b\rho x/Tq$  and a non-degenerate eigenvalue of 1 (see the Appendix for details). This leads to the energy term of the form

$$\ln \det \left( \hat{I} + \frac{2\rho x}{T} \widehat{\Delta\hat{B}} \right) = (q - 1) \ln \left( 1 + \frac{2b\rho x}{qT} \right) \quad (6.28)$$

Note that this term vanishes for the homopolymer ( $q = 1$ ) case. As for the other term of the energy (6.20), it reduces to

$$\left\langle \vec{p} \left| \frac{\rho}{T} \hat{B} \left( \hat{I} + \frac{2\rho x}{T} \widehat{\Delta\hat{B}} \right)^{-1} \right| \vec{p} \right\rangle = \left\langle \vec{p} \left| \frac{\rho}{T} \hat{B} \right| \vec{p} \right\rangle, \quad (6.29)$$

i.e., to the average second virial term (6.27). This term does not contribute to optimization with respect to  $x$ . We find the freezing temperature

$$T_f = \frac{-2b\rho}{q\Xi(2s/[q-1])} \quad (6.30)$$

where  $\Xi(\sigma)$  is given self-consistently by

$$\Xi(\sigma) : \quad \sigma = \ln(1 - \Xi) + \Xi/(1 - \Xi) \simeq \begin{cases} \Xi^2/2 & \text{for } \Xi \ll 1 \\ 1/(1 - \Xi) & \text{for } \Xi \rightarrow 1 \end{cases} \quad (6.31)$$

We see that the freezing temperature decreases with increasing  $q$ . Physically, this corresponds to the fact that in the Potts model, all monomers from differing species interact with each other in the same way, so that the part of the chain without similar monomers is effectively homopolymeric. As  $q$  increases, these homopolymer-like regions increase and the freezing temperature consequently decreases. When  $b$  is negative (positive), we have physical solutions of  $T_f$  for positive (negative)  $\Xi$ . We see from eq. (6.31) that the nature of the  $\Xi$  function is different positive and negative values: there is a singularity at  $\Xi = 1$ , whereas  $\Xi < 0$  is well behaved. Thus, there is a fundamental difference between ferromagnetic-like ( $b < 0$ ) and antiferromagnetic-like ( $b > 0$ ) interactions in terms of the freezing behavior.

Two simplified asymptotic expressions for  $T_f$  can be mentioned, coming from the two asymptotics of the  $\Xi(\sigma)$  function (6.31):

$$T_f \simeq \begin{cases} -(\rho b/\sqrt{s})(\sqrt{q-1}/q) & \text{for effectively flexible chains, } 2s/(q-1) \ll 1 \\ -(2\rho b/q)[1 + (q-1)/2s] & \text{for effectively stiff chains, } 2s/(q-1) \gg 1 \end{cases} \quad (6.32)$$

Recall that the parameter  $s = \ln(a^3/v)$  is related to chain flexibility [Gro94], where  $a$  and  $v$  are the chain spacer size and monomer excluded volume, respectively;  $s$  is small for flexible chains, and large for stiff ones. Note, that the regions of applicability of the two asymptotics in (6.31) are controlled by what can be called the effective flexibility  $\sigma = s/(q-1)$ . Physically, this corresponds again to the specific nature of Potts interactions. Indeed, the main difference between flexible

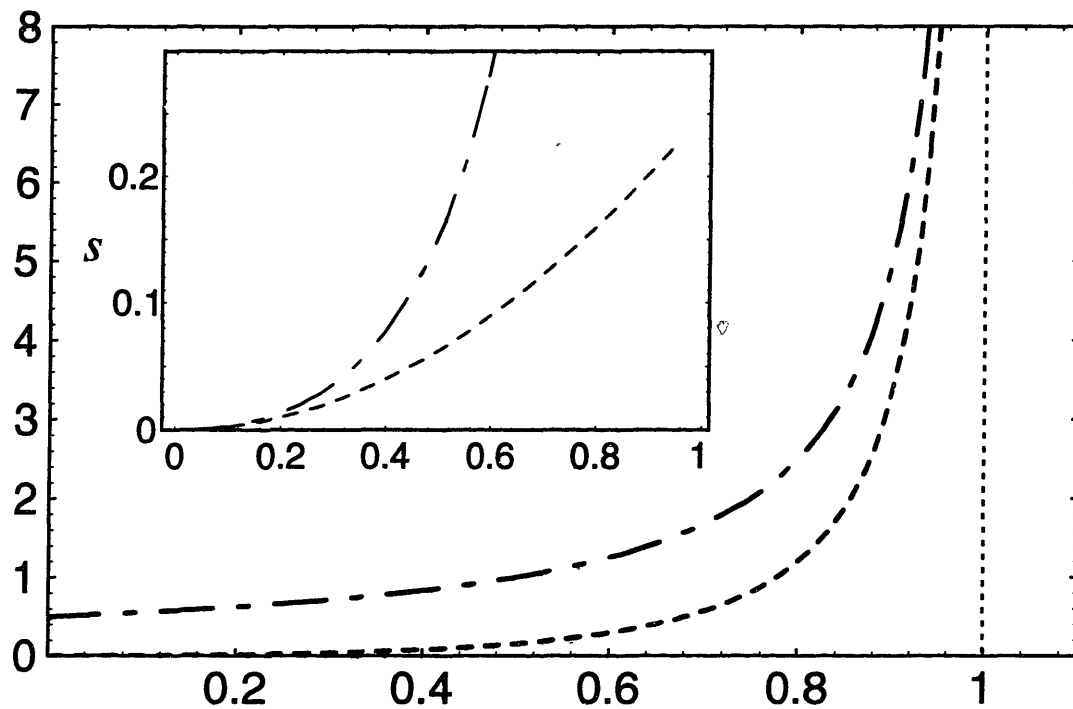


Figure 6-1: Plot of the inverse reduced freezing temperature ( $\Xi$ ) vs the effective flexibility ( $\sigma$ ), with the inset of the graph showing the detail of the small  $s$  vs small  $\Xi$  regime. The important characteristics of this function is that it is described by  $\Xi^2$  for small  $\Xi$  and the existence of a singularity at  $\Xi = 1$ . The solid line denotes the exact solution, the unevenly dashed line denotes the stiff chain expansion, and the evenly dashed line denotes the flexible chain expansion.

and stiff chains is the number of neighbors along the chain in the interaction sphere in space around a given monomer. This number is large for flexible chains and small for stiff chains. As for Potts interactions, what is relevant is how many neighbors along the chain *attract* a given monomer. This number is obviously reduced by factor  $q - 1$ , and this explains the appearance of the effective flexibility  $s/(q - 1)$ .

### $p$ -charge

In the  $p$ -charge model [Gar88b,Sfa94], each monomer has a set of  $p$  generalized charges, which can be  $s_i^k = \pm 1$ . The Hamiltonian is defined to be

$$\mathcal{H} = \sum_{I,J}^N \delta(\mathbf{r}_I - \mathbf{r}_J) \sum_{k=1}^p \chi_k s_I^k s_J^k \quad (6.33)$$

In the interaction matrix, we define each possible combination of charges as a different species. Thus, there are  $q = 2^p$  species in the interaction matrix. For species number  $i$  ( $1 \leq i \leq q$ ), the value of charge  $k$  is given by  $s^k(i) = 2 \left( \left[ \frac{i}{2^k} \right] \bmod 2 \right) - 1$ , where  $[\dots]$  means truncate to the lowest integer. Thus, we have an interaction matrix of the form

$$\widehat{B}_{ij} = \sum_k \chi_k \left[ 2 \left( \left[ \frac{i}{2^k} \right] \bmod 2 \right) - 1 \right] \left[ 2 \left( \left[ \frac{j}{2^k} \right] \bmod 2 \right) - 1 \right] \quad (6.34)$$

The  $\widehat{\Delta\widehat{B}}$  matrix has  $p$  non-zero eigenvalues  $\chi_1, \chi_2, \dots, \chi_p$  and a  $(2^p - p)$ -degenerate eigenvalue of 0. Thus,

$$\ln \det \left( \widehat{I} + \frac{2\rho x}{T} \widehat{\Delta\widehat{B}} \right) = \sum_{i=1}^p \ln \left( 1 + \frac{2\chi_i \rho x}{T} \right) \quad (6.35)$$

and as in the Potts case,

$$\left\langle \vec{p} \left| \frac{\rho}{T} \widehat{B} \left( \widehat{I} + \frac{2\rho x}{T} \widehat{\Delta\widehat{B}} \right)^{-1} \right| \vec{p} \right\rangle = \left\langle \vec{p} \left| \frac{\rho}{T} \widehat{B} \right| \vec{p} \right\rangle \quad (6.36)$$

Thus, the freezing temperature is determined by

$$2s = \sum_{i=1}^p \left[ \ln \left( 1 + \frac{2\chi_i \rho}{T_f} \right) - \frac{2\chi_i \rho / T_f}{1 + 2\chi_i \rho / T_f} \right] \quad (6.37)$$

For the specific case  $\chi_i = \chi$ , we have

$$T_f = -\frac{2\rho\chi}{\Xi(2s/p)}, \quad (6.38)$$

where  $\Xi(\sigma)$  function is defined above by (6.31). As in Potts interactions, the asymmetry of the  $\Xi$  function yields different behavior, depending on the sign of  $\chi$ . Unlike the Potts case, the behavior of the  $p$ -charge model becomes more heteropolymeric, i.e.  $T_f$  increases, with the addition of more species.

The two asymptotics, for flexible and stiff chains, in  $p$ -charge model are

$$T_f \simeq \begin{cases} -\rho\chi(2p/s)^{1/2} & \text{for effectively flexible chain, } s/p \ll 1 \\ -2\rho\chi(1 + p/2s) & \text{for effectively stiff chain, } s/p \gg 1 \end{cases} . \quad (6.39)$$

Note, that effective flexibility is given by  $\sigma = s/p$  for the  $p$ -charge model, i.e. it is again reduced by the number of species.

We note that our result (6.38) reproduces automatically what is trivially expected for the homopolymer case ( $T_f = 0$ , i.e., no freezing, when  $p = 0$ ) and also at  $p = 1$  agrees with our previous result (6.30) at  $q = 2$  in the case of two letter Ising heteropolymer. On the other hand, our equation (6.38), or its asymptotics in the first line of eq (6.39), agrees with earlier results of the work [Sfa94] in the opposite extreme of  $p \gg 1$ , i.e., in the region of applicability of that work.

### 6.3.3 Reduction Theorems

There are several cases in which the same physical system can be depicted in terms of formally different interaction matrices  $\hat{B}$  and/or composition vectors  $\vec{p}$ . Clearly, the expression for the freezing temperature, as well as for any other real physical quantity, must not depend on any arbitrary choice.

For example, there might be some monomer species which are formally included in the list, and in the interaction matrix, but they are not physically presented in the chain, as the corresponding  $p$  vanishes, say,  $p_q = 0$ . It is easy to check, that in this case eq (6.24) is reduced to smaller list of  $q - 1$  monomer species with  $(q - 1) \times (q - 1)$  interaction matrix.

Another example is when there are duplicate species, say, species labeled  $q$  and  $q - 1$  are physically identical, i.e. they interact in identical ways to all other species. Physically, we would expect that this problem is identical to the  $q - 1$  species case, except with the new composition  $p'_{q-1} = p_{q-1} + p_q$ . Even though we skip the proof, eq (6.24) indeed gives this expected reduction.

These two statements, which we call “reduction theorems”, are not only a good check of consistency of our result (6.24), but they will be also important in further discussion.

### 6.3.4 Freezing Temperature: General Consideration

We return to the general analysis of the equation (6.24) for the freezing temperature, and we will show how to implement in the general case both the limits of stiff and flexible chains, similar to how those cases appear in the exact solutions for the Potts and  $p$ -charge models.

We first perform an expansion in powers of  $\xi = 2\rho x/T$ . For example,

$$\widehat{B}(I + \xi\widehat{\Delta}\widehat{B})^{-1} = \widehat{B} - \xi\widehat{B}\widehat{\Delta}\widehat{B} + \xi^2\widehat{B}\widehat{\Delta}\widehat{B}\widehat{\Delta}\widehat{B} + \dots \quad (6.40)$$

Note, that any term  $\widehat{B}(\widehat{\Delta}\widehat{B})^k$ , where  $k$  is a positive integer, is independent of  $\langle B \rangle$ , and therefore is purely heteropolymeric. The matrix product  $(\widehat{\Delta}\widehat{B})_{i_1 i_2} (\widehat{\Delta}\widehat{B})_{i_2 i_3} \dots (\widehat{\Delta}\widehat{B})_{i_{k-1} i_k}$  can be interpreted as the propagation of heteropolymeric interactions from monomer species  $i_1$  to  $i_2$ , from  $i_2$  to  $i_3$ , etc., up to  $i_k$ . As we suppose from the very beginning that all of the heterogeneity comes from the second virial coefficient only, so that all higher order virial terms of the original Hamiltonian are in a sense homopolymeric, all heteropolymeric interactions are simply pair collisions of monomers. Each



monomer takes part, of course, in a variety of pair collisions during a very long time, i.e., in thermodynamic equilibrium. Those collisions are weighted with the corresponding energies, and they form chains of collisions, described by  $\widehat{B}(\widehat{\Delta}\widehat{B})^k$  terms. Depending on both the  $B_{ij}$  interaction matrix and the species occurrence probabilities  $p_i$ , some of those chains might be more or less favorable than others, and this determines freezing transition in the system.

To employ the expansion (6.40), we first rewrite (6.24) by noting that  $\ln \det \widehat{A} = \text{Tr} \ln \widehat{A}$  and  $\langle \vec{p} | \widehat{A} | \vec{p} \rangle = \text{Tr} \widehat{P} \widehat{A}$ , where  $\widehat{P}_{ij} = p_i p_j$ :

$$2s = \text{Tr} \left\{ \ln \left( \widehat{I} + \xi \widehat{\Delta} \widehat{B} \right) - \xi \widehat{\Delta} \widehat{B} \left[ \widehat{I} + \xi \widehat{\Delta} \widehat{B} \right]^{-1} + \xi^2 \widehat{P} \widehat{B} \widehat{\Delta} \widehat{B} \left[ \widehat{I} + \xi \widehat{\Delta} \widehat{B} \right]^{-2} \right\} . \quad (6.41)$$

Now we are in a position to perform the expansion over the powers of  $\xi$ , yielding

$$2s = \sum_{k=2}^{\infty} \xi^k \langle B^k \rangle_m , \quad (6.42)$$

where

$$\langle B^k \rangle_m = \frac{k-1}{k} \text{Tr} \left[ \left( -\widehat{\Delta} - k \widehat{P} \right) \widehat{B} \left( -\widehat{\Delta} \widehat{B} \right)^{(k-1)} \right] . \quad (6.43)$$

The values  $\langle B^k \rangle_m$  can be considered as moments of  $\widehat{B}$  matrix produced by a given  $\widehat{\Delta}$  matrix. In fact, we can make the substitution  $b_{ij} = B_{ij} - \sum_{kl} p_k p_l B_{kl}$ , i.e. remove the ‘‘homopolymer’’ mean from the interaction matrix, and the moments can be rewritten exactly with the exchange  $B_{ij} \rightarrow b_{ij}$ . A consequence of this symmetry is that these moments vanish in the homopolymer case ( $b_{ij} = 0$ ).

We now pass to analysis of two opposite extremes in the equation (6.41).

### Freezing Temperature: Stiff Chain Limit

As we are instructed by the examples of Potts and  $p$ -charge models, what is important in high  $s$  limit is the singularity of the right hand side of (6.41). This is obviously governed by high  $k$  terms of power series, which are basically related to  $(-\widehat{\Delta}\widehat{B})^k$ . This is reminiscent of the standard problems of 1D statistical physics, such as the 1D Ising model, the ideal polymer, or other Markovian processes, where

$(-\widehat{\Delta}\widehat{B})$  plays the role of the transfer matrix. It is well known that highest eigenvalue of the transfer matrix is only relevant in  $k \rightarrow \infty$  limit (“ground state dominance principle”). In this limit,  $\xi \simeq 1/\lambda_{\max}$ , where  $\lambda_{\max}$  is highest eigenvalue of  $(-\widehat{\Delta}\widehat{B})$  matrix, and thus

$$T_f \simeq 2\rho\lambda_{\max} . \quad (6.44)$$

To find the next terms in asymptotic formula for  $T_f$ , we note that the most divergent term in eq (6.42) comes from the last term in (6.41) and is due to  $k\widehat{P}$  term in (6.43), it diverges as  $(1 - \xi\lambda_{\max})^{-2}$ . We know, however, that this term vanishes for both Potts and  $p$ -charge models. Moreover, we can show, that it vanishes also for many other models with some regularities, producing cancellation of correlations and anti-correlations between matrix elements of  $\widehat{B}$ . For this reason, we keep next to the highest singularity, thus obtaining

$$2s \simeq \frac{c}{(1 - \xi\lambda_{\max})^2} + \frac{c'}{(1 - \xi\lambda_{\max})} , \quad (6.45)$$

with  $c$  and  $c'$  being the constants solely defined by  $\widehat{B}$  and  $\widehat{\Delta}$ .<sup>2</sup> This gives finally

$$T_f \simeq 2\rho\lambda_{\max} \left[ 1 + \frac{c' + \sqrt{c'^2 + 4cs}}{2s} \right] \simeq \begin{cases} 2\rho\lambda_{\max} [1 + \sqrt{c/s}] & \text{for } cs \gg c' \quad (c \neq 0) \\ 2\rho\lambda_{\max} [1 + c'/s] & \text{for } cs \ll c' \quad (c = 0) \end{cases} . \quad (6.46)$$

Note that  $\lambda_{\max}$ , as an eigenvalue, depends strongly on the arrangement of matrix elements. Therefore freezing transition for stiff chains is very dependent on the pattern of interactions, not only on their overall heterogeneity. This has clear physical meaning. In case of stiff chains, real monomers represent the physical units of interaction. In other words, quasi monomers almost coincide with monomers. In terms of propagation, or chains of collisions (see above), it is clear that highest eigenvalue of  $(-\widehat{\Delta}\widehat{B})$  matrix corresponds to the lowest (because of the sign) energy of interaction, while the corresponding eigenvector, in terms of the obvious quantum mechanical analogy, is the linear combination of monomers which realizes this lowest

---

<sup>2</sup>Let  $\lambda$  and  $|\psi\rangle$  be the eigenvalue and corresponding eigenvector of the  $(-\widehat{\Delta}\widehat{B})$  operator, respectively. We find that  $c = \langle \mathbf{p} | \psi \rangle \langle \psi - \widehat{B}/\lambda | \psi \rangle$

energy and thus controls the freezing temperature.

### Freezing Temperature: Flexible Chain Limit

The examination of the small  $s$  case may be on the first glance questionable, as our approach is entirely mean-field in nature and, therefore, it might be applicable for large enough  $s$  only. We have seen, however, in the examples of Potts and  $p$ -charge models, that the applicability of the flexible chain limit is controlled by the *effective* flexibility, which is considerably smaller than  $s$  itself. We therefore consider formally the small  $s$  limit, leaving the analysis of applicability for each particular case.

In small  $s$  limit, only the first term with  $k = 2$  is relevant in the series (6.42). Omitting all higher order terms, we obtain the remarkably simple result

$$T_f = \frac{2\rho}{\sqrt{s}} \langle \hat{B}^2 \rangle_c^{1/2} \quad (6.47)$$

where the second cumulant (variance) is defined as  $\langle \hat{B}^2 \rangle_c \equiv \langle [\hat{B} - \langle \hat{B} \rangle]^2 \rangle$  and matrix averages are defined by

$$\langle \hat{B} \rangle \equiv \sum_{ij} p_i p_j B_{ij} . \quad (6.48)$$

Unlike the stiff chain limit, in the flexible chain case at hand, the freezing transition is controlled mainly by overall heterogeneity of interaction energies  $B_{ij}$ . Thus, if one started with an interaction matrix with independent elements and shuffled the matrix elements (even though it is hard to think of real physical experiment of this kind), this transformation would not change the freezing temperature for flexible chains. This is qualitatively a very natural result, as the nature of flexible chains is such that for any given monomer, many of the neighbors in space are neighbors along the chain. In other words, the interaction units are quasi monomers, which are substantially different from the monomers and represent clouds of monomers, where the individuality of each monomer species (with different patterns of energetical preferences to other species) is lost.

In the case of the Potts and  $p$ -charge models, the variance of the interaction matrix yields the flexible chain limits for both the Potts (6.32) and  $p$ -charge (6.38) models. Thus, the solution (6.47) for  $T_f$  in this limit is remarkably simple and powerful. To demonstrate this, we show some particular examples.

### 6.3.5 Independent Interaction Model

In the Independent Interaction Model, all  $B_{IJ}$  are taken independently from Gaussian distribution

$$P(B_{IJ}) = \left(\frac{\tilde{B}^2}{2\pi}\right)^{1/2} \exp\left[-\frac{(B_{IJ} - B_0)^2}{\tilde{B}^2}\right] \quad (6.49)$$

(recall that capital  $I$  and  $J$  are related to monomer numbers along the chain and not to species). From the physical point of view, the independence of, say,  $B_{IJ}$  and  $B_{JK}$  can be realized if and only if the total number of different species is very large, i.e., in the  $q \rightarrow \infty$  limit. The effective stiffness in this limit is small, and we have to use the expression (6.47) for the freezing temperature. Therefore,  $T_f = 2\rho\tilde{B}/\sqrt{s}$ . This indeed coincides with original result of the work [Sha89a].

### 6.3.6 Random Sequences of Real Amino Acids

It is of special interest to examine the freezing transition for polymers comprised of real amino acids, i.e., of constituents of real proteins. This can be done using the matrix of interaction energies derived for amino acids by Miyazawa and Jernigan [Mia85]. We are in a position to examine the freezing transition for random sequences (even though real protein sequences might not be random [Pti86,Pan94c]). In the work [Mia85], interaction energies are given in some conventional  $T = T_{MJ}$  units.<sup>3</sup> In some rough approximation, we identify the MJ matrix with our  $\rho\hat{B}$ .

---

<sup>3</sup>To understand the origin of  $T_{MJ}$ , recall the way that the MJ matrix was derived in [Mia85]. The protein 3D structures data bank was employed such that if there were  $\mathcal{M}_{ij}$  contacts between amino acids labeled as  $i$  and  $j$  in the data bank, and the total number of contacts was  $\mathcal{M}$ , then the ratio  $\mathcal{M}_{ij}/\mathcal{M}$  was interpreted as a probability governed by some effective Boltzmann distribution  $\mathcal{M}_{ij}/\mathcal{M} = \exp[-U_{ij}/T_{MJ}]$ , thus yielding the MJ matrix of energies,  $U_{ij}$ . In the later work [Fin93], it was shown that the ratio  $\mathcal{M}_{ij}/\mathcal{M}$  obeys indeed Boltzmann type formula if proteins do match the random energy model, and then the parameter of distribution,  $T_{MJ}$ , is nothing but the freezing

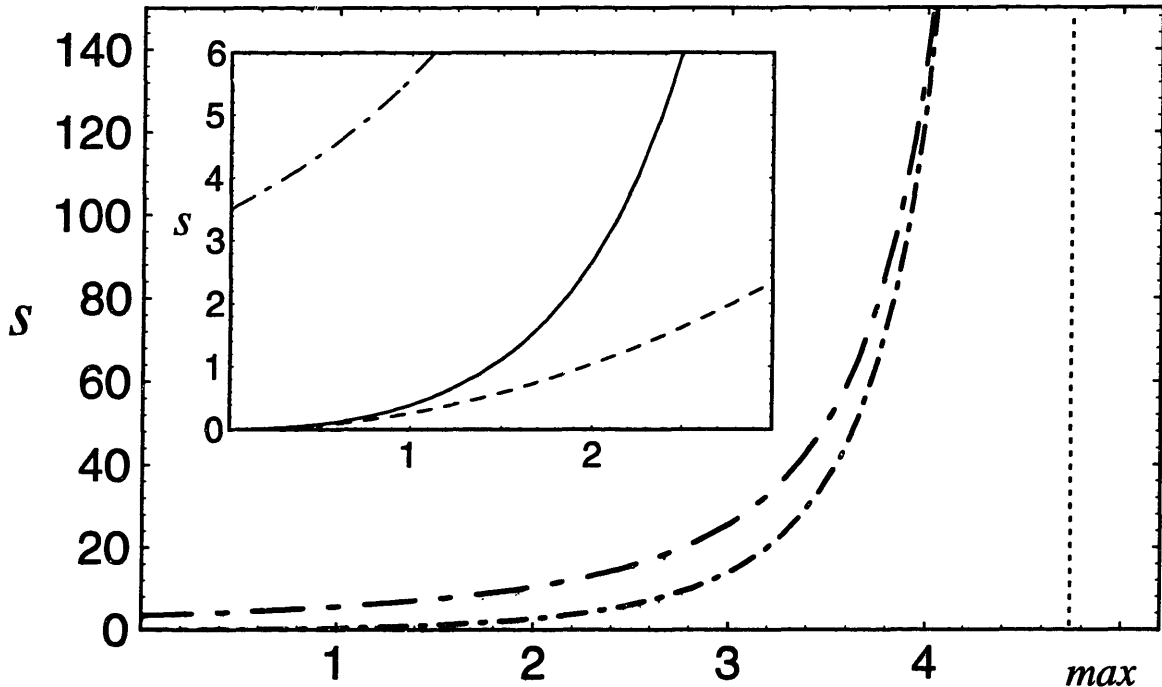


Figure 6-2: For the Miyazawa and Jernigan matrix of amino acid interactions, we plot the flexibility ( $s$ ) vs the reduced inverse freezing temperature ( $\xi$ ), with the inset of the graph showing the detail of the small  $s$  vs small  $\xi$  regime. Qualitatively, this curve is similar to  $\Xi(\sigma)$ . Further note, however, that any physical polymer will be described by the small  $\xi$  regime. The solid line denotes the exact solution, the unevenly dashed line denotes the stiff chain expansion, and the evenly dashed line denotes the flexible chain expansion.

To avoid rewriting of the eq (6.24), we substitute the MJ matrix into (6.24) instead of  $\hat{B}$ , meaning that now  $\xi = 2T_{MJ}/T_f$ . We assume also equal composition  $p_i = 1/q = 1/20$ . We can then numerically calculate the  $\xi$  vs  $s$  dependence. The result is shown in Figure 1. Note the qualitative similarity of the graph of  $\xi$  vs  $s$  for the MJ matrix and  $\Xi$  vs  $s$  given by (6.31).

Given the realistic value of  $s \approx 1.4$  ( $v/a^3 \approx 0.25$ ) for polypeptide chain, we obtain from the Figure 1 the estimate  $\xi \approx 1.6$ , or  $T_f \approx 1.25T_{MJ}$ . By taking more realistic uneven composition, we arrive at  $\xi \approx 1.75$ , or  $T_f \approx 1.14T_{MJ}$ . Note that

---

temperature,  $T_f$ . (We are indebted to A.Gutin for the discussion of this point.) From that logic, we expect thus  $T_f = T_{MJ}$ . Our result is slightly higher. We conclude thus, that there is a reasonable agreement between the works [Mia85], [Fin93] and our results.

for real amino acids system the relevant solution is generally in the high flexibility regime.

## 6.4 Conclusion

Starting from a sequence-model Hamiltonian in which interactions between *species* of monomers is expressed in terms of some arbitrary symmetric matrix  $\widehat{B}$ , we have derived a formalism with which to examine the freezing transition of random heteropolymers. As monomer species interactions are given by some matrix, this formulation is the most general form, assuming that interactions are short range and that heteropolymeric contributions come primarily from two-body interactions.

First, we have related the freezing temperature to the interaction matrix self-consistently. This self-consistent equation can be solved exactly for certain specific systems. For example, models such as the  $q$ -Potts and  $p$ -charge models are important as they describe interesting physical cases, but with only a minimal amount of complexity in their solutions. It is especially interesting that these two simple models have radically different freezing behavior with respect to the number of species. Clearly simply adding new and different monomer species does not necessarily enhance the freezing transition.

Taking another approach, we can trade the accuracy of an exact result for the generality of the assumption of only some arbitrary symmetric interaction matrix. To this end, we solved the exact self-consistent equation perturbatively. Due to the nature of the  $\Xi$  function, there are two regimes of interest: small  $s$  (high effective flexibility) where  $\Xi \rightarrow 0$  and  $s \rightarrow \infty$ , where  $\Xi$  approaches a singularity in the self-consistent formulation. Expanding at these two limits, we found

$$T_f \simeq \begin{cases} (2\rho/\sqrt{s}) \langle \widehat{B}^2 \rangle_c^{1/2} & \text{for small } s \\ \rho\lambda_{\max} & \text{for large } s \end{cases}$$

where  $\lambda_{\max}$  is the largest eigenvalue of the  $-\widehat{\Delta}\widehat{B}$  matrix.

The equation above quantitatively details certain descriptions of what one could

qualitatively call the “heteropolymeric character” of the interaction matrix  $\widehat{B}$  and the species composition  $\vec{p}$ . Specifically, for flexible chains, one would expect that the physical unit of interactions, or quasi monomers, consist of several monomers. The variance of the interaction matrix gives, in a sense, the heteropolymeric width of interactions. If these interaction energies are ordered in the interaction matrix, however, the correlations between monomer species interactions reduces the heteropolymeric nature of the system, and thus reduces the freezing temperature.

In the limit of stiff chains, quasi monomers consist generally individual monomers. Thus, the specific nature of interactions are of paramount importance. In this limit, one can imagine the interactions in space (i.e. not necessarily along the chain) as interactions propagating through the pairwise interactions of monomer species. This chain of interactions, in the stiff polymer limit, becomes very long and thus the system shares characteristics with other one-dimensional systems, such as the 1D Ising model. Specifically, here the freezing temperature is proportional to the largest eigenvalue, which dominates in the long interaction chain limit, of the transfer matrix  $-\widehat{\Delta}\widehat{B}$ .

In conclusion, for models with “heteropolymeric character,” i.e., the interaction matrix and probability distribution cannot be reduced to that of a homopolymer, our theory predicts a freezing transition. Our formalism facilitates the calculation of specific models of interactions, but perhaps most importantly, the direct relationship between the interaction matrix and the freezing transition is demonstrated.

## 6.5 Appendix: Proof of equation (20)

1. Consider first the auxiliary problem of some  $x \times x$  matrix  $\widehat{q}^{(x)}$  with diagonal elements  $\tilde{q}$  and off diagonal elements  $q$ . This matrix has a  $(x - 1)$ -fold degenerate eigenvalue  $\lambda = \tilde{q} - q$ , corresponding to the eigenvectors  $(1 \ -1 \ 0 \ 0 \ \dots \ 0)$ ,  $(1 \ 0 \ -1 \ 0 \ \dots \ 0)$ ,  $\dots$ ,  $(1 \ 0 \ \dots \ 0 \ -1 \ 0 \ \dots \ 0)$ ,  $\dots$ ,  $(1 \ 0 \ 0 \ \dots \ 0 \ -1)$ , and a non-degenerate eigenvalue of  $\lambda = \tilde{q} + (x - 1)q$ , corresponding to the eigenvector  $(1 \ 1 \ 1 \ \dots \ 1)$ . Of course, there are other ways of choosing eigenvectors, in

particular, we can built up orthonormal basis by choosing

$$\mathcal{R}_{\alpha\beta} = \frac{1}{\sqrt{x}} \exp \left[ \frac{2\pi i}{x} (\alpha - 1)(\beta - 1) \right] ; \quad 1 \leq \alpha, \beta \leq x . \quad (6.50)$$

Here  $\alpha$  numerates eigenvectors, while  $\beta$  numerates components of the given eigenvector (or vice versa). We can interpret  $\widehat{\mathcal{R}}^{(x)} = \mathcal{R}_{\alpha\beta}$  as the unitary operator transforming  $\widehat{q}^{(x)}$  to diagonal form,  $\widehat{\mathcal{R}}\widehat{q}\widehat{\mathcal{R}}^{-1} = \widehat{\lambda}^{(x)} \equiv \lambda_\alpha \delta_{\alpha\beta}$ , with the eigenvalues  $\lambda_\alpha$  given above. <sup>4</sup> We will be particularly interested in the case  $q = \bar{q} = 1$ . In this case, the non-degenerate eigenvalue is  $\lambda = 1$ , while all the others are zero.

2. Consider now some general properties of the “direct product” operation for matrices. We repeat the definition:  $\widehat{A}^{(r)} \otimes \widehat{B}^{(s)}$  is  $rs \times rs$ , built up by substitution of  $s \times s$  block  $A_{uv}\widehat{B}^{(s)}$  instead of each matrix element of  $\widehat{A}^{(r)}$ .

1. By matrix row and column operations, it is easy to show that the rule is commutative, i.e.

$$\widehat{A}^{(r)} \otimes \widehat{B}^{(s)} = \widehat{B}^{(s)} \otimes \widehat{A}^{(r)} . \quad (6.51)$$

2. *Block matrix multiplication rule:* it is well known that the operation of block matrix multiplication is carried out in the same scheme as normal matrix multiplication, except the multiplication of elements is replaced by the matrix multiplication of blocks. This can be written as

$$\left( \widehat{A}^{(r)} \otimes \widehat{B}^{(s)} \right) \cdot \left( \widehat{A}'^{(r)} \otimes \widehat{B}'^{(s)} \right) = \left( \widehat{A}^{(r)} \widehat{A}'^{(r)} \right) \otimes \left( \widehat{B}^{(s)} \widehat{B}'^{(s)} \right) . \quad (6.52)$$

3. Commutation of  $\widehat{A}^{(r)} \otimes \widehat{B}^{(s)}$  and  $\widehat{A}'^{(r)} \otimes \widehat{B}'^{(s)}$  depends on commutation of *both* pairs  $\widehat{A}^{(r)}$  &  $\widehat{A}'^{(r)}$  and  $\widehat{B}^{(s)}$  &  $\widehat{B}'^{(s)}$  (this directly follows from previous.)

4. The determinant of a block diagonal matrix equals to the product of deter-

---

<sup>4</sup>For completeness, we write also the inverse of  $\widehat{q}^{(x)}$ : it has diagonal elements  $(\bar{q} - q)^{-1} - q\{(\bar{q} - q)[\bar{q} + (x - 1)q]\}^{-1}$  and off diagonal elements  $-q\{(\bar{q} - q)[\bar{q} + (x - 1)q]\}^{-1}$ .



minants of the diagonal blocks. In particular,

$$\det (\widehat{A}^{(r)} \otimes \widehat{I}^{(s)}) = (\det \widehat{A}^{(r)})^s \quad (6.53)$$

5. The definition of direct product can be trivially generalized for non-square matrices and, in particular, for vectors <sup>5</sup>. For example,  $|\widehat{\rho}^{(nq)}\rangle = \widehat{\rho}^{(n)} \otimes \widehat{\rho}^{(q)}$ .

6. *Matrix operation with a vector:*

$$\widehat{A}^{(r)} \otimes \widehat{B}^{(s)} |\vec{a}^{(r)} \otimes \vec{b}^{(s)}\rangle = \widehat{A}^{(r)} |\vec{a}^{(r)}\rangle \otimes \widehat{B}^{(s)} |\vec{b}^{(s)}\rangle. \quad (6.54)$$

7. *Scalar product of vectors:*

$$\langle \vec{a}^{(r)} \otimes \vec{b}^{(s)} | \vec{a}'^{(r)} \otimes \vec{b}'^{(s)} \rangle = \langle \vec{a}^{(r)} | \vec{a}'^{(r)} \rangle \langle \vec{b}^{(s)} | \vec{b}'^{(s)} \rangle \quad (6.55)$$

The proof of all the above mentioned properties is straightforward.

3. Let us return now to the expression of energy (6.19). We have to address the matrix  $[\widehat{I}^{(qn)} + \frac{2\rho}{T}\widehat{q}^{(n)} \otimes \widehat{\Delta}^{(q)}\widehat{B}^{(q)}]$ . We know (or we assume) that  $\widehat{q}^{(n)}$  is comprised of  $n/x$   $\widehat{q}^{(x)}$  blocks along the diagonal, with  $\widehat{q} = q = 1$ , that is  $\widehat{q}^{(n)} = \widehat{I}^{(n/x)} \otimes \widehat{q}^{(x)}$ . First, this form of  $\widehat{q}^{(n)}$  matrix allows us to factor the matrix of our interest:

$$\left[ \widehat{I}^{(qn)} + \frac{2\rho}{T}\widehat{q}^{(n)} \otimes \widehat{\Delta}^{(q)}\widehat{B}^{(q)} \right] = \widehat{I}^{(n/x)} \otimes \left[ \widehat{I}^{(qx)} + \frac{2\rho}{T}\widehat{q}^{(x)} \otimes \widehat{\Delta}^{(q)}\widehat{B}^{(q)} \right]. \quad (6.56)$$

This means physically that replicas of different groups are not coupled, they do not interact to each other.

The remainder (in the square brackets in the right hand side of (6.56)) can be diagonalized via the rotation operator  $\widehat{\mathcal{R}}^{(xq)} = \widehat{\mathcal{R}}^{(x)} \otimes \widehat{I}^{(q)}$ . Indeed, using properties 2 and 3 above, we have:

$$(\widehat{\mathcal{R}}^{(xq)})^{-1} \left[ \widehat{I}^{(xq)} + \frac{2\rho}{T}\widehat{q}^{(x)} \otimes \widehat{\Delta}^{(q)}\widehat{B}^{(q)} \right] \widehat{\mathcal{R}}^{(xq)} = \widehat{I}^{(xq)} + \frac{2\rho}{T}\widehat{\lambda}^{(x)} \otimes \widehat{\Delta}^{(q)}\widehat{B}^{(q)}. \quad (6.57)$$

---

<sup>5</sup> $\widehat{A}^{(r \times r')} \otimes \widehat{B}^{(s \times s')}$  is generally the matrix  $rs \times r's'$

Recall that there is only one non-zero  $\lambda$ , and therefore the last matrix has one  $q \times q$  block  $(2\rho/T) \widehat{\Delta}^{(q)} \widehat{B}^{(q)}$  in the upper-left corner, it has 1 down this block on the main diagonal, and all other matrix elements are 0.

We are now in a position to simplify the first term of energy (6.19). First, we apply the rule 4 to this energy term, then we note that determinant does not change upon rotation (6.57), while the determinant of the right hand side of (6.57) is trivially computed, yielding

$$\ln \det \left[ \widehat{I}^{(qn)} + \frac{2\rho}{T} \widehat{q}^{(n)} \otimes \widehat{\Delta}^{(q)} \widehat{B}^{(q)} \right] = \frac{n}{x} \ln \det \left[ \widehat{I}^{(q)} + \frac{2\rho x}{T} \widehat{\Delta}^{(q)} \widehat{B}^{(q)} \right] \quad (6.58)$$

As for the second term in (6.19), we first apply the rule 7 to get

$$\begin{aligned} & \left\langle \widehat{\rho}^{(nq)} \left| \frac{1}{T} \widehat{B}^{(q)} \otimes \widehat{I}^{(n)} \left[ \widehat{I}^{(qn)} + \frac{2\rho}{T} \widehat{q}^{(n)} \otimes \widehat{\Delta}^{(q)} \widehat{B}^{(q)} \right]^{-1} \right| \widehat{\rho}^{(nq)} \right\rangle = \\ & = \frac{n}{x} \left\langle \widehat{\rho}^{(xq)} \left| \left( \frac{1}{T} \widehat{B}^{(q)} \otimes \widehat{I}^{(x)} \right) \left[ \widehat{I}^{(qx)} + \frac{2\rho}{T} \widehat{q}^{(x)} \otimes \widehat{\Delta}^{(q)} \widehat{B}^{(q)} \right]^{-1} \right| \widehat{\rho}^{(xq)} \right\rangle. \end{aligned} \quad (6.59)$$

We then use the rotation (6.57) and note that  $\widehat{B}^{(q)} \otimes \widehat{I}^{(x)}$  and  $\widehat{\mathcal{R}}^{(xq)}$  do commute to each other due to the rule 3. This yields the form

$$\frac{n}{x} \left\langle \widehat{\rho}^{(xq)} \left| \left( \widehat{\mathcal{R}}^{(x)} \otimes \widehat{I}^{(q)} \right)^{-1} \left( \frac{1}{T} \widehat{B}^{(q)} \otimes \widehat{I}^{(x)} \right) \left[ \widehat{I}^{(qx)} + \frac{2\rho}{T} \widehat{\lambda}^{(x)} \otimes \widehat{\Delta}^{(q)} \widehat{B}^{(q)} \right]^{-1} \left( \widehat{\mathcal{R}}^{(x)} \otimes \widehat{I}^{(q)} \right) \right| \widehat{\rho}^{(xq)} \right\rangle. \quad (6.60)$$

We consider, therefore, the rotation of density vector  $|\widehat{\rho}^{(xq)}\rangle$ . First, we note that  $\widehat{\rho}^{(xq)} = \widehat{\rho}^{(x)} \otimes \widehat{p}^{(q)}$ . Second, the density, as the physical quantity, is the same for all replicas and does not depend on replica indices. To write it formally, let us define two  $x$ -dimensional vectors  $\vec{i}^{(x)} = (1 \ 1 \ 1 \ \dots \ 1)$  and  $\vec{j}^{(x)} = (1 \ 0 \ 0 \ \dots \ 0)$ . Then we see by direct implementation of formula (RHeq:orthonormbasis)  $\widehat{\mathcal{R}}^{(x)} |\vec{i}^{(x)}\rangle = \sqrt{x} \vec{j}^{(x)}$ . On the other hand,  $\widehat{\rho}^{(x)} = \rho \vec{i}^{(x)}$ . Therefore, according to the rule 5, we have  $\widehat{\mathcal{R}}^{(xq)} |\widehat{\rho}^{(xq)}\rangle = \rho \sqrt{x} \vec{j}^{(x)} \otimes \widehat{p}^{(q)}$ . This yields the energy term in the form

$$\frac{n}{x} \left\langle \rho \sqrt{x} \vec{j}^{(x)} \otimes \widehat{p}^{(q)} \left| \left( \frac{1}{T} \widehat{B}^{(q)} \otimes \widehat{I}^{(x)} \right) \left[ \widehat{I}^{(qx)} + \frac{2\rho}{T} \widehat{\lambda}^{(x)} \otimes \widehat{\Delta}^{(q)} \widehat{B}^{(q)} \right]^{-1} \right| \rho \sqrt{x} \vec{j}^{(x)} \otimes \widehat{p}^{(q)} \right\rangle \quad (6.61)$$

As  $\vec{j}^{(x)}$  has only one non-zero-component, and  $\hat{\lambda}^{(x)}$  has also only one non-zero matrix element, corresponding to the same direction in vector  $x$ -dimensional space, we have

$$\left[ \hat{I}^{(qx)} + \frac{2\rho}{T} \hat{\lambda}^{(x)} \otimes \hat{\Delta}^{(q)} \hat{B}^{(q)} \right]^{-1} |\vec{j}^{(x)} \otimes \vec{p}^{(q)}\rangle = \vec{j}^{(x)} \otimes \left[ \hat{I}^{(q)} + \hat{\Delta}^{(q)} \hat{B}^{(q)} \right]^{-1} \vec{p}^{(q)} + (\vec{i}^{(x)} - \vec{j}^{(x)}) \otimes \vec{p}^{(q)}. \quad (6.62)$$

The last step is to implement the scalar product rule 7, yielding

$$\frac{n}{x} \rho^2 \langle \vec{p}^{(q)} | \left( \frac{1}{T} \hat{B}^{(q)} \right) \left[ \hat{I}^{(q)} + \frac{2\rho}{T} \hat{\Delta}^{(q)} \hat{B}^{(q)} \right]^{-1} | \vec{p}^{(q)} \rangle. \quad (6.63)$$

Combining (6.58) with (6.63), we arrive at (6.20).



# Chapter 7

## Designed Heteropolymers

Using the formalism of the previous chapter, we are able to examine a heteropolymer chain which consists of an arbitrary set of monomers with short range interactions. We show that phase diagram of polymer chain prepared by Imprinting, besides random and frozen globular phases, also includes a third globular phase, which we call the target phase. The random globule is comprised of a vast number of compact conformations, and although the frozen globule is dominated by one or few conformations, these are not under any control and generally do not possess any desirable properties. On the other hand, the target phase is dominated by the desirable conformation. We discuss crude prescriptions for the experimental realization of the target phase regime.

### 7.1 Introduction

It is well known that the equilibrium conformation of proteins is of paramount importance to its biological activity. The equilibrium conformation for a given protein is determined by the linear sequence of monomers and the interactions between them. The relationship between the heteropolymer sequence and its equilibrium conformation is still a mystery. Furthermore, while order-disorder transitions are a common theme in statistical physics, it is at first sight unclear why the equilib-

rium conformation should consist of one (or very few) conformations. However, there are some physical properties of this system which can immediately yield some qualitative insight. First and foremost, the nature of the polymer is fundamentally different than the behavior, for example, of its disconnect constituent monomers, due to the polymeric bonds. These connections restrict the phase space of monomer arrangements and fundamentally change the physical system by introducing frustrations: the nature of the free energy landscape for the polymer system has many local minima due to the constraints of the polymeric bonds. In other words, the polymeric bonds, in addition to a rich variation of monomer interactions (the *heteropolymeric* properties), differentiates the free energy of the conformations. This differentiation combined with the restricted phase space of monomer arrangements allows the possibility of a unique ground state. Thus, upon reducing the acting temperature on the polymer, we can induce a “freezing” transition, where the equilibrium conformation is dominated by this ground state.

With this physical principle in mind, the freezing transition was first investigated for random chains in terms of phenomenological models [Bry87]. Using the principle of “minimal frustration,” the freezing transition was shown to be similar to that of the Random Energy Model (REM) [Der80]. The REM transition was later derived directly from a microscopic Hamiltonian in which the interactions between each two monomers were assumed to be random independently taken from a gaussian distribution [Sha89a]. However, this model did not explicitly include polymer sequence. The polymer sequence was later directly incorporated into the models [Sfa93, Gar88b, Sfa94]. In fact, a freezing transition was shown to exist for random sequences as long the nature of the interactions were heteropolymeric [Pan95a].

However, all of these works refer only to random sequences. It is believed that protein sequences differ in some degree from random sequences [Bry87, Pan94c]. In Shakhnovich-Gutin model [Sha93b], the “design” of sequences is considered to be performed by evolution. Specifically, mutations cause changes in the heteropolymer sequence. Assuming that the fitness is related to the folding properties and therefore the polymer energy, evolution should lead to sequences which have sequences

annealed to minimize the energy when in a particular target conformation  $\star$ .

Recently, we proposed the “Imprinting” model, which is a method to *in vitro* create sequences which can renature and recognize a given target molecule [Pan94b,Pan94d,Pan94e]. As shown in Figure 7-1, the general scheme is to allow the monomers to equilibrate in space prior to polymerization and then polymerize the monomers in such a way that the monomers equilibrium positions remain unchanged. We expect that the minimization of energy of the monomer solution should lead to the minimization of the polymer energy, and therefore, the polymer would renature to the polymerization conformation, as shown in Figure 7-2. If a particular target molecule is placed in the monomer solution, then the monomers should also equilibrate around the target molecule and the resulting polymer should have a complementary site capable of recognizing the particular target molecule. In this sense, an imprinted heteropolymer should be able to specifically recognize the target molecule, thus acting much like an artificial antibody.

In fact, to the level of mean field, these two models are indistinguishable: in both cases, the polymer sequence resulting from the design procedure folds to a particular target conformation  $\star$ . In this work, we concentrate on Imprinting, but we note that the general formalism derived here is applicable to the Shakhnovich-Gutin model as well.

Since this chapter combines the formalism developed in the previous two chapters, to avoid redundancy, we will omit many of the steps involved and quickly move to the results.

## 7.2 Development of the Model

### 7.2.1 Disordered Short-Range Two-Body Interactions

In this system, there are two fixed quantities: the heteropolymer sequence and the nature of interaction between monomer species. We model both aspects explicitly. We calculate the energy of interaction between two monomers based upon the energy

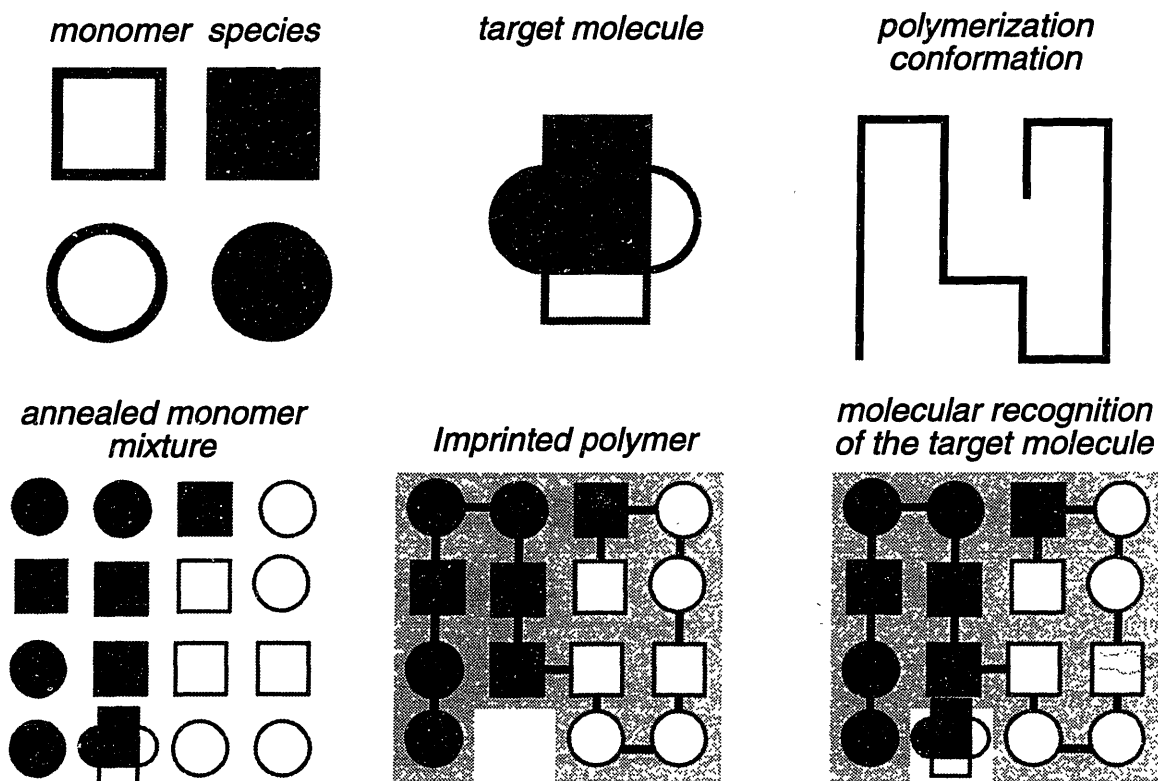


Figure 7-1: The Imprinting process. Even though our analytic treatment is general, we schematically depict Imprinting an example of Imprinting in two dimensions, for monomers which interact as in the  $p$ -charge model (there are energetic preferences toward neighbors which have the same shape (square versus circle) and color (black versus white)). We include a target molecule which allow to interact differently on each side, in this case with the four sides representing all possible monomer species. We place this target molecule in the presence of monomers prior to polymerization and allow this “monomer soup” to equilibrate, leading to an annealed monomer mixture. We model polymerization by choosing some conformation randomly and threading the monomers in the soup along the path of the polymerization conformation in order to define the sequence of the Imprinted polymer. The optimization of the monomer arrangement in the monomer soup leads to an Imprinted polymer which can renature to the polymerization conformation. Furthermore, the polymerization conformation includes a pocket, or “active site,” allowing specific complementary interactions with respect to the target molecule. Thus, Imprinted heteropolymers have the protein-like properties of renaturability and specific molecular recognition.



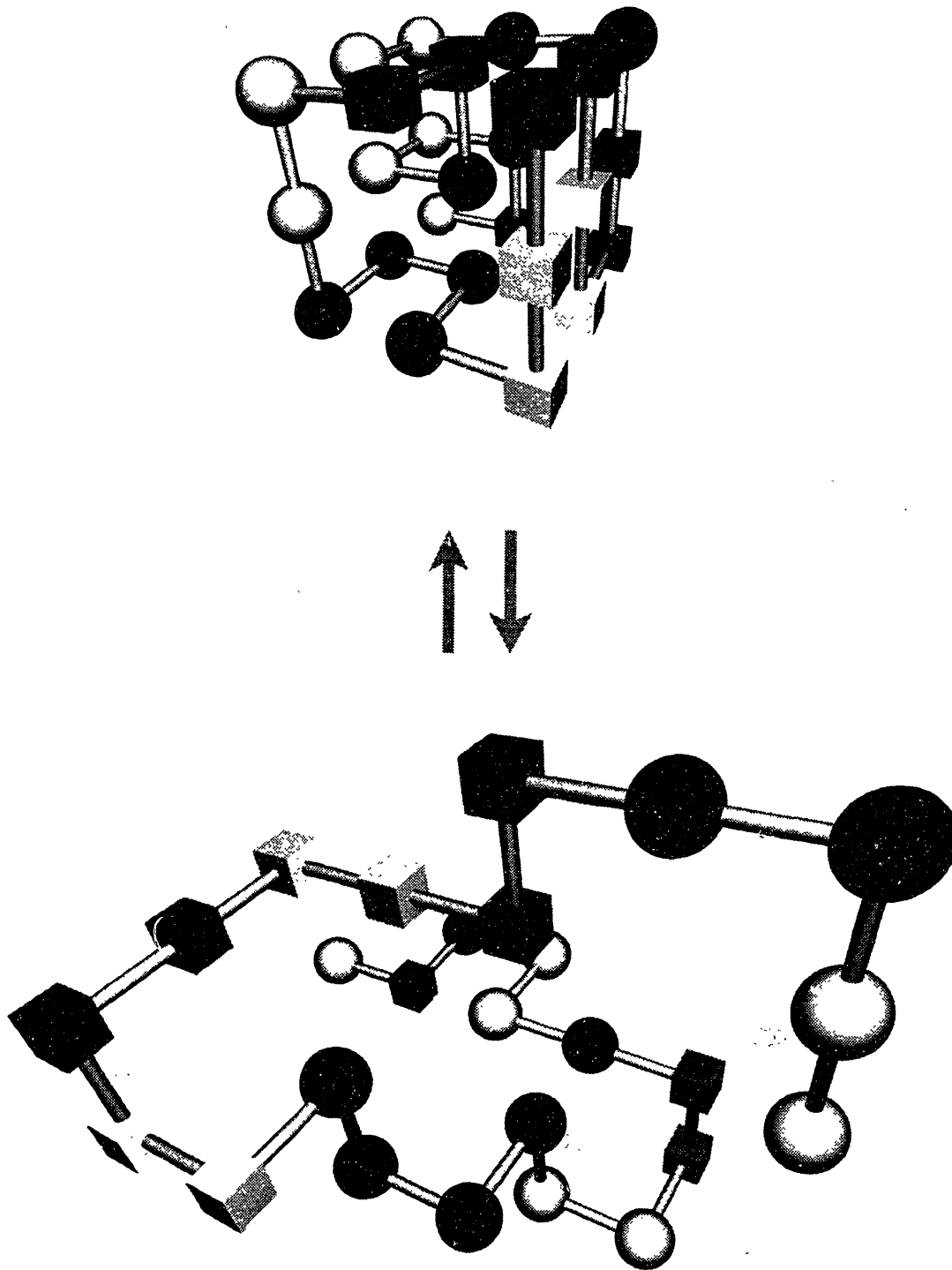


Figure 7-2: Renaturation of an Imprinted heteropolymer. As an example, we employ  $p$ -charge interactions as in Figure 1, except now in three dimensions. The optimization of the spatial arrangement of the monomers prior to polymerization selects sequences which have a low energy when the polymer is arranged in the polymerization conformation. This leads to the ability of the polymer, after denaturation, to successfully fold to the polymerization conformation (renaturation).

of interactions of the respective species of monomers and whether they are in the proximity for interaction:

$$\mathcal{H} = \sum_{i,j}^q \sum_{I,J}^N B_{ij} \delta(\mathbf{r}_I - \mathbf{r}_J) \delta(s_I, i) \delta(s_J, j) + \mathcal{H}' \quad (7.1)$$

where  $B_{ij}$  is the interaction energy between monomer *species*  $i$  and  $j$ ,  $s_I$  is the species of monomer at position  $I$  along the chain, and  $\mathbf{r}_I$  is the position of monomer  $I$ . In this chapter, we will use the following notation: upper case Roman letters relate to monomer numbers along the chain, lower case Roman letters relate to monomer species, and lower case Greek letters denote replica indices.  $\mathcal{H}' = C\rho^2 + \dots$  is the excluded volume virial expansion. This term is purely homopolymeric in nature. Thus, we assume that heterogeneity solely comes from pairwise interactions and all high order interactions contribute primarily to the excluded volume effect.

Note that Hamiltonian (7.1) depends on both conformation (through the set of monomer coordinates  $\mathbf{r}_I$ ), and the sequence (through  $s_I$ ). We formally express that by writing  $\mathcal{H} = \mathcal{H}(\text{conf}; \text{seq})$ .

## 7.2.2 Self-Averaging over the Sequences

As we take a statistical approach, we can only analyze properties of the ensemble, in this case — the ensemble of designed sequences. In each realization, the sequence is fixed, or quenched. Thus, in principle, each particular chain is characterized with the sequence-dependent free energy

$$F(\text{seq}) = -T \ln Z(\text{seq}) . \quad (7.2)$$

In fact, however, free energy is believed to be a self-averaging value, which means that free energy almost does not depend on realization of the sequence, given that composition is fixed and overall length is long enough; therefore the sequence-dependent free energy is in practice almost sequence-independent and, therefore, coincides practically with the mean free energy averaged over the ensemble of se-

quences:

$$F(\text{seq}) \simeq F^* \equiv \langle F(\text{seq}) \rangle_{\text{seq}}^* = -T \langle \ln Z(\text{seq}) \rangle_{\text{seq}}^* . \quad (7.3)$$

Leaving aside for a moment the difficult technical question how to average  $\ln Z$ , let us discuss first the logical aspect: we have to average over the set of all possible  $q^N$  sequences, with a weighting based upon the probability of each particular sequence to appear in the ensemble of designed sequences. To specify this ensemble, recall that our polymerization scheme implies two steps: (i) prearrangement of the set of disconnected monomers in space, governed by the same monomer-to-monomer interactions involved in the Hamiltonian (7.1), at some polymerization temperature,  $T_p$ ; (ii) formation of strong polymeric bonds between prearranged monomers along some independently chosen backbone,  $\star$ , so that newly prepared chain appears in the conformation  $\star$ . Strictly speaking, we have to consider a new ensemble of designed sequences for each preparation conformation  $\star$ ; this is why we keep the superscript  $\star$  throughout the equation (7.3). Doing so, we average as  $\langle \dots \rangle_{\text{seq}}^* = \sum_{\text{seq}} \dots P_{\text{seq}}^*$ , and we identify probability distribution  $P_{\text{seq}}^*$  with the Boltzman weight associated with the Hamiltonian (7.1) at the temperature  $T_p$ . Indeed, in our polymerization procedure, each monomer is assigned with the number along the chain, thus fixing the  $\{s_I\}$  variables; also, as the monomer positions are kept unchanged while polymerizing, vectors  $\mathbf{r}_I$ , related to conformation  $\star$ , are at the same time the coordinates of monomers immediately prior to polymerization. For this reason, on the mean field level the energy of pre-polymerized monomer mixture is indistinguishable from the energy of polymer in the preparation conformation  $\star$ . Therefore,

$$P_{\text{seq}}^* = \exp \left[ -\frac{1}{T_p} \mathcal{H}(\text{conf} = \star; \text{seq}) + \frac{1}{T_p} \sum_{I=1}^N \mu_{s_I} \right] , \quad (7.4)$$

The composition of monomer mixture prior to polymerization is maintained by equilibrium with the surrounding reservoir, where  $\mu_s$  is chemical potential of the component  $s$ . Note that we have not explicitly included any normalization for  $P^*$ . However, any such normalization is just a constant factor on the partition function and is therefore irrelevant.

### 7.2.3 Self-Averaging over Preparation Conformation

In fact, there is a second level of frozen disorder; what is now quenched in the system is the information about preparation conformation  $\star$  used to create the ensemble of sequences. At this level, we can repeat the logic of equation (7.3). Indeed, there are equal grounds to believe in self-averaging with respect to  $\star$ , just as one step before, with respect to sequences. We write therefore

$$F(\text{seq}) \simeq F^\star \simeq F \equiv \left\langle \langle F(\text{seq}) \rangle_{\text{seq}}^\star \right\rangle_\star = -T \langle \ln Z(\text{seq}) \rangle, \quad (7.5)$$

where the last average over both sequence and preparation conformation  $\star$  is performed as

$$\langle \dots \rangle \equiv \sum_\star \sum_{\text{seq}} \dots \exp \left[ -\frac{1}{T_p} \mathcal{H}(\text{conf} = \star; \text{seq}) + \frac{1}{T_p} \sum_{I=1}^N \mu_{s_I} \right]. \quad (7.6)$$

As to the above mentioned question of averaging  $\ln Z$ , we employ the well-known replica trick, in which one solves the simpler problem of averaging  $Z^n$  with positive integer  $n$ , such that all the difficulties appear at the moment of analytic continuation to  $n \rightarrow 0$ :

$$F(\text{seq}) \simeq F = \langle F(\text{seq}) \rangle = -T \langle \ln Z(\text{seq}) \rangle = -T \lim_{n \rightarrow 0} \frac{\langle Z^n(\text{seq}) \rangle - 1}{n}, \quad (7.7)$$

As the appearance of a particular sequence is governed by the same Hamiltonian involved in monomer interactions, we can write the expression for the  $n$ -replica partition function as

$$\langle Z^n(\text{seq}) \rangle = \sum_{\text{seq}} \mathcal{P}_{\text{seq}} \sum_{C_0, C_1, \dots, C_n} \exp \left[ -\sum_{\alpha=0}^n \frac{1}{T_\alpha} \mathcal{H}(C_\alpha, \text{seq}) \right], \quad (7.8)$$

where the following notations are used:  $\alpha = 0, 1, \dots, n$  are the numbers of replicas;  $C_\alpha = C_1, \dots, C_n$  stand for conformations of replica number  $\alpha$ ; replica  $\alpha = 0$  is attributed to the target conformation  $\star$ , that is,  $C_0 = \star$ ;  $T_\alpha = T$  is the acting temperature for  $\alpha \neq 0$ ;  $T_\alpha = T_p$  is the ‘‘polymerization temperature’’ (i.e. the selec-

tive temperature involved in the design procedure) for  $\alpha = 0$ ; and  $\mathcal{P}_{\text{seq}} = \prod_{I=1}^N p_{s_I}$ , where  $p_i = \exp[\mu_i/T_p] [\sum_i \exp(\mu_i/T_p)]^{-1}$  (the fraction of monomers of species  $i$ ). Note, that for brevity we do not write explicitly all of the normalization factors. We will take care of all of them at the very end.

## 7.2.4 Manipulations with Replicas

For further notational simplification, we introduce density distributions of all species for each conformation and replica as

$$m_i^\alpha(\mathbf{R}) = \sum_{I=1}^N \delta(s_I, i) \delta(\mathbf{r}_I^\alpha - \mathbf{R}) ; \quad \{m_i^\alpha(\mathbf{R})\} \equiv \vec{m}^{(q(n+1)\infty)}. \quad (7.9)$$

Then write in terms of those definitions

$$\langle Z^n(\text{seq}) \rangle_{\text{seq}} = \sum_{\text{seq}} \mathcal{P}_{\text{seq}} \sum_{C_0, \dots, C_n} \exp \left\{ - \left\langle \vec{m} \left| \hat{B}^{(q)} \otimes (\hat{T}^{(n+1)})^{-1} \otimes \hat{T}^{(\infty)} \right| \vec{m} \right\rangle^{(q(n+1)\infty)} \right\}, \quad (7.10)$$

where  $\hat{T} = T_\alpha \delta_{\alpha\beta}$ , and  $\langle |\dots| \rangle^{(q(n+1)\infty)}$  means scalar product in which all vectors and operators are supposed to have dimensionality as indicated ( $q(n+1)\infty$  in this case). We use here the operation of direct product  $\otimes$ , in the following sense (identical to what was in [Pan95a]): if there are two matrices (or operators) of different dimensionalities  $r$  and  $s$ , say  $\hat{A}^{(r)}$  and  $\hat{B}^{(s)}$ , then  $\hat{A}^{(r)} \otimes \hat{B}^{(s)}$  is the matrix of dimensionality  $rs$  obtained by mapping of the matrix  $A_{uv} \hat{B}^{(s)}$  onto each matrix element  $(u, v)$  of  $\hat{A}^{(r)}$  matrix. Operator  $\hat{T}^{(\infty)}$  is the identity operator with respect to real coordinate space, meaning that it has the kernel  $\delta(\mathbf{R}_1 - \mathbf{R}_2)$ . Note that for brevity, we have not included the homopolymeric term  $\mathcal{H}'$ ; it does not participate in any heteropolymeric effects and is therefore just a multiplicative constant insofar as the replica heteropolymeric calculations are concerned. We shall take care of this term, along with normalization constants, at the end of calculations.

We now arrive at the formulation which is rather similar to what has been considered in [Pan95a], except the appearance of additional target replica 0 and matrix  $\hat{T}^{(n+1)}$ . We repeat briefly what was done in [Pan95a]. As monomers interact to

each other, the corresponding monomer variables are coupled, which makes averaging over sequences difficult. We trade coupling of monomers for coupling of replicas by introducing fields by the Hubbard-Stratonovich transformation of the form

$$\begin{aligned} \langle Z^n(\text{seq}) \rangle_{\text{seq}} &= \sum_{C_0, \dots, C_n} \int \mathcal{D}\{\phi\} \exp \left\{ \frac{1}{4} \langle \vec{\phi} | (\hat{B}^{-1})^{(q)} \otimes \hat{T}^{(n+1)} \otimes \hat{I}^{(\infty)} | \vec{\phi} \rangle^{(q(n+1)\infty)} \right\} \\ &\times \sum_{\text{seq}} \mathcal{P}_{\text{seq}} \exp \left\{ \langle \vec{\phi} | \vec{m} \rangle^{(q(n+1)\infty)} \right\}. \end{aligned} \quad (7.11)$$

where  $\{\phi_i^\alpha(\mathbf{R})\} = \vec{\phi}^{(q(n+1)\infty)}$  are the fields conjugated to the corresponding densities. We skip the normalization factor which comes from integration over  $\phi$ ; we will take care of this factor below.

Thus, the sum over sequences involves only uncoupled monomers in the last “source” term of the partition function above. This facilitates the summation over the sequences:

$$\begin{aligned} \exp \{\text{source term}\} &= \sum_{\text{seq}} \mathcal{P}_{\text{seq}} \exp \left\{ \langle \vec{\phi} | \vec{m} \rangle^{(q(n+1)\infty)} \right\} \\ &= \prod_{I=1}^N \sum_{i=1}^q p_i \exp \left\{ \sum_{\alpha=0}^n \int d\mathbf{R} \phi_i^\alpha(\mathbf{R}) \delta(\mathbf{r}_I^\alpha - \mathbf{R}) \right\} \end{aligned} \quad (7.12)$$

The relevant order parameters are extracted by expansion over the powers of the fields  $\phi$  (high temperature expansion) up to  $\mathcal{O}(\phi^2)$  (see the condition of applicability below):

$$\begin{aligned} \text{source term} &= \sum_{i=1}^q \sum_{\alpha=0}^n \int d\mathbf{R} \rho^\alpha(\mathbf{R}) p_i \phi_i^\alpha(\mathbf{R}) \\ &+ \frac{1}{2} \sum_{i,j=1}^q [p_i \delta_{ij} - p_i p_j] \sum_{\alpha,\beta=0}^n \int d\mathbf{R}_1 \int d\mathbf{R}_2 \phi_i^\alpha(\mathbf{R}_1) Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) \phi_j^\beta(\mathbf{R}_2) \end{aligned}$$

where we use the standard definitions [Gar88a, Sha89a, Pan94e]

$$Q_{\alpha_1, \dots, \alpha_k}(\mathbf{R}_1, \dots, \mathbf{R}_k) = \sum_{I=1}^N \prod_{\kappa=1}^k \delta(\mathbf{r}_I^{\alpha_\kappa} - \mathbf{R}_\kappa), \quad (7.14)$$

Note that  $\rho^\alpha(\mathbf{R})$ , which in equilibrium does not depend on replica number and which

within the large globule does not depend on  $\mathbf{R}$  either, is the total density of the polymer chain. Following the standard interpretation, replicas are associated with the pure states of the polymer chain [Mez84,Sha89a,Pan94e]. The  $k$ -replica order parameter  $Q_{\alpha_1,\dots,\alpha_k}$  is interpreted as the overlap between replicas  $\alpha_1, \dots, \alpha_k$ . Therefore, a transition to unique structure corresponds to the equilibrium configuration where all replicas overlap, e.g.  $Q_{\alpha\beta} = \rho$ .

Using the definition of the overlap order parameter, we can write the  $(n+1)$ -replica partition function in a simple form:

$$\begin{aligned} \langle Z^n(\text{seq}) \rangle_{\text{seq}} &= \sum_{C_0, \dots, C_n} \int \mathcal{D}\{\phi\} \\ &\times \exp \left\{ \langle \vec{\phi} | \frac{1}{4} B_{ij}^{-1} \delta(\mathbf{R}_1 - \mathbf{R}_2) \mathcal{T}_{\alpha\beta} + \frac{1}{2} Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) \Delta_{ij} | \vec{\phi} \rangle^{(q(n+1)\infty)} + \langle \vec{\rho} | \vec{\phi} \rangle^{(q(n+1)\infty)} \right\} \end{aligned} \quad (7.15)$$

where

$$\Delta_{ij} = p_i \delta_{ij} - p_i p_j \quad \text{and} \quad \vec{\rho}^{(q(n+1)\infty)} \equiv \rho_i^\alpha(\mathbf{R}) = p_i \sum_{I=1}^N \delta(\mathbf{r}_I^\alpha - \mathbf{R}) . \quad (7.16)$$

Note that  $\vec{\rho}^{(q(n+1)\infty)} = \vec{\rho}^{(n+1)\infty} \otimes \vec{p}^{(q)}$ . We are left with the Gaussian integral (7.15) for the  $(n+1)$ -replica partition function, which can, of course, be evaluated. The result, however, is remarkably simplified by the argument given in [Sha89a,Sfa93,Pan94e], which shows that the  $\mathbf{R}$ -dependence of  $Q_{\alpha\beta}$  is of  $\delta$ -type, so that

$$Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) = \rho q_{\alpha\beta} \delta(\mathbf{R}_1 - \mathbf{R}_2) , \quad (7.17)$$

where all the diagonal elements of the new matrix  $\hat{q}^{(n+1)}$  are 1, while its off diagonal elements are either 1 or 0. This means physically that two replicas  $\alpha$  and  $\beta$  might be either uncorrelated (independent), so that  $Q_{\alpha\beta} = 0$ , or they may be correlated so that one repeats the 3D fold of the other down to the microscopic length scale, so that  $Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) = \rho \delta(\mathbf{R}_1 - \mathbf{R}_2)$ . We do not repeat this argument here, as it is explained elsewhere (see the argument presented in [Pan94e] which is slightly different from the original one [Sha89a]).

Now that the  $\widehat{Q}$  matrix has been simplified, we perform Gaussian integration over all  $\phi^\alpha$  variables. At this moment, it is very important to take care of normalizing factors everywhere. We find that the normalization constants generated by the Hubbard-Stratonovich transformation are canceled by the Gaussian integration, yielding

$$\langle Z^n(\text{seq}) \rangle_{\text{seq}} = \sum_{C_0, \dots, C_n} \exp[-NE\{Q\}] \quad (7.18)$$

with the energy per one particle of the form

$$E = \left\langle \bar{\rho}^{(n+1)q} \left| (\widehat{T}^{(n+1)})^{-1} \otimes \widehat{B}^{(q)} \left[ \widehat{T}^{(q(n+1))} + 2\rho\widehat{q}^{(n+1)}(\widehat{T}^{(n+1)})^{-1} \otimes \widehat{\Delta}^{(q)}\widehat{B}^{(q)} \right]^{-1} \right| \bar{\rho}^{(n+1)q} \right\rangle + \frac{1}{2} \ln \det \left[ \widehat{T}^{(q(n+1))} + 2\rho\widehat{q}^{(n+1)}(\widehat{T}^{(n+1)})^{-1} \otimes \widehat{\Delta}^{(q)}\widehat{B}^{(q)} \right] \quad (7.19)$$

where the rule  $\det(\widehat{A}) \cdot \det(\widehat{B}) = \det(\widehat{A}\widehat{B})$  has been used for simplification. The simplest way to deal with normalization is to incorporate some additive constant to the expression for energy (7.19) such that it yields the desirable zero level in some trivial case, for example, in the case of homopolymer, when  $B_{ij} = B$  does not depend on monomer species. This is what we have done in eq. (7.19). Indeed, for the homopolymer case equation (7.10) yields simply the second virial term  $\rho B [n/T + 1/T_p]$ . Exactly so does the equation (7.19), because, as is easy to check,  $\widehat{\Delta}^{(q)}\widehat{B}^{(q)} = 0$  in the homopolymer extreme.

Recall that the elements of the  $\widehat{q}$  matrix are either  $q_{\alpha\beta} = 0$  or 1, which corresponds to no or complete overlap between the conformations of replicas  $\alpha$  and  $\beta$ , respectively. This implies that a single step Parisi replica symmetry breaking (RSB) ansatz is appropriate [Par80]. In fact, by re-arranging the matrix by row and column operations (under which the free energy is invariant), one can write the  $q$  matrix in terms of groups of overlapping replicas. Furthermore, the free energy can be calculated as a function of the nature of the grouping as well. Optimization of the free energy with respect to the nature of the grouping of replicas in fact yields the single step Parisi RSB scheme. For the sake of simplicity, however, we omit the derivation, thus, formally employing the ansatz. Specifically, we assume that the  $q$



matrix is composed of one  $(y + 1) \times (y + 1)$  “target block,” i.e. a group of  $y$  replicas which overlap with the  $\star$  conformation, and  $(n - y)/x$  groups in which  $x$  replicas overlap. (See [Pan94e] for more details on this point.)

## 7.2.5 Free Energy of Replica System

With this form of  $\hat{q}$  matrix in mind, we can apply matrix operations detailed in the Appendix to simplify the energy to

$$E = \frac{1}{2} \ln \det \left[ \hat{I} + \left( \frac{y}{T} + \frac{1}{T_p} \right) 2\rho \widehat{\Delta} \widehat{B} \right] + \left( \frac{y}{T} + \frac{1}{T_p} \right) \left\langle \vec{p} \left| \rho \widehat{B} \left[ \hat{I} + \left( \frac{y}{T} + \frac{1}{T_p} \right) 2\rho \widehat{\Delta} \widehat{B} \right]^{-1} \right| \vec{p} \right\rangle \\ + \frac{n-y}{2x} \ln \det \left[ \hat{I} + x \frac{2\rho}{T} \widehat{\Delta} \widehat{B} \right] + (n-y) \left\langle \vec{p} \left| \frac{\rho}{T} \widehat{B} \left[ \hat{I} + \frac{2\rho x}{T} \widehat{\Delta} \widehat{B} \right]^{-1} \right| \vec{p} \right\rangle \quad (7.20)$$

We have summed through the replica space and now only species space remains. This is why we have dropped the labels for the dimensionality of the vectors and operators, which are now assumed to be in species space, i.e. dimensionality  $q$ .

To comment on the equation (7.20), which is technically the most important result of the work, we note first, that it looks rather similar to what was found in [Pan95a] for random sequences. Recall that the expression (7.20) represents the energy of interaction between replicas. Replicas of different groups do not interact to each other — this is the sense of grouping. The total energy is, therefore, the sum of each independent group’s contributions. We have found in [Pan95a], that the group of  $x$  replicas has the energy

$$\frac{1}{2} \ln \det \left[ \hat{I} + \frac{2\rho x}{T} \widehat{\Delta} \widehat{B} \right] + \left\langle \vec{p} \left| \frac{\rho x}{T} \widehat{B} \left[ \hat{I} + \frac{2\rho x}{T} \widehat{\Delta} \widehat{B} \right]^{-1} \right| \vec{p} \right\rangle . \quad (7.21)$$

This is what equation (7.20) gives for  $(n - y)/x$  groups with spontaneous replica symmetry breaking. As to the target group of  $y + 1$  replicas, its energy is also almost of the same form, except  $y/T + 1/T_p$  appears instead of  $x/T$ , because one replica has in a sense different temperature.

Now, all interesting heteropolymeric properties are seen in the nature of the

phase transition of the new order parameters  $x$  and  $y$ .

We must now estimate the entropy due to the overlap of replicas. To this end, we repeat the arguments of the works [Sha89a,Sfa93,Pan94e]. The density of states per monomer is given by the ratio of the available volume to place additional monomers  $a^3$ , where  $a$  is the spacing between monomers, divided by the excluded volume per monomer  $v$ . For every additional replica in a group, we must consider this entropic reduction, and therefore the entropy per monomer for  $w$  replicas in a single group is therefore  $\ln(a^3/v)(w-1)$ . For  $(n-y)/x$  groups of size  $x$  and one target group of size  $y$ , the entropy is therefore

$$S = Ns \left[ \frac{n-y}{x}(x-1) + y \right], \quad (7.22)$$

where  $s = \ln(a^3/v)$ .

We thus have the free energy

$$\begin{aligned} F &= \frac{1}{2} \ln \det \left[ \hat{I} + \left( \frac{y}{T} + \frac{1}{T_p} \right) 2\rho \widehat{\Delta} \widehat{B} \right] + \left( \frac{y}{T} + \frac{1}{T_p} \right) \left\langle \vec{p} \left| \rho \widehat{B} \left[ \hat{I} + \left( \frac{y}{T} + \frac{1}{T_p} \right) 2\rho \widehat{\Delta} \widehat{B} \right]^{-1} \right| \vec{p} \right\rangle \\ &+ \frac{n-y}{2x} \ln \det \left[ \hat{I} + x \frac{2\rho}{T} \widehat{\Delta} \widehat{B} \right] + (n-y) \left\langle \vec{p} \left| \frac{\rho}{T} \widehat{B} \left[ \hat{I} + \frac{2\rho x}{T} \widehat{\Delta} \widehat{B} \right]^{-1} \right| \vec{p} \right\rangle + s \left[ \frac{n-y}{x}(x-1) + y \right] \end{aligned}$$

Recall that  $x$  is the number of replicas in a non-target group. Therefore  $x$  varies from  $x=1$ , the replica symmetric case, to  $x=n$ , where all of the replicas are in the same (non-target) group. Thus, we first optimize the free energy with respect to  $x$ . As  $x$  must be in between of 1 and  $n$ , optimization yields

$$x = \begin{cases} \xi_f T / 2\rho & \text{for } \xi_f T / 2\rho \leq 1 \\ 1 & \text{for } \xi_f T / 2\rho > 1 \end{cases}, \quad (7.24)$$

where  $\xi_f$  is the solution of the equation

$$2s = \ln \det \left( \hat{I} + \xi_f \widehat{\Delta} \widehat{B} \right) - \text{Tr} \left[ \xi_f \widehat{\Delta} \widehat{B} \left( \hat{I} + \xi_f \widehat{\Delta} \widehat{B} \right)^{-1} \right] + \left\langle \vec{p} \left| \xi_f^2 \widehat{B} \widehat{\Delta} \widehat{B} \left( \hat{I} + \xi_f \widehat{\Delta} \widehat{B} \right)^{-2} \right| \vec{p} \right\rangle \quad (7.25)$$

We find that this result is independent of any design parameters, such as  $y$  or  $T_p$ ,

and is exactly the same as was found in [Pan95a] for random sequences. Physically,  $x$  corresponds to the number of replicas which group due to *spontaneous* symmetry breaking, not the field which draws replicas to the target replica. We interpret that  $x \rightarrow n$  corresponds to maximal freezing of random sequences, while  $x \rightarrow 1$  means transition between frozen to random globular state, where frozen and random globule are characterized with few and vast number of relevant conformations, respectively, even though they are of the same overall density.

As to the other order parameter,  $y$ , it is specifically related to the design procedure, as it represents the number of replicas in the target group (excluding replica 0); thus,  $y$  varies from 0 to  $n$ , when either none or all of the replicas are in the target group. Upon performing the  $n \rightarrow 0$  limit, the interpretation of  $y$  becomes somewhat obscured, as it varies from 0 to  $n = 0$ . It can be shown (see [Pan94e] for more details) that when  $n$  is arbitrarily small but still positive ( $0 < n < 1$ ), there is no optimum of free energy (7.23) with respect to  $y$  within the interval  $0 < y < n$ . This means that the optimum is reached at the boundary of the interval, i.e. either at  $y = 0$  or at  $y = n$ , depending simply on the slope of  $F$  vs  $y$  dependence. Physically, this means, that we predict a first order phase transition from the state with no memory of the target state ( $y = 0$ ) to the other state with strongly memorized target conformation ( $y = n$ ). Therefore, the threshold between these two phases is given by the condition when the slope of the free energy in  $y$  vanishes. This is determined by

$$2s = \ln \det (\widehat{I} + \xi \widehat{\Delta} \widehat{B}) - \text{Tr} \left[ \xi \widehat{\Delta} \widehat{B} (\widehat{I} + \xi_p \widehat{\Delta} \widehat{B})^{-1} \right] \\ + \left\langle \vec{p} \left| -\xi \widehat{B} (\widehat{I} + \xi_p \widehat{\Delta} \widehat{B})^{-1} + \xi_p \xi \widehat{B} \widehat{\Delta} \widehat{B} (\widehat{I} + \xi_p \widehat{\Delta} \widehat{B})^{-2} + \xi \widehat{B} (\widehat{I} + \xi \widehat{\Delta} \widehat{B})^{-1} \right| \vec{p} \right\rangle \quad (7.26)$$

where  $\xi_p = 2\rho/T_p$  and  $\xi = 2x\rho/T$ .

Before passing to the discussion of the results obtained, we comment on the physical meaning of the operator  $\widehat{\Delta}$ , which appears throughout of our formulae. In brief, this operator removes the mean interactions of each species  $i$  from interaction matrix  $B_{ij}$ . Indeed, according to the definition (7.16),  $\widehat{\Delta} \widehat{B} = p_i (B_{ij} - \sum_k p_k B_{ik}) =$

$p_i (B_{ij} - \langle B_{ij} \rangle_j)$ . A more detailed discussion of the  $\widehat{\Delta}$  operator can be found elsewhere [Pan95a].

Also,  $x$  and  $y$  are only coupled to terms involving the  $\widehat{\Delta}$  operator, whereas  $\rho$  is also coupled to  $\langle \vec{p} | \widehat{B} | \vec{p} \rangle$ . Thus, for  $B_{ij} = b_{ij} + B_0$  and  $|B_0| \gg |b_{ij}|$ , we can optimize the free energy with respect to density  $\rho$  independently of optimization with respect to  $x$  and  $y$ . As for optimization with respect to  $\rho$ , this of course includes the homopolymeric terms in  $\mathcal{H}'$ . Physically, density is balanced due to the competition between the attractive two body interactions described by the average second virial coefficients and the primarily repulsive three and higher body interactions described by  $\mathcal{H}'$ . This is simply a homopolymeric effect. A constant density can be realized by a large ensemble of globular conformations, the heteropolymeric effect is the selection of one conformation from this ensemble; this is described by the  $x$  and  $y$  order parameters and will be systematically described below.

## 7.3 Discussion

### 7.3.1 Phase Diagram

To summarize the findings of the previous section, we have shown that there are three macroscopic phases in the globule of designed heteropolymer. One is called a “random globule,” as it is comprised with vast number of conformations; the second phase is called “frozen” as the chain freezes down to a few relevant conformations, but the choice of those conformations remains out of control of the design procedure; and the third phase is called “target,” as in this phase, the chain undergoes freezing to target conformation  $\star$ . These findings are organized in the form of phase diagram, Fig. 1, in the variables acting temperature  $T$  vs polymerization temperature  $T_p$ . We stress, that every vertical line on the diagram represents another physical sample of heteropolymers, which has been prepared at the given temperature  $T_p$  and is now examined at different temperatures  $T$ . So, vertical motion along the diagram means experimentally simply heating or cooling of the system, while horizontal

motion means passing from one sample to another.

The lines of phase transitions are given by equations (7.25) and (7.26). As to the transition between the random globule and frozen phase, the corresponding temperature does not depend on  $T_p$ , thus being represented by horizontal line on the phase diagram; the transition temperature is given by  $T_f = 2\rho/\xi_f$ , where  $\xi_f$  is the solution of equation (7.25). As to the transition between the target state and any other state, the corresponding conditions are given by the equation (7.26). This equation should be treated independently for  $T > T_f$  and  $T < T_f$  regions, as it contains  $x$ -dependence (through  $\xi = 2\rho x/T$ ), and  $x = 1$  at  $T > T_f$  and  $x = T/T_f$  at  $T < T_f$  (see [Pan94e]). Combining (7.25) and (7.26), we have

$$\begin{aligned} & \left\langle \vec{p} \left| \xi \widehat{B} (\widehat{I} + \xi_p \widehat{\Delta} \widehat{B})^{-1} - \xi \widehat{B} (\widehat{I} + \xi \widehat{\Delta} \widehat{B})^{-1} \right| \vec{p} \right\rangle + \\ & + \left\langle \vec{p} \left| \xi_f^2 \widehat{B} \widehat{\Delta} \widehat{B} (\widehat{I} + \xi_f \widehat{\Delta} \widehat{B})^{-2} - \xi_p \xi \widehat{B} \widehat{\Delta} \widehat{B} (\widehat{I} + \xi_p \widehat{\Delta} \widehat{B})^{-2} \right| \vec{p} \right\rangle = \\ & \text{Tr} \left\{ \xi_f \widehat{\Delta} \widehat{B} (\widehat{I} + \xi_f \widehat{\Delta} \widehat{B})^{-1} - \xi \widehat{\Delta} \widehat{B} (\widehat{I} + \xi_p \widehat{\Delta} \widehat{B})^{-1} + \ln \left[ (\widehat{I} + \xi \widehat{\Delta} \widehat{B}) (\widehat{I} + \xi_f \widehat{\Delta} \widehat{B})^{-1} \right] \right\} \end{aligned}$$

at  $T > T_f$ , where

$$\xi = 2\rho/T ; \quad \xi_f = 2\rho/T_f ; \quad \xi_p = 2\rho/T_p , \quad (7.28)$$

and

$$\begin{aligned} & \left\langle \vec{p} \left| \xi_f \widehat{B} (\widehat{I} + \xi_p \widehat{\Delta} \widehat{B})^{-1} - \xi_f \widehat{B} (\widehat{I} + \xi \widehat{\Delta} \widehat{B})^{-1} \right| \vec{p} \right\rangle + \\ & + \left\langle \vec{p} \left| \xi_f^2 \widehat{B} \widehat{\Delta} \widehat{B} (\widehat{I} + \xi_f \widehat{\Delta} \widehat{B})^{-2} - \xi_p \xi_f \widehat{B} \widehat{\Delta} \widehat{B} (\widehat{I} + \xi_p \widehat{\Delta} \widehat{B})^{-2} \right| \vec{p} \right\rangle = \\ & \text{Tr} \left\{ \xi_f \widehat{\Delta} \widehat{B} (\widehat{I} + \xi_f \widehat{\Delta} \widehat{B})^{-1} - \xi_f \widehat{\Delta} \widehat{B} (\widehat{I} + \xi_p \widehat{\Delta} \widehat{B})^{-1} \right\} \end{aligned} \quad (7.29)$$

at  $T < T_f$ .

The last equation (7.29) has the obvious general solution  $\xi_p = \xi_f$ , or  $T_p = T_f$ . Physically, this means that the line of transition between frozen and target states is vertical on the phase diagram. In other words, this transition cannot be caused by acting temperature change. This is perfectly clear, because both frozen and target

states are comprised of one or few conformations, thus having negligible entropies, and therefore their interconversion cannot be temperature-controlled.

The only part of phase diagram remaining to be clarified is random globule to target phase transition at  $T > T_f$ . We call the temperature of the transition  $T_{\text{tar}}$  and the transition line on the phase diagram is therefore  $T_{\text{tar}}(T_p)$ . This is described by equation (7.27). In fact, the line  $T_{\text{tar}}(T_p)$  is the most important issue of this work, because we will speculate below that the region immediately under this line on the phase diagram (where  $T_f < T < T_{\text{tar}}$  and  $T_p < T_f$ ) is the most promising for experimental realization of the target state. Thus, the elucidation of this region is very important. As equation (7.27) is quite cumbersome, we address some specific cases.

### 7.3.2 An Exactly Solvable Model: The Generalized Potts Model

The  $Q$ -Potts model of interactions assumes  $Q$  types of monomer species, with interaction energy between similar and different monomers of  $b$  and  $0$ , respectively. On the other hand, the  $p$ -charge model, suggested in [Gar88b] and studied in [Sfa94,Pan95a], models the presence of  $p$  different physical short range interactions (an abstraction of Coulomb, van-der-Waals, hydrophobic, etc. interactions in real chemical systems). Each monomer is depicted in this model with a set of  $p$  generalized “charges,” each taking one of two possible values, say  $0$  or  $1$ .

We introduce a generalized Potts model, which generalizes both the  $Q$ -Potts and  $p$ -charge models. In this model, each monomer has  $p$  different charges,  $s^1, \dots, s^k, \dots, s^p$ , and we allow the values of each charge  $s^k$  to range from  $0$  to  $Q - 1$ . Furthermore, we define the interaction between charges of monomers  $I$  and  $J$  to be of Potts form: if the value of the charges  $s_I^k$  and  $s_J^k$  are the same, then the interaction energy is  $b_k$ , otherwise it is zero. The total interaction energy between monomers  $I$  and  $J$  is given as the sum of the interaction energy of the charges of the monomers.

The Hamiltonian is therefore defined to be

$$\mathcal{H} = \sum_{I,J}^N \delta(\mathbf{r}_I - \mathbf{r}_J) \sum_{k=1}^p b_k \delta(s_I^k, s_J^k) + \mathcal{H}' \quad (7.30)$$

where  $s_I^k$  is the value of the charge  $k$  of monomer  $I$ . In the interaction matrix, we define each possible combination of charges as a different species. Thus, there are  $q = \prod_{k=1}^p Q_k$  species in the interaction matrix. For species number  $i$  ( $1 \leq i \leq q$ ), the value of charge  $k$  ( $0 \leq k < p$ ) is given by  $s^k(i) = \lfloor i/(Q_k)^k \rfloor \bmod Q_k$ , where  $\lfloor \dots \rfloor$  means truncate to the lowest integer and  $a \bmod b = a - b \lfloor a/b \rfloor$ .

Thus, we have an interaction matrix of the form

$$\hat{B}_{ij} = \sum_k b_k \delta\left(\left\lfloor \frac{i}{2^k} \right\rfloor \bmod Q_k, \left\lfloor \frac{j}{2^k} \right\rfloor \bmod Q_k\right) \quad (7.31)$$

Note that the  $q$ -Potts model is recovered for  $p = 1$ , the  $p$ -charge model is recovered for  $Q_k = 2$ , and the Ising model is recovered for  $p = 1$  and  $Q_k = 2$ .

For simplicity, we consider the case of an even population of all monomer species, i.e.  $p_i = 1/q$ . In this case, the  $\widehat{\Delta\hat{B}}$  matrix has a  $(q - p - \sum_k^p Q_k)$ -degenerate eigenvalue of 0 and for each  $k \in 0, 1, \dots, p-1$ , a  $(Q_k - 1)$ -degenerate eigenvalue of  $b_k/Q_k$ . The energy terms of (7.23) involving determinants can be simplified by

$$\ln \det \left( \hat{I} + \frac{2\rho a}{T} \widehat{\Delta\hat{B}} \right) = \sum_k^p (Q_k - 1) \ln \left( 1 + \frac{2b_k \rho a}{Q_k T} \right) \quad (7.32)$$

where  $a$  is some constant (either  $x$  or  $y + T/T_p$ ). The other energy terms are drastically simplified since

$$\left\langle \vec{p} \left| \frac{\rho}{T} \hat{B} \left( \hat{I} + \frac{2\rho a}{T} \widehat{\Delta\hat{B}} \right)^{-1} \right| \vec{p} \right\rangle = \left\langle \vec{p} \left| \frac{\rho}{T} \hat{B} \right| \vec{p} \right\rangle \quad (7.33)$$

Thus, the freezing temperature is determined by

$$\sum_k^p \left[ \ln \left( 1 + \frac{2b_k \rho}{T_f} \right) - \frac{2b_k \rho / T_f}{1 + 2b_k \rho / T_f} - \frac{2s}{p(Q_k - 1)} \right] = 0 \quad (7.34)$$

For the specific case  $b_k = b$  and  $Q_k = Q$ , we have

$$T_f = -\frac{2\rho b}{Q\Xi[2s/p(Q-1)]}, \quad (7.35)$$

where  $\Xi(\sigma)$  is given self-consistently by

$$\Xi(\sigma) : \quad \sigma = \ln(1 - \Xi) + \Xi/(1 - \Xi) \simeq \begin{cases} \Xi^2/2 & \text{for } \Xi \ll 1 \\ 1/(1 - \Xi) & \text{for } \Xi \rightarrow 1 \end{cases}. \quad (7.36)$$

The freezing temperature for the two asymptotics in flexibility are

$$T_f \simeq \begin{cases} -(\rho b/Q)\sqrt{p(Q-1)/s} & \text{for effectively flexible chain, } s/p(Q-1) \ll 1 \\ -(2\rho b/Q)[1 + (Q-1)p/2s] & \text{for effectively stiff chain, } s/p(Q-1) \gg 1 \end{cases}. \quad (7.37)$$

Note that the validity of these asymptotics is determined by the *effective* flexibility  $\sigma = 2s/p(Q-1)$ . In particular, the regime which we call the “flexible chain limit” (first line of (7.37)) is valid for even rather stiff polymers (eg.  $s > 1$ ) if it has sufficient diversity of monomer species ( $pQ \gg 1$ ). Also note that in the limits  $Q_k = 2$  and  $p = 1$  we recover the results previously derived for the  $p$ -charge and  $Q$ -Potts models [Pan95a].

For the target transition, we have the relation

$$\sum_k^p \left[ \sigma^{(k)} - \frac{\Xi_t^{(k)}}{1 - \Xi_p^{(k)}} - \ln(1 - \Xi_t^{(k)}) \right] = 0 \quad (7.38)$$

where  $\sigma^{(k)} = 2s/p(Q_k - 1)$ ,  $\Xi_t^{(k)} = -2b_k\rho/Q_k T_{\text{tar}}$ , and  $\Xi_p^{(k)} = -2b_k\rho/Q_k T_p$ . For the specific case  $b_k = b$  and  $Q_k = Q$ , we have

$$T_p = \begin{cases} \frac{2b\rho}{Q} \left[ 1 - \frac{2b\rho}{QT_{\text{tar}}} \left( \ln \left[ \frac{1-2b\rho/QT_f}{1-2b\rho/QT_{\text{tar}}} \right] + \frac{2b\rho/QT_f}{1-2b\rho/QT_f} \right)^{-1} \right]^{-1} & \text{when } T_{\text{tar}} > T_f \\ T_f & \text{otherwise} \end{cases} \quad (7.39)$$

This equation is of the same form as was derived in [Pan94e] for the 2-letter Potts polymer, except the value of the interaction constant  $B$  (from two letter Hamiltonian



$\mathcal{H}_2 = B \sum_{IJ} s_I s_J \delta[\mathbf{r}_I - \mathbf{r}_J]$  is replaced by  $b/Q$ . Note that there is no  $p$  dependence other than through the freezing temperature.

### 7.3.3 Expansion around the triple point

To find the behavior for an arbitrary interaction matrix, we trade exact solvability for generality by performing an expansion in the vicinity of  $T_{\text{tar}} = T_f$ ,  $T_p = T_f$  point. We find

$$\frac{T_{\text{tar}} - T_f}{T_f} \simeq \sqrt{\frac{T_f - T_p}{T_f}} + \kappa \frac{T_f - T_p}{T_f} \quad (7.40)$$

Note that the first (square-root) term of this expansion is universal, as neither  $\widehat{B}$  nor  $\vec{p}$  enter in it, other than through the value of  $T_f$ . The properties of a particular polymer, such as  $\widehat{B}$  and  $\vec{p}$ , determine the slope  $\kappa$  of the second (linear) term. In general form, the slope  $\kappa$  is given by

$$\kappa = 1 - \frac{4}{3} \xi_f \frac{\text{Tr} \left[ (\widehat{\Delta B})^3 [\widehat{I} + \xi_f \widehat{\Delta B}]^{-3} \right] + 3 \langle \vec{p} | \widehat{B} (\widehat{\Delta B})^2 [\widehat{I} + \xi_f \widehat{\Delta B}]^{-4} | \vec{p} \rangle}{\text{Tr} \left[ (\widehat{\Delta B})^2 [\widehat{I} + \xi_f \widehat{\Delta B}]^{-2} \right] + 2 \langle \vec{p} | \widehat{B} \widehat{\Delta B} [\widehat{I} + \xi_f \widehat{\Delta B}]^{-3} | \vec{p} \rangle} \quad (7.41)$$

This expression is rather cumbersome, but will be simplified below while considering the limiting cases of high and low flexibility.

Clearly, the expansion (7.40) is well applicable close to the triple point, such that  $(T_f - T_p)/T_f < 1/\sqrt{\kappa}$ .

### 7.3.4 Flexible chain limit

As it was shown in [Pan95a], we can expand (7.25) by powers of  $\xi$ , which leads to the equation for the freezing temperature for random sequences in the form

$$2s = \sum_{k=2}^{\infty} \xi_f^k \langle B^k \rangle_m, \quad (7.42)$$

where “moments” of the matrix are defined as

$$\langle B^k \rangle_m = \frac{k-1}{k} \text{Tr} \left[ (-\widehat{\Delta} \widehat{B})^k \right] - (k-1) \langle \vec{p} | \widehat{B} (-\widehat{\Delta} \widehat{B})^{k-1} | \vec{p} \rangle . \quad (7.43)$$

Recall that these moments do not depend on any constant (homopolymeric) contributions to the interactions: the moments are the same for  $B_{ij}$  and any  $B_{ij} + B_0$ . In particular, we can subtract the mean interaction defining

$$b_{ij} \equiv B_{ij} - \langle B \rangle , \quad \langle B \rangle = \sum_{ij} p_i p_j B_{ij} , \quad (7.44)$$

and write definition (7.43) in terms of  $\widehat{b}$  by simply substituting  $\widehat{b}$  instead of  $\widehat{B}$ .

If chain is flexible and  $s$  is small enough, we can neglect all the terms but first one, yielding [Pan95a]

$$T_f = \rho \sqrt{\frac{2}{s} \langle B^2 \rangle_m} ; \quad \langle B^2 \rangle_m \equiv \sum_{ij} p_i p_j b_{ij}^2 \equiv \langle (B - \langle B \rangle)^2 \rangle . \quad (7.45)$$

Note that  $\langle B^2 \rangle_m$  is simply the variance of the *elements* of the  $\widehat{B}$  matrix, irrespective of their position in the matrix, and thus  $T_f$  for flexible chains is defined mainly by the overall heterogeneity of the interaction matrix [Pan95a].

If we apply the same expansion for the target phase transition (7.27), we get

$$\sum_{k=2}^{\infty} \langle B^k \rangle_m \xi_f^k \left[ 1 + \frac{1}{k-1} \left( \frac{\xi}{\xi_f} \right)^k - \frac{k}{k-1} \left( \frac{\xi}{\xi_f} \right) \left( \frac{\xi_p}{\xi_f} \right)^{k-1} \right] = 0 . \quad (7.46)$$

The simplest approximation for flexible chain, similar to eq. (7.45), means truncation of the series to the first non-vanishing term:

$$\xi_f^2 + \xi^2 - 2\xi\xi_p = 0 \quad \text{or} \quad \frac{T_{\text{tar}}}{T_f} = \frac{T_f}{T_p} \left[ 1 + \sqrt{1 - \left( \frac{T_p}{T_f} \right)^2} \right] \quad (7.47)$$

Note, that in this approximation neither of the properties of particular polymer, such as  $\widehat{B}$  and  $\vec{p}$ , enter to the shape of transition line (7.47), except for the freezing

temperature,  $T_f$ , and, therefore, except for overall heterogeneity of interactions.

More delicate properties of interactions are seen to become important for not so flexible chains, i.e., in the next approximation with respect to  $s$ . In particular, already to second to lowest order, we find freezing temperature in the form

$$T_f = \rho \sqrt{\frac{2}{s} \langle B^2 \rangle_m} \left[ 1 + \sqrt{\frac{s}{2}} \frac{\langle B^3 \rangle_m}{\langle B^2 \rangle_m^{3/2}} \right] \quad (7.48)$$

and, for the target transition in the vicinity of triple point, we get (7.40) with the slope

$$\kappa = 1 + \sqrt{2s} \frac{\langle B^3 \rangle_m}{\langle B^2 \rangle_m^{3/2}}, \quad (7.49)$$

where third moment of interaction matrix

$$\langle B^3 \rangle_m \equiv \sum_{ijk} p_i b_{ij} p_j b_{jk} p_k b_{ki}, \quad (7.50)$$

unlike the second one, is determined by matrix arrangement of the elements  $B_{ij}$  and not only by their overall heterogeneity. In other words, correlations become important between interaction energies of given monomer species to different other species.

### 7.3.5 Stiff Chain Limit

For stiff chains, when  $s$  is large, the main contribution in (7.25) comes from divergence of  $[\hat{I} + \xi_f \widehat{\Delta} \widehat{B}]^{-1}$  term, which is governed by largest eigenvalue of  $(-\widehat{\Delta} \widehat{B})$  operator. We call this eigenvalue and the corresponding eigenvector  $\lambda$  and  $|\vec{\psi}\rangle$ , respectively. It was shown in [Pan95a], that in this case the freezing transition temperature is controlled by the most attractive “mixture” of monomers (represented by  $|\vec{\psi}\rangle$ ), where “mixing” is understood in the sense similar to quantum mechanics. This transforms (7.25) into

$$2s \simeq \frac{c}{(1 - \xi_f \lambda)^2} + \frac{1}{1 - \xi_f \lambda}; \quad c = \langle \vec{p} | \vec{\psi} \rangle \langle \vec{\psi} | -\widehat{B} / \lambda | \vec{p} \rangle \quad (7.51)$$

where we kept the second (less divergent) term because the numerator of the first one vanishes for many particular cases with some regularities in  $\widehat{B}$  matrix, such as, for example, in Potts model,  $p$ -charge model, and some others. In the main approximation, (??) yields  $\xi_f \simeq 1/\lambda$ . More accurately, we obtain

$$\xi_f \simeq \begin{cases} (1/\lambda) [1 - \sqrt{c/2s}] & \text{for } c \neq 0 \\ (1/\lambda) [1 - (1/2s)] & \text{for } c = 0 \end{cases} . \quad (7.52)$$

A similar approach can be applied for the target transition, i.e. for equations (7.27) and (7.41). In particular, for the vicinity of triple point we get

$$\kappa \simeq \begin{cases} \sqrt{8s/c} & \text{for } c \neq 0 \\ 8s/3 & \text{for } c = 0 \end{cases} . \quad (7.53)$$

In both cases,  $\kappa$  is rather large for stiff chain, and thus the region of vicinity of triple point (in the sense of applicability of the regime (7.40)) is small ( $\sim 1/\sqrt{\kappa}$ ). Outside of this region, we can analyse the most singular terms of (7.27), yielding  $T_{tar} \simeq T_p [(1 - \xi_p \lambda) / (1 - \xi_f \lambda)]^2$ , which is almost a vertical line on the phase diagram.

### 7.3.6 Miyazawa-Jernigan Matrix

It is of special interest to consider the imprinting phase diagram for polymers comprised of amino acids. An interaction matrix for amino acids was derived in [Mia85] based upon protein statistics. The phase diagram for imprinted sequences, numerically calculated for the Miyazawa-Jernigan (MJ) matrix, is shown in the Figure 7-3.

It is instructive to see the significance of the particular placement of the matrix elements in the interaction matrix  $\widehat{B}$ , which enter the higher order moments, such as  $\langle B^3 \rangle_m$ , and thus govern, in particular, the slope of the target transition curve near the triple point (7.40). To illustrate this, we also examined the curve  $T_{tar}(T_p)$  for an artificial interaction matrix consisting of the elements of the MJ matrix in a random (symmetric matrix) arrangement. This randomized version of MJ matrix leads to a smaller region of target phase above  $T_f$  (i.e.,  $T_f < T < T_{tar}$  and  $T_p < T_f$ ).



## 7.4 Conclusion

In general, the goal of this work was to examine the effect of a particular type of monomer species interactions on the nature of the freezing phase transformation to the target conformation, and in particular, to study whether such a transformation exists for all heteropolymers. To this end, we wrote an interaction Hamiltonian which assumed only that heterogeneity comes from two-body interactions and interactions are short range: an arbitrary matrix of monomer-species interactions was considered. We were able to calculate the heteropolymeric properties of the freezing and target transition temperature explicitly in terms of the interaction matrix. Thus, the freezing properties of *any* heteropolymer model in which heterogeneity comes solely from binary interactions of monomers can be solved using our formalism, simply by determining the interaction matrix between species and examining properties of this matrix.

A polymer is considered a heteropolymer if it is composed of differing monomeric species, mathematically expressed by  $\widehat{\Delta}\widehat{B} \neq \widehat{0}$ . All interaction matrices of this form lead to a finite freezing temperature for random and designed sequences. Thus, the particular details of the interaction matrix are vital to neither the *existence* of the freezing and target transformations nor the qualitative aspects of the phase diagram (Figure 7-3).

Moreover, in addition to the aspects common to all heteropolymeric interaction matrices, we can address which region of the phase diagram is the most promising from the standpoint of experimental implementation of the Imprinting model. We find that the region of target phase between the freezing and target transitions ( $T_f < T < T_{\text{tar}}$  and  $T_p < T_f$ ) is the optimal region. Indeed, for  $T > T_{\text{tar}}$ , there is no unique structure. On the other hand, at  $T_p > T_f$ , i.e., when the sequences are almost random, the chain freezes to some state which generally has nothing in common with target conformation. Finally, for  $T < T_f$ , some conformations other than target conformation  $\star$  become thermodynamically stable as well, since the designed sequence act much like a random sequence in a conformation other than

target conformation  $\star$ . Indeed, thermodynamic stability or metastability of some conformation means that some additive (proportional to  $N$ ) energy is needed to leave that state once system is there. Therefore, the kinetic self-assembly of the target conformation, even though it is thermodynamically stable, is very problematic at  $T < T_f$ . These arguments are equivalent to the recently formulated criteria of reliable folding kinetics in terms of the gap in the spectrum of energies of the heteropolymer chain [Sali94b].

Thus, the region of phase diagram immediately below the target transition line, but above the freezing temperature, is very important because the equilibrium conformation is the designed conformation  $\star$  and folding to  $\star$  is fast and reliable. Moreover, we can conclude that the design of an experiment should probably include the choice of set of monomers which interact in such a way that to maximize the width of this region on the phase diagram. Also, it is vital to polymerize a dense monomer mixture; therefore, perhaps certain exotic polymerization schemes such as emulsion polymerization should be employed.

Finally, we address the applicability of our theory. First, since we have truncated the series (7.13) to  $\mathcal{O}(\phi^2)$ , we cannot describe any physical properties of the system due to phase transitions in the average value of  $\phi$ , such as phase separation of the monomers. However, these transitions are not found in all interaction matrices; for example they are present in ferromagnetic interaction matrices ( $B_{ij} = -\delta_{ij}$ ) [Sfa93, Fre91] but absent in anti-ferromagnetic interaction matrices ( $B_{ij} = \delta_{ij}$ ). long range Coulomb interactions, one must also consider the screening due to counter ions and polyions. For systems with a large degree of screening (large concentration of ions), the characteristic length of interactions is short and we recover delta function like short range interactions. For a small degree of screening, the interactions cannot discriminate between the placement of polyions within the characteristic length of interactions (Debye length) and thus there is no possibility for freezing to the microscopic length scale, which is the region of interest of the current study. A possible case of interest would be the freezing of a polymer which consists of monomers which interact with short range as well as long range interactions;

however, this is currently beyond the scope of this chapter. solvent effects can be included by the appropriate redefinition of the interaction matrix [Mia85].

In conclusion, the fact that the specific nature of the interaction matrix is not vital to the existence of the target transition may help experimentalists in the implementation of the Imprinting procedure, as one may ignore the details of the interactions chosen. It is also interesting to consider the Imprinting model as a possible scheme of prebiotic evolution: monomers polymerize in a conformation capable of recognizing a given target molecule. Therefore, the diminished role of the specific form of the interaction matrix may also have helped the development of prebiotic evolution.

## 7.5 Appendix: Rotation of Replica Space

In order to simplify both terms in the energy (7.19), we perform the following matrix operations. Consider the structure of  $\widehat{Q}^{(n+1)} \equiv \widehat{q}^{(n+1)}(\widehat{T}^{(n+1)})^{-1}$ . It is a block diagonal matrix. We can label these blocks: the target block of size  $(y+1) \times (y+1)$  is called block 0 and the  $g$  blocks of size  $x \times x$  are labeled from 1 to  $g$ , where  $g = (n-y)/x$ ; we will employ the convention that capital Greek letters label replica blocks. Consider the operators  $\widehat{\mathcal{R}}_{\Gamma}^{(b)}$  which diagonalize the  $\Gamma$  block of  $\widehat{Q}^{(n+1)}$ , i.e.  $\widehat{\mathcal{R}}_{\Gamma}^{(b)} \widehat{Q}_{\Gamma}^{(b)} (\widehat{\mathcal{R}}_{\Gamma}^{(b)})^{-1} \equiv \widehat{\Lambda}_{\Gamma}^{(b)}$  is a diagonal matrix, where  $b$  is the dimensionality of the  $\Gamma$  block.

We then define the block matrix in replica space  $\widehat{\mathcal{R}}^{(n+1)}$  as the diagonal block matrix whose  $g$  diagonal blocks are  $\widehat{\mathcal{R}}_{\Gamma}^{(b)}$ ,  $\Gamma = 0, \dots, g$ ; finally, we extend this operator into species space by  $\widehat{\mathcal{R}}^{((n+1)q)} = \widehat{\mathcal{R}}^{(n+1)} \otimes \widehat{I}^{(q)}$ . Thus, the  $\widehat{\mathcal{R}}^{((n+1)q)}$  operator diagonalizes the block matrix  $\widehat{Q}^{(n+1)} \otimes \widehat{A}^{(q)}$  in replica space, while rendering the species dimensions  $\widehat{A}^{(q)}$  unchanged. Thus,

$$(\widehat{\mathcal{R}}^{((n+1)q)})^{-1} [\rho \widehat{Q}^{(n+1)} \otimes \widehat{\Delta}^{(q)} \widehat{B}^{(q)}] \widehat{\mathcal{R}}^{((n+1)q)} = \rho \widehat{\Lambda}^{(n+1)} \otimes \widehat{\Delta}^{(q)} \widehat{B}^{(q)} \quad (7.54)$$

The eigenvalues  $\Lambda_{\alpha\alpha}$  and eigenvectors of the  $\widehat{Q}^{(n+1)}$  matrix can be calculated by



elementary means.<sup>1 2</sup> For the  $x \times x$  non-target blocks  $\widehat{Q}_\Gamma^{(n+1)}$ , where  $0 > \Gamma \geq g$ , we have a  $(x - 1)$ -degenerate eigenvalue  $\lambda = 0$  and a single non-zero eigenvalue  $\lambda = x$ . For the  $(y + 1) \times (y + 1)$  target block  $\widehat{Q}_0^{(n+1)}$ , we have a  $y$ -degenerate eigenvalue  $\lambda = 0$  and a non-degenerate eigenvalue  $\lambda = y + \tau_p$ , where  $\tau_p = T/T_p$ . Thus, for the whole  $\widehat{Q}^{(n+1)}$  matrix we have a  $[y + (n - y)(x - 1)/x]$ -degenerate eigenvalue  $\lambda = 0$ , a  $[(n - y)/x]$ -degenerate non-zero eigenvalue  $\lambda = x$ , and a non-degenerate eigenvalue  $\lambda = y + T/T_p$ . Thus, we have

$$\begin{aligned} \ln \det \left[ \widehat{I}^{(q(n+1))} + \frac{\rho}{T} \widehat{Q}^{(n+1)} \otimes \widehat{\Delta}^{(q)} \widehat{B}^{(q)} \right] = \\ \ln \det \left[ \widehat{I}^{(q)} + (y + \tau_p) \frac{\rho}{T} \widehat{\Delta}^{(q)} \widehat{B}^{(q)} \right] + \frac{n - y}{x} \ln \det \left[ \widehat{I}^{(q)} + x \frac{\rho}{T} \widehat{\Delta}^{(q)} \widehat{B}^{(q)} \right] \end{aligned} \quad (7.55)$$

As for the second term in (7.19), we again use the technique of rotation in replica space in order to bring this term into block diagonal form. For each block, we have a complete set of orthonormal eigenvectors of the form

$$\widehat{\mathcal{R}}_{\alpha\beta}^{(y)} = \begin{cases} \exp [(2\pi i/z)(\alpha - 1)(\beta - 1)] z^{-1/2} & \text{for } \alpha, \beta > 0 \\ 0 & \text{for } \alpha = 0, \beta > 0 \\ \tau_p(\tau_p^2 + z)^{-1/2} & \text{for } \alpha = \beta = 0 \\ (\tau_p^2 + z)^{-1/2} & \text{for } \alpha > 0, \beta = 0 \end{cases} \quad (7.56)$$

where  $z = x$  and  $0 < \alpha, \beta \leq x$  for the non-target block; and  $z = y$  and  $0 \leq \alpha, \beta \leq y$  for the target block.

<sup>1</sup>For a  $n \times n$  matrix  $M_n$  with diagonal elements  $\tilde{m}$  and off diagonal elements  $m$ , we have a  $(n - 1)$ -fold degenerate eigenvalue  $\lambda = \tilde{m} - m$  and a non-degenerate eigenvalue of  $\lambda = \tilde{m} + (n - 1)m$ . Its eigenvectors are of the form  $R_{\alpha\beta} = \exp [(2\pi i/n)(\alpha - 1)(\beta - 1)] n^{-1/2}$ . The inverse of  $M_n$  has diagonal elements  $\det M_{n-1} / \det M_n = 1/(\tilde{m} - m)$  and off diagonal elements  $-m(\tilde{m} - m)^{n-2} / \det M_n = -m/\{(\tilde{m} - m)[\tilde{m} + (n - 1)m]\}$ .

<sup>2</sup>Consider a  $n \times n$  matrix  $\widehat{M}$  which has one column with elements  $M_{i1} = a$  and all other elements  $M_{ij} = 1, j > 1$ .  $\widehat{M}$  has a  $(n - 1)$ -degenerate eigenvalue  $\lambda = 0$  and a non-degenerate eigenvalue  $\lambda = (n - 1) + a$ , and eigenvectors  $(-1/a, 1, 0, \dots, 0)$ ,  $(-1/a, 0, 1, \dots, 0)$ ,  $\dots$ ,  $(-1/a, 0, 0, \dots, 1)$ , and  $(1, 1, 1, \dots, 1)$ .

Using these eigenvectors, we get the second term in eq (7.20):

$$\begin{aligned}
& \left\langle \widehat{\rho}^{(n+1)} \widehat{\mathcal{R}}^{(n+1)} \left| \frac{\rho}{T} (\widehat{\mathcal{T}}^{(n+1)})^{-1} \otimes \widehat{B}^{(q)} \left[ \widehat{\Gamma}^{(n+1)q} + \frac{\rho}{T} \widehat{\Lambda}^{(n+1)} \otimes \widehat{\Delta}^{(q)} \widehat{B}^{(q)} \right]^{-1} \right| (\widehat{\mathcal{R}}^{(n+1)})^{-1} \widehat{\rho}^{(n+1)} \right\rangle^{(n+1)} \\
& = (y + \tau_p) \frac{\rho}{T} \widehat{B}^{(q)} \left[ \widehat{\Gamma}^{(q)} + (y + \tau_p) \frac{\rho}{T} \widehat{\Delta}^{(q)} \widehat{B}^{(q)} \right]^{-1} + (n - y) \frac{\rho}{T} \widehat{B}^{(q)} \left[ \widehat{\Gamma}^{(q)} + \frac{\rho x}{T} \widehat{\Delta}^{(q)} \widehat{B}^{(q)} \right]^{-1}
\end{aligned}$$

## Chapter 8

# Design and renaturation with different interactions

Using replica mean field theory, we examine the freezing transition of heteropolymers whose sequences are selected to optimize the energy of a given “designed” conformation interacting through a given conformation and then are renatured with a second matrix of interactions. We find that the possibility of folding to the designed conformation is controlled by the correlations of the elements of the design and renaturation interaction matrices. Therefore, in computer simulations of protein folding, for example, one need not exactly reproduce the potentials found in nature (which were used to design proteins), but rather use a matrix which is strongly correlated with the true matrix of interactions; a simple conservative analysis of the permissible error for normally distributed error from the true interaction matrix indicates that even a 30% error in the interaction energy should still yield correct renaturation.

## 8.1 Introduction

### 8.1.1 What is this work about?

The native state of a protein is in a sense “written” in the sequence using the “language” of physical interactions between monomers. In this work, we examine the effects of “misunderstandings” and “misspellings” of this language.

A somewhat related question was recently discussed by Bryngelson [Bry94]. He considered heteropolymer chains with *random* sequence and estimated the probability that its lowest energy conformation will be correctly detailed by the model with noisy distorted potentials of volume interactions between monomers. The result is that the probability,  $p$ , diminishes with noise amplitude,  $\eta$ , as  $p \sim 1 - \text{const} \cdot \eta N^{1/2}$ , for sufficiently long chain, or in thermodynamic limit, there is no chance to compute equilibrium conformation given that some mistakes in the determination of energies are inevitable.

By contrast, we consider here heteropolymer chains with sequences that are *not random*, but rather “designed” [Sha93b], or “imprinted” [Pan94d], or “selected” [Yue92], or, in other words, obey the so-called principle of minimal frustration [Bry87]. We show, that for these chains the situation is dramatically different, and there is finite probability of successful recovery of thermodynamically stable conformation, even in thermodynamic limit ( $N \rightarrow \infty$ ) and for finite non-vanishing  $\eta$ .

### 8.1.2 Sequence design and folding are governed by *different interactions*

As the sequence design is based on energy optimization, it employs physical interactions between monomers. It is however possible, and, moreover, almost inevitable, that these interactions are somewhat different from those governing folding. Apart from speculations on the interactions that governed the “design” of modern proteins by evolution, we mention three illustrations of our thesis:

1. When one tries to find theoretically or computationally the native state for the chain with given sequence (direct protein folding problem); one can say that nature details the interactions used in the design of protein sequences and man-made potentials are used as substitutes in the simulations of renaturation.
2. Similarly, when one is looking for a sequence to fold into a given conformation, one is essentially trying to design the sequence using artificial potentials in such a way, that this sequence under real natural interactions will fold in a desirable way.
3. Speaking of the attempt to reproduce protein-like properties in the man made heteropolymer via the Imprinting procedure [Pan94d], we have to acknowledge some difference between interactions of monomers in the soup prior to polymerization and interactions of the links of polymer.
4. One can consider the renaturation of a protein in a solvent different than that used during “design” also as an experiment in which the interactions during design and renaturation are different.

If there are, say,  $q$ , different monomeric species involved in our polymer ( $q = 20$  for protein), interactions between species  $i$  and  $j$  can be described in terms of the  $q \times q$  matrix  $B_{ij}$ . In general, there are two different matrices,  $B_{ij}^p$  and  $B_{ij}$ , first is governing design and second is governing folding behavior of the already prepared chain.

To have two *different* interaction matrices for design and renaturation is somewhat similar to writer and reader who use different languages. Naprimer, my nadeemsa, chto nash chitatel' schitaet etot tekst napisannym po-angliiski i poetomu vryadli poimet etu frazu.<sup>1</sup> Clearly, such a venture has a chance if and only if those languages are not completely different, but merely dialects of one language.

---

<sup>1</sup>This Russian sentence says: “For example, to the best of our hope, our reader considers this text to be written in English and thus unlikely understands this sentence”.

Similarly, infinitesimally small changes to the interaction matrix should not have any significant ramifications, while on the other hand, a radically different matrix structure should lead to completely different folding behavior. Using the terminology of frozen and target phases, we can ask if the chain designed with some matrix  $B_{ij}^p$  will freeze to target state when governed by another matrix  $B_{ij}$ ? In other words, if we want to get the target phase, how accurate should we be in choosing matrices  $B_{ij}^p$  and  $B_{ij}$ ? Other interesting aspect of the question is which properties of  $B_{ij}^p$  and  $B_{ij}$  matrices are important, that is to which of them the chain behavior is sensitive? And what measure do we use to define the proximity of interaction matrices?

Previous treatments have addressed certain aspects of these questions. We specifically mention here the works [Sha91,Bry94], which we discuss in more details later in this chapter.

## 8.2 The Model

We start from a heteropolymer chain Hamiltonian in which interactions are described in terms of the energy of interaction of species

$$\mathcal{H} = \sum_{I,J}^N B_{s_I s_J} \delta(\mathbf{r}_I - \mathbf{r}_J) \quad (8.1)$$

where  $B_{ij}$  is the interaction energy between monomer *species*  $i$  and  $j$  ( $i, j \in \{1 \dots q\}$ ),  $s_I$  is the species of monomer at position  $I$  along the chain,  $N$  is the number of monomers, and  $\mathbf{r}_I$  is the position of monomer  $I$ . We use the convention that lower case roman letters label species space, upper case roman letters label monomer number along the chain, and lower case greek letters label replicas.

We do not explicitly include in the Hamiltonian (8.1) anything leading to the overall collapse of the chain. We do imply, however, the existence of some strong compressing factor, such as overall homopolymeric-type poor solvent effect (expressed with  $\mathcal{H}' = B\rho^2 + C\rho^3$  with species independent  $B$  and  $C$  and strongly negative  $B$ ) or an appropriate external field (such as the small rectangular box

from which the chain is not allowed to go). We stress that this is of vital importance for the entire approach that the chain is maintained in the globular compact state (compare [Yue95], where the design scheme failed to work just because the requirement of overall collapsed state was relaxed).

Since the heteropolymer sequence does not change during folding, we immediately encounter the technical problem that sequences are a quenched quantity and thus we average the free energy over all sequences (with a particular weighting due to design) rather than the partition function. This leads directly to the replica approach. The details of the corresponding calculation are similar to what is presented elsewhere [Pan95b]. Here we briefly outline the main steps. The replicated partition function can be symbolically written as

$$\langle Z^n \rangle = \sum_{\text{sequence}} \mathcal{P}_{\text{sequence}} \sum_{\{\text{conformations}\}} \exp \left[ - \sum_{\alpha=1}^n \mathcal{H}(\text{sequence}, \text{conformation}_\alpha) / T \right] \quad (8.2)$$

where we explicitly mention the dependence of the Hamiltonian (8.1) on both sequence, which is the same for all replicas  $\alpha \in 1 \dots n$ , and conformation, which is different for different replicas. Probability distribution over the set of sequences,  $\mathcal{P}_{\text{sequence}}$  is defined by the preparation process and thus in our case can be written as

$$\begin{aligned} \mathcal{P}_{\text{sequence}} &\sim [p_{s_1} \cdot p_{s_2} \cdot \dots \cdot p_{s_N}] \times \\ &\times \sum_{\text{target conformation}} \exp [\mathcal{H}^p(\text{sequence}, \text{target conformation}) / T_p] \end{aligned} \quad (8.3)$$

where we drop normalization factor. In the equation (8.3),  $p_s$  is the probability of appearance of the monomer species  $s$  (which is normally controlled by the chemical potentials of components in the monomer soup surrounding the preparation bath),  $\mathcal{H}^p$  is Hamiltonian of the form (8.1) except with the “preparation” matrix  $\hat{B}^p$  instead of  $\hat{B}$  which controls folding through equation (8.2). Accordingly,  $T_p$  is the temperature at which preparation process is performed.

We stress that our approach is not restricted to any particular target conforma-

tion. By contrast, we do average over all possible (compact) target conformations (see equation (8.3)), and thus our scheme picks up not just the good sequences, but the pairs “target conformation - good sequence,” where both terms are well adjusted to each other (see also the discussion in [Yue95]). This is a good match for Imprinting, since we assume that some external field chooses sequence-conformation pairs based upon matching with the field [Pan95c]. Indeed, this may be analogous to protein evolution, in which nature chooses sequence-conformation pairs not for any specific nature of the conformation or sequence but for its functionality; this can be viewed in physical terms as some external field effecting the selection of sequence and conformation [Pan95c].

### 8.3 Free Energy of the Model

Inspection of the equations (8.2, 8.3) indicates that we can formally express the weight corresponding to the design process as an additional replica labeled 0 [Ram94,Pan94e,Pan9

$$\langle Z^n \rangle = \sum_{\text{sequence}} \prod_{I=1}^N p_{s_I} \sum_{\{\text{conformations}\}} \exp \left[ \sum_{\alpha=0}^n \sum_{I \neq J=1}^N B_{s_I, s_J}^{\alpha} \delta(\mathbf{r}_I^{\alpha} - \mathbf{r}_J^{\alpha}) / T_{\alpha} \right], \quad (8.4)$$

where  $B_{ij}^{\alpha=0} \equiv \widehat{B}^p$  is the matrix which expresses the interactions used for the chain preparation (i.e. replica  $\alpha = 0$ ) and  $B_{ij}^{\alpha>0} \equiv B_{ij}$  is the interaction matrix which governs folding or renaturation. Hereafter, conformations are given in terms of position vectors  $\mathbf{r}_I^{\alpha}$  for each monomer number  $I$  and each replica  $\alpha$ . By the sum over conformations we mean the sum in which the condition of chain connectivity is strictly obeyed (technically this can be done either in continuous form as Edwards [Doi86] or in discrete form like Lifshits [Lif78]).

To facilitate averaging over the sequences, we define the densities

$$\rho_i^{\alpha}(\mathbf{R}) = \sum_I^N \delta(s_I, i) \delta(\mathbf{r}_I^{\alpha} - \mathbf{R}), \quad (8.5)$$



then rewrite the exponent in equation (8.4) as

$$\sum_{I \neq J=1}^N \frac{B_{s_I, s_J}^\alpha}{T_\alpha} \delta(\mathbf{r}_I^\alpha - \mathbf{r}_J^\alpha) = \int d\mathbf{R}_1 d\mathbf{R}_2 \sum_{i,j}^q \rho_i^\alpha(\mathbf{R}_1) \frac{B_{i,j}^\alpha}{T_\alpha} \delta(\mathbf{R}_1 - \mathbf{R}_2) \rho_j^\alpha(\mathbf{R}_2) \quad (8.6)$$

and perform a Hubbard-Stratonovich transformation on the quantity  $\rho_i^\alpha(\mathbf{R})$ , thus introducing the conjugate field  $\phi_i^\alpha(\mathbf{R})$ . We average over the sequence and truncate the resulting exponent to  $\mathcal{O}(\phi^2)$ , which yields (see the details in [Pan95b]):

$$\langle Z^n \rangle = \sum_{\text{conformations}} \int \mathcal{D}\{\phi(\mathbf{R})\} \exp \left\{ \int d\mathbf{R} \sum_{\alpha=0}^n \sum_i [p_i \rho^\alpha(\mathbf{R}) \phi_i^\alpha(\mathbf{R})] + \int d\mathbf{R}_1 d\mathbf{R}_2 \sum_{\alpha, \beta=0}^n \sum_{ij} \left[ \frac{1}{4} \left( \frac{B_{ij}^\alpha}{T_\alpha} \right)^{-1} \delta^{\alpha\beta} \delta(\mathbf{R}_1, \mathbf{R}_2) + \frac{1}{2} \Delta_{ij} Q^{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) \right] \phi_i^\alpha(\mathbf{R}_1) \phi_j^\beta(\mathbf{R}_2) \right\} \quad (8.7)$$

where we define the overall density  $\rho_\alpha(\mathbf{R}) = \sum_I^N \delta(\mathbf{r}_I^\alpha - \mathbf{R}) = \sum_{i=1}^q \rho_i^\alpha(\mathbf{R})$  and the replica overlap order parameter  $Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) = \sum_I^N \delta(\mathbf{r}_I^\alpha - \mathbf{R}_1) \delta(\mathbf{r}_I^\beta - \mathbf{R}_2)$ . Since the density is a single replica quantity and we assume the chain as a whole is compressed, that is, density is constant throughout the globule, we simply take  $\rho_\alpha(\mathbf{R}) \equiv \rho$ . Furthermore, using a variational argument, it was shown [Sfa93, Pan94e] that freezing occurs down to microscopic length scales, thus allowing to take  $Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) = \rho q_{\alpha\beta} \delta(\mathbf{R}_1 - \mathbf{R}_2)$ , where the form of the conformation correlator  $q_{\alpha\beta}$  is found to be that of a Parisi matrix with one step symmetry breaking, with either complete overlap ( $q^{\alpha\beta} = 1$ ) or no overlap ( $q^{\alpha\beta} = 0$ ). (This directly corresponds with the Random Energy Model [Der80] introduced directly in previous heteropolymer models [Bry87].) This facilitates Gaussian integration over the  $\phi$  fields. To write the result in even simpler form, we can also include a conformation-independent constant by the transformation  $\phi_i^\alpha \rightarrow 2 \sum_j (\rho \widehat{B}^\alpha / T_\alpha)_{ij}^{1/2} \phi_j^\alpha$  to get

$$\langle Z^n \rangle = \sum_{\text{conformations}} \left[ \int d\{\phi\} \exp \left\{ \sum_{\alpha=0}^n \sum_{ij} \left[ (2\rho \widehat{B}^\alpha / T_\alpha)_{ij}^{1/2} p_i \phi_j^\alpha \right] + \sum_{\alpha, \beta=0}^n \sum_{ij} \left[ \delta_{ij} \delta^{\alpha\beta} + 2 \left( (\widehat{B}^\alpha / T_\alpha)^{1/2} \widehat{\Delta} (\widehat{B}^\beta / T_\alpha)^{1/2} \right)_{ij} \rho q^{\alpha\beta} \right] \phi_i^\alpha \phi_j^\beta \right\} \right]^N \quad (8.8)$$

where we use a hat to indicate that the object is matrix in species space (i.e.  $\hat{A} = A_{ij}$ ). We evaluate this Gaussian integral, yielding the free energy

$$\langle Z^n \rangle = \sum_{\{\text{conformations}\}} \exp[E(q)] , \quad (8.9)$$

where the effective energy of the  $n$  replica system is given by

$$\begin{aligned} \frac{E(q)}{N} = & \frac{1}{2} \ln \det \left[ \hat{I} \delta^{\alpha\beta} + 2\rho \left( \hat{B}^\alpha / T_\alpha \right)^{1/2} q^{\alpha\beta} \hat{\Delta} \left( \hat{B}^\beta / T_\beta \right)^{1/2} \right] + \\ & \frac{1}{\rho} \sum_{\alpha\beta} \left\langle \vec{\rho} \left| \left( \hat{B}^\alpha / T_\alpha \right)^{1/2} \left[ \hat{I} \delta^{\alpha\beta} + 2\rho \left( \hat{B}^\alpha / T_\alpha \right)^{1/2} q^{\alpha\beta} \hat{\Delta} \left( \hat{B}^\beta / T_\beta \right)^{1/2} \right]^{-1} \left( \hat{B}^\beta / T_\beta \right)^{1/2} \right| \vec{\rho} \right\rangle \end{aligned} \quad (8.10)$$

$\langle |\dots| \rangle$  denotes the scalar product over species space, the determinant in the first term is over species and replica space, and the vector  $\vec{\rho}$  is given by  $\vec{\rho}_i^\alpha = p_i \rho$ . Note that the only remaining dependence on conformations come through conformational correlators  $q_{\alpha,\beta}$ . Given the particular structure of  $q^{\alpha\beta}$ , effective energy (8.10) can be expressed directly in terms of the number of replicas which overlap with the target group  $y$  and the size of a group  $x$  for the remaining  $n - y$  replicas divided into  $(n - y)/x$  groups. Thus, we can simplify the expression for effective energy (8.10) by removing replica dimensionalities, as is performed in Appendix A. This also allows one to write the entropy of the macrostate with given  $x$  and  $y$ , as it is associated simply with grouping of replicas  $S = Ns[y + (n - y)(x - 1)/x]$ <sup>2</sup> [Sha89a,Sfa93]. This allows conversion from the sum over conformation to a functional integral over  $Q^{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2)$ , and even further, to conventional integral over  $x$  and  $y$ , which, in the mean field approximation, can be further simplified to optimization of the effective  $n$ -replica free energy

$$\begin{aligned} \frac{F(x, y)}{N} = & \frac{n - y}{2x} \left\{ \ln \det \left[ \hat{I} + 2x \hat{\Delta} \hat{B} / T \right] + \left\langle \vec{p} \left| 2x \hat{B}^1 / T \left[ \hat{I} + 2x \hat{\Delta} \hat{B} / T \right]^{-1} \right| \vec{p} \right\rangle \right\} \\ & + \frac{1}{2} \ln \det \left[ \hat{I} + 2 \hat{\Delta} \hat{B}^p / T_p + 2y \hat{\Delta} \hat{B} / T \right] \end{aligned}$$

<sup>2</sup>As we group two replicas such that they have identical conformation down to microscopic scale related to the volume  $v$ , there is an entropy loss of  $s \equiv \ln(a^3/v)$  per monomer, where  $a$  is the distance between monomers and  $v$  is the excluded volume.

$$\begin{aligned}
& + \left\langle \vec{p} \left| \left[ \widehat{B}^p / T_p + y \widehat{B} / T \right] \left[ \widehat{I} + 2 \widehat{\Delta} \widehat{B}^p / T_p + 2y \widehat{\Delta} \widehat{B} / T \right]^{-1} \right| \vec{p} \right\rangle \\
& - s[y + (n - y)(x - 1)/x]
\end{aligned} \tag{8.11}$$

## 8.4 Analysis of the Free Energy and Phase Diagram

The expression (8.11) is rather similar to what we had in the work [Pan95b] while considering the model with identical interactions for design and folding and, of course, it is exactly reduced to the corresponding equation of that work [Pan95b] when  $\widehat{B} = \widehat{B}^p$ . Furthermore, this expression implies the same structure of phase diagram, with the same three globular phases: random, frozen, and target. (We remind the reader, that overall collapse of the chain is the necessary pre-condition of our approach, and thus globule-to-coil phase transition falls outside of the framework of the present study). To see the structure of phase diagram, we first look at the allowed variations of the order parameters  $x$  and  $y$ .

For simplicity, we consider here only small  $s$  regime. In this case, freezing transitions, which are the main topic of our interest here, occur when  $B$  is (in a reasonable sense) also small. Indeed, freezing phase transitions result physically from the competition between energetic and entropic parts of free energy (8.11), where energetic part favors gathering of replicas into groups while entropic part favors diversity of replicas. For energy to be competitive to an entropy when  $s$  is small,  $B$  must be small as well. This allows one to simplify equation (8.11) truncating it to quadratic order in  $B$ .

As  $y$  is the number of replicas whose conformation coincides with the target conformation, this value must be in between of 0 and  $n$ . What is relevant in replica approach is  $n \rightarrow 0$  limit, and, moreover, only the terms which are linear in  $n$  are to be considered (because higher order terms disappear in the main equation  $\langle \ln Z \rangle = \lim_{n \rightarrow 0} (\langle Z^n \rangle - 1) / n$ ). Accordingly, since  $0 \leq y \leq n$ , we must linearize the free energy in  $y$  as well [Pan94e, Pan95b]. This leads to further simplification of

(8.11):

$$\begin{aligned}
F = & \text{Tr} \left[ (n - y) \left\{ \widehat{\Delta} \widehat{B} / T - x \widehat{\Delta} \widehat{B} \widehat{\Delta} \widehat{B} / T^2 + \widehat{P} \widehat{B} / T - 2x \widehat{P} \widehat{B} \widehat{\Delta} \widehat{B} / T^2 \right\} \right. \\
& + \widehat{\Delta} \widehat{B}^p / T_p + y \widehat{\Delta} \widehat{B} / T - 2y \widehat{\Delta} \widehat{B}^p \widehat{\Delta} \widehat{B} / T T_p - \widehat{\Delta} \widehat{B} \widehat{\Delta} \widehat{B} / T_p^2 \\
& \left. + \widehat{P} \widehat{B}^p / T_p + y \widehat{P} \widehat{B} / T - 2 \left( \widehat{P} \widehat{B}^p \widehat{\Delta} \widehat{B}^p / T_p^2 + y \widehat{P} \widehat{B} \widehat{\Delta} \widehat{B}^p / T T_p + y \widehat{P} \widehat{B}^p \widehat{\Delta} \widehat{B} / T T_p \right) \right] \\
& - T N s [y + (n - y)(x - 1)/x] \tag{8.12}
\end{aligned}$$

While  $y$  describes breaking of the symmetry between  $n$  replicas due to their attraction to the target replica labeled 0,  $x$  describes spontaneous symmetry breaking. When we have integer number of replicas,  $n$ , clearly,  $1 \leq x \leq n$ :  $x$  cannot be smaller than unity, because it is the number of replicas in the group. When  $n \rightarrow 0$ , the logic about the number of replicas in the group is not applicable any more, but it is natural to think that formal inequalities for  $x$  just simply flip signs:  $n \leq x \leq 1$ . With this in mind, we optimize free energy (8.12) with respect to  $x$  yielding the equation which determines  $x$ :

$$s = \frac{x^2}{T^2} \text{Tr} \left[ \widehat{\Delta} \widehat{B} \widehat{\Delta} \widehat{B} + 2 \widehat{P} \widehat{B} \widehat{\Delta} \widehat{B} \right] \tag{8.13}$$

Note, that this equation does not involve either  $T_p$  or  $\widehat{B}^p$  and thus it does not depend on preparation process. This has clear physical meaning. Namely, this reflects the behavior similar to that of REM, because the designed sequence behaves precisely as a random one in almost all the conformations except for the target conformation.

At this point, it is useful to introduce the following matrix ‘‘cumulants’’:

$$\begin{aligned}
\langle B^p B \rangle_c & \equiv \sum_{i,j} p_i p_j B_{ij}^p B_{ij} - \langle B^p \rangle \langle B \rangle = \text{Tr} \left( \widehat{\Delta} \widehat{B}^p \widehat{\Delta} \widehat{B} + 2 \widehat{P} \widehat{B}^p \widehat{\Delta} \widehat{B} \right) \\
\langle B^p \rangle & \equiv \sum_{i,j} p_i p_j B_{ij}^p = \text{Tr} \left( \widehat{P} \widehat{B}^p \right)
\end{aligned}$$

From the above, we can easily find the equation for the freezing temperature for random sequences. Indeed, freezing occurs when replicas start to group, thus spontaneously breaking the permutation symmetry. This happens when  $x = 1$ . Therefore,

freezing temperature is given via the relation

$$T_f^2 = \langle B^1 B^1 \rangle_c / s \quad (8.14)$$

In other words, the freezing temperature is given by the variance of the renaturation interaction matrix [Pan95a]. Note that this is a transition to a unique ground state which is not necessarily (and most likely not) the target conformation: we call this phase the *frozen phase* and we call the high temperature disordered phase in which there is no form of freezing, i.e. many conformations dominate equilibrium, the *random phase*.

To examine freezing to the target conformation, we must examine the conditions at which  $y > 0$ . Since  $y$  varies from 0 to  $n$ , what has physical meaning in the  $n \rightarrow 0$  limit is only the linear in  $y$  term of free energy. Therefore, free energy optimum corresponds to either  $y = 0$  (non-target phase), or to  $y = n$  (target phase). To find the corresponding critical temperature, we must examine the slope of the free energy at the point  $y = 0$  to determine whether  $y = 0$  or  $y = n$  is the stable solution [Pan95b]. The condition “slope” = 0 yields the relationship:

$$\begin{aligned} s &= \text{Tr} \left[ x \left( \frac{\widehat{\Delta B} \widehat{\Delta B}^p}{T} + 2 \frac{\widehat{P B} \widehat{\Delta B}^p}{T} + \frac{\widehat{\Delta B}^p \widehat{\Delta B}}{T_p} + 2 \frac{\widehat{P B}^p \widehat{\Delta B}}{T} \right) - \frac{x^2}{T^2} (\widehat{\Delta B} \widehat{\Delta B} + 2 \widehat{P B} \widehat{\Delta B}) \right] \\ &= 2x \left\langle \frac{B^p B}{T_p T} \right\rangle_c - x^2 \left\langle \frac{B^2}{T^2} \right\rangle_c \end{aligned} \quad (8.15)$$

This equation defines the phase boundary of the *target phase*, in which the system freezes to the target conformation.

We combine eqns (8.13-8.15) to get the boundary of the target phase (i.e. the preparation temperature  $T_p$  which separates the target phase from the random and frozen globule phases):

$$T_p^c = \begin{cases} 2gT/(1 + T^2/T_f^2) & \text{for } T \geq T_f \\ gT_f & \text{for } T \leq T_f \end{cases} \quad (8.16)$$

This is the previously obtained result for the transition to the target phase [Pan95b],

except with the inclusion of a factor  $g \equiv \langle B^p B \rangle_c / \langle B^2 \rangle_c$ , which by definition must be  $-1 \leq g \leq 1$ . This factor gives the degree of correlation between the elements of the two matrices. If the two matrices are the same (i.e. completely correlated)  $g = 1$ . For a lesser degree of correlation,  $0 \leq g < 1$ , the effective freezing temperature is proportional to  $g$ . Note that there is no freezing transition for  $g < 0$ ; in this case, the matrices are anticorrelated and there is no chance for renaturation to the designed state. Thus, the correlation between matrices  $g$  is the measure of the proximity between interaction matrices.

## 8.5 Discussion

By performing explicit calculations for the freezing transition of heteropolymers with different matrices for design and renaturation, we have found three phases: *random*, in which many conformations dominate equilibrium; *frozen*, where the polymer freezes to a single conformation other than the target conformation; and *target*, in which the polymer freezes to the target conformation. In the flexible chain limit, for the case where the design and renaturation matrices are different, the effective critical selective temperature for renaturation to the target phase becomes modified by a factor from the normalized correlation between the matrices ( $T_p^c \rightarrow T_p^c g$ ). For complete correlation,  $g = 1$ . For differences in the design and renaturation matrix ( $g < 1$ ), the selective temperature must be lowered in order to keep the system in the target phase; otherwise, there is no renaturation to the target conformation.

One can get a rough idea of the meaning of this effect by examining the case where the renaturation matrix is the design matrix with some noise:  $B_{ij} = B_{ij}^p(1 + \eta_{ij})$ , where  $\eta_{ij}$  is some noise with a gaussian distribution with a variance  $\sigma^2$ , i.e.  $\mathcal{P}(\eta_{ij}) \propto \exp[-\eta_{ij}^2/\sigma^2]$ . We can average the  $g$  factor over the noise to get

$$\bar{g} = \frac{\langle B_{ij}^p [B_{ij}^p(1 + \eta_{ij})] \rangle}{\left[ \langle [B_{ij}^p(1 + \eta_{ij})]^2 \rangle \right]^{1/2}} = (1 + \sigma^2)^{-1/2} \quad (8.17)$$

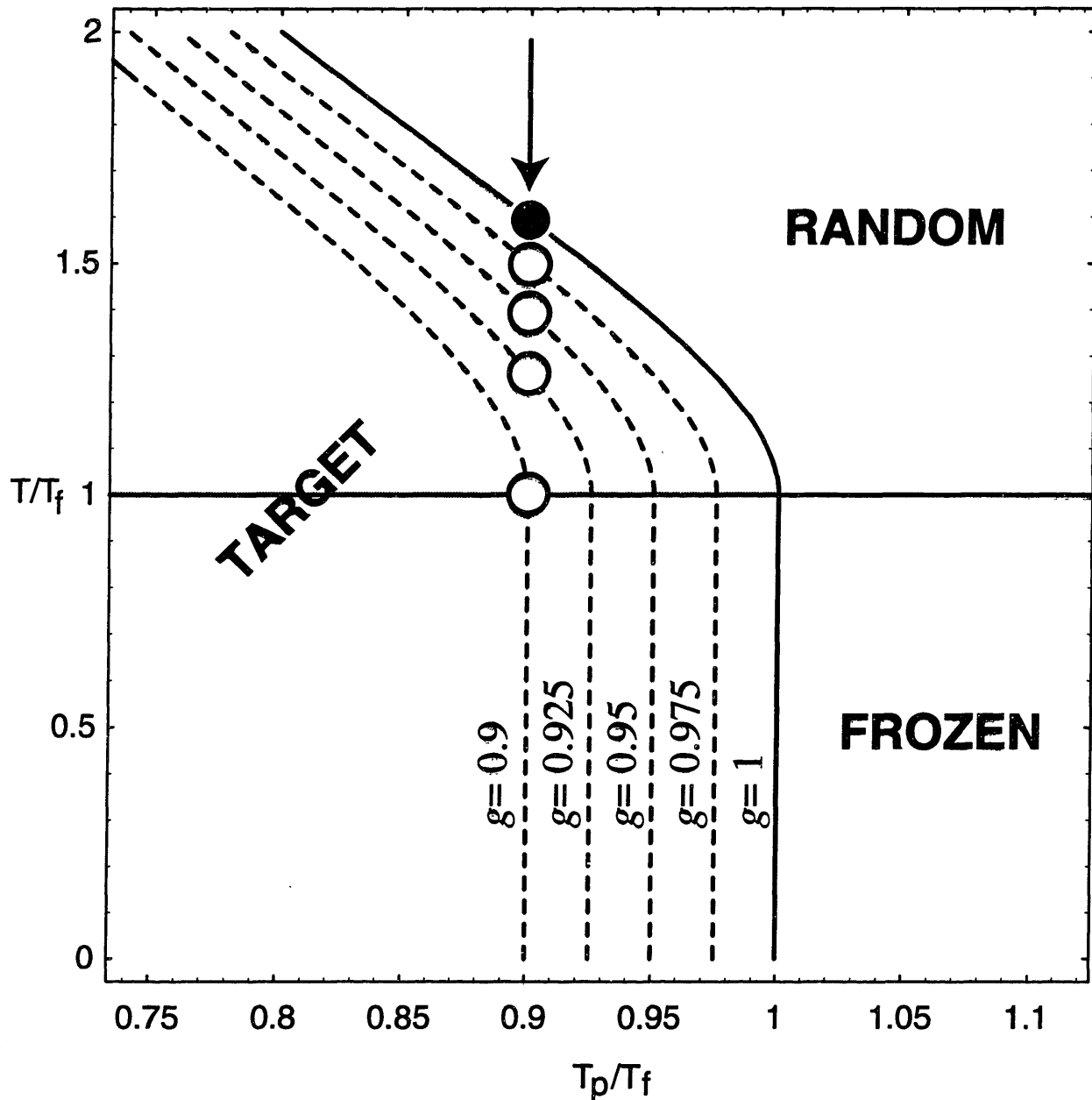


Figure 8-1: Phase diagram for different values of the matrix similarity factor  $g$ . For  $g = 1$ , the design and renaturation matrices are identical and we recover the results previously detailed. For  $0 < g < 1$ , the area of the target phase decreases, since a mismatch of matrices destroys the stability of a weakly designed target conformation. In the text, we argue that proteins are found at the location denoted by the solid circle (●). As the renaturation matrices deviate from the design matrix, the folding temperature for proteins should decrease (○) up to the point  $g = T_p/T_f$  (which we approximate to be 0.9 for proteins); below this point, proteins (at any acting temperature) cannot be in the target phase and only the glassy transition to some random conformation is seen for acting temperatures  $T < T_f$ .

since we have defined  $\langle (B_{ij}^p)^2 \rangle_c = 1$  and have normalized  $B_{ij}$  in the same manner in order to remove the changes due to the noise which only serve to redefine the acting temperature.

To find what value of  $g$  is sufficient to push the system from the target to the random globule phase, we need to know where the system is initially (i.e. for  $g = 1$ , same matrix for design and renaturation). When creating a matrix of species-species energies for proteins, such as the Miyazawa and Jernigan (MJ) matrix [Mia85], what one obtains is actually  $B_{ij}^{\text{MJ}} = B_{ij}/T_p$  (see Appendix B). Since from equation (8.14) we have,  $\langle B_{ij}^2 \rangle_c^{1/2} = s^{1/2}T_f$ , then the variance of the MJ matrix yields  $\langle (B_{ij}^{\text{MJ}})^2 \rangle_c^{1/2} = s^{1/2}T_f/T_p$ ; therefore, with the knowledge of the MJ matrix and the flexibility of proteins  $s$ , we also arrive at  $T_p/T_f \approx 0.9$ . It is also independently hypothesized [Gol92] that the ratio of the folding to the glass temperature should be about  $T_{\text{tar}}/T_f \approx 1.6$ ; using equation (8.16), this leads to a similar estimate for the degree of optimization  $T_p/T_f$ .

Therefore, for a value of  $T_p/T_f \approx 0.9$ , for  $g < 0.9$ , there is no chance for renaturation; as  $g$  approaches 0.9 from 1, one needs a lower  $T_{\text{tar}}$  to make the optimized ground state stable. However, even a conservative estimate of  $g \approx 0.95$  indicates that  $\sigma$  cannot exceed  $\approx 30\%$ . Thus, even a rather conservative minimum correlation factor  $g$  allows a large average error.

Also, note that this error limit is independent of the length of the polymer. Previous calculations [Bry94] have made estimates which are directly based upon  $N$  (i.e. the error must be small compared with  $1/\sqrt{N}$ ); these calculations were performed without design and furthermore detail the phase transition between the dominance of different conformations in the frozen phase, while our treatment is a comparison between the types of freezing (to the target or some random conformation). This is therefore independent of the length of the polymer chain and essentially of a different nature than that of Ref [Bry94]. Furthermore, within our formalism, the transition in  $y$ , the number of replicas in the target group, is first order; therefore, one cannot discuss aspects of renaturation with a given percentage of correct contacts: either there is renaturation to the target conformation or to some entirely



different conformation.

Within the framework of our formalism, the Independent Interaction Model can be recovered by addressing the limit  $q \rightarrow N$  and assuming that  $B$  is a normally distributed matrix; in this case eq (8.14) agrees with the results of more direct calculations of this model [Sha89a]. The error limits in this approximation are derived in exactly the same manner as (8.17). This is not surprising as, in fact, the validity of the approximation of taking the free energy to  $\mathcal{O}(B^2)$  in eq (8.12) is similar to that of the Independent Interaction Model [Pan95b]: we assume that the effective flexibility  $s$  of the polymer is small. However, our treatment allows corrections to this approximation to be systematically derived.

In conclusion, starting from the most general Hamiltonian involving short range binary heteropolymeric interactions, we have derived what measure is used to compare differences in interaction potentials and the limits in which renaturability to the target conformation is still allowed. Simple estimates of normally distributed error indicates that even conservative estimates leave room for 30% error in potentials. Using our formalism, one can make a more informed estimate based upon more precise knowledge of the form of errors involved, i.e. the correlations of errors in the matrix.

## 8.6 Simplification of Equation (7)

We will use a slightly different notation from the rest of the chapter to facilitate calculations: we eliminate indices and simply give the dimensionality of the operators explicitly, eg. we label  $\Delta_{ij}$  as  $\widehat{\Delta}^{(q)}$  since it is a  $q \times q$  dimensional matrix.

We perform the simplification of the elimination of replicas through several steps:

1.  $\widehat{q}$  is of well-known one-step replica symmetry breaking shape, with one distinct group of  $y + 1$  replicas and  $(n - y)/x$  groups of  $x$  replicas each.
2.  $\widehat{M} = \widehat{I} + 2\rho\widehat{q} \otimes \widehat{\Delta} \widehat{B}$  can be viewed as  $(n + 1) \times (n + 1)$  block matrix in replica space, with each matrix element being  $q \times q$  matrix in species space.

- This block matrix is of the same structure as  $\widehat{q}$ , with one  $(y + 1) \times (y + 1)$  super-block and  $(n - y)/x$  of  $x \times x$  super-blocks.
3. The determinant in the first term in free energy is decomposed into the product of determinants of super-blocks.
  4. Vector  $\vec{\rho}$  is composed of  $n + 1$  "blocks"  $p_i$ , thus making the second term in free energy the sum of independent contributions from the groups of replicas. Along with previous, this means that different groups of replicas do not interact and this is why they contribute independently to the free energy.
  5. Effective replica energy  $E$  is now presented in the form

$$\frac{E(x, y)}{N} = \epsilon_y + \frac{n - y}{x} \epsilon_x, \quad (8.18)$$

where  $\epsilon_y$  and  $\epsilon_x$  are the (independent) contributions from the corresponding groups of replicas. (Note, that replica entropy is also of the same form).

6. Both  $\epsilon_y$  and  $\epsilon_x$  have almost the same form as  $E$  (8.10), except simpler matrix  $\widehat{q}$ , with all matrix elements 1, appears instead of  $\widehat{q}$ :

$$\begin{aligned} \epsilon_z = & \frac{1}{2} \ln \det \left[ \widehat{I}^{(zq)} + 2\rho \widehat{q}^{(z)} \otimes \widehat{\Delta}^{(q)} \widehat{B}^{(zq)} \right] + \\ & + \frac{1}{\rho} \left\langle \vec{\rho} \left| \widehat{B}^{(zq)} \left[ \widehat{I}^{(zq)} + 2\rho \widehat{q}^{(z)} \otimes \widehat{\Delta}^{(q)} \widehat{B}^{(zq)} \right]^{-1} \right| \vec{\rho} \right\rangle, \end{aligned} \quad (8.19)$$

where  $z$  is either  $x$  or  $y + 1$ , i.e., the number of replicas in the group.

7. To simplify first term (with determinant), we define rotation unitary operator

$$\widehat{\mathcal{R}}_{\alpha\beta}^{(z)} = \frac{1}{\sqrt{z}} \exp \left[ \frac{2\pi i}{z} (\alpha - 1)(\beta - 1) \right] \quad 1 \leq \alpha, \beta \leq z. \quad (8.20)$$

It is easy to check that this operator transforms  $\widehat{q}^{(z)}$  into diagonal form, where one diagonal matrix element is 1, while all others are 0:

$$\widehat{\mathcal{R}}^{(z)} \widehat{q}^{(z)} \left( \widehat{\mathcal{R}}^{(z)} \right)^{-1} = \widehat{\lambda}^{(z)}, \quad \text{where} \quad \widehat{\lambda}_{\alpha\beta} = z \delta_{\alpha 1} \delta_{1\beta}. \quad (8.21)$$

We define also  $\widehat{\mathcal{R}}^{(zq)} = \widehat{I}^{(q)} \otimes \widehat{\mathcal{R}}^{(z)}$  and note that the determinant is not changed upon rotation. We write

$$\begin{aligned}
& \det \left[ \widehat{I}^{(zq)} + 2\rho \widehat{q}^{(z)} \otimes \widehat{\Delta}^{(q)} \widehat{B}^{(zq)} \right] = \\
& = \det \left[ \widehat{\mathcal{R}}^{(zq)} \right] \det \left[ \widehat{I}^{(zq)} + 2\rho \widehat{q}^{(z)} \otimes \widehat{\Delta}^{(q)} \widehat{B}^{(zq)} \right] \det \left[ \widehat{\mathcal{R}}^{(zq)} \right]^{-1} \\
& = \det \left[ \widehat{I}^{(zq)} + 2\rho \left( \widehat{\mathcal{R}}^{(zq)} \right) \widehat{q}^{(z)} \otimes \widehat{\Delta}^{(q)} \widehat{B}^{(zq)} \left( \widehat{\mathcal{R}}^{(zq)} \right)^{-1} \right] \\
& = \det \left[ \widehat{I}^{(zq)} + 2\rho \left( \widehat{\mathcal{R}}^{(z)} \right) \widehat{q}^{(z)} \left( \widehat{\mathcal{R}}^{(z)} \right)^{-1} \otimes \widehat{\Delta}^{(q)} \left( \widehat{\mathcal{R}}^{(zq)} \right) \widehat{B}^{(zq)} \left( \widehat{\mathcal{R}}^{(zq)} \right)^{-1} \right] \\
& = \det \left[ \widehat{I}^{(zq)} + 2\rho \widehat{\lambda}^{(z)} \otimes \widehat{\Delta}^{(q)} \left( \widehat{\mathcal{R}}^{(zq)} \right) \widehat{B}^{(zq)} \left( \widehat{\mathcal{R}}^{(zq)} \right)^{-1} \right]. \tag{8.22}
\end{aligned}$$

As  $\widehat{B}^{(zq)}$  is diagonal in replica space,  $\widehat{B}^{(zq)} = \widehat{B}_\alpha^{(q)} \delta_{\alpha\beta}$ , we have

$$\begin{aligned}
& \left( \left( \widehat{\mathcal{R}}^{(zq)} \right) \widehat{B}^{(zq)} \left( \widehat{\mathcal{R}}^{(zq)} \right)^{-1} \right)_{\alpha\beta} = \sum_{\gamma\delta} \widehat{\mathcal{R}}_{\alpha\gamma} \widehat{B}_\gamma^{(q)} \delta_{\gamma\delta} \left( \widehat{\mathcal{R}}^{(zq)} \right)_{\delta\beta}^{-1} = \\
& = \frac{1}{z} \sum_{\gamma} \exp \left[ \frac{2\pi i}{z} (\alpha - \beta)(\gamma - 1) \right] \widehat{B}_\gamma^{(q)}. \tag{8.23}
\end{aligned}$$

Taking into account the simple structure of  $\widehat{\lambda}$  (8.21), we arrive at

$$\widehat{\lambda}^{(z)} \otimes \widehat{\Delta}^{(q)} \left( \widehat{\mathcal{R}}^{(zq)} \right) \widehat{B}^{(zq)} \left( \widehat{\mathcal{R}}^{(zq)} \right)^{-1} = \delta_{\alpha 1} \sum_{\gamma} \exp \left[ \frac{2\pi i}{z} (1 - \beta)(\gamma - 1) \right] \widehat{\Delta}^{(q)} \widehat{B}_\gamma^{(q)} \tag{8.24}$$

8. First consider a non-target group of  $z = x$  replicas. In this group, all the replicas are identical meaning that  $\widehat{B}_\gamma^{(q)} = \widehat{B}^{(q)}$  does not depend on replica index  $\gamma$ . This yields

$$\widehat{\lambda}^{(z)} \otimes \widehat{\Delta}^{(q)} \left( \widehat{\mathcal{R}}^{(zq)} \right) \widehat{B}^{(zq)} \left( \widehat{\mathcal{R}}^{(zq)} \right)^{-1} = x \delta_{\alpha 1} \delta_{1\beta} \widehat{\Delta}^{(q)} \widehat{B}^{(q)} \tag{8.25}$$

and thus

$$\det \left[ \widehat{I}^{(xq)} + 2\rho \widehat{q}^{(x)} \otimes \widehat{\Delta}^{(q)} \widehat{B}^{(xq)} \right] = \det \left[ \widehat{I}^{(q)} + 2\rho x \widehat{\Delta}^{(q)} \widehat{B}^{(q)} \right] \tag{8.26}$$

9. Consider now target group of  $z = y + 1$  replicas. In this case,  $\widehat{B}_\gamma^{(q)} = \widehat{B}_p^{(q)}$  for

$\gamma = 1$  and  $\widehat{B}_\gamma^{(q)} = \widehat{B}^{(q)}$  otherwise. We write therefore

$$\begin{aligned} \widehat{I}^{(zq)} + \widehat{\lambda}^{(z)} \otimes \widehat{\Delta}^{(q)} \left( \widehat{\mathcal{R}}^{(zq)} \widehat{B}^{(zq)} \left( \widehat{\mathcal{R}}^{(zq)} \right)^{-1} \right) = \\ = \widehat{I}^{(q)} \delta_{\alpha\beta} + \delta_{\alpha 1} \widehat{\Delta}^{(q)} \left( \widehat{B}_p^{(q)} - \widehat{B}^{(q)} \right) + (y+1) \delta_{\alpha 1} \delta_{1\beta} \widehat{\Delta}^{(q)} \widehat{B}^{(q)}. \end{aligned} \quad (8.27)$$

This is the block matrix of the peculiar form such that only upper block is non-zero in the first column; for that reason, its determinant is equal to the product of determinants of diagonal blocks (see Lemma 1). Thus,

$$\det \left[ \widehat{I}^{((y+1)q)} + 2\rho \widehat{q}^{(y+1)} \otimes \widehat{\Delta}^{(q)} \widehat{B}^{((y+1)q)} \right] = \det \left[ \widehat{I}^{(q)} + 2\rho \widehat{\Delta}^{(q)} \left( y \widehat{B}^{(q)} + \widehat{B}_p^{(q)} \right) \right]. \quad (8.28)$$

10. As to the second term in  $\epsilon_z$  (8.19), it is easily computed using Lemma 2. Indeed,  $\widehat{B}^{((y+1)q)}$  is block diagonal matrix with one block  $\widehat{B}_p^{(q)}$  and  $y$  others  $\widehat{B}^{(q)}$ . On the other hand,  $\widehat{q}^{(y+1)} \otimes \widehat{\Delta}^{(q)}$  is the block matrix with every block being the same  $\widehat{\Delta}^{(q)}$ . Therefore, the matrix in question,  $\left[ \widehat{I}^{(zq)} + 2\rho \widehat{q}^{(z)} \otimes \widehat{\Delta}^{(q)} \widehat{B}^{(zq)} \right]$ , is exactly of the form  $\widehat{V}_{\widehat{g}, \widehat{h}}^{(z)}$  form, where  $\widehat{g} = \widehat{\Delta}^{(q)} \widehat{B}_p^{(q)}$  and  $\widehat{h} = \widehat{\Delta}^{(q)} \widehat{B}^{(q)}$ . Using block matrix multiplication rule, it is easy to compute  $\widehat{B}^{((y+1)q)} \widehat{V}_{\widehat{e}, \widehat{f}}^{(y+1)}$  (see Lemma 2) and then to use the result of Lemma 3. This finally gives

$$\begin{aligned} \frac{1}{\rho} \left\langle \vec{\rho} \left| \widehat{B}^{(zq)} \left[ \widehat{I}^{(zq)} + 2\rho \widehat{q}^{(z)} \otimes \widehat{\Delta}^{(q)} \widehat{B}^{(zq)} \right]^{-1} \right| \vec{\rho} \right\rangle = \\ = \rho \left\langle \vec{p} \left| \left( \widehat{B}_p^{(q)} + y \widehat{B}^{(q)} \right) \left[ \widehat{I}^{(q)} + 2\rho y \widehat{\Delta}^{(q)} \widehat{B}^{(q)} + 2\rho \widehat{\Delta}^{(q)} \widehat{B}_p^{(q)} \right]^{-1} \right| \vec{p} \right\rangle. \end{aligned} \quad (8.29)$$

11. Similar expression for a non-target group of  $x$  replicas can be derived from here by formally putting  $\widehat{B}_p^{(q)} \rightarrow \widehat{B}^{(q)}$  and  $y+1 \rightarrow x$ , this gives

$$\frac{1}{\rho} \left\langle \vec{\rho} \left| \widehat{B}^{(zq)} \left[ \widehat{I}^{(zq)} + 2\rho \widehat{q}^{(z)} \otimes \widehat{\Delta}^{(q)} \widehat{B}^{(zq)} \right]^{-1} \right| \vec{\rho} \right\rangle = \rho \left\langle \vec{p} \left| x \widehat{B}^{(q)} \left[ \widehat{I}^{(q)} + 2\rho x \widehat{\Delta}^{(q)} \widehat{B}^{(q)} \right]^{-1} \right| \vec{p} \right\rangle \quad (8.30)$$

### Lemma 1.

Consider an auxiliary problem of the matrix

This is block matrix, where  $\hat{g}$  is  $q \times q$  matrix and  $\hat{I}$  is identity matrix of the same size  $q \times q$ . The question is to find the determinant of this matrix.

It can be shown by expansion over the elements of the first column, then over the elements of the first column of the remaining minor, and by repeating this operation  $q$  times, that

$$\det [\hat{U}_g^{(z)}] = \det \hat{g} \tag{8.31}$$

independently of the blocks placed in the upper-right triangle (shown conventionally with question marks).

**Lemma 2.**

Consider another auxiliary problem of the following block matrix:

$$\begin{array}{c}
 \underbrace{\hspace{10em}}_z \\
 \left. \begin{array}{c}
 \begin{array}{|c|c|c|c|c|c|c|}
 \hline
 l+g & h & h & h & \dots & h & h \\
 \hline
 g & l+h & h & h & \dots & h & h \\
 \hline
 g & h & l+h & h & \dots & h & h \\
 \hline
 g & h & h & l+h & \dots & h & h \\
 \hline
 \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
 \hline
 g & h & h & h & \dots & l+h & h \\
 \hline
 g & h & h & h & \dots & h & l+h \\
 \hline
 \end{array} \\
 \end{array} \right\} N \\
 = V_{g,h}^{(z)}
 \end{array}$$

Here  $\hat{g}$  and  $\hat{h}$  are matrices  $q \times q$ , they generally do not commute to each other.  $\hat{I}$  is identity matrix of the same size  $q \times q$ . Total size of the block matrix  $\hat{V}_{g,h}^{(z)}$  is, therefore,  $zq \times zq$ . The question is to find inverse of the matrix  $\hat{V}_{g,h}^{(z)}$ .

It turns out that this inverse is in fact the matrix of the same structure, namely

$$\left( \hat{V}_{g,h}^{(z)} \right)^{-1} = \hat{V}_{\hat{e},\hat{f}}^{(z)}, \quad \text{where} \\
 \hat{e} = - \left( \hat{I} + (z-1)\hat{h} + \hat{g} \right)^{-1} \hat{g} \quad \text{and} \quad \hat{f} = - \left( \hat{I} + (z-1)\hat{h} + \hat{g} \right)^{-1} \hat{h} \quad (8.32)$$

The result can be easily proved using block matrix multiplication rule.

**Lemma 3.**

Consider an auxiliary problem of the scalar product

$$\left\langle \vec{\rho}^{(qz)} \mid \widehat{W}^{(qz)} \mid \vec{\rho}^{(qz)} \right\rangle,$$

where  $\vec{\rho}^{(qz)} = \vec{p}^{(q)} \otimes \vec{i}^{(z)} = p_i$  (does not depend on replica indices  $\alpha$ ), and  $\widehat{W}^{(qz)}$  is block matrix comprised of blocks  $\widehat{W}_{\alpha\beta}^{(q)}$ . Obviously, this scalar product is reduced to the scalar products of smaller dimensionality  $q$ , that is, purely in species space,

summed over all the blocks of the matrix:

$$\langle \bar{\rho}^{(qz)} | \widehat{W}^{(qz)} | \bar{\rho}^{(qz)} \rangle = \left\langle \bar{p}^{(q)} \left| \left( \sum_{\alpha\beta} \widehat{W}_{\alpha\beta}^{(q)} \right) \right| \bar{p}^{(q)} \right\rangle. \quad (8.33)$$

## 8.7 Relationship between the average number of species-species contacts and the interaction matrix

Note that this relation can be easily derived directly from our formalism as well: The Hamiltonian (8.1) can also be expressed directly in terms of the number of contacts  $n_{ij}$  between monomers of species  $i$  and  $j$ :  $\mathcal{H} = \sum_{ij}^q B_{ij} n_{ij}$ , where we have previously substituted  $n_{ij} = \sum_{I \neq J}^N \delta_{s_I, i} \delta_{s_J, j} \delta(\mathbf{r}_I - \mathbf{r}_J)$ .

Therefore, the average number of contacts can be directly calculated in terms of the derivative of the free energy with respect to  $B_{ij}$ . However, at this point, we must indicate one point in which we have been a bit cavalier in our previous derivation. Specifically, in order to perform the Hubbard-Stratonovich transformation, we have summed over all pairs of monomers  $\sum_{I, J}$  instead of only the different pairs  $\sum_{I \neq J}$ . This overcounting of self-site interaction leads to a spurious term in the free energy  $\widehat{\Delta} \widehat{B}$ . Excluding this term from the free energy, which is equivalent to performing the sum  $\sum_{I \neq J}$ , carrying terms in the free energy to  $\mathcal{O}(B^2)$ , and taking the derivative with respect to  $B_{ij}$  yields

$$\langle n_{ij} \rangle = p_i p_j \left( 1 - \frac{B_{ij}}{T_m} \right) \approx p_i p_j \exp \left( -\frac{B_{ij}}{T_m} \right) \quad (8.34)$$

where  $T_m$  is the matrix selective temperature in the sense of Ref [Fin93], i.e. either  $T_m = T_p$  for chains in the target phase,  $T_m = T_f$  for chains in the frozen phase, or  $T_m = T$  for chains in the random phase.





## Chapter 9

# Designed Heteropolymer in an External Field

In previous chapters, a procedure to create renaturable heteropolymers, “Imprinting,” has been proposed and examined theoretically. The significance of Imprinting is that certain aspects of a heteropolymer’s native conformation may be controlled during the synthesis stage. We examine this possibility theoretically by introducing an external field during the synthesis and renaturation stages of the model. We find that Imprinting in an external field leads to protein-like heteropolymers which can renature to native conformations which are affected by the field, even in the absence of the field during renaturation. We conclude by commenting on the relevance of these results to the biological and prebiological creation of biopolymers, such as proteins, influenced by the analogs to our external field, such as antigens or ligands.

### 9.1 Introduction

Disordered polymers are one of the most important objects in the physics of disordered systems, mainly because of the potential biological applications. Among other disordered polymeric systems, such as branched polymers and knots, two have

acquired the most attention in recent years: heteropolymers, linear chains with an uneven sequence of different links, and homopolymers situated in disordered environment, such as a white-noise external field.

The main physical peculiarity of heteropolymers is the frustration imposed by the conflicting requirements of the segregation of different monomers in space due to monomer-monomer volume interactions and the connection of monomers due to the polymeric bonds. When interactions are strong enough, freezing behavior similar to the one observed in spin glasses is found. The frozen phase of heteropolymers is dominated by one or very few conformations, or chain folds, that are minimally frustrated [Bry87].

For a homopolymer in a disordered medium, there are also several models to be mentioned. The simplest one views an ideal chain (without excluded volume or other volume interaction between monomers) looking for the deepest potential well. This is described in [Edw88,Cat88]. Not surprisingly, the polymer in this model collapses to microscopic size independent of the chain length. In a more realistic model, this pathological indefinite collapse is prevented by the monomer excluded volume, and the corresponding conformations are described in [Obu90,Hon90]. In this case, frustration is also imposed by the linear connections between chain monomers, the conflicting tendencies being the placement of monomers in the deepest possible wells of the potential (to keep polymer density below the densely packed maximum) and the maintenance of prescribed distances between monomers.

Our aim in this chapter is to consider a generalized model, where both types of disorder are presented simultaneously: a heteropolymer with frozen sequence of links in the disordered external field. The appealing property of this model is that heterogeneity enters twice into the system, first because volume interactions of monomers are of a heteropolymeric fashion and second because different monomers feel the external field also in different ways.

For the sake of simplicity, we restrict ourselves with the simplest assumption of a dense packed system. This means that our polymer is closely packed into the box which is supported by some external pressure, so that density of the system remains

spatially uniform, no matter what is the corresponding energy of the external field. For this system, the behavior in one extreme is known: if the external field is negligibly small (as compared with volume interactions), very few conformations will be frozen out at low temperatures because of the normal heteropolymer freezing transition [Pan95b,Sfa93]. On the other hand, a strong external field imposed on the system with weak volume interactions can be also expected to cause freezing of some distinct conformations — the ones that fit best to the field configuration. What is important, however, is that these two small sets of conformations are generally completely different. This means that sufficiently strong external field destroys the freezing of heteropolymer to the conformation dictated by its sequence.

Another important aspect of the problem is which sequences of monomers we are speaking about. This is to be taken seriously, because sequences are responsible for coding functions in biopolymers and therefore the adequacy of random sequences to model real ones is at least questionable. To this end, two ways to model real sequences were recently suggested [Sha93b,Pan94d,Pan95b]. Even though there are important differences between the two, they both employ the idea to form sequences thermodynamically. Speaking now of an external field, we can consider this field effecting sequence formation, or polymer folding with an already formed sequence, or both. All these possibilities are of great interest, as an external field can represent (to a schematic approximation) some target molecules or ligands, which are either used to control some desirable properties of the sequence, such as presence of an appropriate active site in the “native” conformation, or influence renaturation processes, etc.

We believe that the above mentioned models of sequence design are of great interest for the understanding of biopolymers. In this context, the incorporation of the external field in the model allows us to approach various questions related to the design procedure: suppose, we form the sequence under the action of the field; will it be able to renature without the field? Or vice versa — if the sequence is formed without any field, will the field help or destroy the renaturation? Or what happens to renaturation if acting field is opposite to the one presented in the polymerization

process? We will address these questions below.

We examine systematically the model where

1. heteropolymer has two types of monomers (“black-and-white model”) with Ising type interactions;
2. overall polymer density is maintained such that polymer volume fraction is always one;
3. external field is modeled as quenched random  $\delta$ -correlated potential;
4. interactions are considered unchanged at the stages of sequence formation and of chain folding.

## 9.2 The model

We model heteropolymeric monomer-monomer and monomer-field interactions with the Hamiltonian

$$\mathcal{H} = -B \sum_{I,J} s_I s_J \delta(\mathbf{r}_I - \mathbf{r}_J) - h \sum_I s_I \sigma(\mathbf{r}_I) , \quad (9.1)$$

where  $\sigma(\mathbf{x})$  is the external field. All homopolymeric contributions, such as excluded volume virial coefficients, are omitted for brevity, as they do not couple to heteropolymeric contributions. We repeat, however, that polymer density is kept constant.

Because of self-averaging, and since both the sequence and external field are quenched in each existing chain, the relevant free energy is to be averaged over an ensemble of sequences and external fields  $\sigma$ :

$$F = - \int \mathcal{D}\sigma \sum_{\text{seq}} \mathcal{P}[\text{seq}, \sigma] \ln Z[\text{seq}, \sigma] , \quad (9.2)$$

where  $Z[\text{seq}, \sigma]$  is the corresponding partition function. Ignoring for the moment the technical question how to perform this average, we note that the two elements

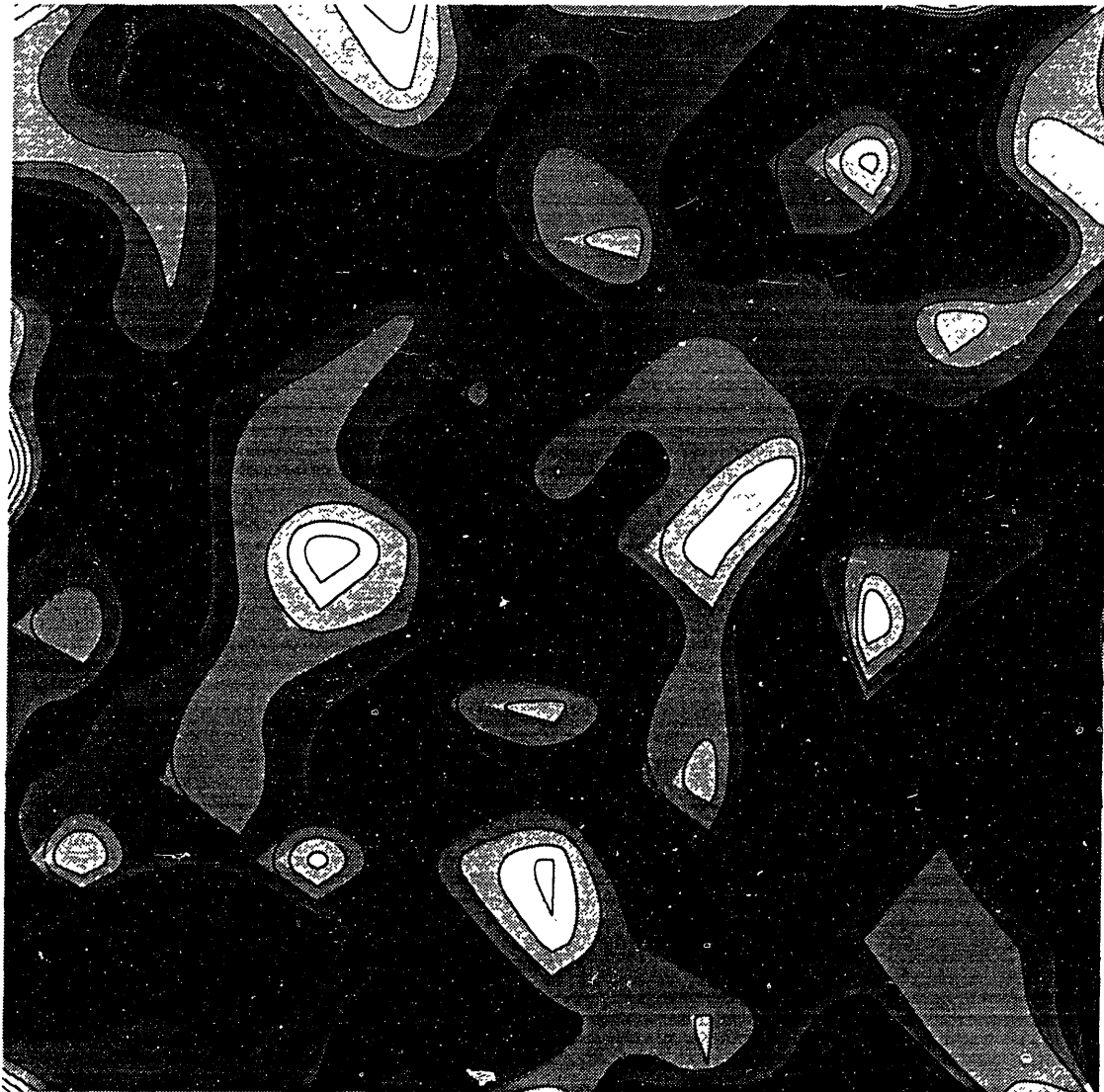


Figure 9-1: Cartoon contour plot of a random field. For a polymer which consists of only two types of monomers (“black” and “white”), we examine Ising interactions for monomer-monomer and monomer-field interactions. Even without monomer-monomer interactions, the polymeric bonds cause frustration since they prevent the monomers from matching “colors” with the field.

of frozen disorder presented here, the sequence and the field, play a considerably different role. Indeed, we are considering the model in which the sequence is formed by a special design procedure and therefore may or may not be dependent on the external field. Moreover, the external field which acts during the chain polymerization may generally be different from the field acting on the already prepared chain. To take care of this fact, we write

$$\mathcal{P}[\sigma, \text{seq}] = \int \mathcal{D}\sigma_p \mathcal{P}[\sigma, \sigma_p] \times \mathcal{P}_{\{\sigma_p\}}[\text{seq}] , \quad (9.3)$$

where  $\mathcal{P}_{\mathcal{B}}[\mathcal{A}]$  stands for *conditional* probability of  $\mathcal{A}$  under the condition  $\mathcal{B}$ .  $\mathcal{P}_{\{\sigma_p\}}[\text{seq}]$  is really the distribution of sequences, and it is dictated by the design procedure. As both of the known procedures to design sequences are based on the equilibration of either the monomer soup in the real space [Pan94d, Pan95b] or the polymer in the sequence space [Sha93b], the distribution of sequences is given as corresponding Boltzmann distribution, and it is therefore proportional to

$$\mathcal{P}_{\{\sigma_p\}}[\text{seq}] = Z_p[\text{seq}, \sigma_p] , \quad (9.4)$$

where  $Z_p[\text{seq}, \sigma_p]$  is the partition function of the polymerization system. Hereafter, we omit all irrelevant normalization constants.

We now employ the replica trick

$$\langle \ln Z \rangle = \lim_{n \rightarrow 0} \frac{\langle Z^n \rangle - 1}{n} \quad (9.5)$$

to perform the average in (9.2). Collecting equations (9.2) through (9.5) together, we get

$$F = \lim_{n \rightarrow 0} \frac{1}{n} \left( \int \mathcal{D}\sigma \mathcal{D}\sigma_p \mathcal{P}[\sigma, \sigma_p] \sum_{\text{seq}} Z^n[\text{seq}, \sigma] Z_p[\text{seq}, \sigma_p] - 1 \right) . \quad (9.6)$$

The structure of this expression allows to consider preparation state as an additional  $n + 1$  replica, albeit with its own temperature and some of Hamiltonian parameters.

To see it, we write

$$Z_r[\text{seq}, \sigma_r] = \sum_{\text{conformations}} \exp \left\{ -\frac{1}{T_r} \mathcal{H}_r[\text{seq}, \sigma_r, \text{conformation}] \right\}, \quad (9.7)$$

where index  $r$  may be absent for replicas  $1, \dots, n$  and it stands for  $p$  (“polymerization”) for additional replica 0.

For simplicity, we do not consider various cases of statistical interdependencies of the fields  $\sigma$  and  $\sigma_p$ . Furthermore, we consider both to be  $\delta$ -correlated white noise, such that

$$\langle \sigma_r(\mathbf{R}) \rangle = 0; \quad \langle \sigma_r(\mathbf{R}) \sigma_r(\mathbf{R}') \rangle = w^2 \delta(\mathbf{R} - \mathbf{R}'). \quad (9.8)$$

As we assume that the value of the field at one position is uncorrelated with its value at a different position, the probability distribution of  $\sigma$  is gaussian:

$$\mathcal{P}[\sigma, \sigma_p] = \delta[\sigma(\mathbf{R}) - \sigma_p(\mathbf{R})] \cdot \exp \left\{ -\int d\mathbf{R} w^{-2} \sigma(\mathbf{R})^2 \right\} \quad (9.9)$$

where  $w$  controls the width of the probability distribution of the external field. Even though we consider  $\sigma$  and  $\sigma_p$  to be strongly correlated, we can examine several physical situations by choosing various combinations of  $h$  and  $h_p$  in the Hamiltonians  $\mathcal{H}$  and  $\mathcal{H}_p$  in the equation (9.7):

1. If the field effects chain design and folding in the same way, we take  $h = h_p$ ;
2. If the field is presented during design only, we take  $h = 0, h_p \neq 0$ ;
3. By contrast, if the field is presented for the existing prepared chain only, when the chain folds, then  $h \neq 0, h_p = 0$ ;
4. The field can affect system during folding stage in the *opposite direction* compared to design stage, in this case  $h = -h_p \neq 0$ .

Thus, to gain physical insight into the system, it should be enough to consider the simplest probability distribution for the external field (9.9), but taking into account general situation with respect to different  $h$  and  $h_p$ .

Since the interactions in the monomer soup are the same as those found in the polymer, the parts of the Hamiltonians  $\mathcal{H}$  and  $\mathcal{H}_p$  describing interactions should be identical, namely, there should be the same  $B$ .

Thus, the  $(n + 1)$ -replica partition function has the following form

$$Z^{n+1} = \sum_{\text{conf}} \exp \left\{ \int d\mathbf{R}_1 d\mathbf{R}_2 \sum_{\alpha=0}^n \sum_{I,J}^N \frac{B}{T_\alpha} s_I s_J \delta(\mathbf{r}_I^\alpha - \mathbf{R}_1) \delta(\mathbf{r}_J^\alpha - \mathbf{R}_2) \delta(\mathbf{R}_1 - \mathbf{R}_2) \right. \\ \left. + \int d\mathbf{R} \sum_{\alpha=0}^n \sum_I^N \frac{h_\alpha}{T_\alpha} s_I \sigma(\mathbf{R}) \delta(\mathbf{r}_I^\alpha - \mathbf{R}) - \int d\mathbf{R} w^{-2} \sigma(\mathbf{R}) \right\} \quad (9.10)$$

where  $h_\alpha$  and  $T_\alpha$  is defined according to

$$h_\alpha = \begin{cases} h_p & \text{for } \alpha = 0 \\ h & \text{for } \alpha > 0 \end{cases} \quad \text{and} \quad T_\alpha = \begin{cases} T_p & \text{for } \alpha = 0 \\ T & \text{for } \alpha > 0 \end{cases} \quad (9.11)$$

As everywhere in this chapter, we drop all the normalization constants.

We go from spins to fields by performing the Hubbard-Stratonovich transformation on the quantity  $\sum_I^N s_I \delta(\mathbf{r}_I^\alpha - \mathbf{R})$  and average over the sequences and external field to get

$$\langle Z^{n+1} \rangle = \sum_{\text{conf}} \int \mathcal{D}\{\phi\} \mathcal{D}\{\sigma\} \exp \left\{ - \int d\mathbf{R} \left[ w^{-2} \sigma(\mathbf{R})^2 + \sum_{\alpha=0}^n \frac{T_\alpha}{4B} \phi_\alpha(\mathbf{R})^2 \right] \right. \\ \left. + \sum_I \ln \cosh \left[ \sum_{\alpha=0}^n \int d\mathbf{R} \delta(\mathbf{r}_I^\alpha - \mathbf{R}) \left( \phi_\alpha(\mathbf{R}) + \frac{h_\alpha}{T_\alpha} \sigma(\mathbf{R}) \right) \right] \right\} \quad (9.12)$$

We can expand the  $\ln \cosh$  to  $\mathcal{O}(\phi^2, h^2)$  to get

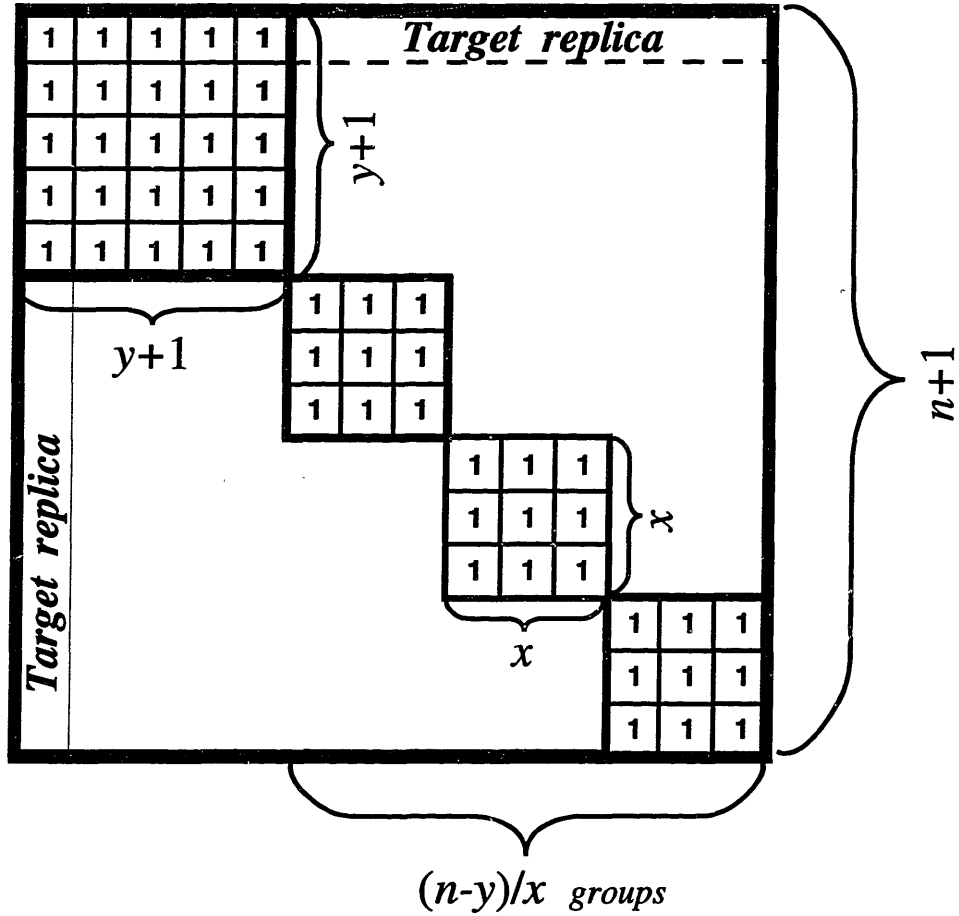
$$\langle Z^{n+1} \rangle = \sum_{\text{conf}} \int \mathcal{D}\{\phi\} \mathcal{D}\{\sigma\} \times \\ \times \exp \left\{ - \int d\mathbf{R}_1 d\mathbf{R}_2 \sigma(\mathbf{R}_1) \sigma(\mathbf{R}_2) \left[ w^{-2} \delta(\mathbf{R}_1 - \mathbf{R}_2) - \frac{1}{2} \sum_{\alpha,\beta=0}^n \frac{h_\alpha}{T_\alpha} Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) \frac{h_\beta}{T_\beta} \right] \right. \\ - \int d\mathbf{R}_1 d\mathbf{R}_2 \sum_{\alpha,\beta=0}^n \phi_\alpha(\mathbf{R}_1) \phi_\beta(\mathbf{R}_2) \left[ \frac{T_\alpha}{4B} \delta_{\alpha\beta} \delta(\mathbf{R}_1 - \mathbf{R}_2) - \frac{1}{2} Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) \right] \\ \left. + \int d\mathbf{R}_1 d\mathbf{R}_2 \sum_{\alpha,\beta=0}^n Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) \phi_\alpha(\mathbf{R}_1) \frac{h_\beta}{T_\beta} \sigma(\mathbf{R}_2) \right\} \quad (9.13)$$



where  $Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) \equiv \sum_I \delta(\mathbf{r}_I^\alpha - \mathbf{R}_1) \delta(\mathbf{r}_I^\beta - \mathbf{R}_2)$  is the conformation correlator between replicas [Sha89a, Pan95b]. This expression (9.13) is rather cumbersome, but it can be substantially simplified by noting that for a polymer in  $3D$  the one step replica symmetry breaking scheme is valid, as it was first noted in [Sha89a]. This result holds true for the case at hand, where an external field is presented, as can be easily shown by reproducing arguments of [Sha89a] in the form of [Pan95b]. In the one step replica symmetry breaking, the free energy is minimized for the correlator  $Q$  such that two replicas either have complete overlap or do not overlap at all [Sha89a, Pan95b]. Thus, this corresponds to the form

$$Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2) = \rho q_{\alpha\beta} \delta(\mathbf{R}_1 - \mathbf{R}_2) , \quad (9.14)$$

where  $\rho$  is the density of the system and  $q_{\alpha\beta}$  is a  $(n + 1) \times (n + 1)$  matrix of the single step replica symmetry breaking form:



There are two ways in which replica symmetry is broken: 1) spontaneously, in which frustrations lead to certain conformations which have differing energies and 2) due to the the selection procedure which explicitly breaks replica symmetry. We parameterize  $q_{\alpha\beta}$  in terms of the number of replicas  $y$  which overlap with replica 0 (and therefore have the polymerization conformation), and the  $(n - y)/x$  groups of replicas which each overlap with  $x$  replicas due to spontaneous replica symmetry breaking.

For further simplification, it is useful to substitute  $\phi_\alpha \rightarrow \phi_\alpha 2\sqrt{\rho B/T_\alpha}$  (still omitting the irrelevant factors in front of the integral) and to use bra ket vector and matrix notations [Pan95b] (where the dimensionality of the vector space is  $n + 1$ ):

$$\langle Z^{n+1} \rangle = \sum_{\text{conf}} \left[ \int d^{(n+1)} \vec{\phi} d\sigma \times \right.$$

$$\times \exp \left\{ -\sigma^2 \left[ \frac{1}{\rho w^2} - \frac{1}{2} \langle \vec{h} | \hat{q} | \vec{h} \rangle \right] - \langle \vec{\phi} | \hat{I} - 2\rho B \hat{T}^{-1} \hat{q} | \vec{\phi} \rangle + 2\sqrt{\rho B} \langle \vec{\phi} | \hat{T}^{-1/2} \hat{q} | \sigma \vec{h} \rangle \right\}^N \quad (9.15)$$

We evaluate this Gaussian integral, which yields  $\langle Z^{n+1} \rangle = \exp(-E + S)$ , where

$$E = \frac{1}{2} \ln \det \left[ \hat{I} - 2\rho B \hat{q} \hat{T}^{-1} \right] + \frac{1}{2} \ln \left[ \frac{1}{\rho w^2} - \left\langle \vec{h} \left| \hat{q} B \rho \hat{T}^{-1} \left( \hat{I} - 2\rho B \hat{q} \hat{T}^{-1} \right)^{-1} \hat{q} + \frac{1}{2} \hat{q} \right| \vec{h} \right\rangle \right], \quad (9.16)$$

where we have taken into account that  $\hat{T}$  and  $\hat{q}$  do commute to each other.  $S$  is the entropy due to the transformation between the sum over conformations and functional integration over  $Q_{\alpha\beta}(\mathbf{R}_1, \mathbf{R}_2)$ .

To simplify further, we need the eigenvalues and eigenvectors of the  $\hat{M} = \hat{I} - 2\rho B \hat{q} \hat{T}^{-1}$  matrix. These have been previously calculated [Pan95b]. These calculations are facilitated by the fact that all of the matrices are block matrices and the associated scalar products can be calculated for each block, then summed. In terms of  $x$  and  $y$ , we find

$$\begin{aligned} \left\langle \vec{h} \left| \hat{q} B \rho \hat{T}^{-1} \left( \hat{I} - 2\rho B \hat{q} \hat{T}^{-1} \right)^{-1} \hat{q} \right| \vec{h} \right\rangle &= B\rho \left( \frac{y}{T} + \frac{1}{T_p} \right) \left( \frac{yh}{T} + \frac{h_p}{T_p} \right)^2 \left[ 1 - 2B\rho \left( \frac{y}{T} + \frac{1}{T_p} \right) \right]^{-1} \\ &\quad + \frac{n-y}{x} \left[ \frac{(B\rho x/T)(hx/T)^2}{1 - 2B\rho x/T} \right] \end{aligned} \quad (9.17)$$

One can easily show that

$$\langle \vec{h} | \hat{q} | \vec{h} \rangle = \left[ \left( \frac{h_p}{T_p} \right)^2 + 2y \left( \frac{hh_p}{TT_p} \right) + \left( \frac{yh}{T} \right)^2 + \frac{n-y}{x} \left( \frac{xh}{T} \right)^2 \right] \quad (9.18)$$

and it has been previously shown that [Pan95b]

$$\ln \det \left[ \hat{I} - 2\rho B \hat{q} \hat{T}^{-1} \right] = \frac{n-y}{x} \ln \left[ 1 - \frac{2B\rho}{T} x \right] + \ln \left[ 1 - 2B\rho \left( \frac{y}{T} + \frac{1}{T_p} \right) \right] \quad (9.19)$$

We have now written the energy entirely in terms of the new scalar order pa-

rameters  $x$  and  $y$ . We can do the same for the entropy [Sfa93,Pan95b]:

$$S = Ns \left[ \frac{n-y}{x}(x-1) + y \right] \quad (9.20)$$

where  $s = \ln(a^3/v)$ . Thus, we can now write the free energy entirely in terms of  $x$  and  $y$ :

$$\frac{F}{N} = \frac{1}{2} \ln \left[ \frac{1}{\rho w^2} - \zeta \right] + \frac{1}{2} \frac{n-y}{x} \ln \left[ 1 - \frac{2B\rho}{T} x \right] + \frac{1}{2} \ln \left[ 1 - 2B\rho \left( \frac{y}{T} + \frac{1}{T_p} \right) \right] + s \left[ n - \frac{n-y}{x} \right] \quad (9.21)$$

where

$$\begin{aligned} \zeta = & B\rho \left( \frac{y}{T} + \frac{1}{T_p} \right) \left( \frac{yh\rho}{T} + \frac{h_p\rho}{T_p} \right)^2 \left[ 1 - 2B\rho \left( \frac{y}{T} + \frac{1}{T_p} \right) \right]^{-1} + \frac{n-y}{x} \left[ \frac{(B\rho x/T)(h\rho x/T)^2}{1 - 2B\rho x/T} \right] \\ & + \frac{1}{2} \left[ \left( \frac{yh\rho}{T} + \frac{h_p\rho}{T_p} \right)^2 + \frac{n-y}{x} \left( \frac{h\rho}{T} \right)^2 \right] \end{aligned} \quad (9.22)$$

To find the temperature at which the system freezes into random conformations, we optimize  $F$  with respect to  $x$ . Note that the  $n \rightarrow 0$  limit must be kept in mind during these calculations, i.e. the free energy should be linearized in terms that are of  $\mathcal{O}(n)$  (i.e.  $n$  and  $y$ ). We find a solution similar to the zero field case [Sfa93,Pan95b]

$$x = \begin{cases} \xi_f T / \rho & \text{for } \xi_f T / \rho \leq 1 \\ 1 & \text{for } \xi_f T / \rho \geq 1 \end{cases} \quad (9.23)$$

where  $\xi_f$  is the solution to the equation

$$2s = \ln(1 - 2B\xi_f) + \frac{2B\xi_f}{1 - 2B\xi_f} + \frac{(h\xi_f)^2}{\Gamma(1 - 2B\xi_f)^2}, \quad \Gamma = \frac{2}{\rho w^2} - (h_p \xi_p)^2 - \frac{2Bh_p \xi_p^2}{1 - 2B\xi_p}, \quad (9.24)$$

and  $\xi_p = \rho/T_p$ . If we expand for small  $\xi$  and  $\xi_p$ , we get  $T_f = \rho \sqrt{(B^2 + h^2 \rho w^2 / 4) / s}$ .

Thus, there are two sources for freezing: the external field and the polymer interaction. Thus, even in the case where the chain is a homopolymer with respect to volume interactions ( $B = 0$ ), but has heteropolymeric interactions with the external field ( $h \neq 0$ ), freezing occurs due to the desire to place monomers in low energy

positions with respect to the field and the polymeric bonds which frustrate this goal. Moreover, in the limit in which there is no field during Imprinting ( $h_p = 0$ ) and polymer-polymer interactions are negligible compared to the external field ( $h \gg B$ ), we can exactly find the freezing temperature  $T_f = h\rho^{3/2}w(4s)^{-1/2}$ ; it is clear that the external field contributes to frustrations and therefore leads to freezing.

We now examine the transition to target group. As there are no extrema within the region  $y = 0 \dots n$ , the free energy is maximized at the boundary, i.e. either at  $y = 0$  (no replicas overlap with the target replica) or  $y = n$  (all replicas overlap with the target replica and therefore the polymer renatures to the designed conformation). To find which value of  $y$  maximizes the free energy, we linearize the free energy in  $y$  and examine the condition where the slope of the free energy with respect to  $y$  changes sign:

$$2s = \ln(1 - 2B\xi) + \frac{2B\xi}{1 - 2B\xi_p} - \frac{1}{\Gamma} \left[ h^2\xi^2 - 2hh_p\xi\xi_p + \frac{2Bh^2\xi^3}{1 - 2B\xi} + \frac{2B\xi\xi_p(2Bh_p\xi_p - h - h_p)}{1 - 2B\xi_p} \right] \quad (9.25)$$

If we expand for small  $\xi$ 's, which corresponds to the small  $s$  (flexible chain) limit, we get the relation

$$\xi^2 + \xi_f^2 = 2g\xi\xi_p, \quad \text{where } g = \frac{4B^2 + 2B(h + h_p) + hh_p}{4B^2 + h^2} \quad (9.26)$$

Thus, we have

$$T_p^c = \begin{cases} 2gT/(1 + T^2/T_f^2) & \text{for } T \geq T_f \\ gT_f & \text{for } T \leq T_f \end{cases} \quad (9.27)$$

To this order, we find that  $T_p^c$  is simply the zero field ( $h = h_p = 0$ ) case multiplied by  $g$  and with a modified  $T_f$  (since  $T_f$  is a function of  $h$  to this order). Note that in the limit  $B \rightarrow 0$ , (9.26) and (9.27) are exact.

### 9.3 Discussion

We have found that in the flexible chain limit, the external field case is a simple generalization of the zero field case, simply with a newly defined freezing temperature

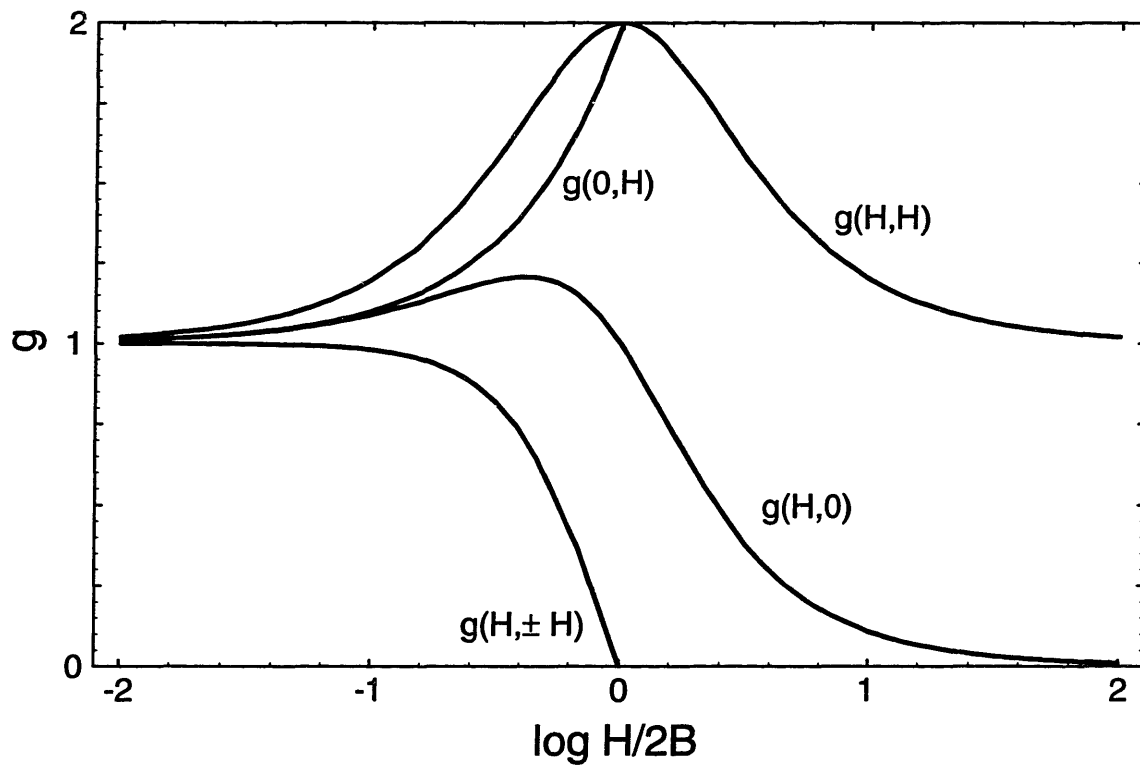


Figure 9-2: In the flexible chain limit, the effect of the external field during design ( $h_p$ ) and/or during renaturation ( $h$ ) enters into the theory as a rescaling of the polymerization temperature  $T_p$ . The effective polymerization temperature is given by  $T_p \rightarrow T_p/g(h, h_p)$ . We plot  $g$  for the 4 cases addressed in the discussion.

$T_f$  and a factor  $g(h, h_p)$  in the definition of the threshold polymerization temperature  $T_p^c$ . To examine these modifications, we must first look at the behavior of  $g(h, h_p)$ , shown in Fig. 9-2.

We examine the following cases:

1. Imprinting and Renaturation in the presence of the field ( $h = h_p$ )

We see that for the same field strength during Imprinting and renaturation, freezing to the target conformation is enhanced, as  $g(h, h) > 1$ , and therefore the threshold polymerization temperature  $T_p^c$  is greater than the  $h = h_p = 0$  case. In fact, for  $h = h_p = 2B$ , we find that the frustration is maximized, as both the polymer-polymer and polymer-field interactions contribute equally. At very high field ( $h \gg B$ ), the contribution from the polymer-polymer interactions are negligible and there is freezing solely due to the external field; thus, in this limit  $g \rightarrow 1$ .

2. Imprinting with the field, but Renaturation without the field ( $h = 0, h_p \neq 0$ )

We see trivially from (9.26) that in the limit,  $g = 1 + h_p/2B$ . Thus even without the external field during renaturation, the polymer renatures to the polymerization conformation. This is crucial to the molecular recognition ability of Imprinted polymers, as we would require the polymer to fold to its native state in order to recognize the external field. As one would expect, we can see directly from the plot of  $g$  in Fig. 9-2 that  $T_p^c$  for this case must be lower than the case where the polymer is designed *and* renatured with the field ( $h = h_p$ ): the polymer must be better optimized in order to renature without the field originally present during Imprinting. For the high field limit ( $h \gg B$ ), then  $g$  (and therefore  $T_p^c$ ) grows linearly with  $h_p$ . Of course, in this extreme, the effect of  $B$  is unimportant and what must be examined is  $h_p/T_p$ : there is no distinction between lowering the temperature at a fixed field strength and raising the field strength with a fixed temperature.

3. Imprinting without the field, but Renaturation with the field ( $h \neq 0, h_p = 0$ )

In this case, the field acts to destroy the process of renaturation. If the field is sufficiently strong ( $h \gg B$ ),  $g$  approaches zero. For the intermediate field case ( $h \simeq B$ ), there is a maximum in  $g(h, 0)$ , with  $g(h_{\max}, 0) > 1$ ; this is due to the added frustrations due to the competition between the polymer-polymer and polymer-field interactions.

#### 4. Imprinting without the field, but Renaturation with the opposite field ( $h = -h_p$ )

Here, we apply exactly the opposite field which the system wishes to recognize. When the field strength is equal to the strength of the polymer-polymer interactions ( $h = B$ ), the field destroys any possibility of renaturation to the target conformation. For higher field strengths ( $h > 2B$ ),  $g$  becomes negative; this implies that for Imprinting to work, we need a negative polymerization temperature, which simply switches back the sign of  $h_p$  (the switching of the sign of  $B$  is irrelevant in this limit).

In conclusion, Imprinted polymers in an external field display protein-like behavior. For example, they can renature to an Imprinted conformation which has been affected by a given external field without the field present during renaturation. This property is analogous to an antibody renaturing without the antigen present. Also, we have shown that the field can disrupt folding to the polymerization conformation in the cases where either the field was absent or of the opposite sign during Imprinting.

Furthermore, these results are not only applicable to the *in vitro* Imprinting procedure. Indeed, one can consider the optimization of proteins by biological evolution to be selection of sequences which minimizes the energy of the heteropolymer in a particular conformation [Sha93b], which on the level of mean field calculations, is formally identical to Imprinting [Pan94d, Pan95b]. Therefore, these results can be interpreted in terms of possible biological or prebiological evolutionary mechanisms. Indeed, in terms of biological evolution, one can consider many forms of external fields whose effects nature would like to incorporate in the native conformation of a given enzyme or antibody. Furthermore, due to its minimal requirements and simple



design scheme, Imprinting has been proposed as a mechanism for prebiotic evolution [Pan94d,Pan95b]; one may speculate that the monomer soup of the primordial earth was an *in vivo* Imprinting-like experiment, in which primitive ligands acted as external fields, allowing the creating of heteropolymers capable of biological-like functions, such as molecular recognition.



# Chapter 10

## Quenched and Annealed Disorder in the REM

For a distribution of states consistent with the Random Energy Model, we discover a relationship between the replica diagrams in the free energy of quenched and annealed ensembles of polymer sequences. This relationship allows the description of the freezing transition to arbitrary order in the interactions. Furthermore, the elucidation of this formal relationship sheds light on the meaning of the REM approximation and allows a direct derivation of the freezing transition without the use of replicas.

We start with the most general microscopic Hamiltonian, i.e.

$$\mathcal{H} = \sum_{i,j}^q \sum_{I,J}^N B_{ij} \delta(\mathbf{r}_I - \mathbf{r}_J) \delta(s_I, i) \delta(s_J, j) \quad (10.1)$$

where capital Roman numerals label monomer number along the chain, lower case roman numerals label type of monomer species,  $s_I$  is the species of monomer number  $I$  ( $s_I \in \{1 \dots q\}$ ), and  $B_{ij}$  is the matrix of energies of interaction of species  $i$  and  $j$ . This Hamiltonian has the interpretation that monomers number  $I$  and  $J$ , when close in space, interact with an energy based upon the interaction energy of their corresponding species.

We are interested in the case where there is quenched disorder in the sequences. Thus, we must average the free energy over the ensemble of sequences. To facilitate this, we use the replica trick. We make the transformation from spins to fields using the Hubbard-Stratonovich transformation. This yields the replicated partition function

$$\langle Z^n(\text{seq}) \rangle_{\text{seq}} = \mathcal{N} \int \mathcal{D}\{\phi\} \sum_{\text{conf}} \sum_{\text{seq}} P_{\text{seq}} \exp \left[ \frac{T}{4} \langle \vec{\phi} | \hat{B}^{-1} | \vec{\phi} \rangle^{(q n \infty)} + \langle \vec{\phi} | \vec{\rho} \rangle^{(q n \infty)} \right]. \quad (10.2)$$

Summation over sequences only appears in the “source term”  $G$ :

$$\exp \{G\} = \sum_{\text{seq}} P_{\text{seq}} \exp \left\{ \langle \vec{\phi} | \vec{\rho} \rangle^{(q n \infty)} \right\}. \quad (10.3)$$

We sum over sequences to yield

$$G = \sum_{I=1}^N \ln \left\{ \sum_{i=1}^q p_i \exp \left[ \sum_{\alpha=1}^n \int d\mathbf{R} \phi_i^\alpha(\mathbf{R}) \delta(\mathbf{r}_I^\alpha - \mathbf{R}) \right] \right\} \quad (10.4)$$

Series expansion for small  $\phi$  yields the series

$$\begin{aligned} G &= \sum_{\alpha} \sum_i \phi_i^\alpha \Delta_i \left\langle \sum_I \delta(\mathbf{r}_I^\alpha - \mathbf{R}) \right\rangle + \frac{1}{2} \sum_{\alpha, \beta} \sum_{i, j} \phi_i^\alpha \phi_j^\beta \Delta_{ij} \left\langle \sum_I \delta(\mathbf{r}_I^\alpha - \mathbf{R}_1) \delta(\mathbf{r}_I^\beta - \mathbf{R}_2) \right\rangle + \dots \\ &= \sum_{r=1}^{\infty} \frac{1}{r!} \sum_{\alpha_1, \dots, \alpha_r} \sum_{i_1, \dots, i_r} \Delta_{i_1, \dots, i_r} Q^{\alpha_1 \dots \alpha_r}(\mathbf{R}_1 \dots \mathbf{R}_k) \prod_{s=1}^r \phi_{i_s}^{\alpha_s} \end{aligned} \quad (10.5)$$

where  $\Delta_{i_1, \dots, i_r}$  are operators which yield cumulants, i.e.

$$\begin{aligned} \Delta_i &= p_i, \quad \Delta_{ij} = p_i \delta_{ij} - p_i p_j, \quad \Delta_{ijk} = p_i \delta_{ij} \delta_{jk} - 3p_i p_j \delta_{jk} + 2p_i p_j p_k, \\ \Delta_{ijkl} &= p_i \delta_{ij} \delta_{jk} \delta_{kl} - 4p_i p_l \delta_{ij} \delta_{jk} - 3p_i p_k \delta_{ij} \delta_{kl} + 12p_i p_k p_l \delta_{ij} - 6p_i p_j p_k p_l \end{aligned} \quad (10.6)$$

and the overlap order parameters are defined by

$$Q^{\alpha_1 \dots \alpha_k}(\mathbf{R}_1 \dots \mathbf{R}_k) \equiv \sum_I \prod_{i=1}^k \delta(\mathbf{r}_I^{\alpha_i} - \mathbf{R}_i) \quad (10.7)$$

For the case of an annealed heteropolymer,  $n = 1$  and the overlap order param-

eters simply become

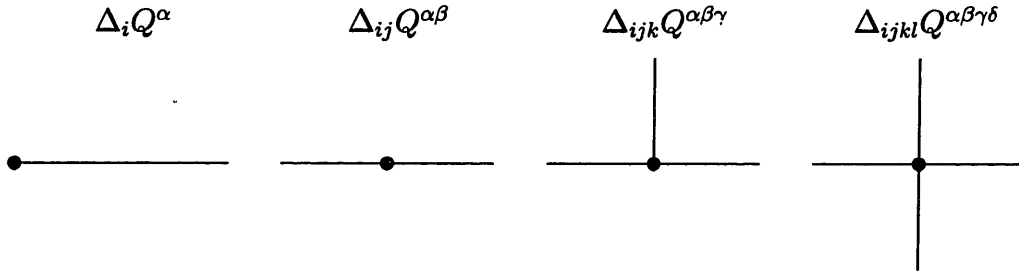
$$Q^{\alpha_1 \dots \alpha_k}(\mathbf{R}_1 \dots \mathbf{R}_k) \rightarrow \rho(\mathbf{R}_1) \prod_{i=1}^{k-1} \delta(\mathbf{R}_i - \mathbf{R}_{i+1}) \quad (10.8)$$

Thus, the free energy is significantly simplified:  $F = F_0 + F_{\text{int}}$ , where

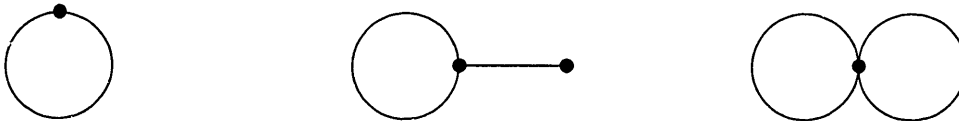
$$\begin{aligned} F_0 &= \frac{1}{4} \int d\mathbf{R}_1 d\mathbf{R}_2 \phi_i(\mathbf{R}_1) B_{ij}^{-1} \delta(\mathbf{R}_1 - \mathbf{R}_2) \phi_j(\mathbf{R}_2) \\ F_{\text{int}} &= \rho \int d\mathbf{R} \left[ \Delta_i \phi_i(\mathbf{R}) + \frac{1}{2} \Delta_{ij} \phi_i(\mathbf{R}) \phi_j(\mathbf{R}) + \frac{1}{3!} \Delta_{ijk} \phi_i(\mathbf{R}) \phi_j(\mathbf{R}) \phi_k(\mathbf{R}) + \right. \\ &\quad \left. \frac{1}{4!} \Delta_{ijkl} \phi_i(\mathbf{R}) \phi_j(\mathbf{R}) \phi_k(\mathbf{R}) \phi_l(\mathbf{R}) + \dots \right] \quad (10.9) \end{aligned}$$

where we have assumed that density is constant and summation over repeated indices.

To facilitate perturbative calculations, we write the Feynman diagrams for the interaction terms:



To calculate the free energy to  $\mathcal{O}(B^2)$ , we need to calculate the sum of all connected diagrams. However, we must not include diagrams which lead to self-energies which should not be in the Hamiltonian in the first place, i.e. we cannot contract a vertex with a line from the same vertex ( $I$  must not be conserved!). For example, graphs which directly contract lines coming from the same vertices do not appear:



Before plunging into a perturbative calculation, we can simplify the free energy by considering the form of the replica overlap order parameters. Within the Random Energy Model (REM) approximation, we say that replica symmetry is broken in

only a single step fashion, i.e. replicas either completely overlap or do not overlap at all. We call the set of replicas which overlap a group of replicas. Thus, we have the form for  $Q$ :

$$Q_{\alpha_1, \alpha_2, \dots, \alpha_k}^{\text{RSB}}(\mathbf{R}_1, \dots, \mathbf{R}_k) = q_{\alpha_1, \alpha_2, \dots, \alpha_k} \rho(\mathbf{R}_1) \delta(\mathbf{R}_1 - \mathbf{R}_2) \times \dots \times \delta(\mathbf{R}_{k-1} - \mathbf{R}_k) \quad (10.10)$$

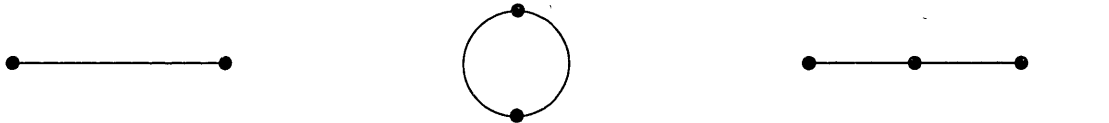
where

$$q_{\alpha_1, \alpha_2, \dots, \alpha_k} = \begin{cases} 1 & \text{for } \alpha_1, \alpha_2, \dots, \alpha_k \text{ in the same group} \\ 0 & \text{otherwise} \end{cases} \quad (10.11)$$

Another limiting case which is useful to consider is the annealed sequence regime, formally denoted by  $n \rightarrow 1$ . In this case, we see that we have the exact relation

$$Q_{\alpha_1, \alpha_2, \dots, \alpha_k}^{\text{ann}}(\mathbf{R}_1, \dots, \mathbf{R}_k) = \rho(\mathbf{R}_1) \delta(\mathbf{R}_1 - \mathbf{R}_2) \times \dots \times \delta(\mathbf{R}_{k-1} - \mathbf{R}_k) \quad (10.12)$$

Let's calculate the diagrams which contribute to the free energy. To  $\mathcal{O}(B^2)$ , we have

$$Q^\alpha Q^\alpha \Delta_i B_{ij} \Delta_j \quad \Delta_{ij} B_{jk} \Delta_{kl} B_{li} Q^{\alpha\beta} Q^{\beta\alpha} \rho_c \quad Q^\alpha Q^{\alpha\beta} Q^\beta \Delta_i B_{ij} \Delta_{jk} B_{kl} \Delta_l$$


For the term  $\mathcal{O}(B)$ , we have no heteropolymeric coupling. In fact, this term is the homopolymeric second virial coefficient, as we couple the density squared with the mean of the interaction matrix.

We would like to relate the diagrams for the quenched case ( $n \rightarrow 0$ ) in the REM approximation to the diagrams in the annealed case ( $n = 1$ ). We see that since the replica space is not coupled to the species space, we can examine the replica overlap parameter contributions independently. To  $\mathcal{O}(B^2)$ , we have contributions of

$$\rho \sum_{\alpha\beta} (Q^{\alpha\beta})^2 = \rho^2 \sum_{\alpha\beta} Q^{\alpha\beta} = \rho^3 \frac{n}{x} x^2 \quad (10.13)$$

In other words, when we sum over the overlap parameter, we have  $n/x$  groups and each group has an area in  $\alpha\beta$  space of  $x^2$ . Also note that we have included factors

of the cutoff density  $\rho_c$ . This factor arises from terms which yield  $\delta(0)$ . Since we are most interested in the regime  $\rho = \rho_c$ , these factors are not important and will be handled just as the total density.

To  $\mathcal{O}(B^3)$  we have the diagrams

$$\Delta_{ij} B_{jk} \Delta_{kl} B_{li} Q^{\alpha\beta} Q^{\beta\alpha} \Delta_{ij} B_{jk} \Delta_{kl} B_{li} Q^{\alpha\beta} Q^{\beta\alpha} \quad Q^\alpha Q^{\alpha\beta} Q^\beta \Delta_i B_{ij} \Delta_{jk} B_{kl} \Delta_l \quad \Delta_{ijk} Q^{\alpha\beta\gamma}$$

Analogously, the contribution of the replica overlap terms in each diagram simply gives  $(n/x)x^3$ , i.e. the number of groups times the *volume* of the group in  $\alpha\beta\gamma$  space.

One can easily demonstrate that the higher order terms behave in the same fashion. Within the REM approximation, we have either replicas completely overlap or do not overlap at all. Thus,  $Q_{\alpha_1\alpha_2\cdots\alpha_k} = Q_{\alpha_1\alpha_2} Q_{\alpha_2\alpha_3} \times \cdots \times Q_{\alpha_{k-1}\alpha_k}$ , and thus for multi-loop diagrams, for  $r + 1$  powers of  $B$ , we matrix multiply  $Q_{\alpha\beta}$   $r$  times and sum over all of the elements yielding the factor  $nx^r$ ; for single-loop diagrams, for diagrams with  $r + 1$  powers of  $B$ , we matrix multiply  $r + 1$   $Q_{\alpha\beta}$  terms and then take the trace yielding  $(n/x)x^{r+1} = nx^r$ . Thus, since every sum over replicas is accompanied by the propagator (and therefore  $B$ ), the perturbative expansion yields the same results as the annealed case, except that each term to order  $r + 1$  in  $B$  gets a factor of  $x^r$ . Thus, within the REM, the free energy for the quenched case is the free energy for the annealed case at an effective temperature  $T \rightarrow T/x$  plus any entropy resulting from the differences between how we organize states in the annealed and quenched systems.

The formal relationship between the annealed and quenched free energies in the REM approximation described above can be explained using physical arguments. Consider the energy spectrum in the REM: a Gaussian distribution of energy states. The annealed case can be considered as the selection of those sequences which optimize the energy for a given conformation while the quenched case consists of the ensemble of conformations which optimize the energy for a given sequence.

To mean field in the REM approximation, these two ensembles are equivalent: the configuration of monomers which minimize the energy can be viewed either as those sequences which optimize the energy for a given conformation or those conformations which optimize the energy of a given sequence.

Before continuing, we outline the general methodology which the previous argument about the relationship between the annealed and quenched diagrams in the REM approximation infers. Specifically, for the polymeric system, we calculate the free energy of the annealed system  $F_{\text{ann}}(T)$  and add the polymeric entropy  $S_{\text{poly}}$  (related to the flexibility of the polymer):  $F(T) = F_{\text{ann}}(T) - TS_{\text{poly}}$ . The entropy of the system is given by  $S(T) = -\partial F/\partial T$ , and freezing occurs at the temperature at which the entropy vanishes, yielding the relationship at the freezing temperature

$$S_{\text{poly}} = \left. \frac{\partial F_{\text{ann}}}{\partial T} \right|_{T_f} \quad (10.14)$$

Thus, we need not introduce replicas, as only the annealed free energy is needed. We emphasize that this result is not “simplified” compared to previous explicit replica calculations and in fact involves exactly the same assumptions and approximations (REM and freezing to a microscopic length scale); however, since the problem is now simplified to the calculation of the annealed case, we can *improve* on previous works using this formalism by carrying out the annealed free energy to all orders in  $\phi$ .

We continue with the explicit calculations for the annealed case. While we can calculate this by summing the appropriate diagrams above, we can much more simply evaluate this series by means of a cumulant expansion. For pedagogical simplification, consider the case of a lattice heteropolymer. We take the condition that  $\rho = \rho_c$  and thus assume that every lattice site is occupied. Furthermore, we ignore self interactions by prohibiting interactions involving the same site. We can write the free energy without introducing the auxiliary field  $\phi$ . We start with the partition function

$$Z = \sum_{s_1} \sum_{s_2} \cdots \sum_{s_N} P_{\text{seq}} \exp[-\sum'_{I,J} B_{s_I s_J}] \quad (10.15)$$



Note that we have replaced the delta function which allows interactions only between neighboring monomers with the analogously weighted sum only over neighboring sites on the lattice  $\Sigma'$ . This is most commonly solved using a cumulant expansion, in which

$$\begin{aligned}
\frac{F}{T} &= -\ln \left\{ \sum_{s_1} \sum_{s_2} \cdots \sum_{s_N} P_{\text{seq}} \exp[-\Sigma'_{I,J} B_{s_I s_J}] \right\} \\
&= -\ln \left\{ \sum_{ij} p_i p_j \exp[-B_{ij}] \right\} \\
&= -\sum_{r=0}^{\infty} \frac{(-1)^r \langle B^r \rangle_c}{r! T^r}
\end{aligned} \tag{10.16}$$

where  $\langle B^r \rangle_c$  is the cumulant of the elements of the interaction matrix averaged over the composition probability, eg.  $\langle B^2 \rangle_c = \langle B^2 \rangle - \langle B \rangle^2 = \sum_{ij} p_i p_j B_{ij}^2 - (\sum_{ij} p_i p_j B_{ij})^2$ .

For the case of an ensemble of random heteropolymers with quenched sequence, within the REM approximation, for each group of  $x$  replicas there is an entropy loss per monomer of  $(x-1)s$ , where  $s = \ln a^3/v$ ; thus, the total change in entropy for  $n/x$  groups is

$$S = -Ns \frac{n}{x} (x-1) \tag{10.17}$$

We must also include the appropriate factor of  $x^r$  to each graph of order  $B^{r+1}$ . This yields the free energy per particle for quenched disorder:

$$\frac{F}{T} = -\sum_{r=0}^{\infty} \frac{(-1)^r \langle B^r \rangle_c}{r! T^r} x^{r-1} + s \frac{n}{x} (x-1) \tag{10.18}$$

We have now written an expression for the free energy in terms of the replica order parameter  $x$ . The only approximation involved is the REM model (i.e. single step RSB and therefore the introduction of  $x$ ) and that the cumulant series converges (i.e. there is no phase segregation; in the language of the  $\phi$  fields, the mean of  $\phi$  vanishes); we have thus included fluctuations in  $\phi$  to all orders. However, we now proceed to calculate the phase transition in  $x$ . While we could in principle examine the fluctuations in  $x$ , it is not clear that the “soft modes” of fluctuation of the replica overlap parameters are fluctuations in  $x$  rather than perhaps overlap between groups

which do not overlap in the mean field solution. Therefore, we perform mean field calculations. Upon deriving the free energy with respect to  $x$ , we find

$$s = \sum_{r=2}^{\infty} \frac{(-1)^r}{r!} (r-1) \langle B^r \rangle_c \left( \frac{x}{T} \right)^r \quad (10.19)$$

The solution for the equilibrium value of  $x$  from the above yields  $x = T/T_f$ , where  $T_f$  is the freezing temperature, determined by the above when we set  $x = 1$  and  $T = T_f$ . This is a polynomial in  $T_f$  and therefore we expect the possibility of discontinuities in the freezing temperature.

In conclusion, we have shown the relationship between the annealed and quenched free energy in the REM approximation. The simple nature of this relationship allows calculations involving quenched disorder previous involving replica methods to be performed simply in terms of the annealed disorder analog. As an example, we applied this formalism to the case of the freezing transition of a heteropolymer with quenched sequence and short range interactions. However, the relationship between annealed and quenched diagrams is independent of the length scale of the interactions (this just puts a potential function in the propagator); thus, one can also examine the freezing transition of other quenched polymeric systems, such as polyampholytes, using this formalism.

**Part IV**

**Experimental**

f

1

2



.

z

:

2



# Chapter 11

## Protein Correlations

The sequences, or primary structures, of existing biopolymers, in particular — proteins, are believed to be a product of evolution. Are the sequences random? If not — what is the character of this non-randomness? To explore the statistics of protein sequences, we employ the idea of mapping the sequence onto the trajectory of a random walk, originally proposed by Peng et al<sup>1</sup> in their analysis of DNA sequences. Using three different mappings, corresponding to three basic physical interactions between amino-acids, we found pronounced deviations from pure randomness, and these deviations seem directed towards the minimization of the energy of the 3D structure. We consider this result as evidence for a physically driven stage of evolution.

### 11.1 Introduction

From the molecular point of view, biological evolution implies the change of the set of sequences of existing proteins. In the same spirit, pre-biological evolution is also understood as the creation and possibly subsequent change of some primary ensemble of sequences (not necessarily protein sequences). Thus, evolution can be viewed as some walk, search, and optimization in sequence space. This space, however, is astronomically big, since the number of possible sequences is exponential

in the length of polymer chains involved. For this reason, an exhaustive search in sequence space is well known to be prohibitively time consuming and, therefore, at least some element of randomness seems inevitable for any understandable picture of evolution.

It can be shown mathematically, that a random choice of a point in sequence space, with uniform probability distribution over the entire space, is equivalent to a completely random formation of the sequence in a letter-by-letter manner without any correlations. Therefore, delicate deviations of the sequences from pure randomness, or correlations between monomers along the sequences, might be of great importance, as they can yield some fingerprint relating to the process which has created the existing biopolymers.

Similar arguments were used to justify the concept which is imaginatively stated as “proteins are slightly edited random copolymers” [Pti86]. For example, it was shown that the lengths distribution of  $\alpha$ -helices in proteins follows accurately what could be expected for just random sequences [Pti86]. Some other tests can also be found in [Pti86] (and the references therein). We also mention, that the small degree of “editing” is closely related to neutral theory of evolution [Kim83]. In the spirit of the concept of “proteins as edited random copolymers,” we address in this work the aspect in which they are “edited.”

To look for this non-randomness, one has to decode the sequence in an appropriate manner. For example, some peculiar correlations between monomers were recently found in purine-pyrimidine representation of DNA sequences [Peng92]. As for proteins, we expect that this decoding has to be related to the 3D structure and the folding properties of a protein chain. Indeed, the 3D structure of protein is believed to be completely encoded in the sequence. On the other hand, it is exactly the 3D structure which defines all of the aspects of a protein’s functionality and, therefore, the properties of a protein in competition under evolutionary selective pressure. In other words, the relationship between the sequence and the selective promise of the protein is mediated by the 3D structure. Thus, as the 3D structure can be considered to be “written” in the amino acid sequence in the “language” of the

interactions between amino acids, we decode protein sequences according to the role of each particular residue in the determination of the protein's three-dimensional structure. Namely, we consider three ways to decode protein sequence, related to the three most important kinds of volume interactions — Coulomb interaction, hydrophobic/hydrophilic interaction, and hydrogen bonding.

## 11.2 Brownian Bridge Representation for Protein Sequences

Technically, we employ the idea of Peng et al [Peng92] and map protein sequence onto the trajectory of artificial 1D random walker. More precisely, we construct for each sequence a one-dimensional walker which makes steps of size  $\sigma$  up and down at discrete time moments  $i$ ,  $0 \leq i \leq L$ . The walker is required to return to the origin after the entire trip of  $L$  steps, so that the corresponding trajectory is a “Brownian bridge.” A purely random walker, which corresponds to a random sequence, is expected to travel about  $\sigma \cdot \sqrt{L}$  from the origin on mean-square-average. To reach farther, it must go mainly in one direction for the first half-time ( $i < L/2$ ) and mainly back in the second half-time ( $i > L/2$ ) thus approaching the maximal distance of  $\sigma \cdot L/2$ . On the other hand, to keep as close to the origin as possible, it must compensate each step to one direction by a subsequent opposite step. Therefore, persistent types of correlations in protein sequences would be manifested in trajectories which go beyond the random one, while alternating correlations would lead to the trajectories which do not travel as far.

In order to employ this test of non-randomness, we have calculated for each of the amino-acid sequences obtained from the Data Bank [2] the trajectories of three different artificial walkers, each related to a kind of physical interactions between residues — hydrophobic ( $A$ ), hydrogen bonds ( $B$ ), and Coulomb ( $C$ ). The subsequent steps of each walker are given by the numbers  $\{\xi_i\}$  defined as

A.  $\xi_i = +1$  if monomer number  $i$  in the given sequence is highly hydrophilic (Lys, Arg, His, Asp, Glu) or  $\xi_i = -1$  in any other case;

B.  $\xi_i$  may be +1 or -1 for monomers capable (Asn, Gln, Ser, Thr, Trp, Tyr) or not capable (all others) of hydrogen bonding [Dre90];

C.  $\xi_i$  may be +1, -1 or 0 for positively (Lys, Arg, His) or negatively charged (Asp, Glu) and neutral (all others) monomer  $i$ , respectively [Dre90].

In order to look for correlations by comparing the trajectories, we have to exclude the dependencies on protein length, overall composition and the step size of the walker. This is done by the following definition of trajectories:

$$r(\lambda) \equiv \left\langle \left[ \sum_{i=0}^{\lceil \lambda L_p \rceil} \frac{\Delta \xi_i^{(p)}}{\sigma^{(p)}} \right]^2 \right\rangle_p, \quad (11.1)$$

where  $p$  denotes a given protein,  $\langle \dots \rangle_p$  means average over the set of proteins,  $\lceil \dots \rceil$  means take the next highest integer, and  $L_p$  is the total number of amino acids in  $p$ . (i) to exclude  $L_p$ -dependence, we rescale the number of steps taken ( $l$ ) as  $\lambda = l/L_p$ ,  $0 \leq \lambda \leq 1$ ; (ii) to exclude the walker's drift due to the protein overall composition, we subtract the term linear in  $\lambda$  for each protein by  $\Delta \xi_i^{(p)} = \xi_i^{(p)} - \overline{\xi^{(p)}}$ ,  $\overline{\xi^{(p)}} = (1/L_p) \sum_{i=0}^{L_p} \xi_i^{(p)}$  (in this way the trajectory is brought to the bridge shape); (iii) to exclude the step-size dependence, we divide by  $\sigma^{(p)} = \sqrt{\sum_{i=0}^{L_p} [\xi_i^{(p)} - \overline{\xi^{(p)}}]^2}$ . In other words,  $r(\lambda)$  is the distance traveled by the effective walker (i.e. with the mean drift removed) after taking  $\lceil \lambda L_p \rceil$  steps of size  $\sigma$ .

Our procedure to construct the walkers is thus a modification of the original Peng *et al* [Peng92] procedure, in such a way, that (a) we average over an ensemble of different proteins rather than along the chain and (b) all the trajectories are bridges.

The trajectories  $r_A(\lambda)$ ,  $r_B(\lambda)$ ,  $r_C(\lambda)$ , along with the theoretically found trajectory

$$r_{rand}(\lambda) = \frac{1}{\lambda^{-1} + (1-\lambda)^{-1}}. \quad (11.2)$$

for purely random case, are shown in the Figure 1 for a set of globular proteins (those coded as catalysts in the Data Bank). The  $r_A(\lambda)$  and  $r_B(\lambda)$  bridges are clearly over  $r_{rand}(\lambda)$  manifesting pronounced persistent correlations in the distribution of



hydrophobicity. Alternating correlations are found between electrical charges on protein chains because  $r_C(\lambda)$  is definitely under  $r_{rand}(\lambda)$ . This is the main finding of the work.

### 11.3 Brownian Bridges for Some Particular Sets of Proteins

Some developments of this main result are as follows. When we look at early forms of life, such as prokaryotes, we find that the corresponding Brownian bridges shown in Figure 2 fit quite well to a phenomenological scaling generalization of eq (2) of the form

$$r(\lambda) = \frac{L_0^{2\alpha-1}}{\lambda^{-2\alpha} + (1-\lambda)^{-2\alpha}}, \quad (11.3)$$

yielding quantitative results of  $\alpha_A = 0.520 \pm 0.005$ ,  $\alpha_B = 0.520 \pm 0.005$ , and  $\alpha_C = 0.470 \pm 0.005$  for prokaryotes. Clearly,  $\alpha > 1/2$  and  $\alpha < 1/2$  means persistent and alternating type of correlations, respectively. In order to exclude small polypeptides as well as multiglobular proteins, we have examined only proteins with lengths between 110 and 750 amino acids. For simplicity, we take  $L_0 = 110$ , ie the shortest chain in the ensemble, but we have found no special qualitative dependence on  $L_0$ .

We stress here that  $\alpha \neq 1/2$  does not imply any fractal interpretation, contrary to the DNA case, because we average over the ensemble of different sequences rather than over the sliding window in one sequence.

Of course, the statistical errors are greater for smaller subsets of sequences. Nevertheless, the main qualitative finding ( $\alpha_A, \alpha_B > \frac{1}{2}$ ,  $\alpha_C < \frac{1}{2}$ ) remains valid for all of the considered groups of globular proteins. At the same time, we have to mention, that some of the bridges, for example  $r_A(\lambda)$  for enzymes from plants, exhibit clear irregularities and asymmetries, which remain unexplained. For the subset of coil-like proteins (ie denoted to be coiled in a comment or keyword of the database), we found  $\alpha_A, \alpha_B$ , and  $\alpha_C > \frac{1}{2}$ ; this is easily related to the known periodicity of fibrillar protein sequences.

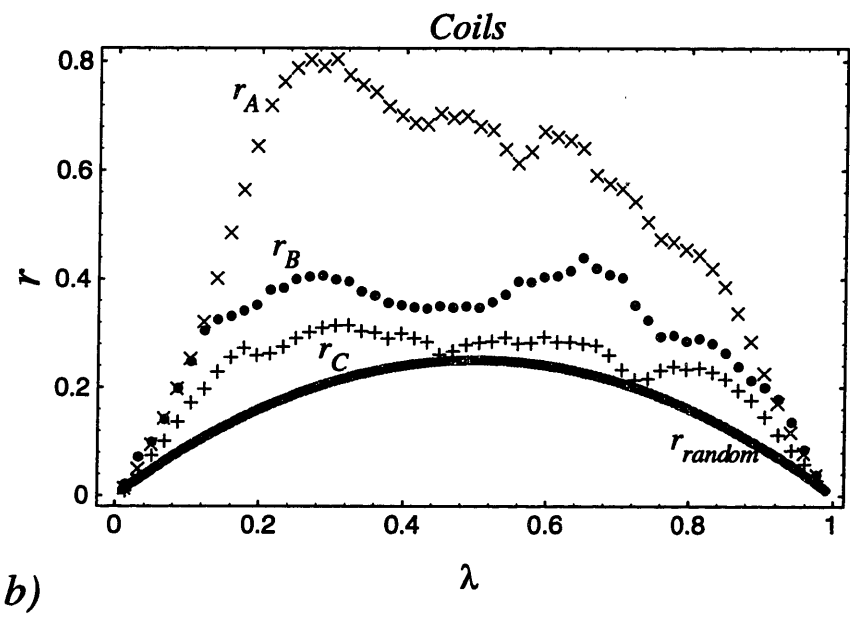
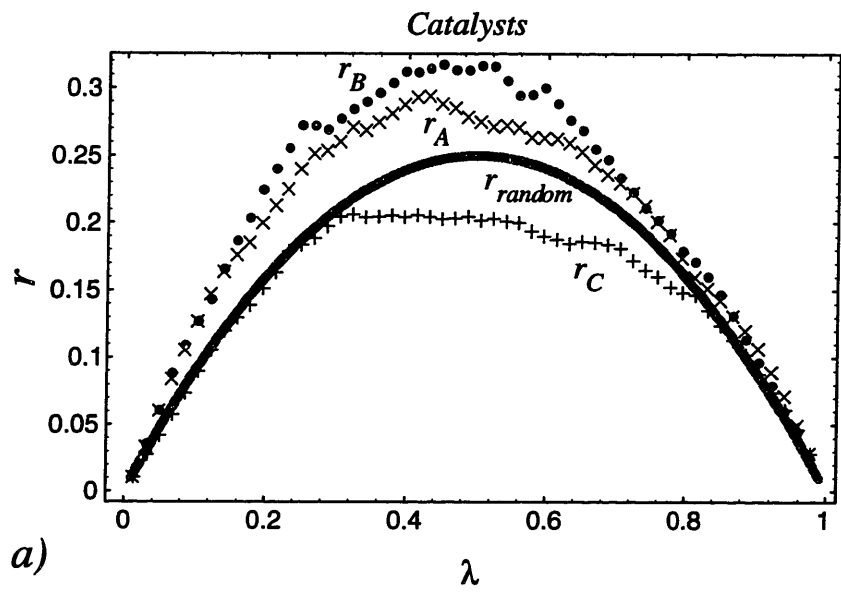


Figure 11-1: Brownian bridges for hydrophilic ( $\times$ ), hydrogen bonding ( $\bullet$ ), and coulomb ( $+$ ) mappings of sequences of proteins with (a) catalytic activity, and therefore globular structure, and (b) coiled structure. *a)* The general qualitative behavior for catalysts ( $\alpha_A > \frac{1}{2}$ ,  $\alpha_B > \frac{1}{2}$ , and  $\alpha_C < \frac{1}{2}$ ) is seen, when compared to the bridge corresponding to an ensemble of random sequences  $r_{rand}$  (thick gray curve), ie  $\alpha = \frac{1}{2}$ . *b)* Persistent correlations are found in all mappings for coils.

### Prokaryote Catalysts

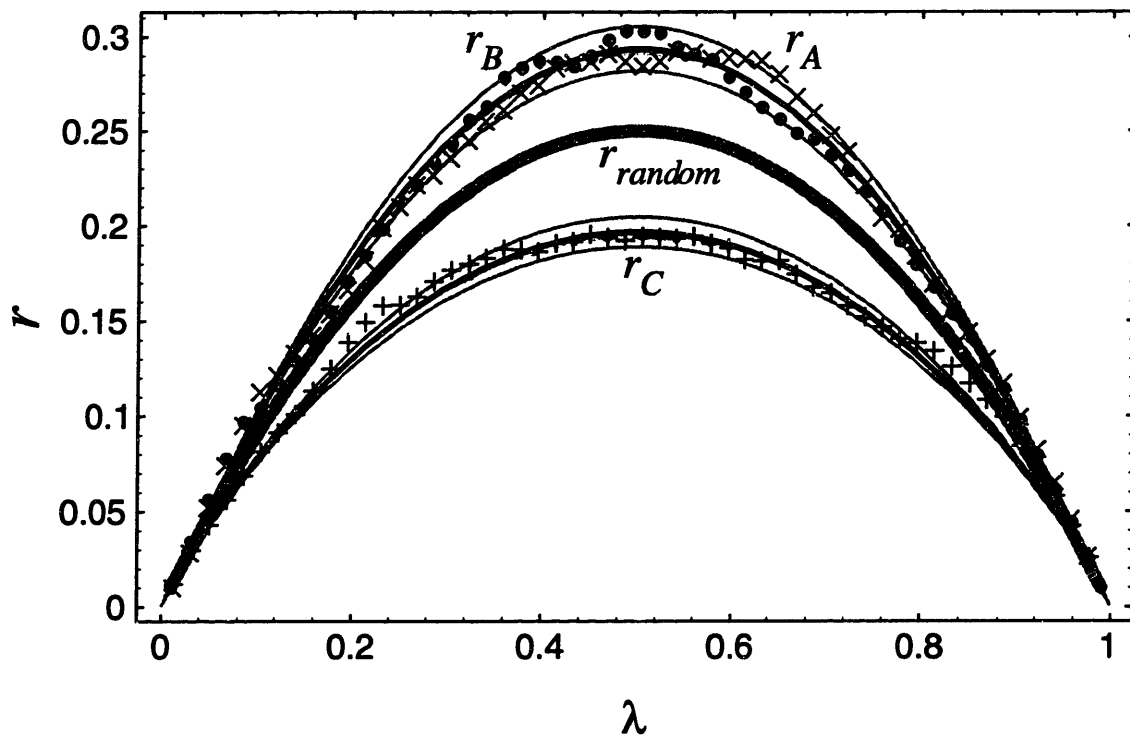


Figure 11-2: Brownian bridges for hydrophilic ( $\times$ ), hydrogen bonding ( $\bullet$ ), and coulomb ( $+$ ) mappings of sequences of prokaryote proteins sequences. We find that these bridges fit well to eq (3) with  $\alpha_A = 0.520 \pm 0.005$ ,  $\alpha_C = 0.520 \pm 0.005$ , and  $\alpha_C = 0.470 \pm 0.005$  ( $L_0 = 110$ ) The thin gray lines bounding a given bridge give the error spread specified above.

In order to insure that these results are not artifacts of the procedure used, we performed several control tests. In particular, artificial shuffling of the units along the chain as well as randomly shuffled versions of the maps  $A$ ,  $B$ , and  $C$  all lead to random sequences ( $\alpha = 0.5 \pm 0.0025$ ).

## 11.4 Discussion

To conclude, we speculate on the possible explanations for the non-randomness of protein sequences. As mentioned in the Introduction, we believe that the deviations from randomness seen are the fingerprints of an evolutionary process, biological or pre-biological. On the other hand, the results  $\alpha_A, \alpha_B > \frac{1}{2}$ ,  $\alpha_C < \frac{1}{2}$  appear to be a manifestation of some process driven by physical interactions between monomers. Indeed, a sequence with a tendency toward alternating signs of charges along the chain ( $\alpha_C < \frac{1}{2}$ ) has, at the same conformation, obviously lower Coulomb energy compared to another hypothetical sequence with blocks of the charges of the same sign. Analogously, hydrophilic monomers energetically prefer to concentrate at the loops which are on the surface of the globule and thus in contact with the solvent. Therefore, there is the coincidence: the set of protein sequences, known to be a product of evolution, looks similar to the result of some physical game with repulsion and attraction of monomers.

What could be the reason for this coincidence? Consider the recent works [Sha93b, Pan94b], where two different procedures were suggested to prepare, or at least to imitate the preparation of heteropolymers with sequences capable of renaturation into a given molecular fold. One of them [Sha93b] is based on annealing of the sequence of the polymer with a chosen target conformation. Another procedure [Pan94b] implies, prior to polymerization, prearrangement of monomers in space due to the interplay of repulsive and attractive interactions. These processes are both driven physically and lead therefore to  $\alpha_A, \alpha_B > \frac{1}{2}$ , and  $\alpha_C < \frac{1}{2}$ . We have analyzed correlations along the artificial sequences produced by our model of polymerization [Pan94b] and found very reasonable agreement with the data for

real proteins (eg. prokaryotes). We conclude from this consideration, that some physically driven process, where the same set of monomer-to-monomer interactions is employed as in the renaturation of the existing proteins, is likely to be one of the stages of evolution, biological or pre-biological.

From this perspective, it might be instructive to compare correlations in different groups of organisms vs evolutionary age. Figure 3 shows the bridges for proteins from several different groups of organisms. As to the Coulomb bridge, an evolutionary trend towards larger  $\alpha_C$ , or less alternating correlations, is clearly seen. On the other hand, our data do not reveal any trend with respect to  $\alpha_A$  and  $\alpha_B$ . This is not at all unexpected, as the Brownian bridges for hydrogen bonding and hydrophilic mappings had greater variation, and therefore errors in  $\alpha$  estimation, than the Coulomb mapping, so that a trend might not be seen even if there was one. If one believes in the trend revealed by Fig. 3a, this implies that biological evolution somehow allows the elimination of the correlations imposed by the prebiological creation of sequences. We must stress, however, that this question remains of much more speculative character than our main finding shown in Fig. 1.

One might consider our main results as only the reflection of physical constraints involved with the formation of heteropolymers with a unique structure (similar to, for example, obvious constraint that the total charge of the chain cannot be too large), i.e. the correlations obtained represents the fact that certain sequences are more favorable due to physical criteria. However, the sheer fact that correlations are seen in the ensemble of proteins, which are assumed to be a product of evolution, is exactly how we understand our statement that at least some stage of biological or pre-biological evolution has selected protein sequences based upon physical criteria.

### Catalysts (Coulomb Mapping)

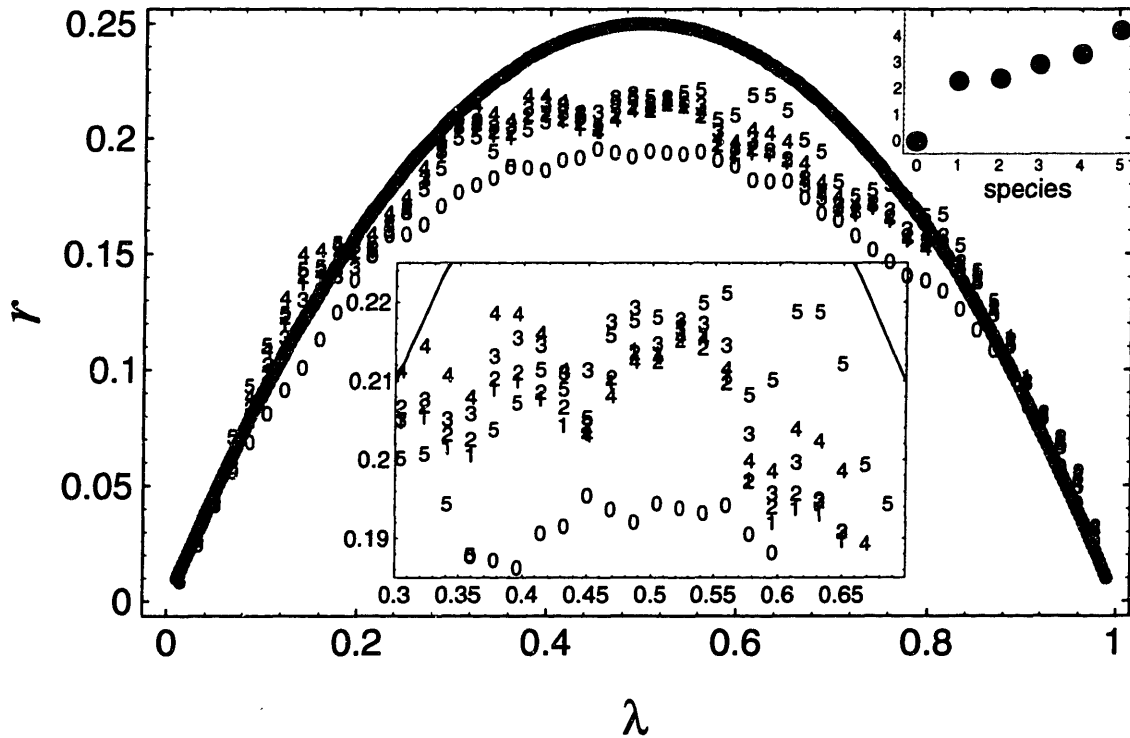


Figure 11-3: Brownian Bridges for a series of evolutionary groups: Coulomb mapping, with a magnified region  $0.3 \leq \lambda \leq 0.7$  in the lower center. There is a clearly seen trend such that the younger (larger label numbers) evolutionary groups have bridges closer to  $r_{rand}$  (thick gray curve). This trend can be characterized by computing the difference ( $\Delta$ ) between the area under the Brownian bridge for a given species and the area under the bridge for random sequences. We have chosen the domain (0.3,0.7) for integration since the error becomes great outside of this range. The result is seen in the upper right hand corner. Another quantitative measure of the evolutionary trend would be to fit each bridge with eq (3) and plot  $\alpha_i$  vs  $i$ ; qualitatively, this leads to the same conclusion, but since individual bridges do not necessarily fit very well to eq (3), except for prokaryotes, this fit introduces some artificial errors. (0=prokaryota, 1=chordata, 2=tetrapoda, 3=metazoa, 4=mammalia, 5=rodentia)

### *Catalysts (Hydrophilic Mapping)*

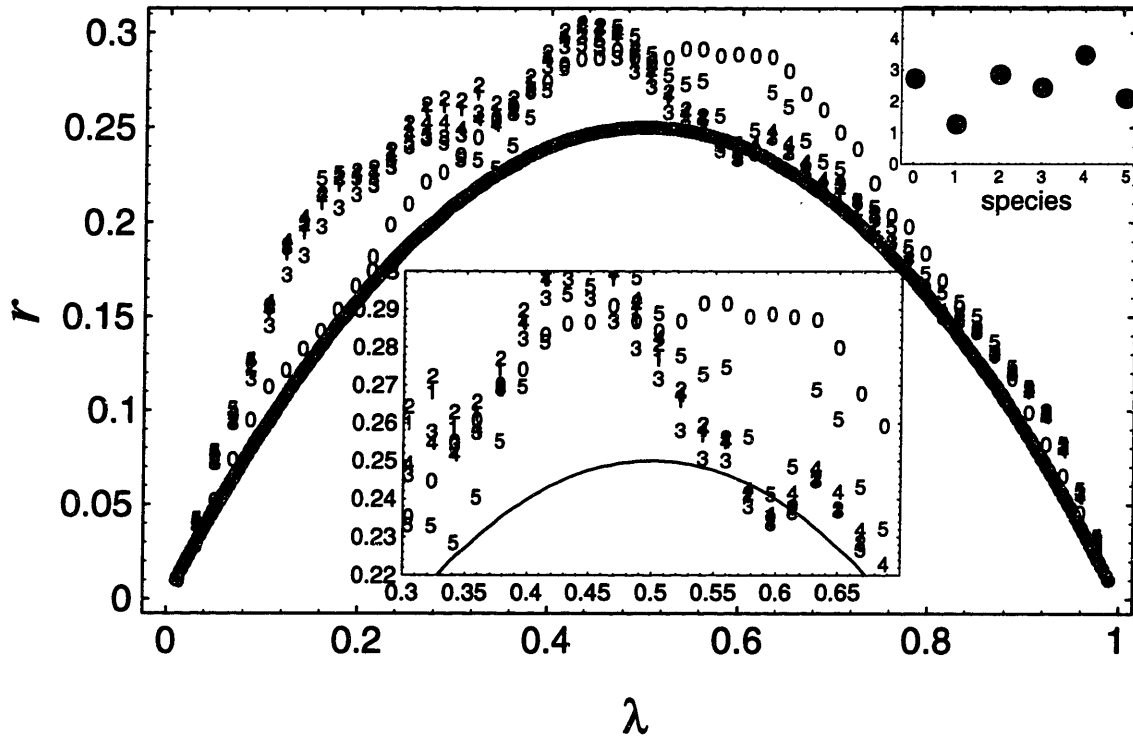


Figure 11-4: Brownian Bridges for a series of evolutionary groups: Hydrophilic mapping. Again the prokaryote bridge fits well to eq (3) with  $\alpha > \frac{1}{2}$ . As in the Coulomb case, the bridges for the other evolutionary groups deviate more from eq (3) than the prokaryote bridge; however, the evolutionary trend found with the hydrophilic mapping is not seen as clearly, as shown in the plot of  $h_i$  vs  $i$  in the upper right hand corner. (0=prokaryota, 1=chordata, 2=tetrapoda, 3=metazoa, 4=mammalia, 5=rodentia)

# Appendix A

## Derivation of Equation (1)

We start with a given ensemble of protein sequences. With the decoded sequence  $\{\xi_1, \xi_2, \dots, \xi_L\}$ , we map it onto the trajectory as

$$x(l) = \sum_{t=1}^l \xi_t. \quad (11.4)$$

The walker defined by Eq (A1) may have a strong drift, so that the leading term in  $x(l)$  might be linear in  $l$ ; this is related simply to the mean composition of the chain considered. Since overall composition is beyond our interest here, we define the reduced trajectory:

$$y(l) = x(l) - (l/L)x(L), \quad (11.5)$$

$L$  being the total number of links in the entire polymer chain. Obviously, the  $y$ -walker returns back to the origin after the entire “trip.” The corresponding trajectory  $y(l)$  is called a “Brownian bridge.”

In principle,  $y$  is expected to scale as  $L^\alpha$  with chain length. For example, we have considered  $y^2(L/2)$  for each protein, and made the log-log plot, where each point corresponds to one particular protein and has coordinates  $L, y^2(L/2)$ . This plots indicate clearly the tendency toward power law dependence of the type  $y^2(L/2) \sim L^{2\alpha}$ . However, because of restricted statistics available and great fluctuations, it is hard to come to the convincing conclusions with this approach.

In order to collect all the data in a comparable form, we have rescaled all the Brownian bridges compensating for different proteins with different lengths and variances of  $\xi$  distribution, by

$$z^2(\lambda) = \frac{y^2}{L(\xi - \bar{\xi})^2} \quad (11.6)$$

where  $\overline{(\dots)}$  = averaging over a given protein sequence (eg  $\bar{\xi} = \frac{1}{L} \sum_{i=1}^L \xi_i$ ) and to exclude  $L$ -dependence, we rescale the number of steps taken ( $l$ ) as  $\lambda = l/L$ , where



$0 \leq \lambda \leq 1$ .

With the rescaled trajectories  $z^2(\lambda)$ , we perform averaging over the ensemble of proteins:

$$r(\lambda) = \langle z^2(\lambda) \rangle_{\text{ensemble}} . \quad (11.7)$$

which, when combined with equations (A1) through (A3), yields eq. (1).

## Appendix B

### Derivation of Equation (3)

A Brownian bridge is generally the trajectory of a random walk which starts and terminates at the same point in space, say, in the origin. Let us consider first the simplest case of a random walk without correlations and let us evaluate the probability distribution for the walker displacement  $z$  as a function of “time”  $l$ ,  $\mathcal{P}_l(z)$ . This can be considered as the probability for two walkers to meet each other at the point  $z$  at the “moment”  $l$ : both of them start from the origin, but the first begins at zero time and walks for the time  $l$  while the second begins at the time  $L$  and walks back in time for the period  $L - l$ . For the uncorrelated process, we have thus

$$\mathcal{P}_l(z) = p_l(z) \cdot p_{L-l}(z). \quad (11.8)$$

Since there are no correlations,  $p_l(z)$  is simply the standard Gaussian distribution

$$p_l(z) = (la\pi)^{-1/2} \exp \left[ -\frac{z^2}{la} \right], \quad (11.9)$$

where  $a$  is a parameter. We see therefore that in this case

$$\mathcal{P}_l(z) = \text{const} \cdot \exp \left[ -\frac{z^2}{a} \left( \frac{1}{l} + \frac{1}{L-l} \right) \right], \quad (11.10)$$

and  $r(l) = \langle z^2(l) \rangle = \int z^2 \mathcal{P}_l(z) dz$  thus obeys equation (2).

We now return to a more general case. Scaling arguments imply that the distribution  $p_l(z)$  is of the form

$$p_l(z) = \text{const} \cdot \exp \left[ - \left( \frac{z}{al^\alpha} \right)^\beta \right], \quad (11.11)$$

where  $\alpha$  and  $\beta$  are critical exponents. Supposing equation (B1) is valid (which is generally may not be true), one easily gets the expression for  $\mathcal{P}_l(z)$  and then for  $r(l) = \langle z^2(l) \rangle$ . At  $\beta = 2$  we recover exactly equation (3). It is clear from the derivation, that applicability of equation (3) is restricted from two sides, namely, the validity of (B1) and the supposition  $\beta = 2$ . Our statistical analysis shows no need in trying other values of  $\beta$  as well as in consideration of any generalization of (B1). The simple variant of equation (3), considered as purely phenomenological, works reasonably well.

To understand the physical meaning of critical exponent  $\alpha$ , one has to look at Equation (B4). In terms of random walk representation, (B4) implies that r.m.s. displacement of the walker scales as  $l^\alpha$  with “time”  $l$ . Certainly, it is analogous to the excluded volume problem in polymer physics, where the size of polymer chain is known to scale as  $l^\nu$  with chain length  $l$ , where  $\nu > 1/2$  ( $3/5$  in classical Flory theory [Flo53]) or  $\nu < 1/2$  ( $1/3$  for dense globule) depending on prevailing of repulsive or attractive monomer-to-monomer interactions, respectively. Therefore,  $\alpha$  is analogous to the critical exponent of correlation radius. It is worthwhile to mention here, that  $\alpha > 1/2$  was found for DNA sequences [Peng92].

## Appendix C: Bridges for Different Species

For completeness, we include the bridges for Coulomb and Hydrophobic/hydrophilic mappings for several different species. Apart from any deviations from  $\alpha = 1/2$  quantitatively measured from fitting to bridges, it is interesting to simply examine the bridges and see how they deviate from the random bridge.

For example, the hydrophobic hydrophilic bridges do not fit our scaling relation for bridges parameterized by  $\alpha$ ; clearly, some generalized function with another degree of freedom to describe these deviations from bridges with a given  $\alpha$  must be employed. Also, the bridges for plants show marked deviations in the early parts of the sequence. This most likely has a biological explanation, which is unfortunately unknown to the author.

We include Table 11.1 to describe which species are involved and some statistics related to the ensemble used. Note that due to the nature of the database, the different species are not equally represented. Therefore, the bridges with fewer members in the ensemble will be much more noisy (for example, consider the bridge for molluscs which only have 2 members, shown here as a control/example). Also, from Table 11.1, one can see which subgroups dominates certain groups. For example, diptera dominates insects (162 out of 191) which in turn dominates arthropods (191 of 201); in fact, we see that diptera dominates arthropods. This is most likely due to the common use of flies in biology and is an example of the sample biases we encounter.

| Label            | Explanation                | Number | (length) |
|------------------|----------------------------|--------|----------|
| all species      | all species                | 6016   | 457      |
| amphibia         | amphibians                 | 25     | 447      |
| angiospermae     | flowering plants           | 775    | 435      |
| animals          | all animals                | 1881   | 471      |
| arthropoda       | insects, spiders, etc.     | 201    | 453      |
| artiodactyla     | pigs, goats, sheep, cattle | 196    | 458      |
| aves             | birds                      | 109    | 461      |
| bacteria         | bacteria                   | 1576   | 412      |
| chordata         | have a notochord           | 1596   | 478      |
| crustacea        | crustaceans                | 10     | 417      |
| dicotyledoneae   | pair seeded angiosperms    | 570    | 436      |
| diptera          | flies                      | 162    | 480      |
| embryophyta      | vascular plants            | 823    | 436      |
| eukaryota        | cells have a nucleus       | 3539   | 475      |
| eutheria         | have a placenta            | 1391   | 490      |
| fungi            | fungi                      | 682    | 519      |
| insecta          | insects                    | 191    | 455      |
| mammalia         | mammals                    | 1393   | 490      |
| metazoa          | multicellular animals      | 1881   | 471      |
| mollusca         | molluscs                   | 2      | 333      |
| monocotyledoneae | single seeded angiosperms  | 205    | 435      |
| phycophyta       | primitive plants           | 52     | 411      |
| pisces           | fish                       | 47     | 322      |
| planta           | all plants                 | 875    | 434      |
| primates         | primates                   | 467    | 505      |
| prokaryota       | cells lack a nucleus       | 2217   | 414      |
| protozoa         | single-celled eukaryotic   | 101    | 600      |
| reptilia         | reptiles                   | 21     | 183      |
| rodentia         | rodents                    | 585    | 486      |
| tetrapoda        | four-limbed vertebrates    | 1548   | 483      |
| vertebrata       | have backbones             | 1595   | 478      |
| viridae          | viruses                    | 260    | 585      |

Table 11.1: Legend for plots of bridges for different species. Label refers to the label used in the following plots, Explanation gives a description of the taxonomic label, Number indicates how many sequences were of this type, and (length) is the mean length of the sequences of the given ensemble.

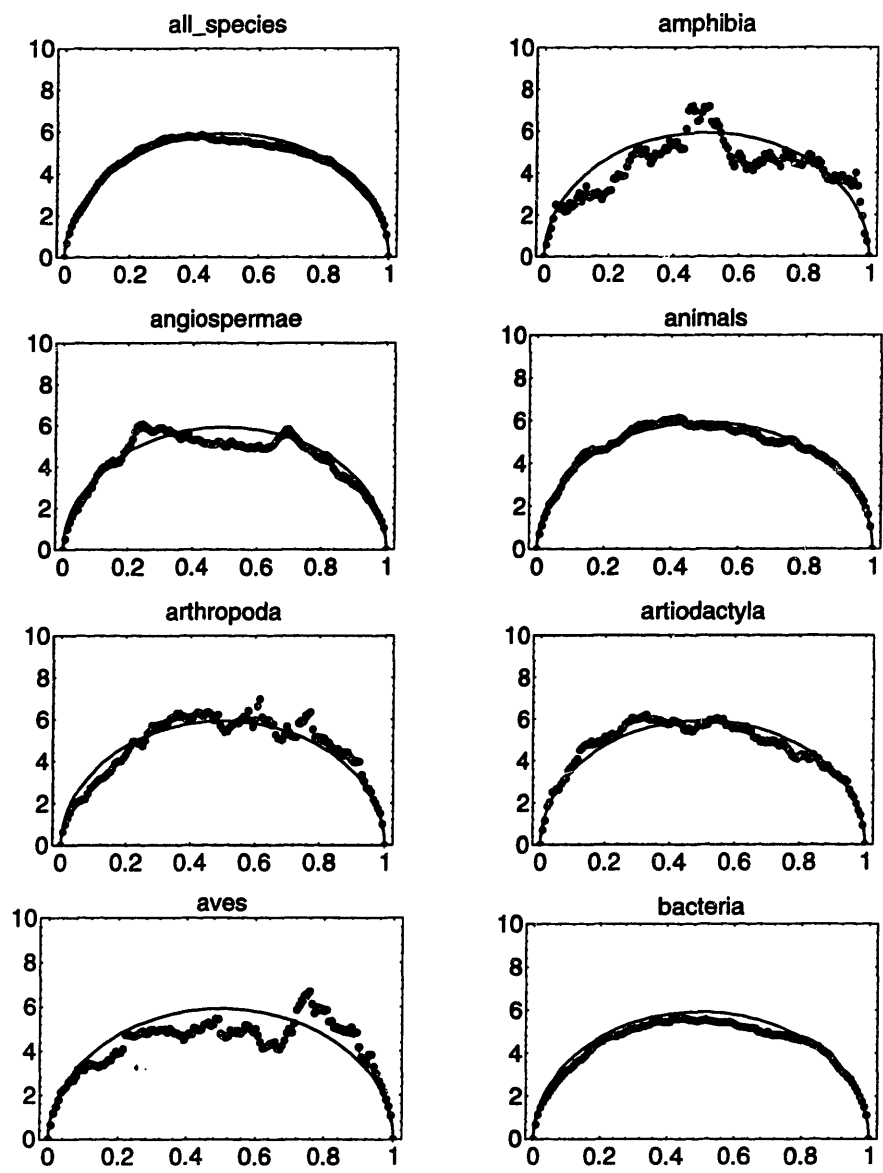


Figure 11-5: Bridges of different species for Coulomb mapping

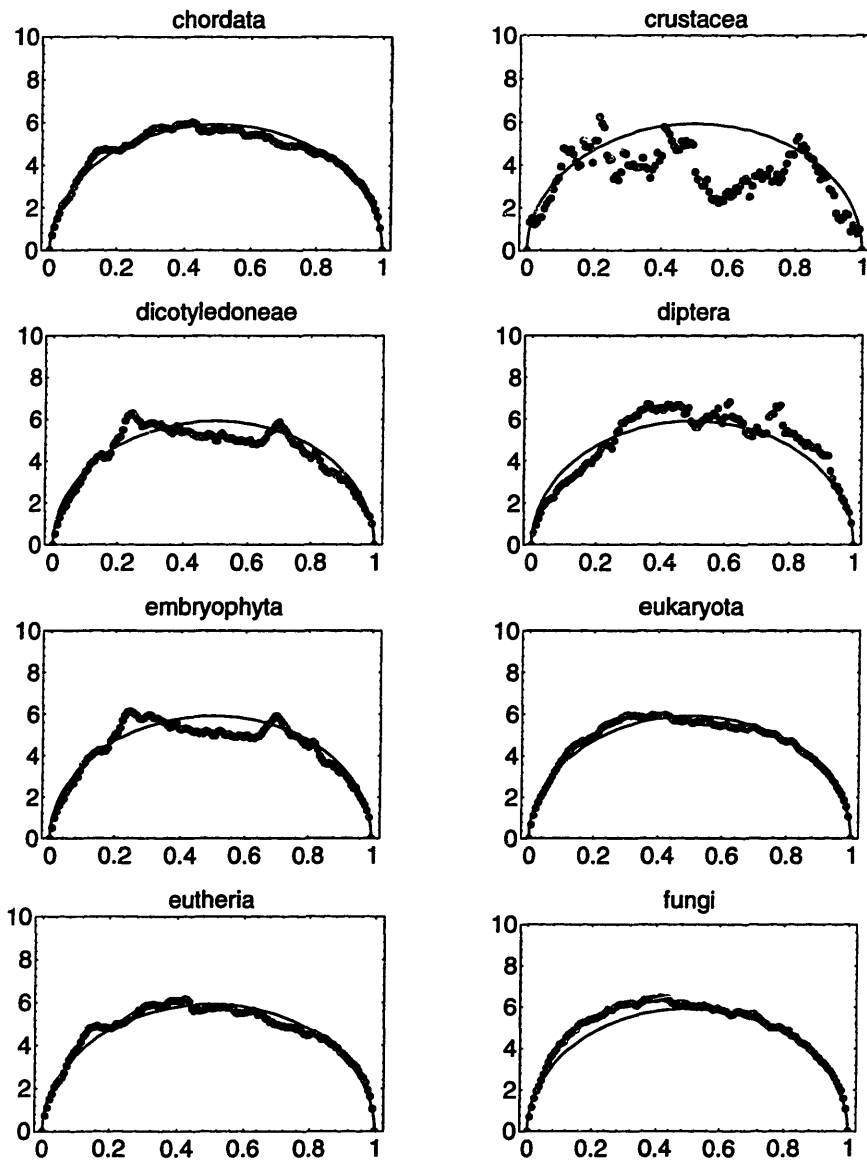


Figure 11-6: Bridges of different species for Coulomb mapping

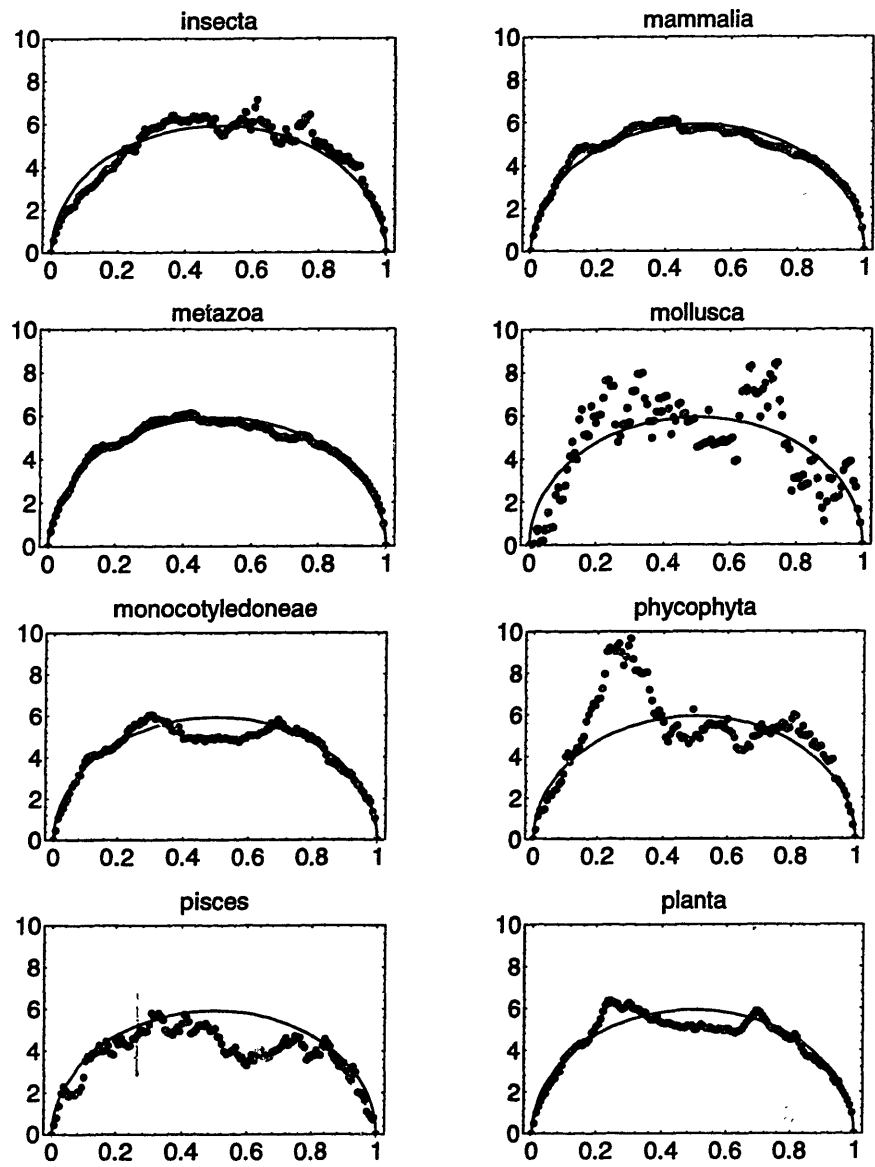


Figure 11-7: Bridges of different species for Coulomb mapping

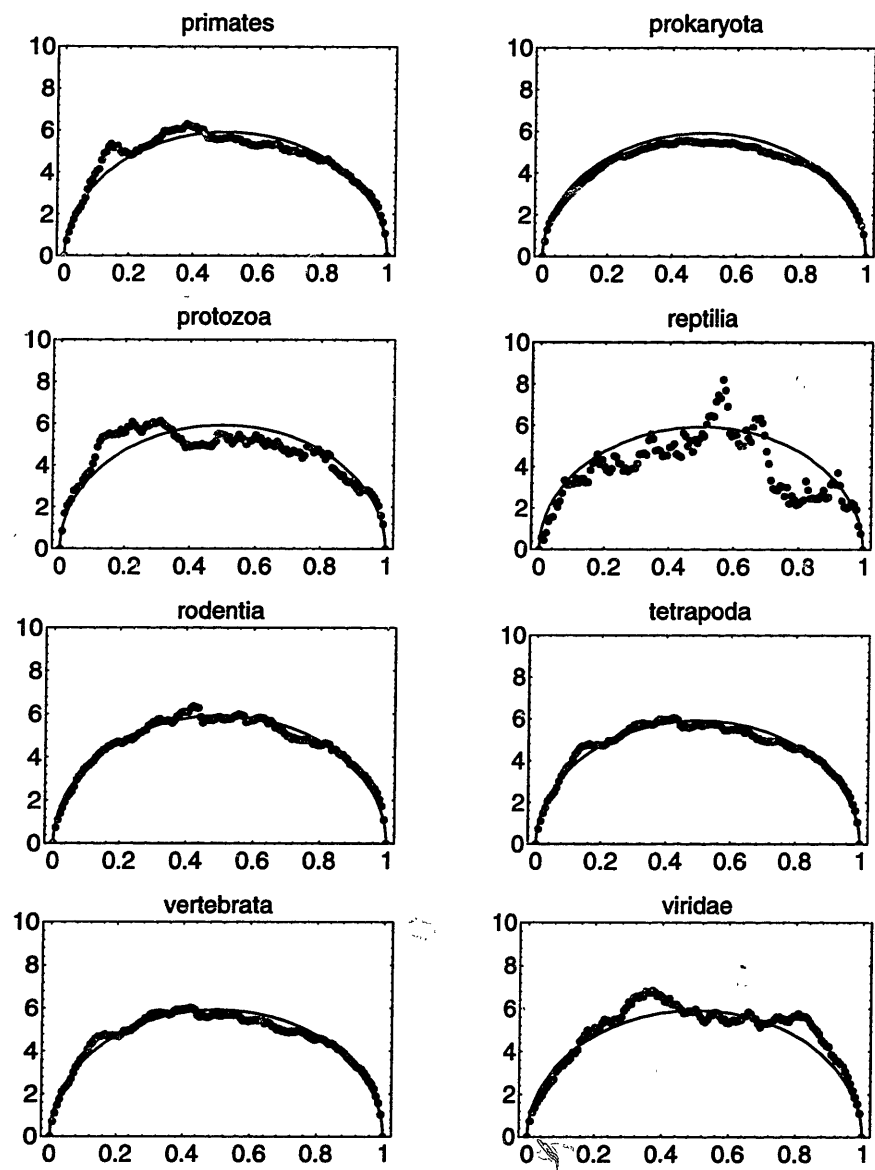


Figure 11-8: Bridges of different species for Coulomb mapping



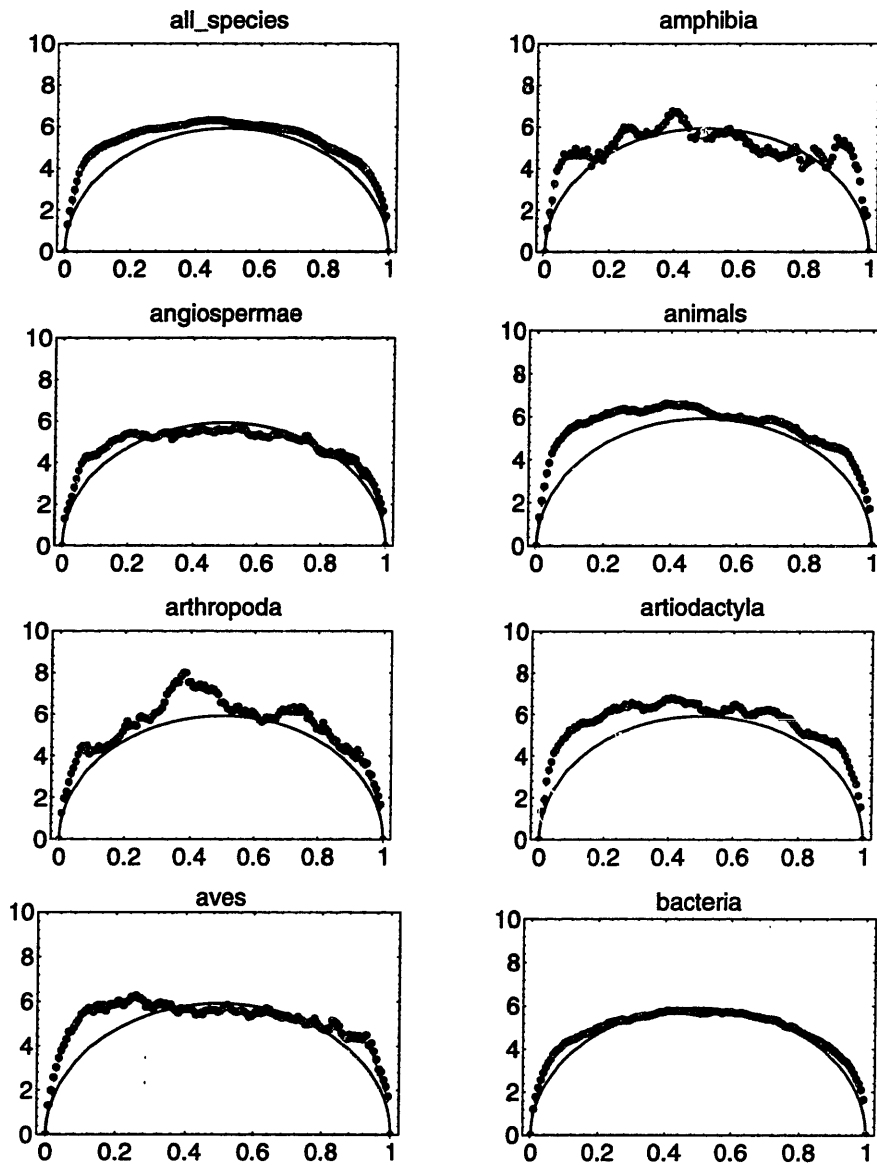


Figure 11-9: Bridges of different species for Hydrophobic/hydrophilic mapping

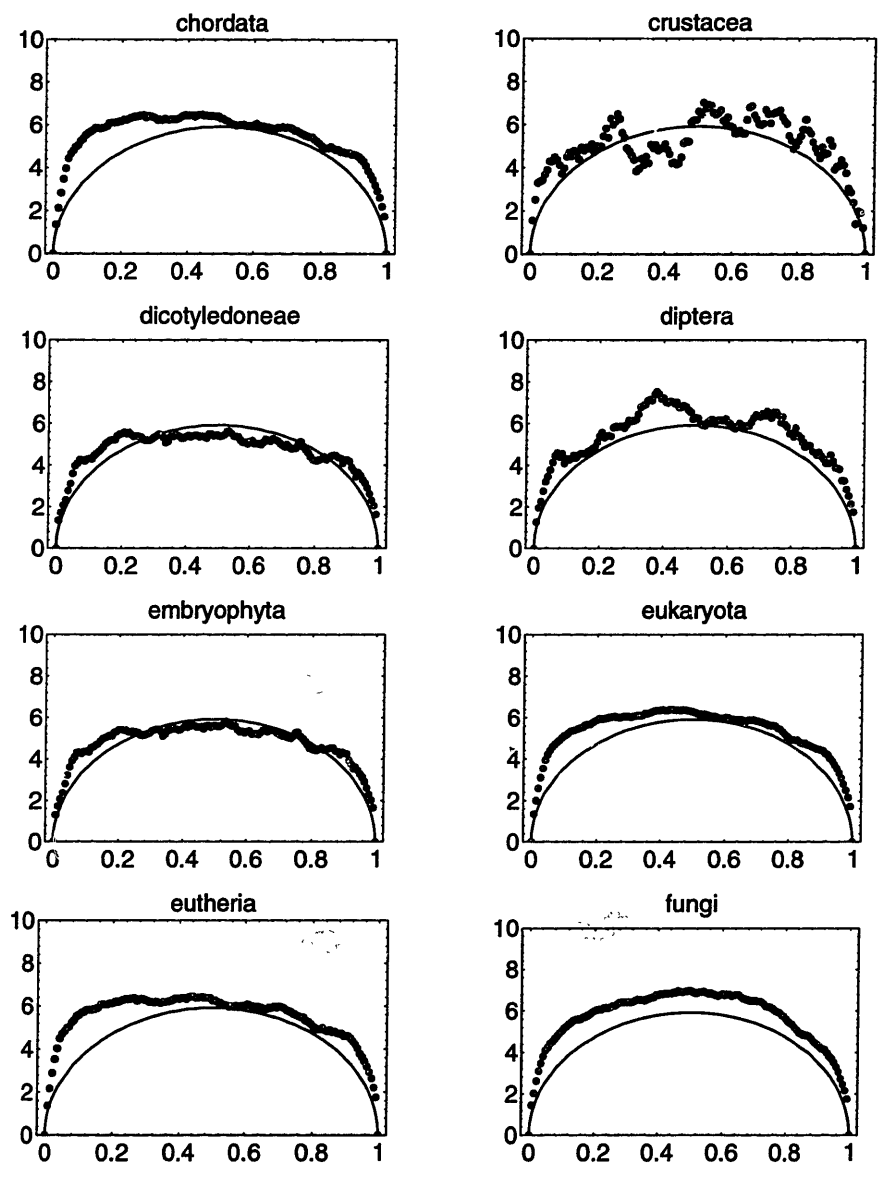


Figure 11-10: Bridges of different species for Hydrophobic/hydrophilic mapping

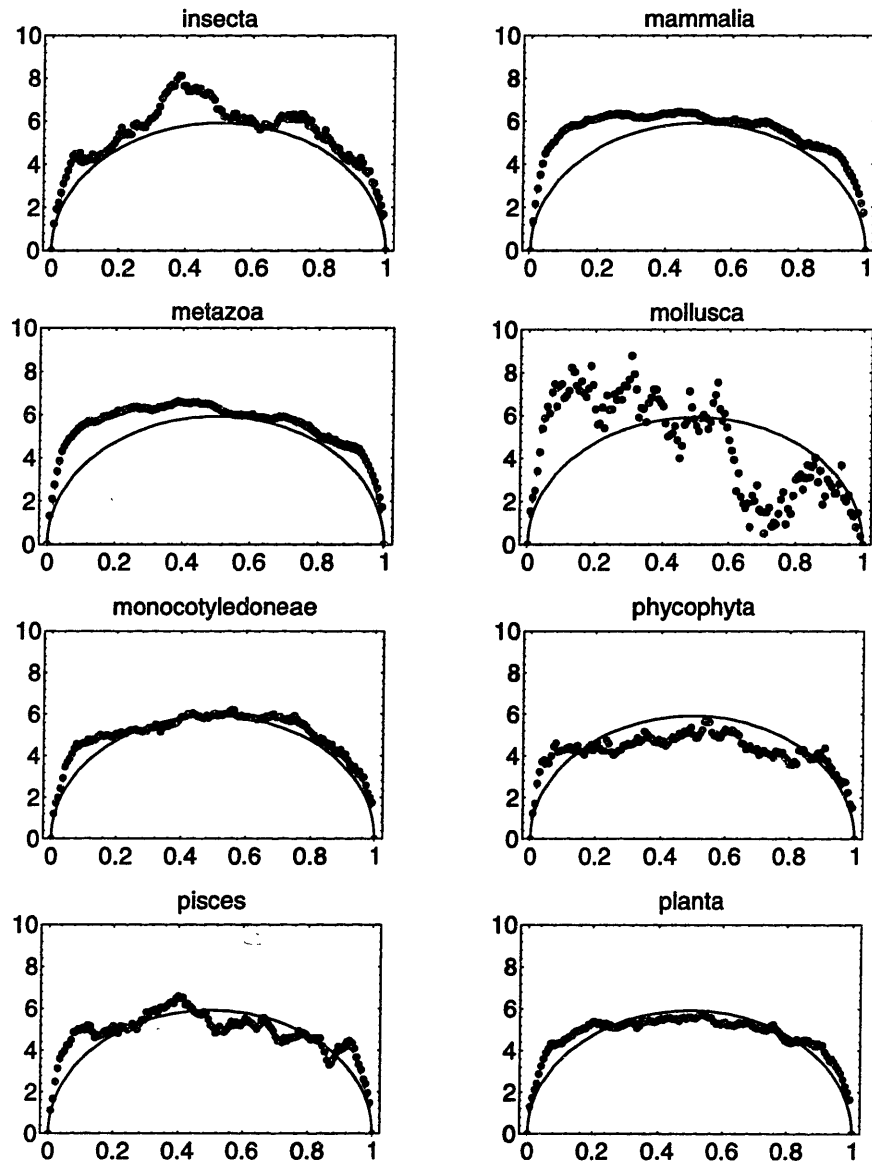


Figure 11-11: Bridges of different species for Hydrophobic/hydrophilic mapping

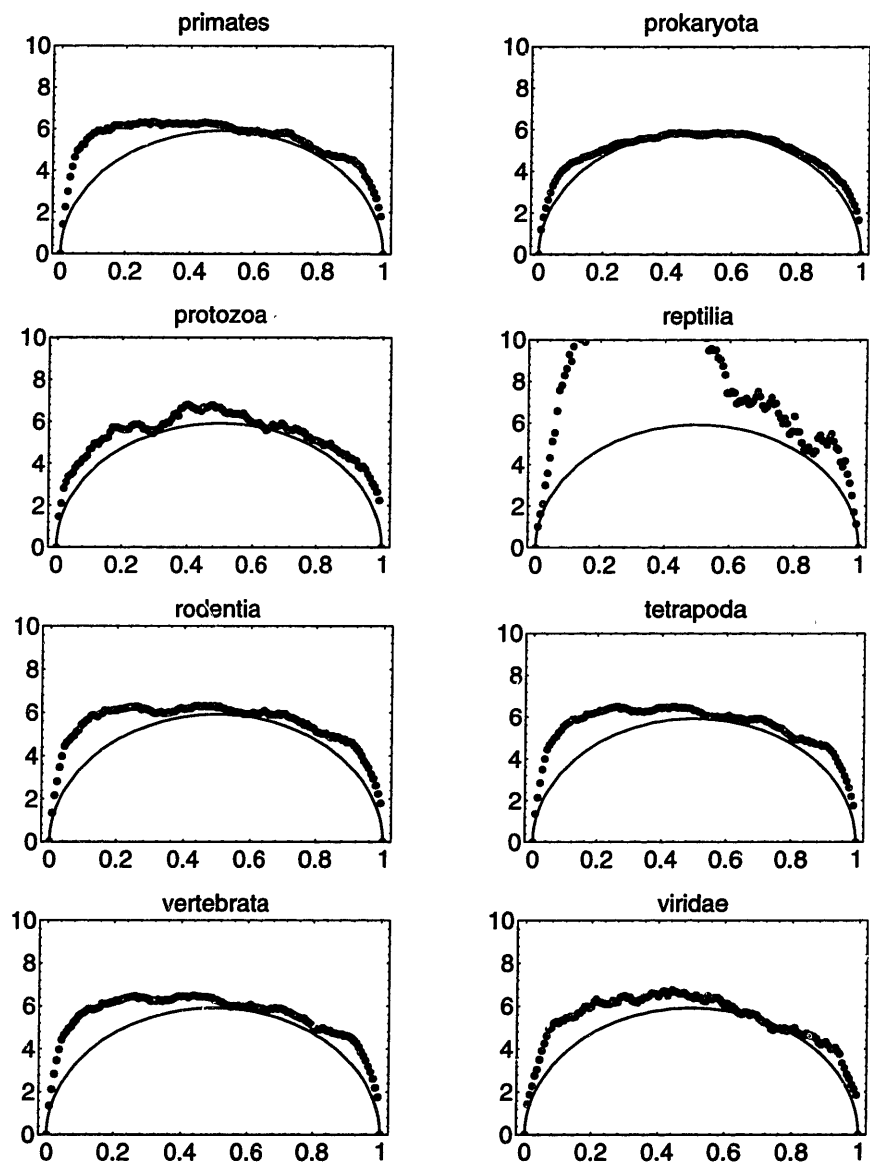


Figure 11-12: Bridges of different species for Hydrophobic/hydrophilic mapping

# Chapter 12

## NMR Analysis

Recently, multiple density phases in heteropolymer gels were discovered. To examine the differences between these phases, Nuclear Magnetic Resonance spectroscopy was performed. For a gel which normally has four phases, it was found that the two collapsed phases were similar and solid-like whereas the two swollen phases were liquid like and different from each other as well as the solid-like phases.

### 12.1 Introduction

The swelling phase transition of gels has been rigorously studied experimentally and theoretically [Shi93]. A reason for this study is the numerous industrial applications as well as a good model of polymer behavior, visible on a macroscopic scale: since the gel is essentially one large macromolecule, physical properties of single polymer chains (such as the coil to globule transition) have gel analogs.

Thus, it is most common that a gel swelling transition will occur between two phases: one in which entropy dominates (swollen for hydrogen bonding ionic interactions; collapsed for hydrophobic interactions) and one in which energy dominates (collapsed for attractive interactions; swollen for hydrophobic interactions). Recently, copolymer gels with *multiple phases* have been discovered [Ann92]; upon varying some external parameter such as pH or temperature, one can obtain

hysteresis curves which have different values of density of a given pH, for example.

The nature of the complicated hysteresis curves for gels with multiple phases is strongly related to the monomer species composition. On the homopolymer limits (gels made with only one of the comonomers), one recovered the standard two phase behavior common to homopolymer gels. As one approaches some mixture between the two types of monomers, the multiple phase behavior appears. Thus, the nature of composition is directly linked to the physical phase behavior.

This is reminiscent of the chapter on the freezing transition of random heteropolymers. We expect that the polymer chains in the gels with multiple phases have essentially random sequences (assuming there are no strange effects due to extreme differences in polymerization reaction rates between different monomer species). Thus, it is not surprising to consider that the physical behavior of the system will be strongly related to the nature of the composition.

Why bother studying this system? The study of bioheteropolymers is very important in molecular biology (eg. the protein folding problem). However, it remains unclear just what effect evolution has had on the selection of protein sequences, for example. Thus, philosophically, it is much "cleaner" to work on a system in which the physical question is clear. In this case, given a heteropolymer gel, what is the effect of the heteropolymeric nature of the system on its phase behavior.

While this has been extensively studied in terms of density measurements [Ann92,Ann93], little else is known. One technique to unlock the secrets of the nature of these phases is to examine the phases using Nuclear Magnetic Resonance (NMR). NMR can yield descriptions of the nature of the phases (whether they may be rigid and solid like or more liquid like) as well as microscopic structure. In this chapter, we will examine the NMR spectra of a copolymer gels previously studied and known to have four phases.

## 12.2 Experimental

The recipe for the gel was identical to that studied by Annaka and Tanaka [Ann92]. Specifically, 480 mM methacryl-amido-propyl-trimethyl-ammonium-chloride (MAP-TAC), 220 mM Acrylic acid (AAc), Bisacrylimide (BIS), Amonium persulfate (APS). The pregel solution was poured into a test tube and gelled at 60° C overnight. Next the gel was removed from the test tube, crushed through a 1mm filter, and then thoroughly washed. Next, the gel was placed on telfon sheets under the fume hood and completely dried. To completely ensure that there was no water in the sample (which would lead to a large water peak in the NMR analysis, swamping out all other peaks), we placed the gel in a vial and heated it at 80° C for a day.

The completely dry gel was rehydrated with D<sub>2</sub>O and the appropriate phase was reached by the appropriate addition of NaOH and HCl. At each of the four phases in the phase diagram, a sample was taken and placed in a 5mm NMR tube.

NMR analysis was performed using the home made 600 MHz instrument at the MIT Magnet Lab. The parameters used are summarized in Figure 12-1.

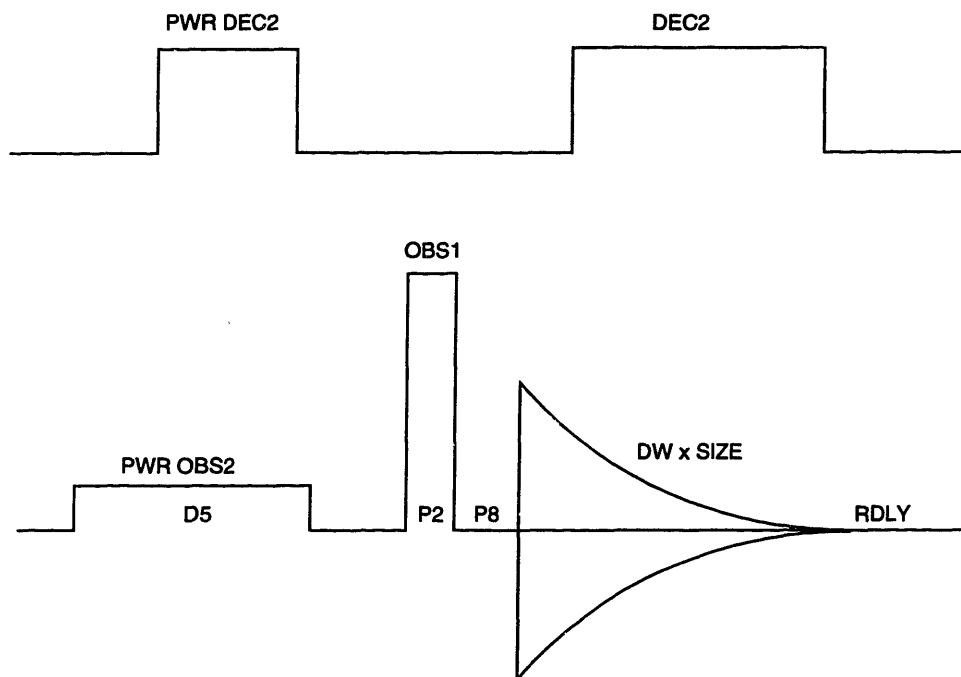
## 12.3 Results and Discussion

The NMR spectra for each phase is shown in Figure 12-2. Immediately, we see that these phases differ not only in density but in their NMR spectra. Interestingly, the spectra for the two most collapsed phases are very similar and are both very different from the two most swollen phases.

To get some idea of what these spectra mean, lets consider the spectra obtained by examining the spectra of solutions of just the monomers ....

Therefore, since the spectra of the gels samples are not just superposition of the spectra for monomers, the phases represent some complicated arrangement (and therefore interactions) of monomer species. Furthermore, the three swollen phases represent radically different monomer arrangements.

We can also make some estimate of the nature of the phases in terms of the



Transmitter Parameters:

OBS1 = 100.0 DB  
 OBS2 = 80.0 DB  
 DEC1 = 80.0 DB  
 DEC2 = 80.0 DB

Receiver Parameters:

GAIN = 40.0 DB  
 SW = 10.0 kHz  
 SIZE = 8192  
 AT = 0.819 sec

Pulse Program Parameters:

Pulses:

P2 = 15.0  $\mu$ sec  
 P8 = 30.0  $\mu$ sec

Delays:

D5 = 2000.0  $\mu$ sec  
 RDLY = 1000.0  $\mu$ sec

Figure 12-1: Pulse sequence and parameters used in the NMR analysis.



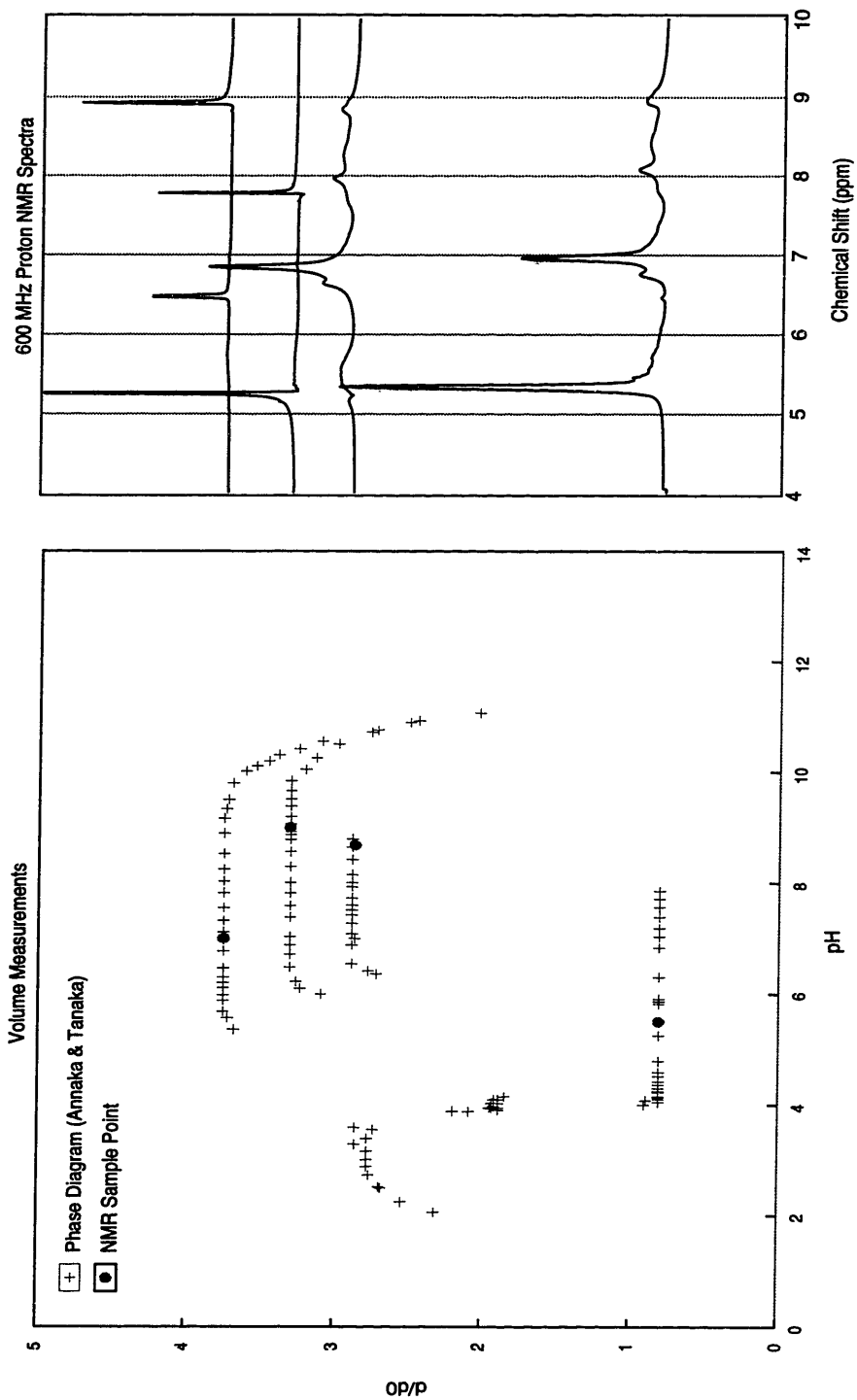


Figure 12-2: NMR Spectra for Acrylic Acid (480) MAPTAC (220) heteropolymer gel.

rigidity of the sample. Specifically, solid-like signals are broader due to the lack of translational narrowing. Therefore, the two collapsed phases appear to be solid-like whereas the two swollen phases are liquid-like. The existence of two liquid-like phases is an interesting situation which has not been addressed either theoretically or experimentally.

In terms of comparing the phenomenon of multiple phases of gels with the theories of heteropolymer freezing discussed in the previous chapters, we must stress that there are several potentially fundamentally different aspects:

1. The previous theories were for single chain behavior and not for gels. Physically, we would expect that this would have a major impact. Specifically, the physical source of freezing in single polymer chains is the quenched sequence and polymeric bonds which prohibit certain monomer configurations in space, thereby causing frustration. With multiple heteropolymer chains in a gel, the degree of frustration is substantially less since while a single chain may not be able to reach a given low energy monomer configuration, such a configuration may be achieved by many chains; in other words, the fact that the monomers are now quenched on several chains greatly increases the degrees of freedom available thereby reducing the frustration.

Furthermore, it may be fallacious to apply the results for branched heteropolymers with a quenched sequence [Gut93] to gels since the branched heteropolymer case studied had the double frustration of quenched sequence and quenched branching whereas the collapsed gel may not have such constraints.

2. The previous theories involved globular heteropolymers, whereas the interesting aspect of the multiple phases discovered was the numerous degree of swollen phases.
3. Multiple phases have been found in Acrylic acid gels, which may be formally considered a homopolymer, but are an annealed heteropolymers since the monomers can be either neutral, ionized, or hydrogen bonded. Here, there is

no quenched sequence and the only form of frustration in the system is the quenched crosslinks.

4. Finally, it is unclear if there are specific aspects of the polymerization performed to create heteropolymer gels with multiple phases which lead to sequences which are not random and that the nature of the phases represents physical effects due to this non-randomness.

In conclusion, it is clear that there is some complicated behavior involved in the nature of these phases. However, this very preliminary NMR analysis cannot give too much information beyond this characterization.



## **Part V**

# **Conclusions**

1000

2

1000

# Chapter 13

## Summary

From a physicist's point of view, one remarkable property of proteins is the non degeneracy of the ground state (in a coarse grained sense) and the ability to fold quickly and reliably to this state. Understanding the statistical physics of this "freezing" transition will shed light on how proteins have evolved to their present state and how one can create synthetic protein-like heteropolymers capable of functions one desires. In this thesis, we studied the freezing transition using a variety of techniques, including mean field replica analytic treatments, exact thermodynamics by computational enumeration, Monte Carlo kinetics simulations, and NMR analysis.

As it is difficult and not always particularly enlightening to describe the behavior of a particular sequence, we examined ensemble averages. To some zeroth approximation, proteins are no different from an ensemble of random sequences; in fact, it has been previously shown that even random sequences have a freezing transition. Therefore, we investigated the freezing behavior of random sequences, to hopefully gain some understanding of the role of the polymeric frustrations in the freezing transition.

However, it is suspected that proteins are not random, but have been optimized in some manner. First, we quantitatively examined the nature of this optimization in the statistics of protein sequences. Next, to model proteins, we considered an ensemble of sequences selected such that they minimize the monomer monomer in-

teraction energy in a particular conformation. This optimization of interactions can also be considered in terms of what we call “Imprinting”: specifically, consider a laboratory experiment in which monomers are allowed to interact before polymerization at some temperature sufficiently low such that the interaction energy between the monomers leads to some low energy configuration; after rapid polymerization, the resulting polymer conformation is in an optimized, low energy conformation due to the optimization before polymerization. An interesting question to ask is whether the optimization of the monomers before polymerization is sufficient to cause the polymerization conformation to be the ground state. If true, then the placement of perturbing fields in the presence of the monomers prior to polymerization will allow one to make the analogous perturbations in the resulting polymerization conformation and therefore the ground state. For example, if a given “target” molecule is used as the perturbing field by placing this molecule in the monomer soup prior to polymerization, one expects that the resulting polymerization conformation will have a complementary “active site” capable of specifically recognizing the target molecule.

Examining the freezing transition analytically is an interesting physical problem since the heteropolymeric sequences are quenched. Thus, the polymeric bonds cause frustration, much like that found in spin glasses. In fact, one can consider the spin glass case as a quenched lattice with an annealed sequence and the heteropolymer case as a quenched sequence with an annealed lattice (polymer conformation). Therefore, it is not surprising that many of the tools and physical intuition associated with spin glasses can be carried over to the heteropolymer problem. Specifically, to calculate the free energy, we employed the replica trick and examined the conditions for conformation overlap between pure states (replicas). Within the replica framework, we explicitly considered the effect of an ensemble of designed heteropolymers. The resulting phase diagram detailed the relationship between renaturation to the polymerization conformation or some random conformation and the polymerization and acting temperatures. Also, we analytically examined the freezing transition for arbitrary interactions and detail the relationship between the nature



of the interaction matrix and aspects of the freezing transition.

Computationally, one can examine designed or “Imprinted” sequences by enumerating every possible globular conformation. Since the number of conformations grows quickly with the number of monomers, we employed a massively parallel supercomputer and a special work stealing algorithm to carry out the enumeration. Enumeration has the benefit that the entire energy spectrum of a given heteropolymer sequence can be examined and one can guarantee that the ground state is non-degenerate. Furthermore, we examined the connections between the energy spectrum (a purely thermodynamic property) and the folding kinetics. Enumeration confirms that the polymerization conformation of Imprinted sequences is the unique ground state in 60% of the sequences.

Folding kinetics suggest whether a given sequence cannot just renature thermodynamically (i.e., with infinite time), but rather whether the sequence will fold just as proteins do: quickly and reliably. We verified for Imprinted sequences, previous results for evolutionary designed chains that the kinetic renaturation is linked to the nature of the energy spectrum: when there is a large gap in energy between the ground and first excited states, the folding is quick and reliable.



# Chapter 14

## Future Work

At the end of anything, it is natural to ask where do we go from here? There are two aspects of this work, while naturally intertwined in this thesis, will probably diverge in terms of future work.

### 14.1 Experimental Realization of Imprinting

The future of any experimental realization of Imprinting is most likely intimately related to the ability to polymerize monomers in the manner suggested in this work. Indeed, there seems to even be a lack of consensus on the nature of polymerization; some consider it a tumultuous process in which any Imprinting-like prearrangement would be lost, while others feel that with the appropriate conditions (and potentially more exotic polymerization schemes), Imprinting-like polymerization may indeed be feasible. At this point, however, the success of Imprinting shifts from a physical question about the nature of optimization and heteropolymer folding to details of the nature of monomers and methods chosen to facilitate Imprinting-like polymerization.

However, there are already some interesting experiments which indicate that Imprinting-like optimization may be preserved during polymerization. For example, in a recent experiment on copolymer gels, Yu and Tanaka [Yu93] have examined gels consisting of monomers capable of existing in three states: hydrogen bonded,

neutral, or ionized. At low pH, the monomers are not ionized and tend to hydrogen bond. Therefore, gellation at low pH (in spirit equivalent to low polymerization temperature  $T_p$ ) should lead to a greater degree of hydrogen bonding between monomers. Gels were made at low pH (presumably with hydrogen bonds) and high pH (presumably without hydrogen bonds). The swelling behavior of these two gels is consistent with the existence of the preservation of the pre-gellation hydrogen bonding conditions, as the high pH gel swelled to a higher degree than the low pH gel. Therefore, this experiment potentially indicates that Imprinting optimization may indeed be preserved during polymerization.

Also, the theory of Imprinting detailed in this work is for single polymer chains. Unfortunately, single polymer chains are a great experimental challenge, since in order to avoid aggregation one must keep the chain density low, but in order to get some good signal to noise ratio, one must have enough sample. This competition between noise and aggregation may be overcome by grafting single polymer chains to other, more inert bodies, but one must be careful that these new bodies do not contribute either. On the other hand, gels do not have this problem and for this reason have been a good testbed for polymer physics since many single polymer chain phenomena have gel equivalents. However, the delicate nature of freezing, for example the nature of frustration imposed by polymeric bonds, may have startlingly different manifestations in gels versus single polymer chains. Thus, while gels do not have the experimental difficulties associated with single polymer chains, the nature of the corresponding theory for gels is not clear.

## 14.2 Correlations in Protein Sequences

Finally, the work presented here on the correlations in protein sequences is clearly very preliminary. First, the basic result itself could be placed on firmer ground by addressing certain potential flaws in our methodology. The protein ensembles employed were chosen based upon keywords found in the database. Thus, potential problems include bias in proteins studied due to which proteins experimentalists

have chosen to examine as well as potential oversampling of proteins due to the listing of homologues. A careful “cleaning” of the dataset would set this question to rest.

Also, work could be done to push these ideas further. Clearly, we expect the nature of the tertiary folds to be important; indeed, we have hypothesized that the correlations found were remnants of optimization of the energy of the tertiary fold. Incorporation of the protein structure into the analysis of the question “was there a physically driven phase of evolution?” might lead to more insights into the problem.

### 14.3 Solution of the Protein Folding Problem

The solution to the protein folding problem is still unsolved and remains elusive for several reasons. Clearly, a quantum-chemical solution is intractable for a system with easily 2000 atoms. Stepping down the ladder of computational difficulty versus realistic modeling, molecular dynamics also has its limitations for protein folding since it is not completely tractable (still due to long lengths) and must employ approximated potentials. Certainly, brute force methods like these will not be the solution. Perhaps a greater understanding of the physics of protein folding will lead to more sophisticated computational treatments.

One example of such an approach is to develop more realistic potentials based upon our knowledge of the physics of protein folding and the statistics of known protein conformations. Making certain assumptions about the nature of evolutionary optimization in proteins, for example as we did in this thesis, one may be able to make reasonable potentials for use in crude lattice based exhaustive searches or Monte Carlo methods.

In the end, as it is clear that a sophisticated method is not just preferable but required, understanding the physics involved will be of paramount importance.

1/2/20

# Bibliography

- [Abe81] H. Abe, N.Go, *Biopolymers* **20**, 1013 (1981).
- [Amit87] D. J. Amit, H. Gutfreund, and H. Sompolinsky, *Annals of Physics* **173**, 30 (1987).
- [Ann92] M. Annaka and T. Tanaka, *Nature* **355** 430 (1992).
- [Ann93] M. Annaka, D. Berling, J. Robert, and T. Tanaka, *Macromolecules* **26** 3234 (1993).
- [Bai92] A. Bairoch and B. Boeckmann, *Nucleic Acids Res.* **20**, 2019 (1992).
- [Bry87] J. D. Bryngelson and P.G. Wolynes, *Proc. Nat. Acad. Sci., USA* **84**, 7524 (1987).
- [Bry94] J. D. Bryngelson, *J. Chem. Phys.* **100**, 6038 (1994).
- [Cat88] M. E. Cates and R. C. Ball, *J. Phys. (Paris)* **49**, 2009 (1988).
- [Der80] B. Derrida, *Phys. Rev. Lett* **45**, 79 (1980).
- [Diao94] Y. Diao, *J. Stat. Phys.* **74**, 1247 (1994).
- [Doi86] M. Doi, M. and S. F. Edwards, *Theory of Polymer Dynamics* Clarendon Press, Oxford (1986)
- [Dre90] D. Dressler and H. Potter, *Discovering Enzymes*, (Scientific American Library, New York 1990).

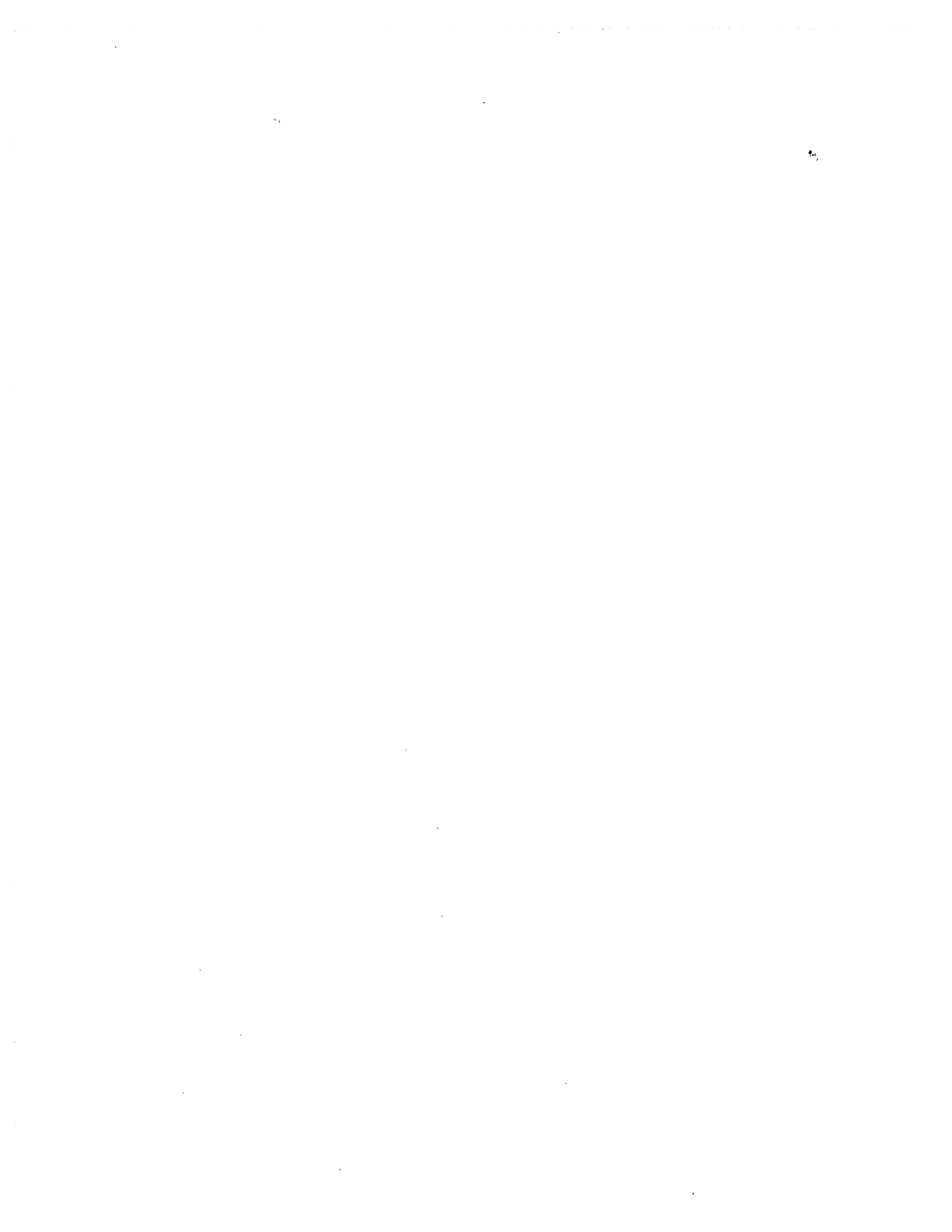
- [Edw88] Edwards, S.F., Muthukumar, M., *J. Chem. Phys.* **89**, 2435 (1988).
- [Fin93] A. V. Finkelstein, A. M. Gutin, and A. Ya. Badretdinov, *FEBS Letters* **325**, 23 (1993).
- [Flo53] P.J. Flory, *Principles of Polymer Chemistry*, Cornell University Press (1953).
- [Fre91] G. H. Fredrickson and S. T. Milner, *Phys. Rev. Lett.*, **67**, 835 (1991).
- [Gar88a] T. Garel and H. Orland, *Europhys. Lett.* **6**, 307 (1988).
- [Gar88b] T. Garel and H. Orland, *Europhys. Lett.* **6**, 597 (1988).
- [Gar94] T. Garel, L. Leibler, H. Orland, *J. Phys. II (Paris)* **4**, 2139 (1994).
- [Gol92] R. A. Goldstein, Z. A. Luthey-Schulten, , P. G. Wolynes, *Proc. Nat. Acad. Sci., USA* **89**, 4918 (1992).
- [Gro94] A. Yu. Grosberg and A. R. Khokhlov, *Statistical Physics of Macromolecules*, (AIP, New York, 1994).
- [Gut93] A. M. Gutin, A. Yu. Grosberg, and E. I. Shakhnovich, *J. Physics* **A26** 1037 (1993).
- [Hal94] M. Halbherr, Y. Zhou, and C. Joerg, *International Workshop on Massive Parallelism: Hardware, Software, and Applications* (1994).
- [Hon90] J. D. Honeycutt and D. Thirumalai, *J. Chem. Phys.* **93**, 6851 (1990).
- [Kim83] M. Kimura, *The Neutral Theory of Molecular Evolution*, Cambridge Univ. Press, New York (1983) .
- [Lau89] K. F. Lau and K. A. Dill, *Macromolecules* **22**, 3986 (1989).
- [Lif78] I. M. Lifshits, A. Yu. Grosberg, A. R. Kokhkolov, *Rev. Mod. Phys.* **50**, 683 (1978).
- [Met53] N. Metropolis, *et al*, *J. Chem. Phys.* **21**, 1087 (1953).



- [Mez84] M.G. Mezard, G. Parisi, N. Surlas *et al.*, *J. Phys. (Paris)* **45**, 843 (1984).
- [Mia85] Miyazawa and Jernigan, *Macromolecules* **18**, 534 (1985).
- [Nis79] I. Nishio, S-T. Sun, G. Swislow, and T. Tanaka, *Nature* **281**, 208 (1979).
- [Obu90] Obukhov, S.P., *Phys. Rev.* **A42**, 2015 (1990).
- [Pan94a] V. S. Pande, C. Joerg, A. Yu. Grosberg, and T. Tanaka, *J. Phys.* **A27**, 6231 (1994).
- [Pan94b] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, *Proc. Nat. Acad. Sci., USA* **91**, 12972 (1994)
- [Pan94c] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, *Proc. Nat. Acad. Sci., USA* **91**, 12976 (1994)
- [Pan94d] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, *J. Chem. Phys.* **101**, 8246 (1994).
- [Pan94e] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, *J. Phys. (Paris)* **4**, 1771 (1994).
- [Pan95a] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, *Phys. Rev.* **E51**, 3381 (1995).
- [Pan95b] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, *Macromolecules* **28**, 2218 (1995).
- [Pan95c] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, *J. Phys. A*, *in press* (1995).
- [Pan95d] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, *J. Chem. Phys.*, *in press* (1995).
- [Pan95e] V. S. Pande, A. Yu. Grosberg, and T. Tanaka, *in preparation* (1995).
- [Par80] G. Parisi, *J. Phys.* **A13**, 1887 (1980) .

- [Pau65] L. Pauling quoted in D. Dressler and H. Potter, *Discovering Enzymes*, (Scientific American Library, New York 1990).
- [Peng92] C.-K. Peng, *et al*, *Nature* **356**, 168 (1992).
- [Pti86] O. B. Ptitsyn and M. V. Volkenstein, *J. of Biomol. Structure and Dynamics* **4**, 137 (1986).
- [Ram94] S. Ramanathan and E. I. Shakhnovich, *Phys. Rev.* **E50**, 3907 (1994).
- [Sali94a] A. Sali, E. I. Shakhnovich, and M. Karplus, *Nature* **369** 248 (1994).
- [Sali94b] A. Sali, E. I. Shakhnovich, and M. Karplus, *J. Mol. Biol.* **235**, 1614 (1994).
- [Sfa93] C. Sfatos, A. Gutin, and E. I. Shakhnovich, *Phys. Rev.* **E48**, 465 (1993).
- [Sfa94] C. Sfatos, A. Gutin, and E. I. Shakhnovich, *Phys. Rev.* **E50**, 2898 (1994).
- [Sha89a] E. I. Shakhnovich, and A. Gutin, *Biophysical Chem.* **34**, 187 (1989).
- [Sha89b] E. I. Shakhnovich and A. M. Gutin, *Studia Biophysica* **132**, 47 (1989).
- [Sha90a] E. I. Shakhnovich, and A. Gutin, *J Chem. Phys.* **93**, 5967 (1990).
- [Sha90b] E. I. Shakhnovich and A.M. Gutin, *Nature* **346**, 773 (1990).
- [Sha91] E. I. Shakhnovich and A. M. Gutin, *J. Theor. Biol* **149**, 537 (1991).
- [Sha92] E. I. Shakhnovich and M. Karplus in *Protein Folding* Ed. T. Creighton (1992)
- [Sha93a] E. I. Shakhnovich, *Phys. Rev. Lett.* **72** 3907 (1994).
- [Sha93b] E. I. Shakhnovich and A. M. Gutin, *Proc. Nat. Acad. Sci., USA* **90**, 7195 (1993).
- [She75] D. Sherington and S. Kirpatrik, *Phys. Rev. Lett.* **35**, 1792 (1975).
- [Shi93] M. Shibayama and T. Tanaka, *Advances in Polymer Sciences* **109**, 1 (1993).

- [Soc94] N. D. Socci and J. N. Onuchic, *J. Chem. Phys.*, **101**, 1519 (1994).
- [Wol91] P. G. Wolynes, in *Spin Glass Ideas in Biology*, edited by D. Stein (World Scientific, Singapore, 1991).
- [Yu93] X-H. Yu, *Polymer Interactions and the Phase Transition of Gels*, MIT Doctoral Thesis (1993).
- [Yue92] K. Yue and K. A. Dill, *Proc. Nat. Acad. Sci., USA* **89**, 4163 (1992).
- [Yue95] K. Yue, K. M. Fiebig, P. D. Thomas, H. S. Chan, K. A. Dill, and E. I. Shakhnovich, *Proc. Nat. Acad. Sci., USA* **92**, 325 (1995).



# THESIS PROCESSING SLIP

FIXED FIELD: ill. \_\_\_\_\_ name \_\_\_\_\_

index \_\_\_\_\_ biblio \_\_\_\_\_

► COPIES: Archives Aero Dewey Eng Hum  
Lindgren Music Rotch Science

TITLE VARIES: ►  \_\_\_\_\_

NAME VARIES: ►  \_\_\_\_\_

IMPRINT: (COPYRIGHT) \_\_\_\_\_

► COLLATION: 132 p

► ADD. DEGREE: \_\_\_\_\_ ► DEPT.: \_\_\_\_\_

SUPERVISORS: \_\_\_\_\_

NOTES:

cat'r: \_\_\_\_\_ date: \_\_\_\_\_  
► DEPT: Phy page: 560  
► YEAR: 1995 ► DEGREE: Ph.D.  
► NAME: PANDE, Vijay Satyanand