Approaches to Determining the Three-Dimensional Structure and Dynamics of Bacterial Chromosomes

by

Matthew A. Wright

B.A. Chemistry, B.M. Music Performance
University of Southern Maine, 1999

Submitted to the Department of Chemistry
in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY
in Physical Chemistry

at the

Massachusetts Institute of Technology

September 2005

Signature of Author ____________________
Department of Chemistry
September 1, 2005

Certified by ____________________
George M. Church
Professor of Genetics, Harvard Medical School
Thesis Supervisor

Accepted by ____________________
Robert W. Field
Chairman, Departmental Committee on Graduate Students

This doctoral thesis has been examined by a Committee of the Department of Chemistry as follows:

Professor Keith A. Nelson________________ ______________________________

Committee Chairman

Professor George M. Church ____

Professor of Genetics, Harvard Medical School

Thesis Supervisor

Professor Leonid Mirny____

Associate Professor of Physics, Massachusetts Institute of Technology

Committee Member

Approaches to Determining the Three-Dimensional Structure and Dynamics of Bacterial Chromosomes

by

Matthew A. Wright

Submitted to the Department of Chemistry
in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in
Chemistry

Abstract

The information in genomes is only partially contained in the linear sequence of their nucleotides. Their folding into dynamic three-dimensional structures creates spatial relationships between loci that likely play important functional roles. Yet so far only the broad outlines of this spatial organization have been discerned.

In chapter 2 of this thesis I describe a general constraint-based framework for defining the configuration space of chromosomes. Analogous to protein structure determination through NMR, such a framework allows the quantitative reduction of the conformation space down to the level of a single structure or an ensemble of structures. It is compatible with both experimentally determined and theoretical constraints, particularly those motivated by evolutionary optimality.

In chapter 3, I describe the first method to search for signals of large-scale three-dimensional structure in genome sequences. The results suggest that there is strong selection for three-dimensional relationships within the chromosome, particularly those related to transcription. The signals generated recapitulate both known structural data from microscopy and functional data on genome-wide transcription levels. Moreover, a detailed analysis of these signals in *E. coli* suggests previously unknown structural features including chromosome-long periodic looping and an axis of high transcriptional activity. There are immediate applications to other bacteria and potentially to eukaryotes.

Thesis Supervisor: George M. Church
Title: Professor of Genetics, Harvard Medical School

Table of Contents

I dedicate this thesis to my parents, Bonnie and Tom Wright, for their love and encouragement, to my grandfather, Robert E. Arnot, and uncle, Robert B. Arnot, who ignited in me a passion for medical science, and to my undergraduate chemistry advisor, Henry J. Tracy, who first showed me the beauty of basic scientific research.

Acknowledgements

It was my grandfather, Robert Arnot, who introduced me to science, through my first scientific tool, a calculator, given to me at age 2. Although it was the buttons that fascinated me, I like to think that a bit of arithmetic wore off during the months that followed while the calculator never left my side. My grandfather was an enormous influence on me, through his example as a fine physician and remarkable human being, and through his passion for science as a means for human good. He brought the mindset of a researcher to the care of patients long after he left his own formal work in psychopharmacology.

It is my father from whom I inherited a love for mathematics. He is among the few people I know who takes on a glow at the mere mention of the words "linear algebra" and can spend an evening discussing abstract multidimensional spaces. For better or for worse I seem to have inherited these traits. Moreover, he has a seemingly insatiable curiosity, evidenced by a constant and ever-changing stack of scientific texts heaped upon his desk. His curiosity serves as a reminder to always inquire.

I thank my mother who has been a solid support for me and for my entire family through many trials; additionally she lovingly guided me from music into science. I thank my piano teacher Laura Kargul who both enriched my life through 7 years of teaching and helped to open me to possibilities outside of music. I thank my family as a whole, my grandmother, Mary Arnot, my siblings,Tom Doug and Emily, and my uncle Robert, an important support. I thank also many friends, in particular Brad Coull, David Goodman, Andrew Butterworth, Tracey Orick, and Janice Sullivan, Cathy O'Neil and Jen Johnson,

close friends and inspiring mathematicians, and Serhan Altunata, an excellent physicist. I thank the entire Sievers family, quite a scientific clan, in particular Kathryn for her quick mind, LeRoy, for his wonderful practicality as a physicist/engineer, and Charlie, for frequent conversations on physics and Fourier analysis.

Stephen MacDonald first introduced me to the beauty of higher mathematics, in a calculus course that I originally took as a core math requirement in college. His brilliant teaching, somehow allowed his students to always see one step ahead in the web of mathematical logic and convinced me to take one, then two, then three more courses with him while still a piano performance major. Ultimately this led me to embrace science as a career.

John Ricci's, patient and clear teaching of general chemistry and then physical chemistry introduced me to the mathematics of chemical structure. And his long talks with me illuminated the possibilities of a life in science.

It was Henry Tracy, though, who first introduced me to scientific research. He inspired in me a true passion for science by his enormously enthusiastic classroom teaching and through his guidance in my first independent research project, a year-long organo-metallic synthesis, which, for him, given my lack of synthetic ability was a true labor of love. Outside of chemistry, he exposed me to a broad range of science, from quantum mechanics to astronomy, which he introduced me to through his own telescope from the fields outside his home in Maine. He became a mentor and a good friend.

In my brief stay at Caltech, I had the privilege to work with a truly great theorist, Rudolf Marcus, and to listen as he worked through the complex problem of ion-transport across interfaces. During this time, I gained an appreciation for the multiple angles a theorist must approach a problem from and the multiple fields a theorist must therefore draw upon. I was also fortunate to work with Niles Pierce and ErikWinfree, both highly interdisciplinary scientists who encouraged me to be broad in choosing problems.

It is George Church and all of the scientists of the Church lab who have been the enablers of my doctoral work and who have been the teachers that have nurtured my growth as a scientist over the past 5 years. I thank everyone I've known in the lab but must mention a few in particular. Tzachi Pilpel, was my first mentor here and offered me important encouragement and guidance. He is a wonderful scientist, with a keen mind, amazing clarity of thought, and an amazing ability to communicate. He continues to serve as a model and inspiration. Kun Zhang, a fabulous experimentalist, has been both a patient teacher and friend. Farren Isaacs, has been an excellent editor and a sharp mind. Xiao Xia Lin has been a great support. Peter Kharchenko, an excellent and critical colleague. Kyriacos Leptos, an avid conversationalist on language and music. Dana Pe'er has been both an inspiring example, and a source of wonderful encouragement. Mark Umbarger, a strong support, and a rigorous critic. Doug Selinger a fellow philosopher and important friend. Yonatan Grad, an inspiration as a scientist, as a humanist, and also as friend. Jay Shendure, has been an important example, a startlingly quick mind which both critically grasps and will not relinquish the important details of a problem but which also views these details within the compass of the whole. His is a mind simultaneously critical and visionary and practical. Nikos Reppas has managed to somehow be both a roommate and a bay-mate for three years. He has been my educator in the lab, fielding an average of 1 question every 5 minutes while I attempted experimental work, a constant editor and a fabulous example of both an amazing work ethic and extraordinary scientific rigor.

Daniel Segrè deserves pages of his own. He has been my closest mentor. Indeed much of the work of this thesis is as much his as it is my own. His mind is a philosopher scientist's and serves as a model of creativity. He fosters a work environment which simultaneously encourages free expression of ideas and hones these ideas critically to maturity. From him I have learned the process of theoretical science and been guided through this process during our chromosome structure work. His like-mindedness has made him both a scientific colleague and an intellectual colleague for talks on differential geometry, language, the vast span of physics, and music. He has become among my closest friends.

And finally there is George Church whose leadership for the lab consists in both his example, and in his creation of a place where it is possible to pursue almost any question of interest from sequencing to quantum computing. He fosters an environment of immense scientific possibility. More than this, he is a true model of the enlightened scientist, aware of the problems of the world and of science's intersection with them, always choosing problems within this broadest of contexts based, not on a narrowing of skills and interests, but on a broadening of them, on a rigorous examination of the current state of the world, identification of the problems of importance, and an incredibly expansive, inclusive, and creative set of approaches to solving them. I can think of no better example of what a scientist should strive to be. And I thank him for this time in his lab, his mentorship, and his inspiration.

Chapter 1

Introduction

Our current ways of thinking about the cell have been enormously influenced by the success of molecular biology and genetics. We think of genomes as sequences of letters with subsequences that can be deleted, inserted, inverted, or mutated. We represent the flow of information from DNA through mRNA through protein by strings where T's move to U's, and triplets are transformed to an alphabet of K's and Y's and W's. We represent interactions among genes, proteins, and metabolites by lines whose colors and ends delineate repression or activation.

Even as biology is broadening from the study of individual cellular components to system-wide measurements and models, our thinking is still heavily influenced by these abstractions. DNA binding sites for transcription factors are represented by linear weight matrices [1] genetic circuits are represented by diagrams similar to electrical circuits, and proteome-wide interaction maps are represented by exploding tangles of lines [2]. Our mathematical models too are colored by these abstractions. Reaction dynamics are still often modeled in the same way they were in the 1960's when we had not yet abandoned the conception of the cell as a homogeneously stirred mixture and thought of reactions as simple bimolecular $A + B \rightarrow AB$. [3]

In many cases these abstractions are extraordinarily useful, allowing us to discern logic that might have escaped our notice in more complex representations. And often these abstractions contain our full knowledge of the system since so much of biology is, of necessity, learned from purifying molecular components and analyzing them outside of their cellular context or from making genetic manipulations with simple phenotypic readouts. Yet still these abstractions are divorced from the notion of the cell as a physical object. We know that the cellular space is exquisitely organized. Macromolecular crowding can be so extreme that the cellular space is closer to liquid crystal than to aqueous solution [4]. DNA natively is not a simple string, but a double helix where subsequent nucleotides are rotated relative to each other.. Proteins are highly convoluted objects with marvelously intricate folds. Without attention to the true nature of this cellular space, we are robbed of important intuition - our intuition about arrows and lettered strings is very different from our intuition about physical objects moving and rotating and interacting. This kind of visual intuition has played a pivotal role in the history of science. Newton, it is said, first conceived of orbital motion by imagining the trajectory of a sphere thrown such that it falls at the same rate its horizontal motion takes it around the body it orbits. Maxwell conceived of electrodynamics by imagining tubes of fluid as the lines of a field. Kekule solved the structure of benzene in a dream about a snake biting its tail. And most famously, Einstein conceived of relativity in a series of thought experiments involving elevators and clocks and traveling on light beams.

Fortunately, the appreciation of cellular spatial organization is growing rapidly. Scientists talk now of interconnected networks of protein "machines" containing tens of proteins [3]. For example, in addition to the ribosome, we now refer to a replisome for DNA replication, large holoenzymes for transcription, and even complexes of metabolic enzymes as machines for metabolic pathways. Some have referred to the entire cell as a massive macromolecular organelle [5]. Advances in structural biology are yielding atomic resolution structures of complexes as large as the ribosome while, simultaneously, imaging from electron microscopy is reaching levels of resolution that allows known atomic domain structures to be fit into density maps of large cellular regions [6]. Such techniques are creating images of the cell components in their true space at particular moments in time, both revealing new complexes and discriminating between the many interaction partners found in techniques such as tandem affinity purification mass spectrometry and yeast two-hybrid analysis .

A change in our understanding of the spatial organization of prokaryotic cells is also underway. *In vivo* fluorescence imaging is revealing that these organisms which were until little more than a decade ago thought to be unstructured, have intricate patterns of protein localization that are carefully choreographed in time [7]. These dynamics have begun to reveal the precise mechanisms underlying processes such as the cell division cycle and the creation of cellular asymmetries [8]. Interestingly, such insights also suggest that the differences between prokaryotes and eukaryotes are much smaller than once thought; the finding that bacteria have cytoskeletal proteins and rapid chromosome segregation systems not unlike the spindle apparatus of eukaroyotic mitosis seems to indicate that many of the mechanisms governing fundamental cell processes are conserved across prokaryotes and eukaryotes [9].

Chromosome Folding

The genomic DNA that we envision as a linear sequence of letters is embedded in this complex cellular space, supercoiled or wound around nucleosomes and folded into higher order domains, compacted thousands of times its contour length. It intrinsically rotates an entire helical turn every ~10.6 base pairs such that contacts within and between binding proteins are affected by changes in the local helical twist. The entire information content of the genome is thus not fully realized without its embedding in this space including the local alterations of structure along the helix and the global coiling and looping that organizes genes and promoters and other loci in three-dimensions. It is this spatial organization that is the topic of this thesis.

The spatial organization of genomes has only recently come under scrutiny. In eukaryotes, the nucleus is beginning to be understood three-dimensionally in terms of distinct functional territories, and chromosomes in terms of domains with non-random spatial distributions [10]. An intricate relationship between chromosome structure and transcription has been revealed, extending from the level of nucleosome positioning, through the level of larger scale loops, to the positioning of entire domains [11]. In human cells, it has been shown that certain alleles on separate chromosomes are "kissing" - clustered together in space such that they regulate each other's transcription [12] – and that the repositioning of chromosomal loci can lead to transcriptional activation or silencing [11]. Within the nucleus there are recognizable structures which play known functional roles, the nucleolus for rRNA synthesis, structures for storage of splicing components, "factories" for transcription, and other structures of unknown function [11, 13]. Additionally, there are known proteins, lamins, that anchor various loci to the nuclear envelope and constrain their movement. All of this is yielding a view of the nucleus as an elaborately scaffolded or marvelously self-assembling entity, and of chromatin as a three-dimensional mesh whose geography is highly regulated and through which binding proteins diffuse and exchange, generating a highly interconnected network. The "codes" specifying nucleosome positioning and the compaction of DNA into condensed heterochromatin are under intense investigation, as are the ways in which larger chromosome loops are formed and regulated [14]. There are even indications of a broad relationship between chromosome spatial organization and disease; chromosome structure plays a role in triplet repeat disorders, malfunction of lamin disorders, and also in cancer [11].

While this global spatial organization is tremendously exciting, it appears that for now it will be difficult to unravel in eukaryotes. We do not yet know the rules governing either the folding or the positioning of genomic loci. Indeed, while the positioning of domains in territories is non-random, it also appears to be not completely deterministic [11]. Although this view may be complicated by cell lines used, and the anecdotal nature of the measurements, for now, until either the rules are better understood or better global means of measuring are found, descriptions of nuclear organization will likely be probabilistic.

Bacterial Chromosome Organization

In bacteria, where the cellular substructures are simpler, transcription and translation are coupled, and no envelope separates chromosome from cytoplasm, we expect the spatial organization of the genome to be particularly strongly related to function and the rules of folding perhaps simpler. Bacteria should therefore serve as a logical testing ground to derive the rules of chromosome folding. Notably, because of the recent commonalities discovered between mechanisms involved in prokaryotic and eukaryotic subcellular architecture, we may expect that at least certain rules governing bacterial chromosome folding may also be shared with eukaryotes. For these reasons, the folding of bacterial chromosome is the focus of this thesis.

The existence of significant spatial organization of the bacterial chromosome fold has been long under-appreciated. The same initial experiments that contributed to views of the bacterial cell as an unstructured space shaped the initial views of the chromosome [15]. In transmission electron microscopy images acquired over thirty years ago, the chromosome appeared as an amorphous mass, identifiable primarily by the exclusion of ribosomes. We now know that the harsh specimen preparation for these images - replacing solvents and soaking in heavy metals - were prone to generating artifacts. Electron microscopic images of chromosomes from lysed cells also showed seemingly disordered structures, notable for large numbers of loops, extending from the lysed membrane. This led to the view that the chromosome was effectively a disordered rosette of loops, extending from a central core. Consistent with this view, much *in vivo* experimental work has involved classification of "topological domains" of supercoiling which were thought to be equivalent to these loops [16]. The properties of these domains are important for both the nature of local chromosome compaction and for its relationship to transcription and replication. Therefore I discuss them in some detail below.

Topological Domains

In vivo, bacterial DNA is, topologically, a negatively supercoiled circle. The twist of the DNA along its helical axis (Tw) is supplemented with a "writhe" (Wr) which causes larger loops of the entire double helix to form [17]. This writhe is constrained within the structure by the joined ends of the circle and thus without a break in one of the DNA strands, the total amount of Tw and Wr is constant (called Lk, the linking number.) Negatively supercoiled DNA is energetically less stable than relaxed double helical DNA and thus, if a single stranded break allowing the DNA to freely rotate about the other strand is introduced, the supercoils will spontaneously be released from the structure. This relaxation should propagate until the entire structure is fully relaxed [17].

In nicking experiments *in vivo*, where a single stranded break is introduced at a particular position along the bacterial chromosome, however, the relaxation of supercoiling does not propagate through the entire structure but rather is constrained to certain regions [18]. These regions are referred to as topological domains because their local topology is insulated from the rest of the structure.

Variations in the local linking number forms a first level of conformational regulation above the level of the double helix. Without any topological change, the total linking number of the chromosome remains constant [17]. However, the local linking number can be changed by transcription, replication, or binding of DNA binding proteins such as nucleoid-structuring proteins, RNA or DNA polymerase, or transcription factors which may constrain local structure or unwind the double helix, thereby transforming twist into writhe. The edges of a topological domain are formed by such constraints and can alter the local linking number by constraining certain amounts of the global linking number within their boundaries to promote various effects. The potential energy of supercoiling can be used, for instance, to locally unwind the positively twisted DNA of the double helix necessary for both transcription and replication [17].

A recent set of discoveries suggest that there is an elaborate interplay between this topological structure of the chromosome and transcription. In particular, the expression level of hundreds of genes are affected by changes in supercoiling [19]. In some genes, the promoter regions are underwound (local Lk too high) so that RNA polymerase and initiating factors are misaligned. In other regions the promoters are overwound (local Lk too low) leading to a similar misalignment. Alteration of local negative supercoiling writhe by the binding of nucleoid-associated proteins like HU, HNS, or FIS can twist the promoter into the proper configuration, aligning the binding faces of the helix and initiating transcription [20] Additional local supercoiling can also contribute to transcription initiation in promoters where the elongation step is energetically disfavored because of local high energy GC sequence content (called discriminator sequences.) [20] Intriguingly, since supercoils can propagate until they reach a topological barrier, formation of a new supercoiling restraint (by a binding event say) can send local supercoiling to a neighboring site, allowing the DNA to act as a sort of telegraph [17]. Amazingly, it seems that the global level of supercoiling (the global linking number which is modulated by a set of topoisomreases capable of removing negative supercoils and a DNA gyrase capable of adding them) is also a global control on transcriptional state. The amount of negative supercoiling appears to be used to precisely titrate the amount of ribosomes produced to meet the requirements of given nutrient conditions [20]. Topological domains have one further important consequence in solving a problem of replication: the isolation of positive supercoils generated by the unwinding of replicating DNA in front of DNA polymerase from the rest of the structure [16]

Since the dynamic nature of topological domains is consistent with the disordered loop structures seen in lysed cells, until very recently the entire chromosome structure was believed to consist of a disordered, dynamic collection of supercoiled domains [21]. However, some intriguing observations suggested otherwise. Chromosomal inversions between certain regions of the chromosome were found to be disallowed [22] Synthetic constructions of these inversions that yielded viable cells suggested that the disallowed inversions were structurally disallowed. Thus these experiments indicated that there was additional structure to the chromosome or that certain regions of topological domains were not totally fluid.

Improvements in the resolution of fluorescence microscopy, however, yielded the most radical reevaluation of bacterial chromosome structure, in a startling set of observations. First, fluorescently labeled origins and termini were observed to occupy reproducible positions along the cellular axis in *E. coli, C. crescentus and B. subtilis* [23]. Moreover, these regions exhibited reproducible spatiotemporal dynamics during replication and cell division. Subsequently, in *E. coli*, Niki *et al.* measured the positions of a set of chromosomal loci between the origin and terminus using fluorescence in situ hybridization (FISH) and found that their positions along the longitudinal axis of the cell corresponded linearly with their distance from the origin along the genome sequence [24]. Finally, in *C. crescentus*, Viollier *et al.* measured a set of 141 different chromosomal loci using both FISH in formaldehyde-fixed cells and GFP-lac fusions *in vivo* and found that this linear relationship held for every locus tested [25]. Furthermore, they followed several loci simultaneously with the origin of replication throughout the cell cycle and observed specific replication dynamics and a rapid movement after replication back to cellular locations occupied before replication initiation. These experiments indicated that the bacterial chromosome is in fact highly organized into a structure that is symmetric about the origin of replication and which compacts the DNA such that it preserves genetic distance from the origin. Additionally, they showed that this structure has tightly controlled dynamics.

Evolutionary Optimality of Chromosome Structure

While fluorescence microscopy indicates in broad outlines a very large-scale structural order to the chromosome, and topological domains illustrate ways in which supercoiling properties affect the local promoter level and suggest a highly dynamic local level structure to the genome, the large-scale folding bridging these two structural regimes is unknown. In other words, how are topological domains of genomic DNA packaged into the symmetric linear arrangement of chromosomal arms observed by microscopy? This is the structural level at which long-range spatial interactions between genes would occur and which would determine the nature of intermediate compaction which may be of profound importance for replication. It is this scale that I focus on in this thesis.

Because of coupled transcription-translation, this folding has the potential to organize groups of functionally related genes such that their protein products are translated and assembled into complexes in the region where they are transcribed, forming a scaffold for the assembly of the many large protein machines which have begun to enter our conception of the cell.. It may likewise organize highly active genes around transcription factories similar to those in the eukaryotic nucleus [26].

During replication this folding must allow segregation without extreme entanglement. Indeed, this is perhaps the reason for the linear correlation of genomic distance from the origin and longitudinal position in the cell viewed by_microscopy; such compaction would help prevent regions yet to be replicated from becoming entangled with newly replicated regions. The paired fork model of replication by which the set of DNA polymerases replicating the chromosome bidirectionally from the origin of replication remain fixed together in space and pull the DNA through, would explain the symmetry observed between chromosome halves about the origin of replication as well [27]; these points must be close together when moving through the polymerase machine. Additionally, since this model extrudes daughter DNA naturally to opposite sides of the cell, it offers an elegant solution to the segregation problem even with some small level of entanglement. By generating a biased movement of DNA in opposite directions, it allows the resolution of entangled DNA strands by topoisomerases (itself unbiased) to ultimately separate the daughter chromosomes [16]

Although the intermediate-scale folding has been inaccessible to direct microscopic measurement, several intriguing observations have been made of long-range positional correlations in transcription both in absolute expression level and in the expression correlation of gene pairs in E. coli. These correlations extend far beyond the ~10kb level of topological domains to 100kb and even to 600 kb [28-30]. Moreover they seem to change as a function of environmental state, thus indicating potentially that there are long-range spatial contacts between various chromosome regions that are modulated by transcriptional state. Positioning of certain transcription factors and their binding sites also have been reported to be periodic.[31] However, it is not yet clear to what extent these correlations and periodicities are confined to particular regions of the chromosome – the 600kb correlation extends the entire chromosome length but some studies have found the smaller periodicities confined to particular regions [28] – and it is not clear how they relate to a global folding of the chromosome.

In this thesis I approach the problem of chromosome folding from the perspective of function. The fundamental ideas are three-fold. First, the three-dimensional spatial organization of the chromosome is likely to have been optimized by the process of evolution. Evolutionary selection on genomes should be working not only at the level of gene content and controlling elements, but also at the level of structure. Thus insight into the structure can be gained by an understanding of the "optimization function" that evolution is optimizing in the same way that understanding the "objective function" for metabolic fluxes or for gene expression levels has yielded accurate predictions of these quantities [32, 33]. Indeed, by understanding even a subset of the optimization criteria or a few constraints we can restrict the space of possible conformations immensely. And a constraint-based framework allows for the testing both of the implications of a set of constraints and for their compatibility. Chapter 2 describes work on constraints and optimization in detail, detailing the theoretical and experimental observations underlying constraints and optimization criteria for chromosome structure. It outlines a general constraint-based method for describing feasible conformations given hard distance bounds which is based on the method of distance geometry and also portions of the likely global optimization functions operating on the chromosome. Furthermore, it descibes a set of montecarlo methods, one parametric, the other non-parametric, for finding optimal configurations within the constrained configuration space.

The second fundamental idea of this thesis is that, as a result of evolutionary optimization, we expect to see signals of selection for three dimensional-spatial relationships recorded in the hundreds of bacterial genomes that have now been sequenced. Chapter 3 describes this work. Within it, I describe the first evidence of evolutionary selection for spatial relationships in bacterial genomes. The signals we find are based on a simple but novel method to identify points that are likely evolutionarily selected for spatial vicinity and a method to extract structural information from them. I describe an analysis of these signals in *E. coli* where they show unambiguous periodicities that span the entire length of the chromosome and a strong position preference for a single phase of the period. The signals strongly suggest a periodically looped possibly helical organization of the E. coli chromosome with a single chromosome-long longitudinal axis along which most of the pairing occurs. The pairing is also strongly correlated with transcriptional level indicating likely functional relevance. The end of this chapter describes a first attempt to elucidate structural features of the *E. coli* chromosome fold based on fitting of the pair data derived from comparative genomics to explicit models. Even fitting at this simple level of complexity reveals interesting functional consequences of structure like the preference of highly expressed genes for a single helical face.

The appendices on complete determination of biological systems and selection of oligonucleotide probes relate to constraint-based determination and potential experimental measurements that could generate a large set of constraints for chromosome folding. Extensions of the oligonucleotide probe work in particular are discussed in the last chapter on conclusions and future directions.

It is the final contention of this thesis that solving the structure of a chromosome fold, whether bacterial or human, should be approached quantitatively using the same mathematical framework used for solving the structure of a protein – instead of minimizing the error between observed and calculated electron densities, we minimize more generally the error between a broad set of constraints gathered from diverse experimental methodologies and a model of the fold at some given level of resolution, whether the 10kb, 1kb, 100bp or atomic scale With enough constraints, it will be possible to solve both the structure and its dynamics as a function of cell state. This 4D chromosome trajectory will undoubtedly contain a wealth of information about cell function and the spatial organization of the nucleus, nucleoid, and the cell.

References

1. Zhu, Z., J. Shendure, and G.M. Church, *Discovering functional transcription-factor combinations in the human cell cycle.* Genome Res, 2005. **15**(6): p. 848-55.
2. Scholtens, D., M. Vidal, and R. Gentleman, *Local modeling of global interactome networks.* Bioinformatics, 2005. **21**(17): p. 3548-57.
3. Alberts, B., *The cell as a collection of protein machines: preparing the next generation of molecular biologists.* Cell, 1998. **92**(3): p. 291-4.
4. Ovadi, J. and P.A. Srere, *Macromolecular compartmentation and channeling.* Int Rev Cytol, 2000. **192**: p. 255-80.
5. Sali, A., et al., *From words to literature in structural proteomics.* Nature, 2003. **422**(6928): p. 216-25.
6. Aloy, P., et al., *Structure-based assembly of protein complexes in yeast.* Science, 2004. **303**(5666): p. 2026-9.
7. Jensen, R.B. and L. Shapiro, *Cell-cycle-regulated expression and subcellular localization of the Caulobacter crescentus SMC chromosome structural protein.* J Bacteriol, 2003. **185**(10): p. 3068-75.
8. Gitai, Z., *The new bacterial cell biology: moving parts and subcellular architecture.* Cell, 2005. **120**(5): p. 577-86.
9. Gitai, Z., M. Thanbichler, and L. Shapiro, *The choreographed dynamics of bacterial chromosomes.* Trends Microbiol, 2005. **13**(5): p. 221-8.
10. Taddei, A., et al., *The function of nuclear architecture: a genetic approach.* Annu Rev Genet, 2004. **38**: p. 305-45.
11. Misteli, T., *Concepts in nuclear architecture.* Bioessays, 2005. **27**(5): p. 477-87.
12. Spilianakis, C.G., et al., *Interchromosomal associations between alternatively expressed loci.* Nature, 2005. **435**(7042): p. 637-45.
13. Phair, R.D., et al., *Global nature of dynamic protein-chromatin interactions in vivo: three-dimensional genome scanning and dynamic interaction networks of chromatin proteins.* Mol Cell Biol, 2004. **24**(14): p. 6393-402.
14. Dekker, J., et al., *Capturing chromosome conformation.* Science, 2002. **295**(5558): p. 1306-11.
15. Travers, A. and G. Muskhelishvili, *Bacterial chromatin.* Curr Opin Genet Dev, 2005.
16. Hardy, C.D., et al., *Disentangling DNA during replication: a tale of two strands.* Philos Trans R Soc Lond B Biol Sci, 2004. **359**(1441): p. 39-47.
17. Hatfield, G.W. and C.J. Benham, *DNA topology-mediated control of global gene expression in Escherichia coli.* Annu Rev Genet, 2002. **36**: p. 175-203.
18. Postow, L., et al., *Topological domain structure of the Escherichia coli chromosome.* Genes Dev, 2004. **18**(14): p. 1766-79.
19. Peter, B.J., et al., *Genomic transcriptional response to loss of chromosomal supercoiling in Escherichia coli.* Genome Biol, 2004. **5**(11): p. R87.

20. Travers, A. and G. Muskhelishvili, *DNA supercoiling - a global transcriptional regulator for enterobacterial growth?* Nat Rev Microbiol, 2005. **3**(2): p. 157-69.
21. Thanbichler, M., P.H. Viollier, and L. Shapiro, *The structure and function of the bacterial chromosome.* Curr Opin Genet Dev, 2005. **15**(2): p. 153-62.
22. Segall, A., M.J. Mahan, and J.R. Roth, *Rearrangement of the bacterial chromosome: forbidden inversions.* Science, 1988. **241**(4871): p. 1314-8.
23. Webb, C.D., et al., *Bipolar localization of the replication origin regions of chromosomes in vegetative and sporulating cells of B. subtilis.* Cell, 1997. **88**(5): p. 667-74.
24. Niki, H., Y. Yamaichi, and S. Hiraga, *Dynamic organization of chromosomal DNA in Escherichia coli.* Genes Dev, 2000. **14**(2): p. 212-23.
25. Viollier, P.H., et al., *Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial DNA replication.* Proc Natl Acad Sci U S A, 2004. **101**(25): p. 9257-62.
26. Cook, P.R., *Predicting three-dimensional genome structure from transcriptional activity.* Nat Genet, 2002. **32**(3): p. 347-52.
27. Dingman, C.W., *Bidirectional chromosome replication: some topological considerations.* J Theor Biol, 1974. **43**(1): p. 187-95.
28. Jeong, K.S., J. Ahn, and A.B. Khodursky, *Spatial patterns of transcriptional activity in the chromosome of Escherichia coli.* Genome Biol, 2004. **5**(11): p. R86.
29. Allen, T.E., et al., *Genome-scale analysis of the uses of the Escherichia coli genome: model-driven analysis of heterogeneous data sets.* J Bacteriol, 2003. **185**(21): p. 6392-9.
30. Carpentier, A.S., et al., *Decoding the nucleoid organisation of Bacillus subtilis and Escherichia coli through gene expression data.* BMC Genomics, 2005. **6**(1): p. 84.
31. Kepes, F., *Periodic transcriptional organization of the E.coli genome.* J Mol Biol, 2004. **340**(5): p. 957-64.
32. Edwards, J.S. and B.O. Palsson, *Metabolic flux balance analysis and the in silico analysis of Escherichia coli K-12 gene deletions.* BMC Bioinformatics, 2000. **1**(1): p. 1.
33. Dekel, E. and U. Alon, *Optimality and evolutionary tuning of the expression level of a protein.* Nature, 2005. **436**(7050): p. 588-92.

Chapter 2

Constraint-based determination of chromosome structure

with **Daniel Segrè**

Abstract

We outline a method to integrate data on the three-dimensional structure of whole chromosomes and genomes. The method is based on evolutionary optimality. We hypothesize that the three dimensional structure of the chromosome is optimal for many of the functions that it performs. These functions determine geometrical and dynamical constraints that can be expressed mathematically. Our goal is to gather enough constraints to create a 4D model of the chromosome. Changes in the structure of the chromosome over time (the 4th dimension) will reflect the varying functional constraints on the genome during different phases of the cell cycle.

How to fold a chromosome

Massive effort has been devoted to studying cellular structure at the nanometer and micrometer scales. Very little success, however, has been achieved in studying structure at intermediate scales. These mesoscopic scales are important in many biological processes. For instance, aggregation of molecular components in specific cellular regions is involved in processes as diverse as mitosis and bacterial chemotaxis, spatial gradients are crucial for the development of many multicellular organisms, and the geometry of cells is exquisitely tailored to their function.

The structure of the chromosome is of particular importance. It forms the spatial framework for transcription and determines the genes that are accessible to the transcription machinery (6). Moreover, its replication and spatial segregation into daughter cells are necessary for successful reproduction. The extended DNA of most organisms is hundreds of times longer than the diameter of the nuclear or cellular membrane and must be folded extensively to fit inside the cell. This fold can be described as a configuration in space with regions of varying flexibility. Some regions may be quite rigid while others may diffuse freely throughout most of the cell. Still other regions may sample the space unevenly, spending most of their time in particular areas. Our goal is to understand and characterize these folds.

A viable fold must allow the DNA to fulfill its functions for the cell: transcription of the necessary genes and replication. We hypothesize that evolutionary pressure has selected those configurations that are optimal for these processes. The functions determine geometrical and dynamical constraints that can be expressed mathematically. Varying constraints on the genome during different phases of the cell cycle and in different environmental conditions change the structure of the chromosome over time (the 4th dimension.) We have developed methods to generate structures based on these constraints.

4D information in the genome?

There are two major factors we expect to determine the folds (structures) of the chromosome: constraints from the physical chemical properties of DNA and optimization from the process of evolution. The energy of a particular structure is determined by chemical properties such as bond lengths, bond angles, and electrostatic interactions. Thus, physics and chemistry define the space of feasible structures of the chromosome. Evolution chooses particular feasible structures from this space, by altering local sequence properties, generating sites for geometry altering binding proteins, or, in part, by changing the linear position of genes along the chromosome which, through their interactions with the transcription machinery and various proteins may mechanically and dynamically shape the chromosome fold .

Since evolution may influences structure by altering the linear sequence of the genome, we expect genome sequences and annotation to contain information about the structure of the chromosome. The situation is analogous to protein folding where the primary amino acid sequence gives information about the three dimensional structure of the protein (7). In chromosome folding, gene function and other experimental data may give information about the three dimensional structure of the chromosome.

Many research efforts have been aimed at predicting the structure of proteins from sequence data. These theories use the physical chemical properties of amino acid sequences to find the most energetically stable structure. (7). With chromosomes, the physical properties of the nucleic acid sequence leave a large amount of flexibility in what structures are formed. Indeed, at mesoscopic scales, the atomic forces are observed mainly in the elastic properties of DNA: its persistence length (the distance within which points on the strand feel the presence of each other and which determines its bendability), and the energy of supercoiling (8). The information from functional genome annotation, which reflects evolutionary selection, however, is potentially far more constraining in that certain folds will be more or less efficient for the cell. Our methods use such functional information to predict aspects of chromosome structure.

Theoretical and Experimental Distance Constraints

The methods we have developed are constraint based. We establish a set of criteria (constraints) and then search for structures that satisfy these criteria. Constraint-based models have been immensely successful in modeling metabolic fluxes and are generally useful in several ways: (i) the degrees of freedom of the constrained system (or the dimensionality of the "feasible" space) provide information about the degree of understanding of the system, (ii) tests of the consistency of multiple constraints offer the possibility of critically revising data or our interpretations of data, (iii) computer simulations or optimization algorithms can be applied within constrained feasible spaces to search for solutions that correspond to specific configurations; optimal configurations are particularly interesting as they may capture important properties of evolutionary adaptation.

The constraints on chromosomes can be motivated by experiment or by theory. Experimental data sources include confocal microscopy, cross-linking, and other visualization techniques. Confocal microscopy can be used to follow particular chromosomal loci in time (9) and measure the distance between these loci. Cross-linking experiments can provide genome scale information about the distances between many chromosomal loci simultaneously. Recently, a cross-linking technique determined the large scale morphology of the entire yeast chromosome III (10). Other visualization techniques, particularly electron tomography promise to give detailed views of the entire chromosome.

The two main categories of theoretical constraints in our current model are transcriptional constraints, and replication constraints. Transcriptional constraints focus on the configurations of the chromosome that are energetically efficient for the cell in terms of the spatial locations of transcribed genes. For complexes to form , for example, the components must diffuse into the same region of space and encounter each other in the proper configuration (11). This probability is increased when the loci of translation (for proteins) or transcription (for RNAs) are closer together. In bacteria, this can be achieved by having the physical locations on the genome where the components are transcribed close together in space (figure 2.1, black squares.) Another possible

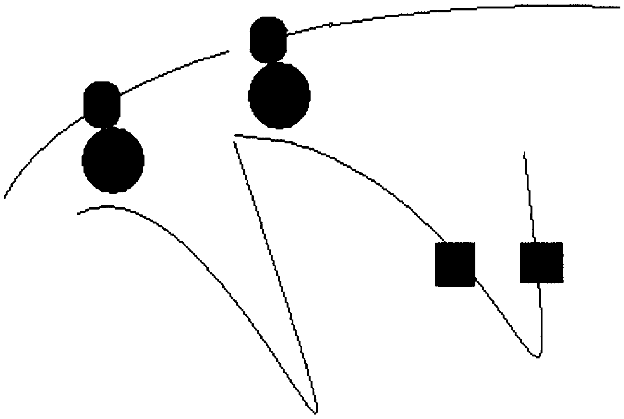


Figure 2.1 Two types of distance constraints motivated by transcription in bacteria: membrane genes (blue) close to the membrane and genes that code for components of a protein complex (black)

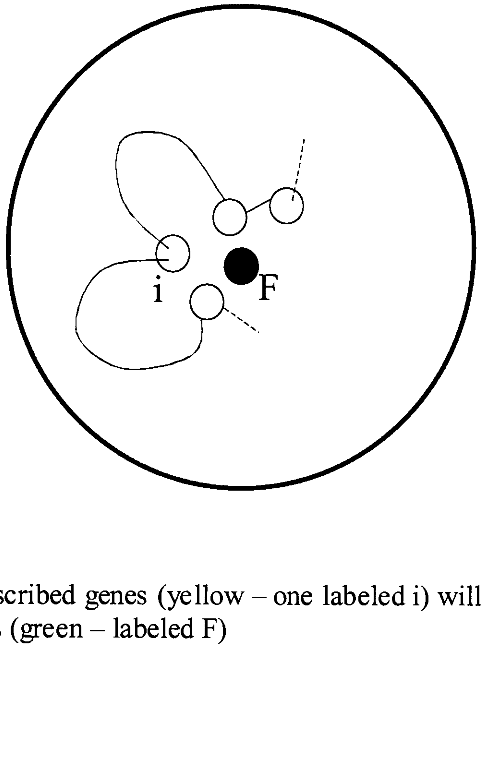


Figure 2.2 Highly transcribed genes (yellow – one labeled i) will be located close to transcriptional factories (green – labeled F)

transcriptional constraint in bacteria is that membrane protein coding genes be close to the membrane where they can be transcribed, translated, and then directly inserted into the membrane in a process called transertion which may itself play a large role in structuring the chromosome (figure 2.1 red circles.) (12). A third transcriptional constraint is that genes with high levels of transcription be located close to regions of high transcription activity ("transcription factories") thought to be located at specific positions in the cell (figure 2.2.) For fixed transcription factories, genes that are highly transcribed will, of necessity, spend much of their time attached to the factories. It is believed that the transcriptional factories are one of the most important organizing factors of chromosome structure (6). There are many other possible transcriptionally motivated constraints such as those that locate transcription factors close to their binding sites and those that optimize the relative spatial positions of enzymes involved in metabolism (13).

It is useful to note two possible ways in which transcriptional constraints could act upon chromosome structure. One is through the evolutionary positioning of genes linearly along the genome and the selection of specific sequences that may control the geometry of small regions of the chromosome (for instance AT tracts cause bends or binding sites for specific chromosome conformation altering proteins can cause regions to take on specific geometries.) Another is the product of the "mechanical forces" between proteins or RNAs that are still attached to their genomic loci. In bacteria, when proteins that are attached through ribosomes and mRNAs to their sites of transcription interact, they may cause these sites to spend more time close together thus enforcing particular structures. We can imagine this happening with ribosomal RNAs in the eukaryotic nucleolus, for instance, where all of the ribosomal RNAs are colocalized and transcribed.

Modes of Bacterial Chromosome Replication

During reproduction, the entire genome must be copied and segregated into the daughter cells. In bacteria, this process is often depicted with simple diagrams using circles to represent the replicating chromosome. However, these diagrams ignore the compactness of the genome inside the cell. The situation is instead one where a circle, folded hundreds of times, is duplicated and segregated. The fact that the genome must be so compact during this process places many constraints on the structure and these constraints can be expressed mathematically given a replication model.

Topological and spatial factors affecting genome replication have been discussed many times in the literature, and theoretical and experimental evidence for a bidirectional paired fork process has accumulated. (8, 14, 15). We have used the paired fork process as a basis for formulating replication constraints for our chromosome model. The paired fork is a set of four linked DNA polymerases, which is probably fixed in space. Replication proceeds in both directions (with two polymerases in one direction and the other two in the opposite directions.) Since the polymerases are linked and fixed, it is the chromosome, not the polymerases that move. The chromosome is pulled through the polymerase complex and daughter DNA emerges on either side. The effect is that while the DNA is replicated, it is separated into different sides of the cell (figure 2.3.)

There are several ways in which replication constraints can be incorporated into models. If the diffusion of the chromosome through space is faster than the rate of replication, we can expect folding of the newly generated chromosome as it grows from the replication forks. However, the constraints on the fold at any given moment will only be those that are derived from the portion of the new chromosome that has been already replicated or its interaction with the other daughter chromosome or the remaining unreplicated chromosome. As the replicated chromosome grows, new constraints will be active, but these new constraints will be acting on a structure that is already partially folded. Thus the space of possible folds that meet the constraints at later times is dependent on the space of possible folds at earlier times. This results in final folds (after replication) that reflect the history of the growing replicated chromosome (figure 2.4.) The effects are analogous to known dynamic effects in protein folding where the final conformation of the protein is affected by folding of the partially translated protein intermediates before translation is completed.

Replication yields constraints related to the ability to segregate and disentangle structures as they go through the replication process. Daughter structures that are more entangled, are more difficult to separate, requiring greater topoisomerase activity. It is possible to quantify entanglement by examining overlap between replicating structures using various methods, for example support vector machines which can determine the optimal surface separating two structures Furthermore, since loci symmetrically opposite the origin of replication are pulled through the polymerase complex simultaneously, it is possible to impose constraints that reflect this symmetry, for instance by enforcing a

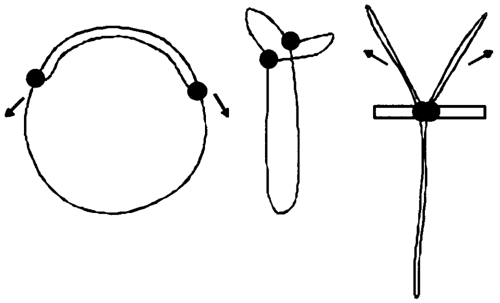


Figure 2.3 A diagram of possible chromosome replication mechanism. Left: the classical view; Right: the paired fork model; Center: an intermediate state, shown to emphasize the transformation from one model to the other.

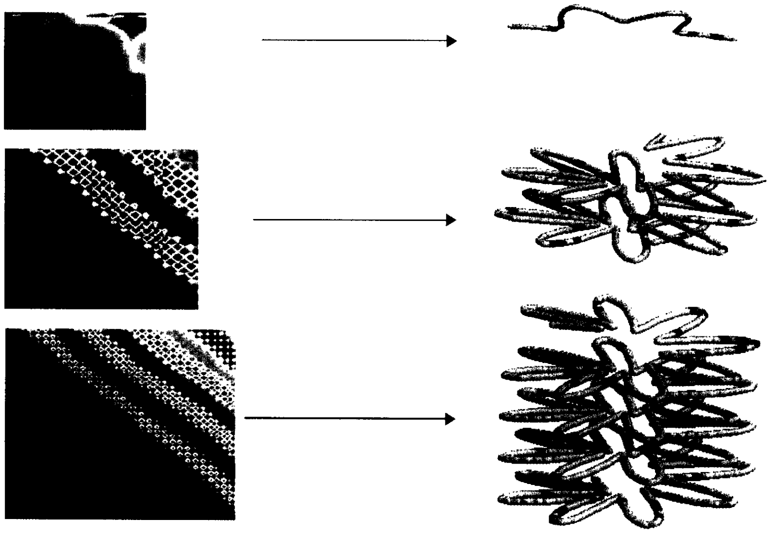


Figure 2.4 At left, representations of the matrices of distances between points on a growing chromosome and at right, the structures.

symmetric flattened circle shape (flattened about the origin and terminus) for the entire chromosome.

Distance Geometry at the Genome Level

The static constraints on chromosome structure can be represented as upper and lower bounds on the distances between chromosomal loci. Proximity of a membrane protein-coding gene to the membrane (modeled as a sphere), for instance, can be enforced by setting a lower bound on the distance between the center of the cell and the membrane protein-coding gene. Structures can then be found which satisfy these constraints.

The constraints do not completely determine the structure of the chromosome, rather, they limit the space of possible conformations. This conformation space effectively defines both the number of distinct structure classes consistent with the constraints and the flexibility of the chromosomal loci in each structure class. It is therefore important to have a means, not only of generating structures consistent with the constraints, but also of performing an unbiased sampling of the conformation space to determine the ensemble of structure classes. Fortunately, nuclear magnetic resonance spectroscopists have faced similar problems determining the structure of proteins from NOE (Nuclear Overhauser Effect) data (16, 17). We have adapted one of their methods based on the mathematics of distance geometry for our models.

Method of Distance Geometry

Mathematically, the upper and lower bounds on distances between loci define the conformation space. However, this space is defined in terms of distances and not in terms of the coordinates necessary to visualize a structure and examine many of its properties. Distance geometry takes the distance bounds and finds coordinates that are consistent with them in three-dimensional space. For an under-constrained structure, the problem of choosing a distance matrix that satisfies the bounds is NP hard (non –polynomial time) but heuristic algorithms exist that are quite efficient at finding solutions.

Our algorithm is an adaptation of the EMBED algorithm of Timothy Havel (16, 17). It consists of four steps. 1. *Bound Smoothing,* 2. *Metrization,* 3. *Projection*, and 4. *Optimization*. The *bound smoothing* step addresses the issue that although the initial distance bounds constrain only a small subset of pairs of loci, Euclidean distances are not independent of each other and thus the bounds contain much more information than the specific distances that they constrain. Bound smoothing makes some of these dependences into explicit bounds. The best known example of a distance dependence which is a property of the geometry of a space is the triangle inequality which restricts the feasible distance between a triplet of points i, j, and k. ($d_{ij} \leq d_{ik} + d_{jk}$)

There are additional higher order distance relations between the distances of 4, 5, 6, …, n points from each other in three-dimensional space. These higher order relations are mathematically complex and computationally time consuming and therefore infeasible to apply iteratively to the problem of distance geometry as existing algorithms require. Our algorithm performs bound smoothing with only the triangular inequality, a good approximation that is standard practice in the field. This bound smoothing serves to significantly increase the number of explicit constraints on the structure.

The second step in the algorithm is to choose a set of distances that are consistent with the bounds. This step, called *metrization*, results in the selection of a random distance matrix that is representative of a single structure. One at a time, distances are chosen randomly between the bounds and after each choice, the algorithm performs an additional step of bound smoothing to incorporate the changes in the constraints that depend on the new distance (figure 2.6)

Metrization results in a distance matrix that represents a single structure. The next step, *projection*, is used to generate a set of coordinates that are consistent with the distance matrix. Since our bound smoothing algorithm only smoothes based on the triangular inequality, the distance matrix generated after *metrization* is guaranteed to be consistent with distances between the n points in n –1 dimensional space not necessarily in three dimensions. The projection step finds coordinates in three dimensions with distances that are closest to the original metrization distances (figure 2.7 and box 2.1).

This projection into three-dimensional space often results in minor violation of the initial constraints. The final step in the algorithm is optimization of the coordinates so that they minimally violate the initial distance bounds. We perform the optimization using a simple error function and the method of steepest descent.

The entire distance geometry process results in an unbiased sampling of the conformation space defined by the initial distance constraints. Additionally, the bound smoothing step produces several useful results. First, it allows the consistency of the constraints to be determined in n-1 dimensions. This is useful in determining if any

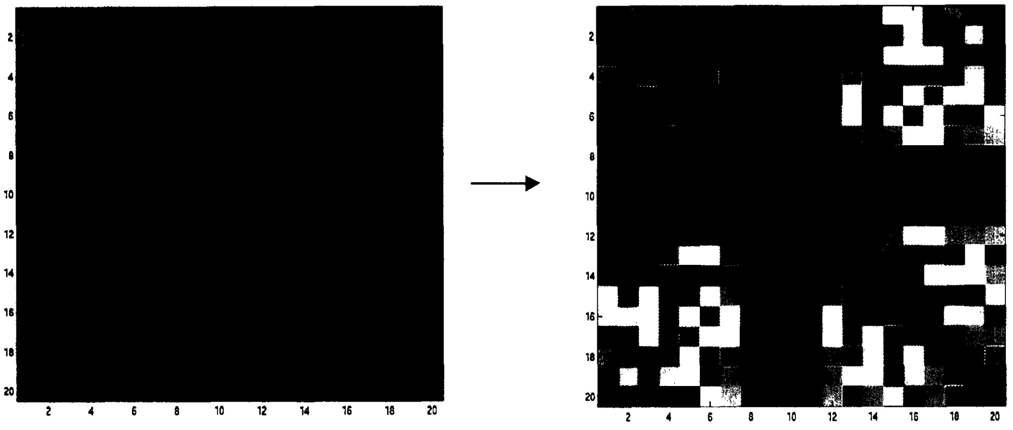


Figure 2.5 The process of distance geometry generates a matrix of distances (pictured on the right) that are consistent with a set of distance bounds (pictured on the left) Blue in the left matrix indicates a distance that is unconstrained (except for along the diagonal where it indicates zero distance from a point to itself.) The other colors represent upper bounds on the distances, red indicating the largest.

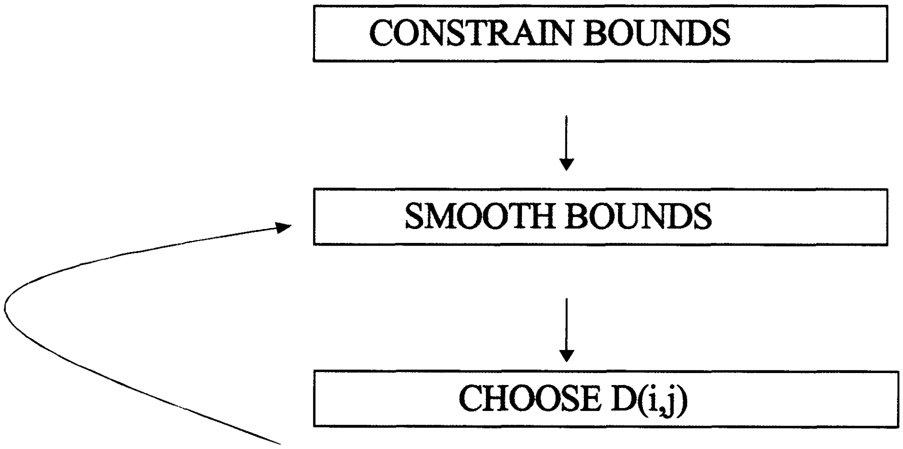


Figure 2.6 Diagram of the process of metrization using distance geometry

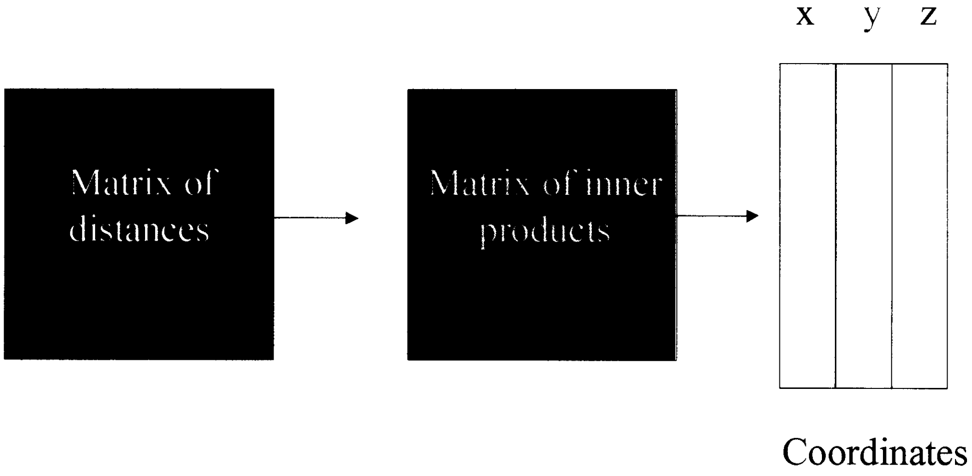


Figure 2.7 The projection process takes a matrix of distances and generates a set of coordinates in three-dimensional space.

Given a matrix of all pairwise distances between the n points, D, one can generate a matrix of inner products M, by introducing an extra point, the origin, which is usually taken to be the centroid of the distance data. This matrix can be factored using SVD or principal component analysis to give the coordinates of the original structure (up to an arbitrary translation and rotation.) This can be expressed mathematically

$$M=(I-U)\,D\,D\,(I-U)$$

Here, U is an n x n matrix of ones, and I is the n dimensional identity matrix.

Factoring M into

$$U\,S\,V' = U\,S^{1/2}\,S^{1/2}\,V' = L\,L^t = X\,X^t$$

where X is an n x 3 matrix of the coordinates. The last equality follows from the fact that $L^* L^t$ is a lower triangular matrix times its transpose, where L^t will have all zeros except for the top 3 rows (if we are talking about a three dimensional structure.) Since there is only one factorization of a matrix into lower and upper triangular matrices and we can choose X coordinates such that X^t is an upper triangular matrix. Then X and L must be identical

Box 2.1: Decomposition of a Matrix of Distances into a Set of Coordinates subsets of the constraints are impossible to simultaneously satisfy. Second, bound smoothing generates a set of lower bounds and upper bounds that take into account all triangular inequalities. These lower and upper bound matrices are themselves independent distance matrices which represent the most expanded and most contracted structures in n-1 dimensions that are consistent with the constraints.

Results from Mycoplasma

We used the distance geometry method to model the chromosome structure of the bacterium *Mycoplasma pneumoniae* (18). It is nearly a minimal cell with a genome that is 816 kbp long and only 688 genes. It has limited metabolism, no known regulation, and very few DNA binding proteins (figure 2.8). Since it is so simple, we expect chromosome structure to be under strong evolutionary selection. We constrained the 110 membrane protein coding genes to be close to a spherical membrane with the diameter of an average *Mycoplasma* cell. We also constrained the 52 annotated ribosomal protein-coding genes to be close to each other. Furthermore, we constrained the origin and terminus of replication to be at opposite poles of the cell, an observation that has been made in many microscopy experiments. Finally, to reflect the geometry of the paired fork model, we modeled the genome as a flattened circle (with origin and terminus of replication as poles.)

This resulted in a model with 165 loci (13,530 distances) and 1547 constraints. With this input, the implementation of our algorithm in MATLAB takes about twenty minutes to generate a structure. It is therefore possible to generate hundreds of structures in a few days time. Below we present some statistical data from analysis of 160 structures generated using the Mycoplasma constraints. Figure 2.9 shows a particular structure. Note that the membrane protein coding genes (in blue) are all close to the membrane and the ribosomal protein coding genes (in red) cluster together. No connecting DNA segments are shown because the chromosomal distance between subsequently modeled loci is sufficient to stretch across the cell. Figure 2.10 shows the distribution of distances between two membrane proteins, a membrane protein and a ribosomal protein, two ribosomal proteins, and between a ribosomal protein and the center of the cell. These distributions show that the constraints have been satisfied and show evidence of clusters of distinct structure classes. For instance, the membrane – membrane protein distribution appears to be bimodal. Further statistical analysis is necessary before deciding if the peaks are significant.

Optimization Algorithms

Distance geometry is extraordinarily useful as a rigorous way of defining the conformational space accessible to a structure given a set of known constraints and also in allowing the implications of these constraints to be derived. However, within the framework of distance geometry, the constraints implemented are hard (they must be satisfied) and, in the case of chromosomes, there are many aspects of a fold that may reflect optimal, desirable properties but not necessarily required properties. For example the spatial colocalization of genomic loci encoding components of large protein complexes may allow for maximally efficient assembly of complexes but such complexes

816 Kbp
90% Coding
688 Genes
110 Membrane Proteins
52 Ribosomal Proteins
No Active Transport
No Regulation
Limited Metabolism
Few DNA Binding Proteins
.5 µm diameter (1470 bp)

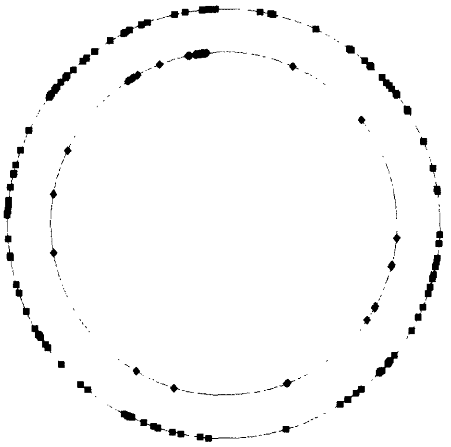


Figure 2.8 Fundamental data about *M. pneumon iae* (left), and a diagram (right) with the positions of membrane and ribosomal genes along the chromosome

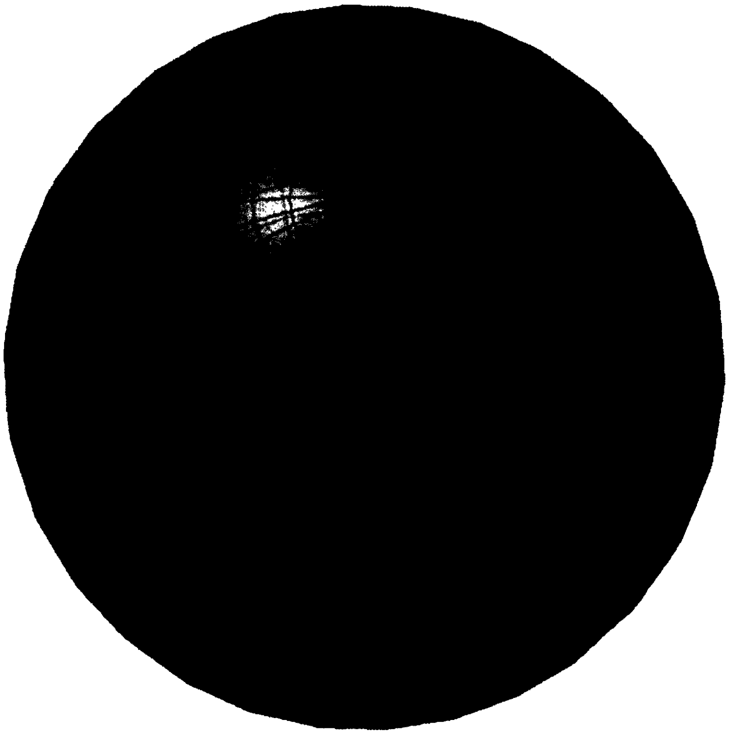


Figure 2.9 A visualization of a particular structure. Green sphere: membrane. Blue dots: membrane protein coding gen es. Red dots: ribosomal genes

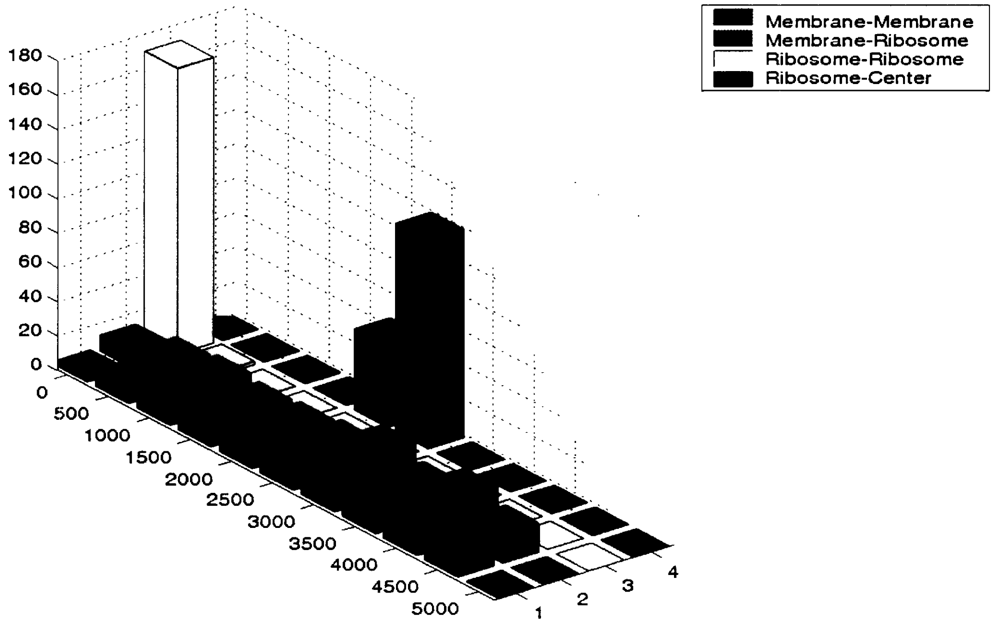


Figure 2.10 Results of the statistical analysis of 160 runs of Distance Geometry and structure refinement programs for the chromosome of *M. pneumon iae*.
Each histogram displays the distribution of distances (in Angstrom) between two chosen points.

might still be assembled less efficiently without spatial clustering of the coding loci. These sorts of properties may be optimized subsequent to a hard constraint method like distance geometry, by constructing a cost function reflecting the degree to which a particular structure satisfies the desired properties and using an optimization process to find structures which minimize the cost. For example we can write the degree to which a conformation spatially colocalizes the ribosome components and places membrane protein coding genes close the membrane in the cost function below.

$$C = w_1 \cdot \sum_{i} \mathrm{d}(M_i, \mu) + w_2 \cdot \sum_{i,j} \mathrm{d}(R_i, R_j)$$

Equation 2

(where $d(M_j,\mu)$ is the distance between transmembrane gene i and the membrane, and $d(R_j, R_j)$ is the distance between ribosomal genes i and j.)

Therefore, we can use distance geometry to produce an ensemble of structures satisfying an initial set of hard constraints and then optimization based on a cost function with high weights maintaining the hard constraints of the distance metrization to find optimal structures within this reduced ensemble. We developed two algorithms to find configurations that represent optima.

Helichrome

Our first algorithm uses a set of six parameters to describe the chromosome structure. Here, the constraints are not derived from an initial distance geometry metrization but rather from the set of parameters that describe the space of feasible structures. The parameters define a supercoiled helix which can assume a diversity of structures from simple helices to complex structures such as the structure pictured in figure 2.11 top right panel. Supercoiled helices are consistent with known chromatin structure in higher organisms and have several other desirable characteristics: they automatically enforce the polarity of the origin and terminus of replication, and add a degree of order to the structure which should solve some problems of entanglement associated with replication. Box 2 (top) lists the six parameters. Large and small refer to the primary helix and supercoil, and the large helix frequency is an extra oscillation that changes the overall radius of the large helix as a function of the arc length. Mathematically these structures are described by the equations for a local helix in the frenet frame defined by the large helix. The equations are listed in vector form in the bottom of box 2.

Using *helichrome* we perform a random walk in the parameter space, beginning with randomly generated initial structures and then randomly change parameters to generate test structures for each iteration. The cost of the test structure is compared with the previous structure. If lower, the new structure is accepted. If higher, it is accepted with probability $e^{-\beta\Delta C}$ where ΔC is the difference in cost between the new structure and the old and β is the 1/kT, a Botzmann factor that varies inversely with the temperature. The temperature is lowered according to an "annealing schedule" beginning with high temperatures at early iterations so that structures of higher energy are frequently accepted and decreasing so that in later iterations the structures "anneal" to a final state.

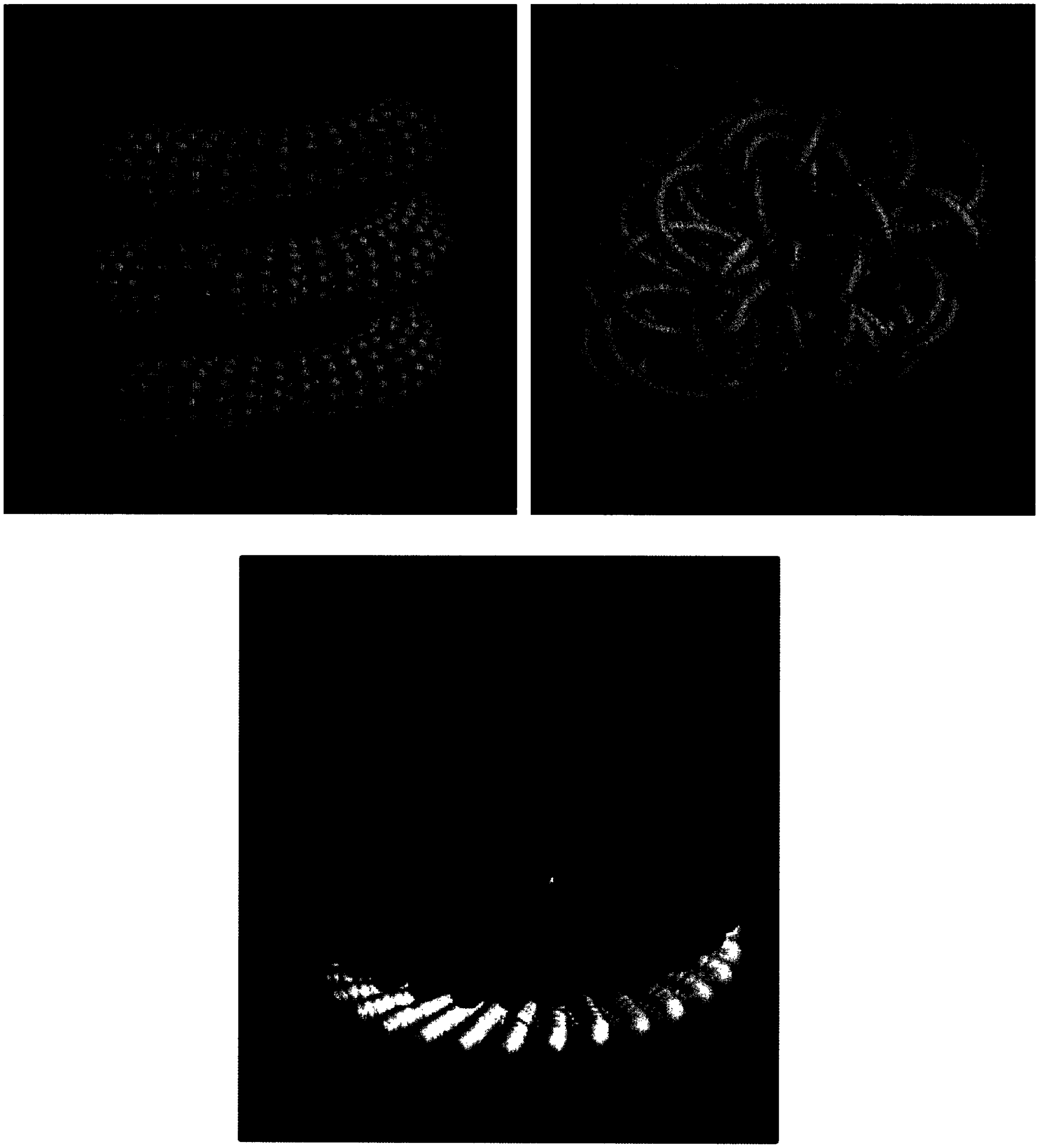


Figure 2.11 Results of Helical Optimization with Helichrome. Top left: Simple supercoil and right complex supercoils. Red indicates ribosomal coding gene, blue membrane protein coding gene. Bottom: Simple helix with added cosine modulation colored in order of distance from replication origin.

Parameter	*Symbol*
Large helix rise	a
Large helix radius	R
Large helix frequency	d
Small helix frequency	w
Small helix sine radius	As
Small helix cosine radius	Ac

$$\vec{x}_{frame} = \begin{pmatrix} R\cos(dt)\cos(t) \\ R\cos(dt)\sin(t) \\ at \end{pmatrix}$$

$$\vec{t} = \frac{\dot{\vec{x}}}{\left\|\dot{\vec{x}}\right\|}$$

$$\vec{n} = \frac{\dot{\vec{t}}}{\left\|\dot{\vec{t}}\right\|}$$

$$\vec{b} = \vec{t} \times \vec{n}$$

$$\vec{x}_{local} = \begin{pmatrix} \vec{t} & \vec{n} & \vec{b} \end{pmatrix} \begin{pmatrix} Ac \cdot \cos(wt) \\ As \cdot \sin(wt) \\ 0 \end{pmatrix}$$

$$\vec{x} = \vec{x}_{frame} + \vec{x}_{local}$$

Box 2 Frenet Frame Equations for the class of Supercoiled Helices used by Helichrome x_{local} is the supercoil and x_{frame} is the large helix. t, n, and b are derived from x_{local}

After many iterations the resulting structures are of significantly lower energies than the initial structure and cluster into a small number of structure kinds (figure 2.12.) These structures visibly meet many of the constraints. A few examples are shown in the figure 2.11 .

Optchrom

Our second algorithm is a simple random walk of a freely jointed chain. We model the chromosome as a series of n points (816 in most of our runs.) At each iteration, we make a random movement of the structure, evaluate the change in cost, and accept or reject the new structure based on a Boltzmann probability. Again, after many iterations the resulting structure have much lower cost than the initial structure. Here, however, there are 2n –1 parameters and the structures have immense freedom.

Figure 2.13 shows a schematic representation or Optchrom. The panel on the left represents part of the freely jointed chain where the pink point is being moved. The panel on the right depicts the random movement generated by choosing a random displacement r and random direction defined by angles θ and φ.

Our primary use for the algorithm has been to refine the structures that result after optimization with Helichrom although it could easily be used to optimize the results of distance geometry metrization. When used in this way, Optchrom allows local optimization and introduction of disorder to the Helichrom structures so that they better meet the constraints. The resulting structures are generally of lower energy than the initial Helichrom structures but maintain, to some extent, the large scale order of the initial structure (figure 2.14).

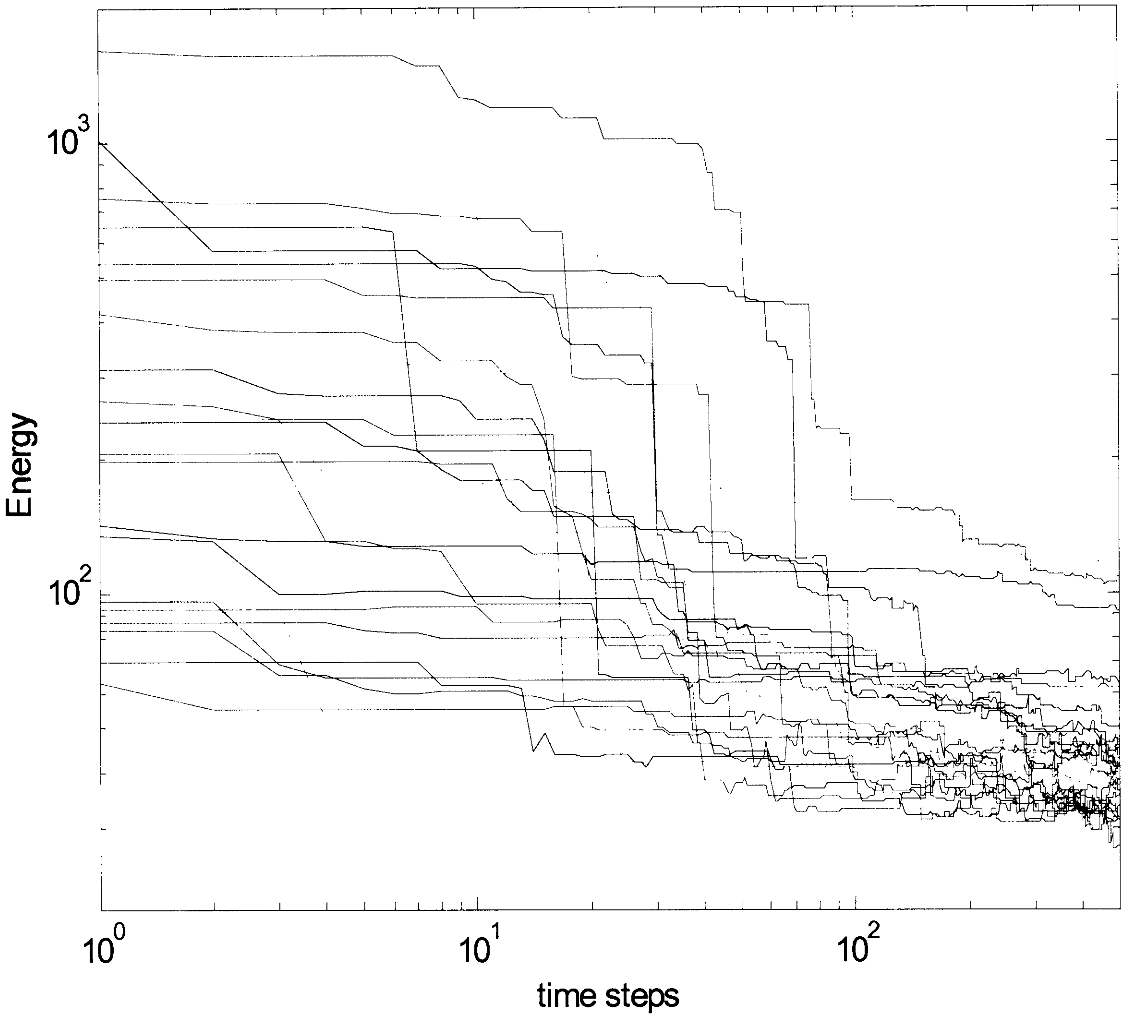


Figure 2.12 Energy of structures as a function of number of iterations of Montecarlo sampling.

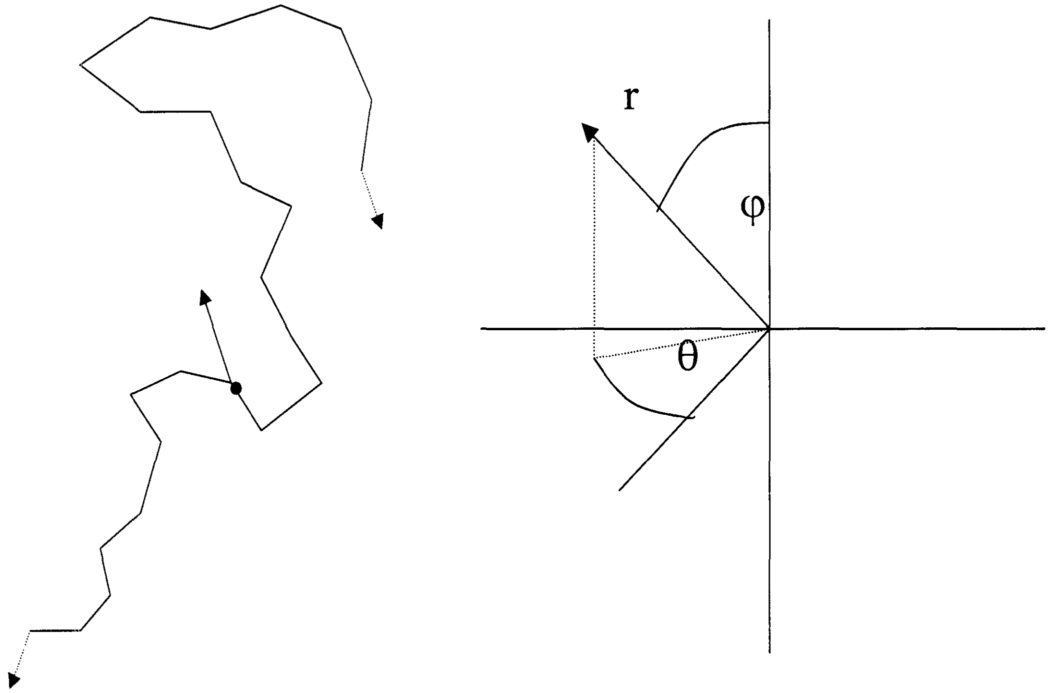


Figure 2.13 Schematic of Freely Jointed Chain Random Walk, Optchrom. Left: Displacement of a point on the structure and Right: Choice of the size and direction of the displacement.

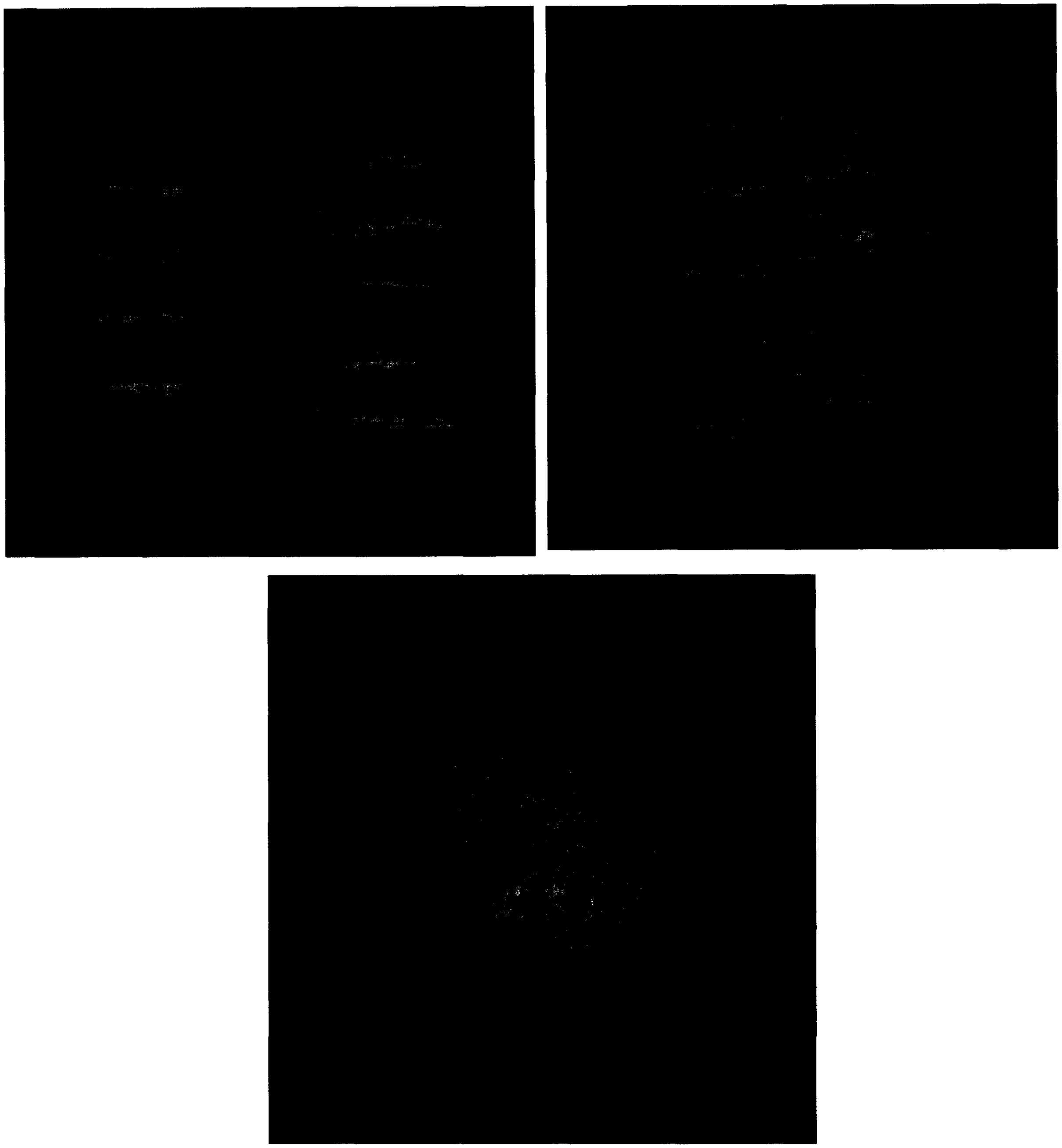


Figure 2.14 The results of Optchrom. Top left: Initial structure. Top right: structure of after relaxation through optchrom viewed from the side Bottom: top view of relaxed structure. In all panels, red indicates ribosomal protein coding gene. Blue indicates membrane protein coding gen e.

Predictions and possible experimental tests

Once we have generated an ensemble of structures that sample the conformation space of a given set of constraints it is possible to make predictions about the outcome of various experiments. We can, for example, predict the maximum or minimum distance between two points given the initial bound constraints and the initial smoothing in distance geometry. Such predictions could be verified by confocal microscopy using fluorescently labeled loci. We can also make global predictions about the distribution of distances that are expected in measurements (i.e. the flexibility of a locus) or the distribution of cross-linking frequencies that we expect in an experiment like chromosome conformation capture (10).

These predictions can be useful in many ways. They can allow us to distinguish between the viability of one particular set of constraints and another. The results of an experiment might allow us to rule out certain theoretical constraints as simply inconsistent with the data. Additionally we may be able to prescribe certain specific measurements that would allow us to distinguish between different structure classes of a given constraint set.

We expect there to be a close relationship between experiment and computation where experiments provide constraints and validation for structural models, and models provide direction for experiments. We believe that this work introduces an important new strategy for modeling large-scale structures in the cell and potentially provides a starting point for future cellular level studies based on evolutionary optimality.

References

1. Tomita, M. (2001) *Trends Biotechnol* **19,** 205-10.
2. Bailey, J. E. (2001) *Nat. Biotechnol.* **19,** 503-4.
3. Covert, M. W., Schilling, C. H., Famili, I., Edwards, J. S., Goryanin, II, Selkov, E. & Palsson, B. O. (2001) *Trends Biochem. Sci.* **26,** 179-86.
4. Jamshidi, N., Edwards, J. S., Fahland, T., Church, G. M. & Palsson, B. O. (2001) *Bioinformatics* **17,** 286-7.
5. Schilling, C. H., Edwards, J. S. & Palsson, B. O. (1999) *Biotechnol. Prog.* **15,** 288-95.
6. Cook, P. R. (2002) *Nat Genet* **32,** 347-52.
7. Simons, K. T., Bonneau, R., Ruczinski, I. & Baker, D. (1999) *Proteins* **Suppl 3,** 171-6.
8. Hearst, J. E., Kauffman, L. & McClain, W. M. (1998) *Trends Genet* **14,** 244-7.
9. Gasser, S. M. (2002) *Science* **296,** 1412-6.
10. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. (2002) *Science* **295,** 1306-11.
11. Bray, D. (1998) *Annu Rev Biophys Biomol Struct* **27,** 59-75.
12. Dworsky, P. (1976) *Biochem J* **154,** 239-41.
13. Danchin, A. & Henaut, A. (1997) *Curr Opin Genet Dev* **7,** 852-4.
14. Postow, L., Crisona, N. J., Peter, B. J., Hardy, C. D. & Cozzarelli, N. R. (2001) *Proc Natl Acad Sci U S A* **98,** 8219-26.
15. Dingman, C. W. (1974) *J Theor Biol* **43,** 187-95.

16. Havel, T. F., Kuntz, I. D. & Crippen, G. M. (1983) *J Theor Biol* **104,** 359-81.

17. Havel, T. F., Crippen, G. M., Kuntz, I. D. & Blaney, J. M. (1983) *J Theor Biol* **104,** 383-400.

18. Dandekar, T., Huynen, M., Regula, J. T., Ueberle, B., Zimmermann, C. U., Andrade, M. A., Doerks, T., Sanchez-Pulido, L., Snel, B., Suyama, M., Yuan, Y. P., Herrmann, R. & Bork, P. (2000) *Nucleic Acids Res* **28,** 3278-88.

Chapter 3

“Just in place” functional organization of bacterial chromosomes

(manuscript in preparation with Daniel Segrè and Peter Kharchenko)

Abstract

Bacterial chromosomes are compacted hundreds of times to fit within the cell[1-5]. In spite of this complexity, in several bacteria the compacted structure has been shown to be highly symmetric and ordered along the long-axis of the cell[6,7]. It is believed that the resulting folding of the genome is important for transcription and replication[8], but the structure remains unclear. Pairs of spatially close chromosomal loci can be used as constraints on the structure[9]. Here, by analyzing computationally 105 bacterial genomes, we identify a large putative set of such pairs and study their distribution along the *E. coli* chromosome. We find that paired genes are regularly spaced at multiples of 117Kb over the whole chromosome length. This pattern, along with the axial order, suggests that each half of the chromosome (or arc) is organized into a 117Kb periodic structure, potentially a helix. The periodic looping of the structure would align most paired genes along a single cell-long axis in each arc, similar to amphipathic α-helical proteins. Additionally, we find that the positions of the pairs are highly correlated with genome-wide expression. Thus the axial regions would correspond to hotspots of intense transcriptional activity, possibly a sign of selective pressure for optimal (or "just in place") localization of highly transcribed genes.

Bacterial chromosomes are organized, at the local level (~10Kb), into small irregular supercoiled domains which play an important role in transcription[1,10,11]. Recent fluorescence microscopy measurements in several bacteria have shown that loci are highly localized and undergo choreographed dynamics[7,12]. In addition, longitudinal position of a locus in the cell in several bacteria has been shown to be linearly related to chromosomal distance from the origin of replication along each arc (Fig. 3.1)[6,7]. It is not known how the topological domains are folded up into this high degree of order[13]. However, several studies showing long-range correlations and periodicities in expression and sequence along the chromosome, suggest that this folding may both be regularly structured and functionally significant[14-18].

We approach this problem from an evolutionary perspective. If the spatial organization of genes due to the folding of the chromosome is functionally important, then it will be under selection pressure during evolution[19]. We therefore expect to find signatures of spatial selection in genome sequences. In particular, we look for gene pairs selected to maintain spatial vicinity on the fold. Such pairs would represent constraints on the structure. This evolutionary perspective, which we address through comparative genomics[20,21], allows us to simultaneously identify potential structural constraints and their functional relevance. We concentrate on a resolution above ~20Kb, beyond the shortest expression correlations[16,17] and the average size of topological domains[1]. In the first part of our work we describe the methods used to identify the close pairs, in the second we analyze their distributions along the chromosome of *E. coli*, and lastly we discuss structural implications.

We computationally identified potential spatially close pairs based on their preference for close chromosomal positions, and their functional dependence (measured by phylogenetic co-occurrence[22]) in several bacterial genomes (see Methods.) We hypothesized that close chromosomal position preference across many genomes together with functional dependence will reflect preference for spatial vicinity. Our search spanned 5,000 genes and their orthologs in 105 bacterial genomes (a total of 12.5 million pairs) and yielded over 34,218 pairs, at a p-value cutoff of 10^{-10}.

In order to analyze the properties of the pairs, we take explicitly into account the symmetry between the two chromosomal arcs[6,7] (Fig. 1a). We therefore divide the pairs into two sets: those that are within a single arc (adjacent-pairs), which could yield information on the folding of each arc (Fig. 1a), and those that are on opposite arcs (across-pairs), which could give information about the relative position of the arcs. For each set we define a chromosomal distance: an adjacent-distance for each adjacent pair, defined as the gene-gene distance along the genome sequence. And an across-distance, for each across pair, defined as the difference in distances of the genes from the origin, (Fig. 3.1). The across-distance reflects the deviation from perfect symmetry. The distributions of these chromosomal distances should bear the signature of different chromosomal folds (Supplementary Fig. 3.S1).

We first examined whether the selected pairs show signs of the observed closeness of symmetric points about the origin of replication (Fig. 3.1 and Refs 6,7). To do this, we

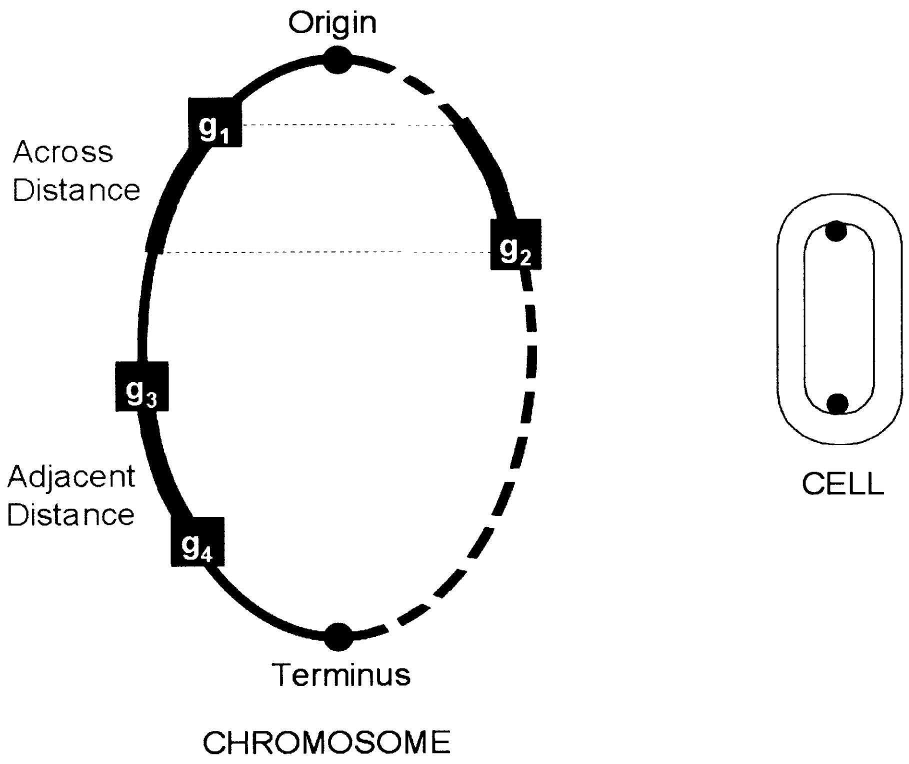


Figure 3.1 Structural features of bacterial chromosomes can be analyzed by using appropriate gene-gene metrics along the genomic sequence.Left: Definition of two types of distance between pairs of genes along a circular chromosome. The Origin and Terminus of replication define a symmetry axis, which divides the chromosome into two arcs (right - dashed, and left - solid). The Adjacent Distance is defined, for gene pairs on the same chromosome arc (g_3 and g_4, adjacent-pairs), as the difference between their genomic sequence positions (green). The Across Distance, for gene pairs on opposite arcs (g_1 and g_2, across-pairs) is the difference in their genomic distance to the origin of replication (blue). Right: In some organisms, it has been observed that the order in which the genes appear on both axes of the chromosome is preserved (arrows) on the physical structure of the folded chromosome (dashed area) in the cell.

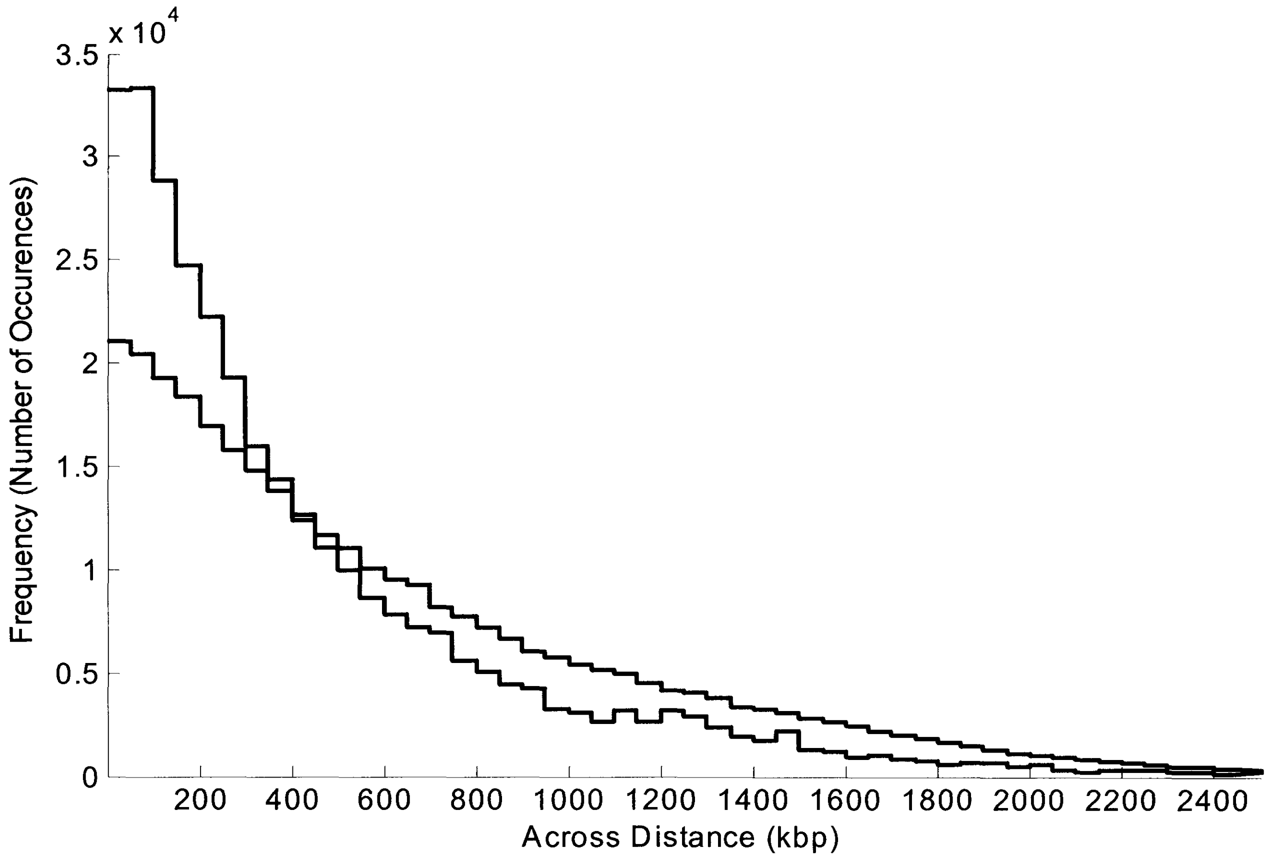


Figure 3.2 Histogram of the across distance for paired genes occurring on different chromosome arms in any of 68 genomes (zero represents perfect symmetry). The distribution of the pairs selected through closeness on the genome and phylogenetic co-occurrence (black line, 318000 total pairs) is compared toa distribution for random pairs (red line).

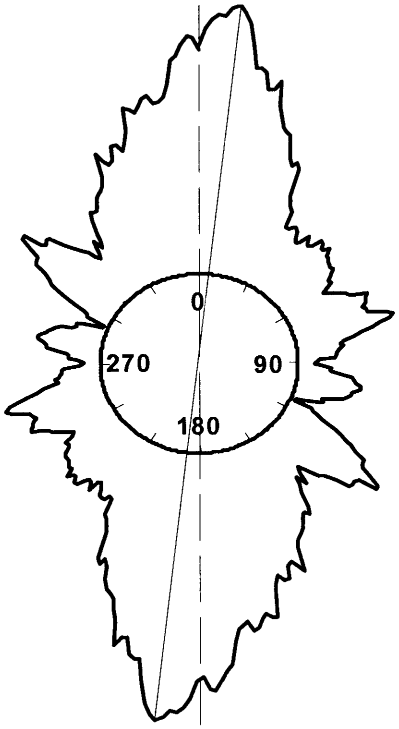


Figure 3.3 Symmetry score as a function of the position of the origin of replication in *E. coli*. The symmetry score (black curve) indicates the number of selected pairs that are on opposite arms of the chromosome for this particular choice of origin position along the circle and have across distance within a symmetry window of 200kb. Increasing distance from the circle center represents increasing symmetry score. Black radial line: position of maximal symmetry. Dashed line: actual origin.

studied the distribution of the across-distances over many genomes (see Methods.) The distribution is significantly different from random ($p<10^{-300}$), displaying a pronounced skew towards high-symmetry values (Fig. 3.2). In many individual genomes, the pair information is significant enough to allow an unbiased prediction of the position of the origin of replication (Fig. 3.3).

Next, we examined the distributions of adjacent and across-distances in detail in *E. coli*. For both distributions, the expectation for a random fold is a curve peaked at zero, which decreases monotonically and linearly with increasing distances (Supplementary Figure 3.S2). Instead, the distribution of adjacent-distances reveals a series of major peaks that are spaced at intervals of 117 kb (Fig. 3.4, and Fourier spectrum inset). This observation is consistent with correlations of 115kb in genome-wide expression data[14,16,17] and periodicities of ~100Kb[23]. The distribution of across-distances displays a similar set of peaks, spaced also at 117Kb (Fig. 3.5). We find similar patterns in other organisms (in particular *C. crescentus*, see Supplementary Fig. 3).

To study whether these preferred distances occur between genes anywhere along the genome or are confined to certain regions, we analyzed the pair density along the chromosome, defined as the number of times a gene is involved in pairs with other genes. The density curves were normalized to correct for possible biases in overall gene density. As shown in Fig. 3.6, the adjacent-pair position distribution is far from uniform, having strong peaks at preferred positions throughout the genome. The across-pair distribution displays the same set of peaks. Notably, the most prominent peaks, or pair clusters, are

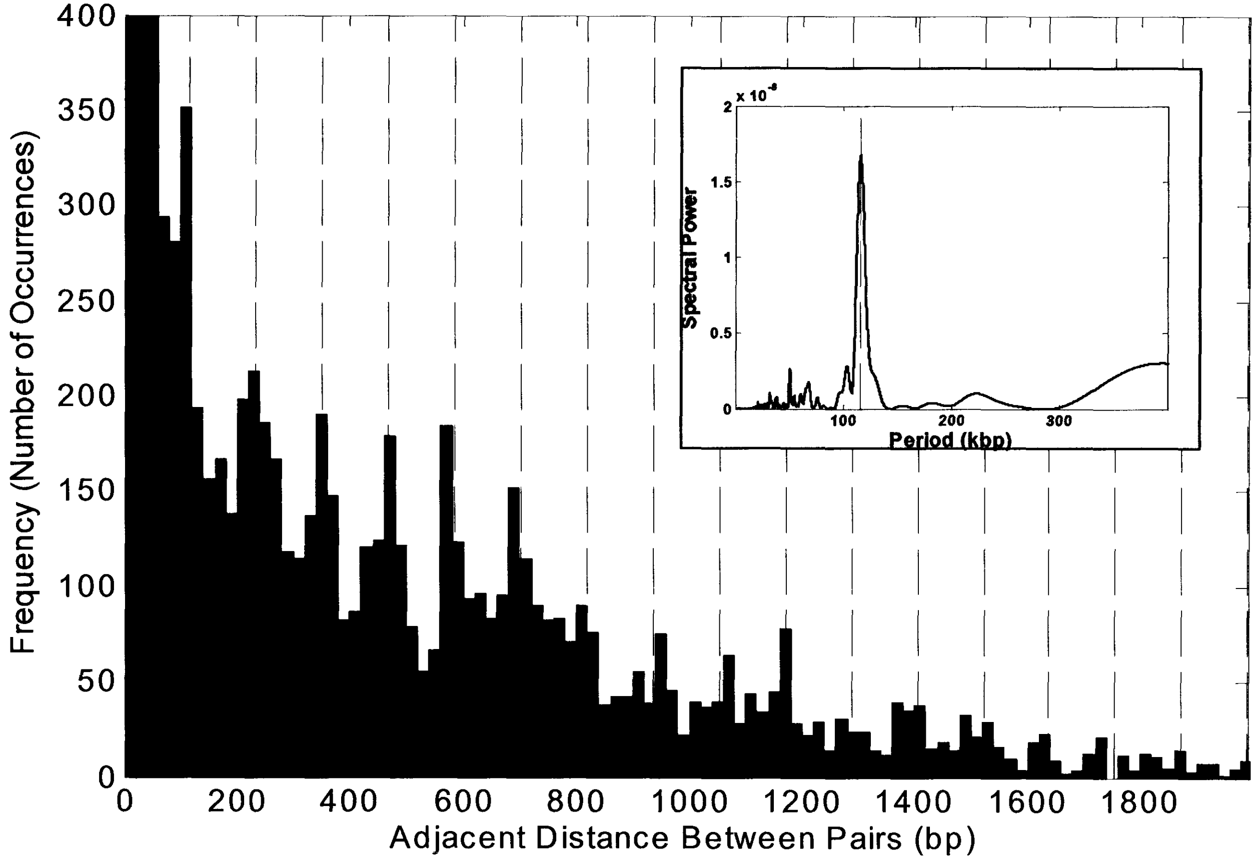


Figure 3.4 Distribution of adjacent distances, between paired genes occurring on the same chromosome arc of the *E. coli* genome (see Figure 1a). The distribution displays periodic peaks, indicating a preferred chromosomal distance (117 Kb) between genes belonging to a pair. A 117 kb –spaced grid is overlapped to the histogram. **Inset**: Fast Fourier Transform analysis of the distribution, displaying a major 117Kb peak.

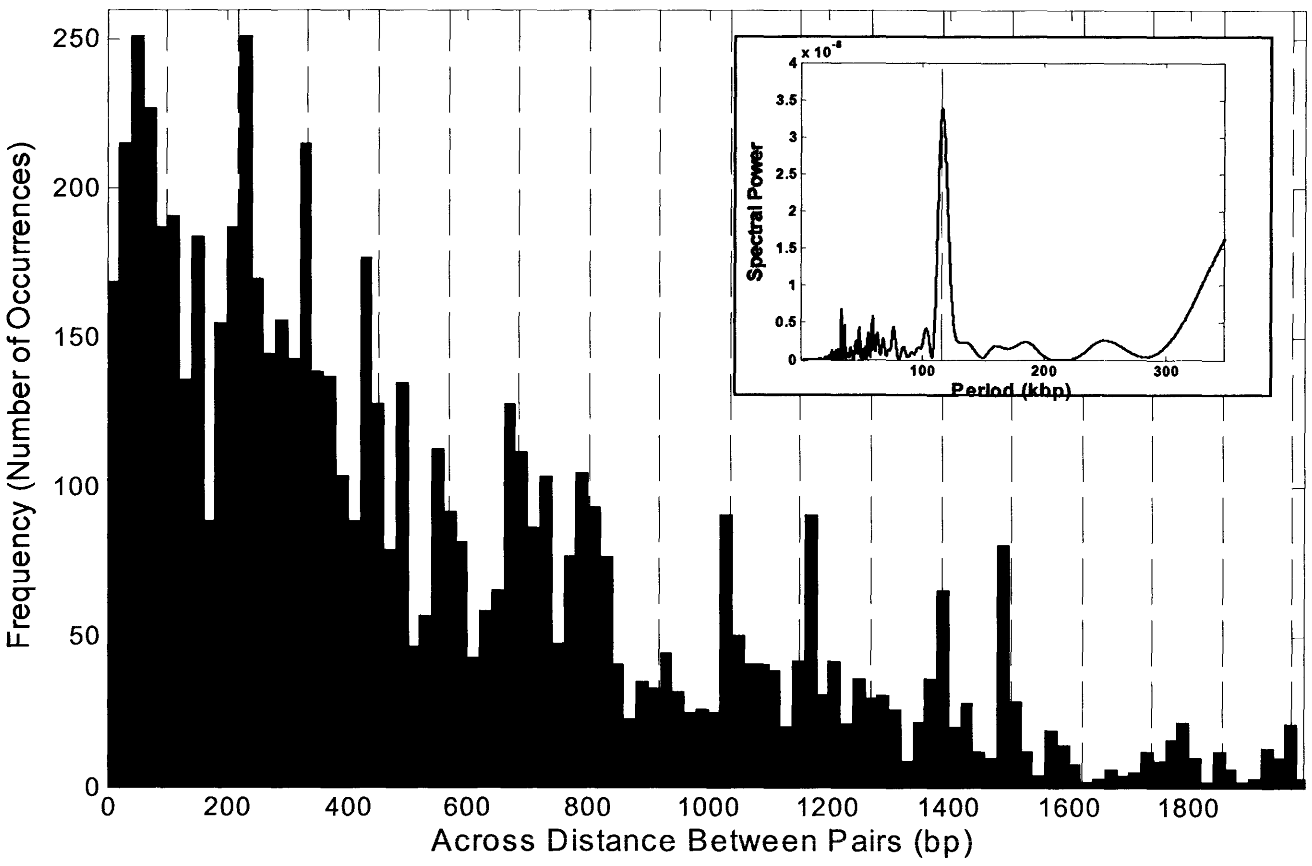


Figure 3.5 Distribution of across-distances between paired genes on opposite chromosomal arcs in the *E. coli* genome. The same periodicity of 117Kb can be observed, and confirmed by Fast Fourier Transform analysis (Inset).

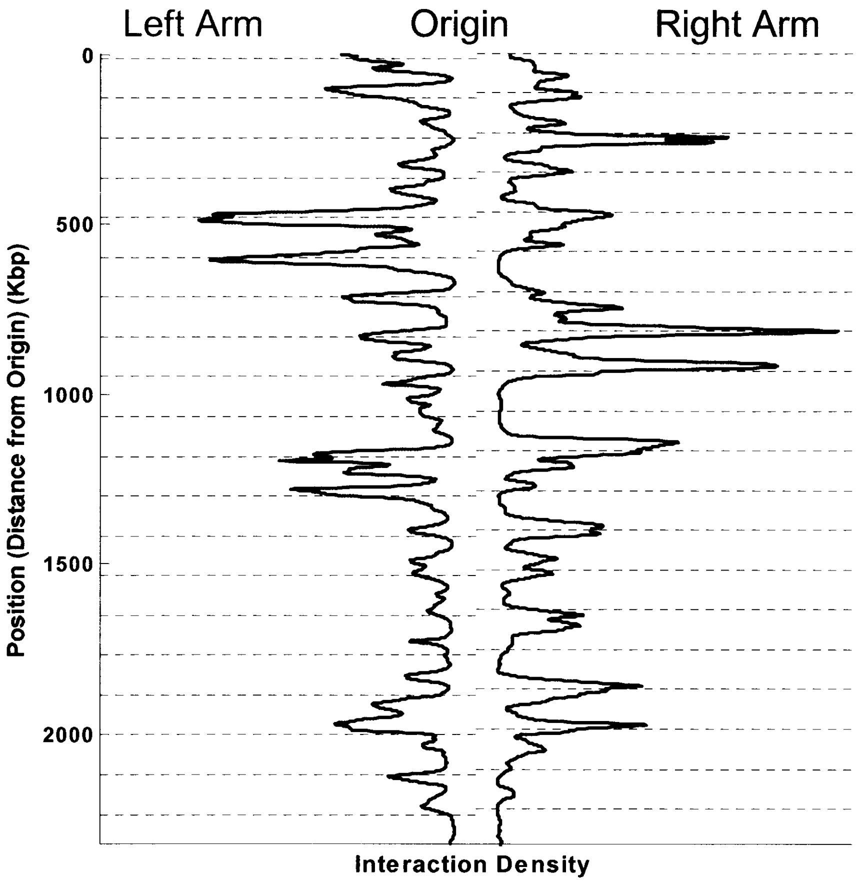


Figure 3.6 Density of paired genes along the chromosome. The distance from the central axis indicates the number of times a gene at this position is in a pair with another gene, i.e. the total number of pairs involving this position. The distributions for the right arm (black) and for the left arc (blue) are placed facing each other to emphasize the symmetry. Peaks occur preferentially in phase with the 117kb-spaced rid (horizontal lines).

arranged on a single grid spaced at the same 117kb period, which spans the entire length of each chromosome arc symmetrically (Fig. 3a).

Thus, we observe a strong periodicity in the distances between the paired genes, and in the positions they occupy, which extends over the entire length of the genome. Since we expect pairing to indicate spatial vicinity, these patterns suggests that each arc is organized into a set of 117Kb loops which bring the pair clusters close together in space. Because of the experimental evidence of linear longitudinal order[7] (Fig. 1a), we propose that the loops in each arc are arranged into a cell-long stack, potentially a helix. The pair clusters may then align along two (possibly coinciding) longitudinal axes or faces (Fig. 3.9). Each axis would be analogous to the hydrophobic face of an amphipathic α-helical protein[24].

While the pair clustering at particular chromosomal locations (and hence along the axes described above), leaves the detailed geometry of the loops unknown, it also prompts the question: is the pairing along a single axis indicative of some benefit to the cell? A first answer to this question comes from the observation that gene pairs contain many transcription and translation-related genes (Supplementary Table). To study the functional implications in more detail, we analyzed genome-wide expression data. Specifically we examined the correlation between pair density and absolute gene transcriptional level derived from log growth expression data for *E. coli* (Ref 14 and Methods). As visualized in Fig. 3.7, the pair density profile recapitulates often in fine details the expression profile along the entire chromosome. This relationship, which is also confirmed by a correlation coefficient of 0.6 (p-val=10^{-32}), indicates that the

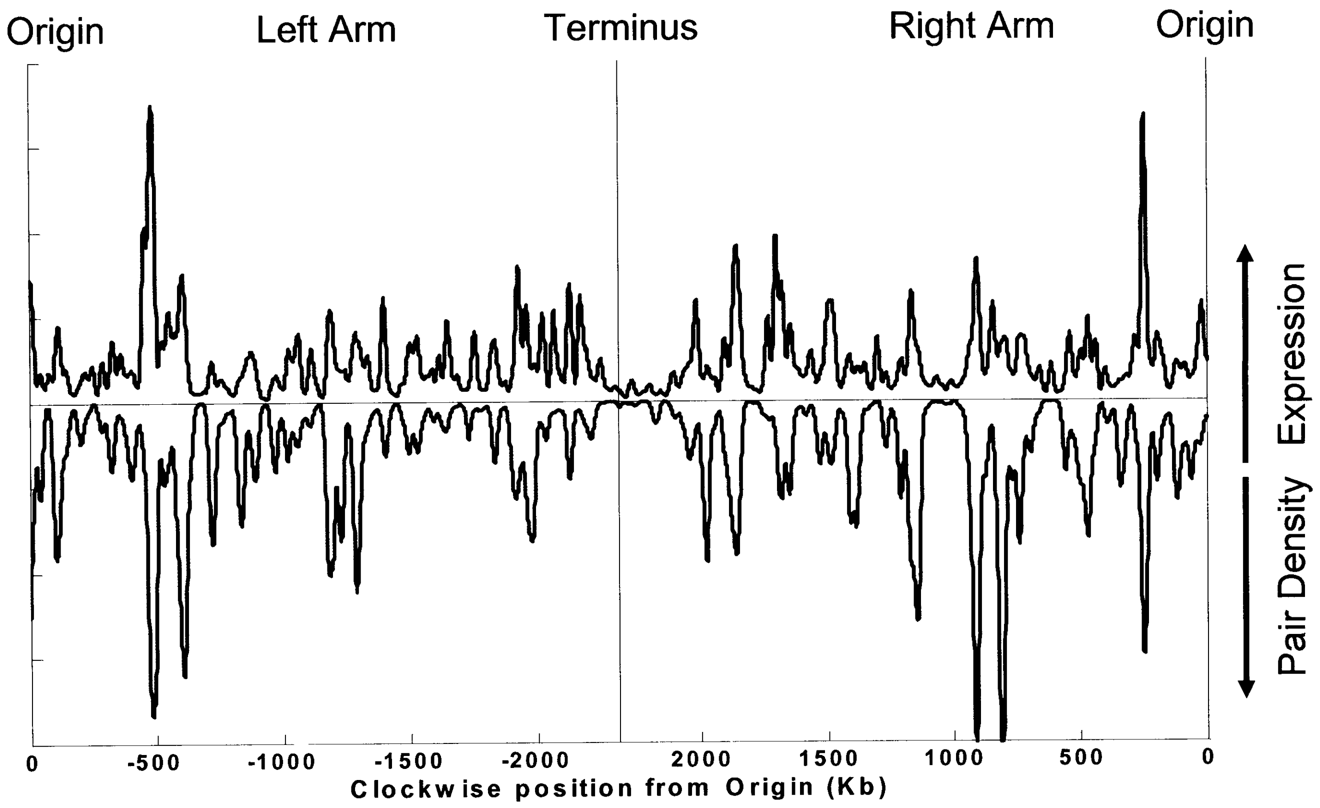


Figure 3.7 Comparison of transcription level with pair density along the chromosome. Absolute expression level (red, on top) during log phase growth is plotted as a function of chromosomal position. Expression is compared with pair density (bottom distribution) along the left (blue) and right (black) arcs of the chromosome. Distance above the horizontal zero indicates increasing expression and below zero it indicates increasing pair density (in arbitrary units). Both distributions are normalized by the density of genes along the chromosome.

periodically positioned clusters of paired genes along the entire chromosome are hotspots of intense transcriptional activity. Intriguingly, this correlation decreases as *E. coli* enters stationary phase (Supplementary Fig. 3.S5), suggesting that the hotspots of transcription are important specifically for log phase growth.

These results can be integrated to objectively evaluate quantitative models of chromosome structure, and understand their biological implications. In particular, by interpreting the pairs as close distance constraints, we can calculate a goodness of fit (similar to RMS in molecular modeling), by ascribing an energy to any given structure. Since the pair cluster data suggest a cell long stack of loops for each arc, we began our analysis with two simple helices, one for each arc. Analyzing the energy as a function of the period (Fig. 3.8a) and the relative rotation of the helices (Fig. 3.8b) shows a narrow optimum (Fig. 3.9 a,b). We also analyzed structures in which consecutive loops along the stack in each arc are rotated relative to each other, thereby maintaining the axial alignment of pair clusters (Fig. 3.10b, Supplemental Fig. 3.S6 and Methods). The ensuing optimal structure (Fig. 3.10b) is a coiled coil, characterized by a 580Kb periodicity in addition to the 117Kb loop organization, consistent with previously observed longer–range periodicities[14,16,17] and longer range periodicities observed in the pair density (Supplementary Fig. 3.S7). Structural models can also be overlapped with functional information, as exemplified by the helical moments[24] shown in Fig. 3.9c. In several optimal structures obtained (Fig. 3.9a,b 3.10b), a pair cluster axis common to both arcs is

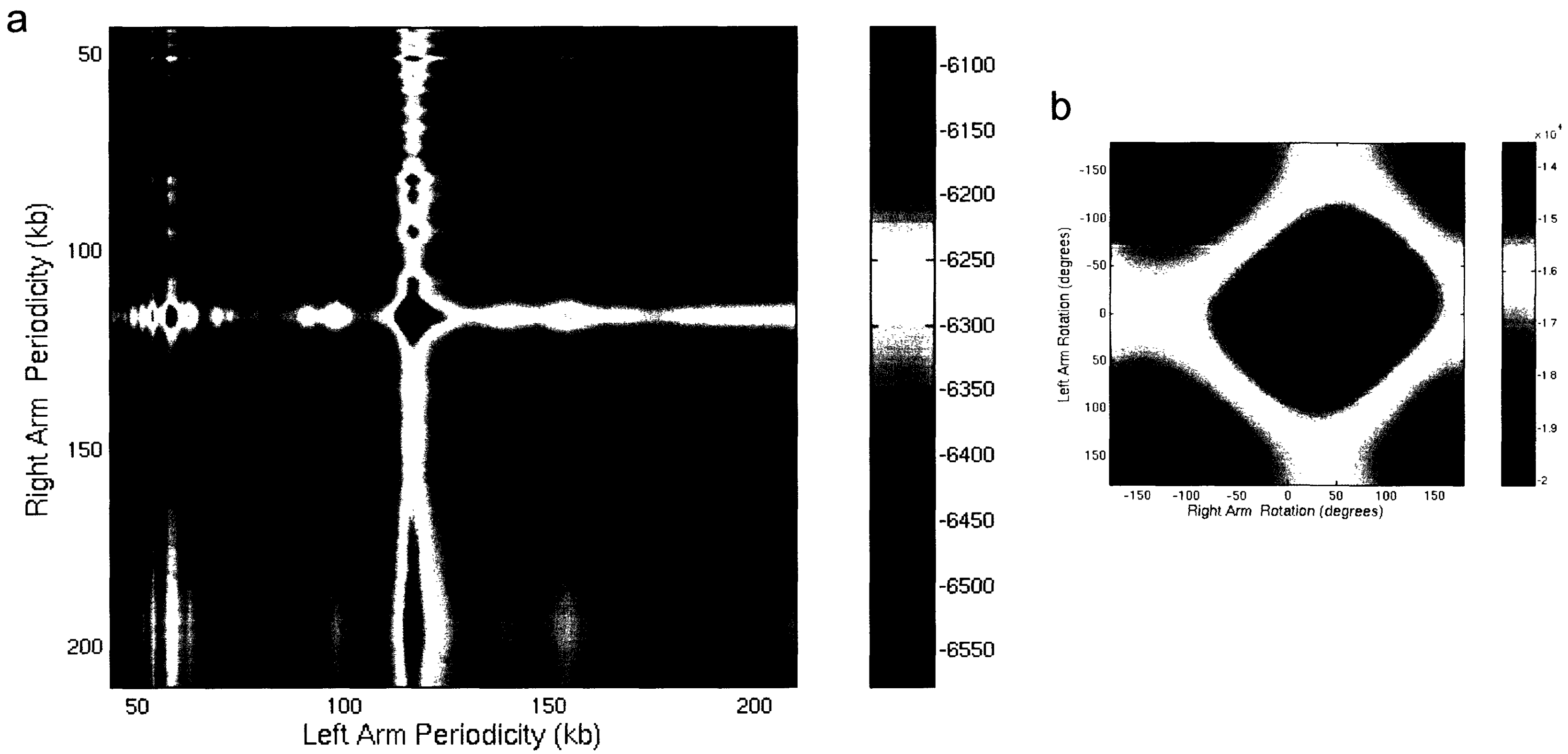


Figure 3.8 a) Landcape for fit of pairs as a function of the size of looping for right (x axis) and left (y axis) chromosome arms as helices. Red indicates best fit to pairs. b) Fitness landscape for rotation of helices about their individual central axes. The period is fixed at the 117 kb (optimum) and the helices are separated so that they touch along a single face. Red indicates best fit to pairs.

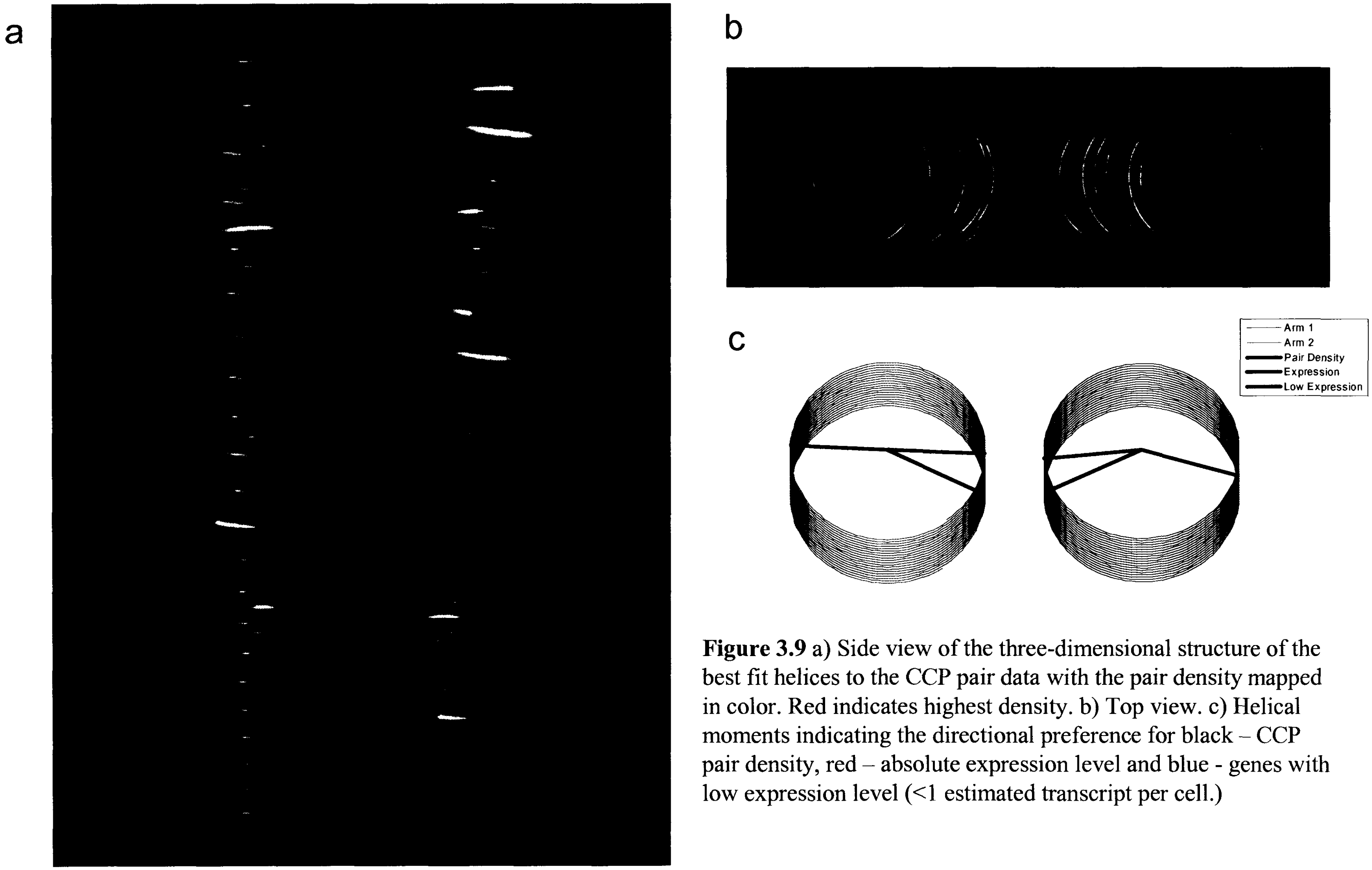


Figure 3.9 a) Side view of the three-dimensional structure of the best fit helices to the CCP pair data with the pair density mapped in color. Red indicates highest density. b) Top view. c) Helical moments indicating the directional preference for black – CCP pair density, red – absolute expression level and blue - genes with low expression level (<1 estimated transcript per cell.)

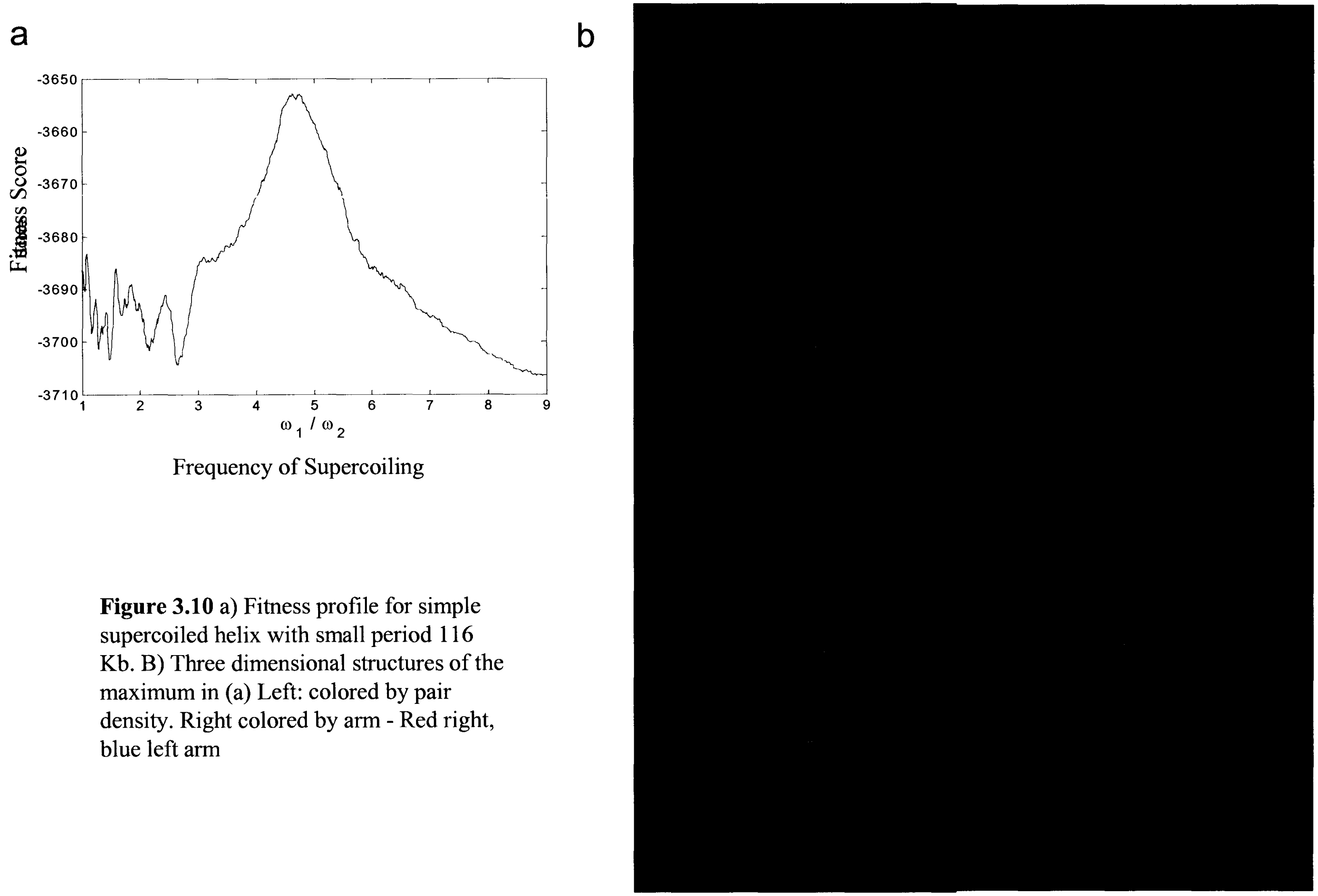


Figure 3.10 a) Fitness profile for simple supercoiled helix with small period 116 Kb. B) Three dimensional structures of the maximum in (a) Left: colored by pair density. Right colored by arm - Red right, blue left arm

positioned in the center of the nucleoid, which correlation with expression indicates is a high transcriptional region. This is consistent with the observation that transcription plays a role in organizing the nucleoid[18], and in particular with the localization of RNA polymerase in foci at the center of the bacterial nucleoid[25-27]. Interestingly, this localization of RNA polymerase to the center of the nucleoid is also related to log phase growth; the foci disperse in other conditions[26] analogous to the way in which the expression correlation with pair density decreases during the time course into the stationary phase.

Our analysis infers, solely on the basis of bacterial genome sequences, a set of strongly symmetric and periodic gene distribution signals. These signals are consistent with observations from microscopy, expression correlation, and absolute expression level. This suggests that these evolutionarily chosen genes pairs are representative of a structural periodicity and that this periodicity is functionally relevant. The interpretation of a periodically looped chromosome fold coherently fits all of the observations and offers a strong biological explanation for the pairing. While potential alternative interpretations may be possible, they would have to explain the observed symmetry, distance and density periodicities, as well as expression correlation.

We do not expect the chromosome to be statically organized. Rather, we expect that the features of the fold we describe are related to log-phase growth and that the structure may exhibit complex dynamics through replication[6,7,12] as well as condition-dependent changes[2,18,19,28]. Additionally, the 117kb loops we describe should be viewed as approximate backbones composed of smaller-scale, possibly irregular, topological domains (Supplementary Fig. 5). In our modeling, we used pair data as structural constraints on the fold. Extensions of these quantitative approaches (e.g. Distance Geometry[29]) and optimization algorithms[13], incorporating a range of experimental, computational and physical constraints may, like in the case of proteins, be used for detailed chromosome folding predictions. Additional experimental data, for example directly measuring the propensity of different chromosomal loci to be spatially close, could help test our predictions, and provide additional useful constraints. Moreover, observations of fitness changes following transpositions of genomic regions containing pair clusters may serve as direct tests of the proposed models.

The high correlation between pair density and genome-wide expression level holds independently of any specific interpretation. However, we interpret this correlation as indicative of an evolutionarily optimized organization which maintains highly transcribed genes in specific chromosomal locations. This "just in place" principle, which extends to the space domain the use of optimality[30,31], may be responsible in general for placing specific genes in positions where they can be maximally efficient: for example membrane proteins near the membrane[32], or components of large macromolecular complexes where they can be cotranslated and assembled. The optimization almost certainly reflects replication and ordered structures, such as the ones we propose (Fig 5a,b), would help minimize entanglement during replication.

A optimization of the chromosome fold for transcription is in line with previous hypotheses[18] of nucleoid organization as well as with emerging understanding of analogous organizational principles in the eukaryotic nucleus[8,18,19,33]. This suggests that comparative genomics approaches similar to ours may be used also in the context of eukaryotes. Ultimately, genome sequences and their structures may turn out to be highly interdependent aspects of a single, finely tuned system.

Methods

Selection of gene pairs according to vicinity on the chromosome and philogenetic co-occurrence

First, evolutionary clustering of genes on the chromosome was calculated based on the null hypothesis that orthologous genes are randomly ordered on the chromosomes. For a pair of genes *x* and *y*, chromosome clustering was evaluated as a probability $P(x,y) = \prod_{g \in G} P_g\left(D \leq d_g(x,y)\right)$, where G is a set of query genomes and $P_g\left(D \leq d_g(x,y)\right)$ is a probability of observing gene order distance D less then or equal than $d_g(x,y)$ - the distance between orthologs of genes *x* and *y* in the organism g. P_o was calculated numerically under the null hypothesis, based on the organism chromosome sizes. The results are based on a set of 105 bacterial and three eukaryotic genomes (*S. cerevisiae, S. pombe, C. elegans*) from Genbank. The set was screened to eliminate closely related species, using ortholog occurrence mutual information threshold of 0.9. Orthology mapping was established using best bi-directional orthologs from KEGG SSDB[34]. Phylogenetic profile co-occurrence association was assessed using hypergeometric probability distribution, as described in Ref. 22. The orthologs were determined using best bi-directional BLASTP hits against NCBI NR protein dataset. Organisms containing orthologs for less than 1% of *S. cerevisiae* genes were excluded from calculations.

Pairs of genes were selected by taking all pairs of orthologs with phylogenetic co-occurrence and chromosome clustering scores of p-value $< 10^{-10}$

Across and adjacent distances on chromosome arcs

The chromosome was separated into two halves (arcs), clockwise from origin to terminus (right arc) and counterclockwise from origin to terminus (left arc) (see Fig. 3.1). A gene pair was considered adjacent if genes were both on the same arc and across if genes were on opposite arcs. Each gene was assigned a coordinate g, representing its distance from the origin, and gene-gene distances were calculated as the difference in these coordinates ($g_1 - g_2$), "across-distances" for across-pairs and "adjacent-distances" for adjacent pairs. The origin positions were taken from the Genome Atlas (http://www.cbs.dk/services/GenomeAtlas)

Distribution of the across-distances over many genomes

The histogram for selected gene pairs was constructed by taking all pairs in a subset of 68 bacteria for which we had mappings through Clusters of Orthologous Groups (COGS) (http://www.ncbi.nlm.nih.gov/COG/) and calculating the across-distance for all pairs in all genomes where the genes were on opposite arms. The p-value was calculated using a Kolmogorov-Smirnov distribution test.

Prediction of the position of the origin of replication

The position of the origin of replication in *E. coli* was computationally rotated to each of 360 equally spaced positions along the chromosome. For each position, the across distance was calculated for all pairs placed on opposite arms of the chromosome for this choice of origin. The symmetry score was calculated as the number of pairs with across distance, d < 200,000 excluding all pairs with genomic positions about the rotated origin or terminus that differ by less than 200,000.

Distributions of Distances and Positions, and Fast Fourier Transform

All pairs in *E. coli* were classified as either across (on opposite arms) or adjacent (on the same arm) as defined above. We constructed a histogram of the genomic distances between the genes in each pair for a particular class, using adjacent distances for adjacent pairs and across distances for across pairs. We estimated the continuous probability density for the distances between the pairs in the across and in the adjacent classes using a Gaussian smoothing window ($\sigma = 4000$bp) with the ksdensity function in MATLAB (smoothing with a Gaussian window and normalizing the total density to 1.) We then took the discrete Fourier transform using a standard Tukey window to taper the ends (ratio of .75 for tapered to untapered length) and a length of 200,000.

The position densities were calculated using the Gaussian smoothing estimate above with $\sigma = 8000$bp. The grid at 117Kb was fit to the center of the largest peak in each half of the chromosome.

Expression correlation

We calculated an average of the absolute transcript level for wild type standard growth conditions (MOPS minimal glucose) measured on 5 affymetrix microarrays in the ASAP database (www.genome.wisc.edu/tools/asap.htm) Data reported in (Allen *et al.*[14]) This data was smoothed using a truncated gaussian $\sigma = 8000$bp (maximum width 16,000bp.) We calculated the Pearson correlation coefficient of the smoothed data with the pair position density, sampling once every 16,000 bp to avoid smoothing artifacts. The p-value was computed using a t-test with n -2 degrees of freedom (n is the length of the data.) The p-value and correlation coefficient did not vary significantly with the choice of sampling phase.

Three-dimensional models and energy landscapes

Simple helix

The coordinates of the helices were calculated according to the helical equations

$$\begin{aligned} x &= r\cos(t) \\ y &= r\sin(t) \\ z &= pt \end{aligned}$$

where r = .30, which is ¾ of the diameter (in microns) of the *E. coli* cell, and the pitch, p was varied to reflect different periods of looping. Genes were assigned to the helices by dividing the total arc length of the helix that fits inside the cell (height or z coordinate 2.5 μm) into segments that each represent 1 kb of a genome arc (2320 segments in E. coli.)

Planar Supercoils

Planar supercoils were constructed according to the equations.

$$\begin{aligned} x &= r(\cos(ft) + \cos(gt)) \\ y &= r(\sin(ft) + \sin(gt)) \\ z &= t \end{aligned}$$

where r is a radius of .15,

$$\begin{aligned} f &= \frac{2\pi}{mk} \\ g &= (k+1)f \end{aligned}$$

Where, m is the period of the small supercoils and k is the number of small rotations per large supercoiling rotation.

Energy Landsapes

To evaluate how well a given structure satisfies the constraints derived from the selected gene pairs, we calculated the sum S of the spatial distances between all pairs (k) on the model structure:

$$S = \sum_{k} D_k$$

where D_k is the Euclidean distance between the positions occupied by the two genes in pair k. The coordinate of each gene along the genome was taken to be the position of its center.

References

1. Postow, L., Hardy, C. D., Arsuaga, J. & Cozzarelli, N. R. Topological domain structure of the Escherichia coli chromosome. *Genes Dev* **18**, 1766-79 (2004).
2. Thanbichler, M., Viollier, P. H. & Shapiro, L. The structure and function of the bacterial chromosome. *Curr Opin Genet Dev* **15**, 153-62 (2005).
3. Lindow, J. C., Kuwano, M., Moriya, S. & Grossman, A. D. Subcellular localization of the Bacillus subtilis structural maintenance of chromosomes (SMC) protein. *Mol Microbiol* **46**, 997-1009 (2002).
4. Teleman, A. A., Graumann, P. L., Lin, D. C., Grossman, A. D. & Losick, R. Chromosome arrangement within a bacterium. *Curr Biol* **8**, 1102-9 (1998).
5. Seto, S. & Miyata, M. Partitioning, movement, and positioning of nucleoids in Mycoplasma capricolum. *J Bacteriol* **181**, 6073-80 (1999).
6. Niki, H., Yamaichi, Y. & Hiraga, S. Dynamic organization of chromosomal DNA in Escherichia coli. *Genes Dev* **14**, 212-23 (2000).
7. Viollier, P. H. et al. Rapid and sequential movement of individual chromosomal loci to specific subcellular locations during bacterial DNA replication. *Proc Natl Acad Sci U S A* **101**, 9257-62 (2004).
8. Cook, P. R. The organization of replication and transcription. *Science* **284**, 1790-5 (1999).
9. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306-11 (2002).

10. Travers, A. & Muskhelishvili, G. DNA supercoiling - a global transcriptional regulator for enterobacterial growth? *Nat Rev Microbiol* **3**, 157-69 (2005).

11. Hatfield, G. W. & Benham, C. J. DNA topology-mediated control of global gene expression in Escherichia coli. *Annu Rev Genet* **36**, 175-203 (2002).

12. Webb, C. D. et al. Bipolar localization of the replication origin regions of chromosomes in vegetative and sporulating cells of B. subtilis. *Cell* **88**, 667-74 (1997).

13. Wright, M. A., Segrè, D. & Church, G. M. in *ICSB 2002, 3rd International Conference on Systems Biology* 229-230 (Stockholm, Sweden, 2002).

14. Allen, T. E. et al. Genome-scale analysis of the uses of the Escherichia coli genome: model-driven analysis of heterogeneous data sets. *J Bacteriol* **185**, 6392-9 (2003).

15. Kepes, F. Periodic transcriptional organization of the E.coli genome. *J Mol Biol* **340**, 957-64 (2004).

16. Jeong, K. S., Ahn, J. & Khodursky, A. B. Spatial patterns of transcriptional activity in the chromosome of Escherichia coli. *Genome Biol* **5**, R86 (2004).

17. Carpentier, A. S., Torresani, B., Grossmann, A. & Henaut, A. Decoding the nucleoid organisation of Bacillus subtilis and Escherichia coli through gene expression data. *BMC Genomics* **6**, 84 (2005).

18. Cook, P. R. Predicting three-dimensional genome structure from transcriptional activity. *Nat Genet* **32**, 347-52 (2002).

19. Chakalova, L., Debrand, E., Mitchell, J. A., Osborne, C. S. & Fraser, P. Replication and transcription: Shaping the landscape of the genome. *Nat Rev Genet* (2005).

20. Yanai, I. & DeLisi, C. The society of genes: networks of functional links between genes from comparative genomics. *Genome Biol* **3**, research0064 (2002).

21. Iyer, L. M., Makarova, K. S., Koonin, E. V. & Aravind, L. Comparative genomics of the FtsK-HerA superfamily of pumping ATPases: implications for the origins of chromosome segregation, cell division and viral capsid packaging. *Nucleic Acids Res* **32**, 5260-79 (2004).

22. Bowers, P. M. et al. Prolinks: a database of protein functional linkages derived from coevolution. *Genome Biol* **5**, R35 (2004).

23. Kepes, F. Periodic epi-organization of the yeast genome revealed by the distribution of promoter sites. *J Mol Biol* **329**, 859-65 (2003).

24. Pilpel, Y., Ben-Tal, N. & Lancet, D. kPROT: a knowledge-based scale for the propensity of residue orientation in transmembrane segments. Application to membrane protein structure prediction. *J Mol Biol* **294**, 921-35 (1999).

25. Lewis, P. J., Thaker, S. D. & Errington, J. Compartmentalization of transcription and translation in Bacillus subtilis. *Embo J* **19**, 710-8 (2000).

26. Cabrera, J. E. & Jin, D. J. The distribution of RNA polymerase in Escherichia coli is dynamic and sensitive to environmental cues. *Mol Microbiol* **50**, 1493-505 (2003).

27. Liu, M. et al. Global transcriptional programs reveal a carbon source foraging strategy by Escherichia coli. *J Biol Chem* **280**, 15921-7 (2005).

28. Danchin, A. & Henaut, A. The map of the cell is in the chromosome. *Curr Opin Genet Dev* **7**, 852-4 (1997).

29. Havel, T. F., Kuntz, I. D. & Crippen, G. M. The combinatorial distance geometry method for the calculation of molecular conformation. I. A new approach to an old problem. *J Theor Biol* **104**, 359-81 (1983).

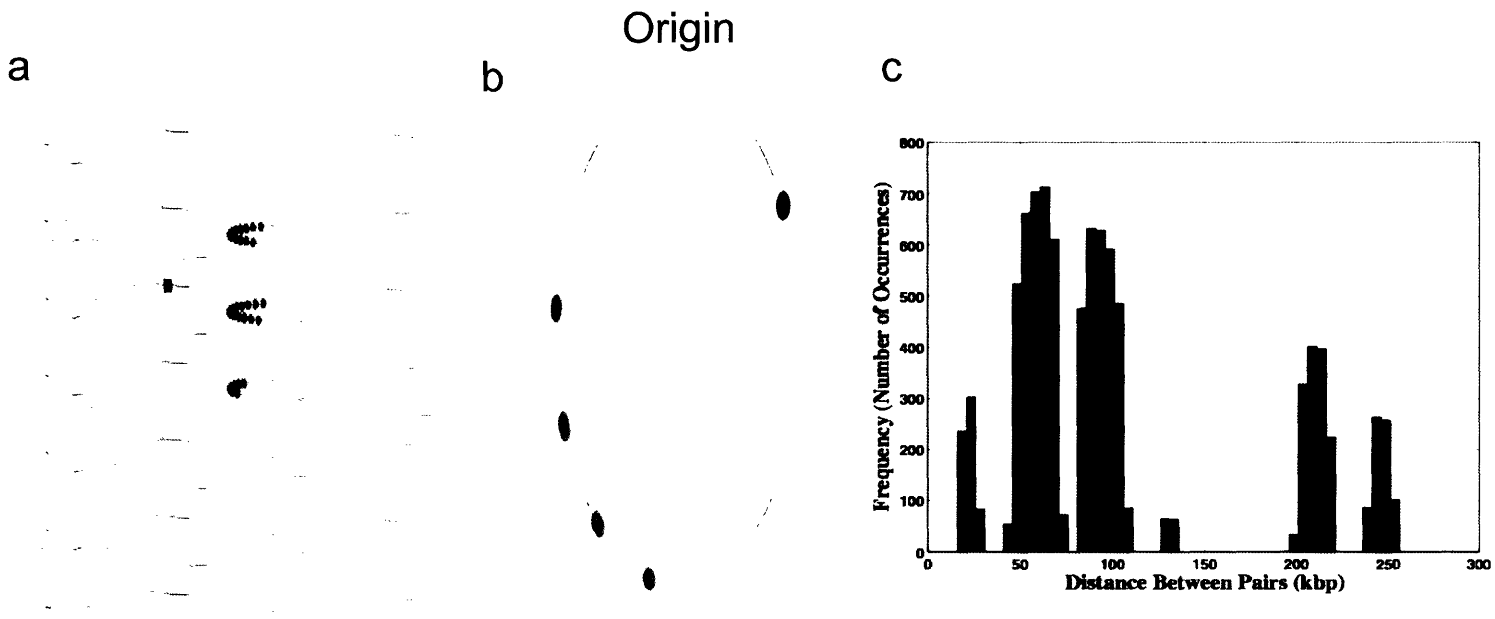
30. Zaslaver, A. et al. Just-in-time transcription program in metabolic pathways. *Nat Genet* **36**, 486-91 (2004).

31. Schuster, S. & Heinrich, R. Time hierarchy in enzymatic reaction chains resulting from optimality principles. *J Theor Biol* **129**, 189-209 (1987).

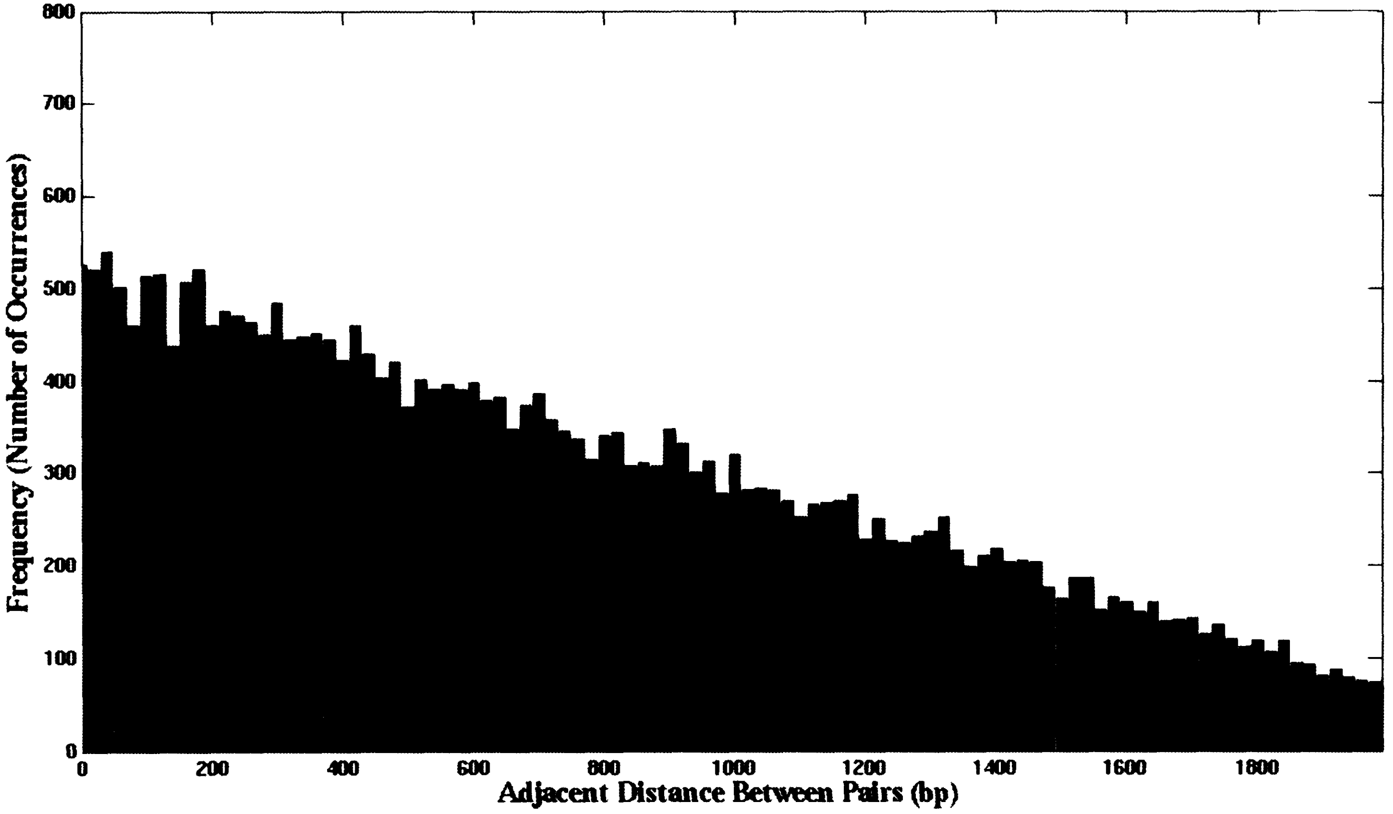
32. Woldringh, C. L. The role of co-transcriptional translation and protein translocation (transertion) in bacterial chromosome segregation. *Mol Microbiol* **45**, 17-29 (2002).

33. Taddei, A., Hediger, F., Neumann, F. R. & Gasser, S. M. The function of nuclear architecture: a genetic approach. *Annu Rev Genet* **38**, 305-45 (2004).

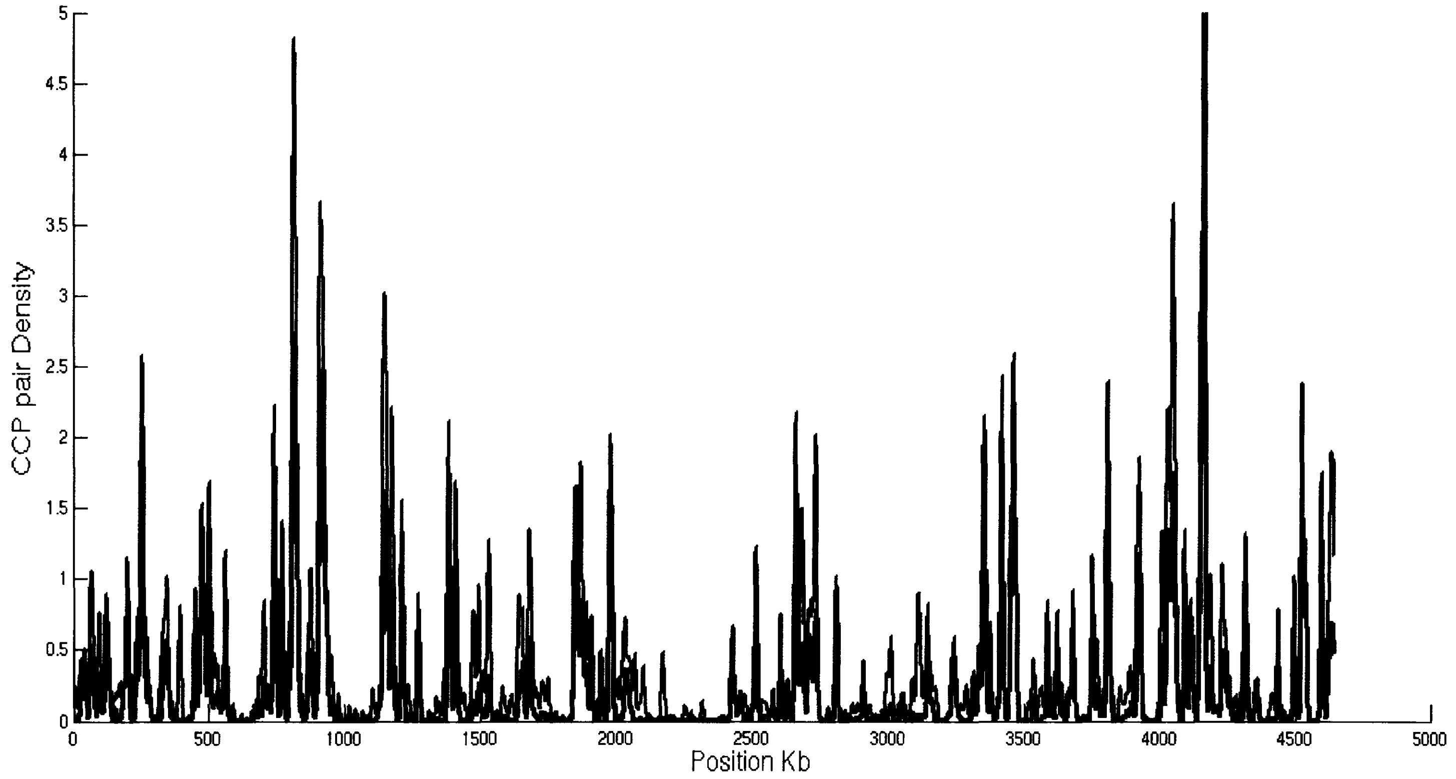
34. Itoh, M., Akutsu, T. & Kanehisa, M. Clustering of database sequences for fast homology search using upper bounds on alignment score. *Genome Inform Ser Workshop Genome Inform* **15**, 93-104 (2004).



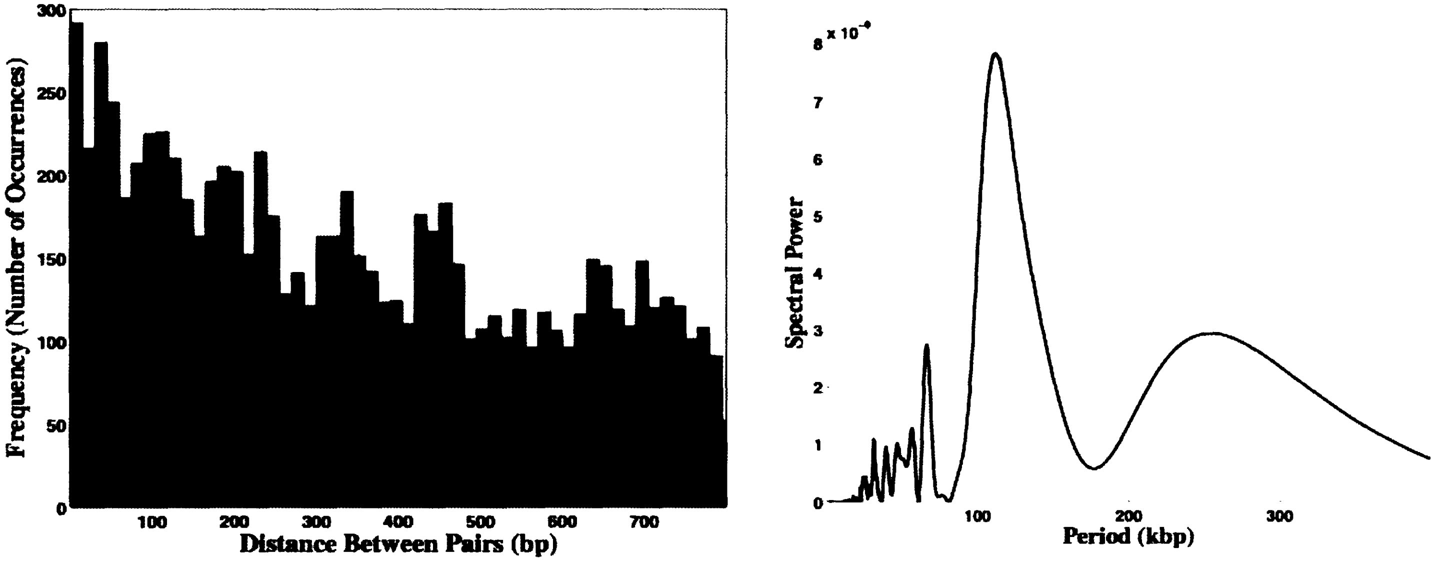
Supplemental Figure 3.S1 Characteristic signatures of structure from pairs of points that are close in space. a) A hypothetical chromosome fold with a single locus colored red and all of the loci on the opposite chromosome arm within a distance threshold colored in green. b) The positions of the red a green labeled loci as they would appear in the genome sequence. c) The distribution of symmetric across distances (as defined in Figure 3.1) between loci within the same spatial distance threhold as in a and b.



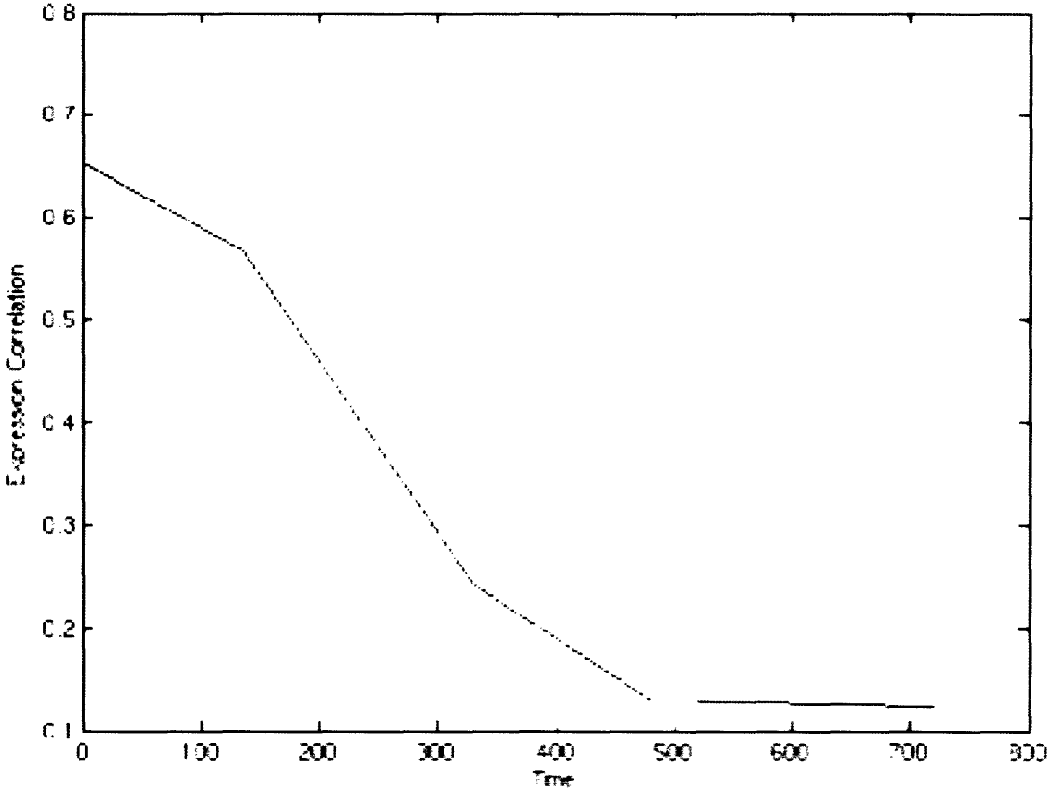
Supplementary Figure 3.S2 Random distribution of distance between pairs.



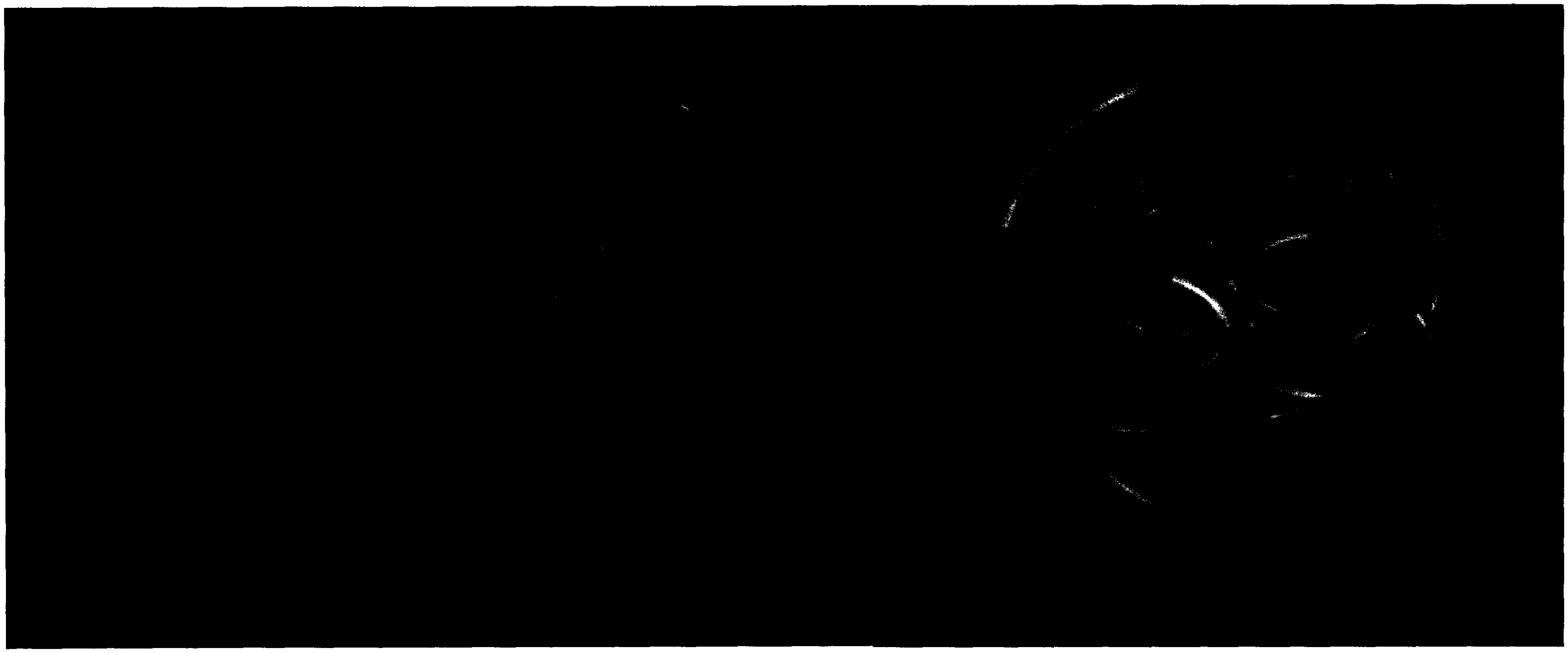
Supplemental Figure 3.S3 CCP pair position density distribution for across and adjacent pairs. Black line - position density for adjacent. Red line- position density for across



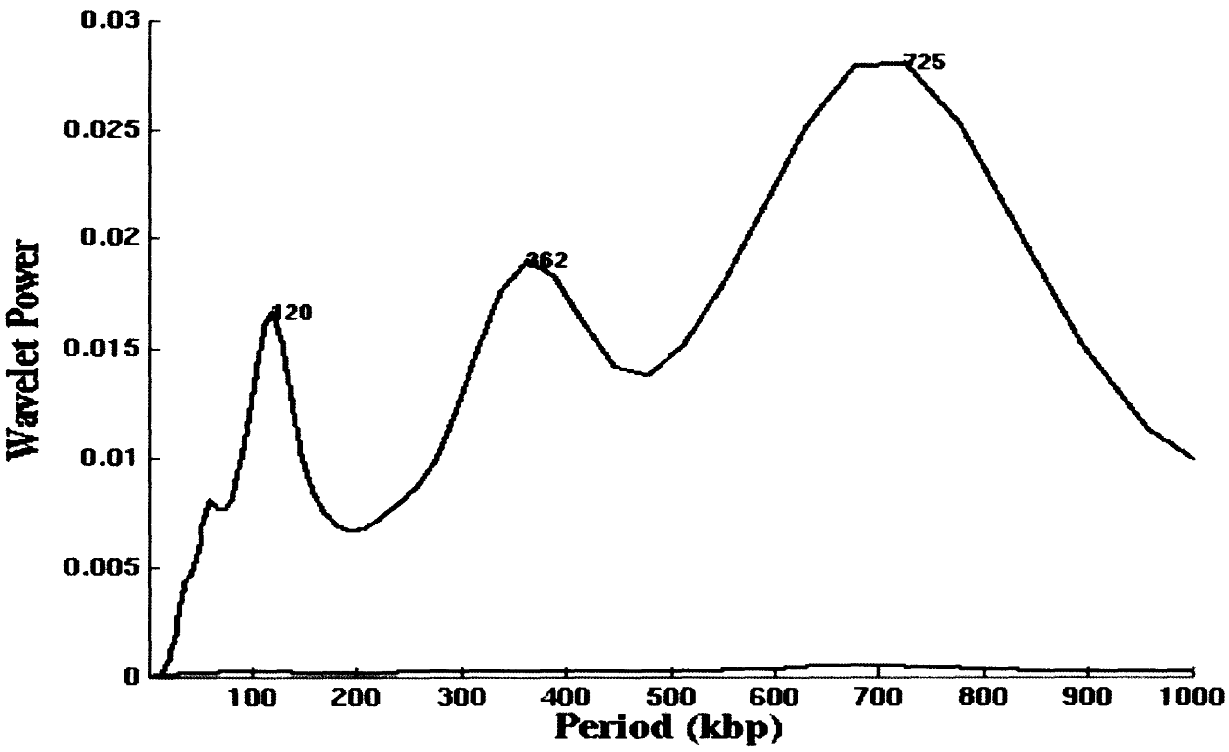
Supplemental Figure 3.S4 Caulobacter adjacent distribution and its Fourier Spectrum. Black lines - grid at 117Kb. Dashed line 117 Kb.



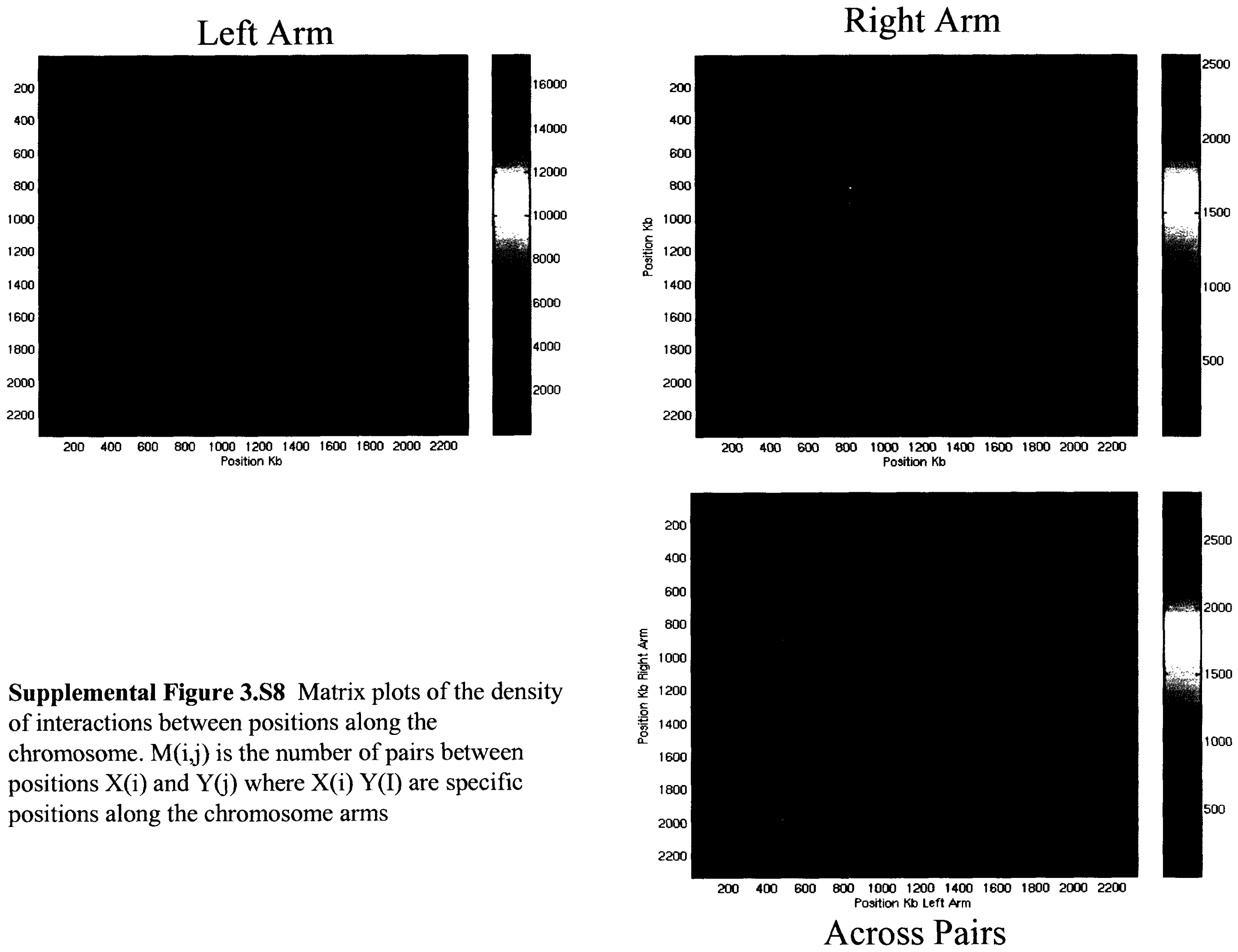
Supplemental Figure 3.S5 Correlation between density of pairs and expression decreases along the time course of *E. coli* cells going from log to stationary phase.



Supplemental Figure 3.S6 A helix with an additional planar supercoil at a frequency of twice the small loop period generating a double looped structure that places the high pair density region (red) along a central axis. Left side view. Right top view



Suuplementary Figure 3.S7 Sum of the wavelet power spectrum of the CCP pair position density



Supplemental Figure 3.S8 Matrix plots of the density of interactions between positions along the chromosome. M(i,j) is the number of pairs between positions X(i) and Y(j) where X(i) Y(I) are specific positions along the chromosome arms

Supplementary Material

Supplementary Methods

Wavelet Analysis

We calculated the 2D wavelet spectrum of the CCP pair position density using a morlet wavelet with wavenumber $\omega_\psi = 5$ according to the equation

$$\Psi_\sigma(t) = c_\sigma \pi^{-\frac{1}{4}} e^{-\frac{1}{2}t^2} \left(e^{i\sigma t} - \kappa_\sigma\right)$$

where

$$\kappa_\sigma = e^{-\frac{1}{2}\sigma^2}$$

and the normalization is

$$c_\sigma = \left(1 + e^{-\sigma^2} - 2e^{-\frac{3}{4}\sigma^2}\right)^{-\frac{1}{2}}$$

the relationship between the wavenumber and the scale is given by

$$(\omega_\Psi - \sigma)^2 - 1 = (\omega_\Psi^2 - 1)e^{-\sigma\omega_\Psi}$$

We summed the $|\psi_\sigma|^2$ across all positions t, for each scale, σ to get the total wavelet intensity independent of position at each scale.

Supplemental Table 1

Gene No.	Kb from Origin	No. of Pairs	Name
Peak Number 1			
3972	247	118	UDP-N-acetylenolpyruvoylglucosamine reductase
3973	248	34	biotin-[acetylCoA carboxylase] holoenzyme synthetase and biotin operon repressor
3981	252	28	preprotein translocase
3982	252	228	component in transcription antitermination
3983	253	120	50S ribosomal subunit protein L11
3984	253	144	50S ribosomal subunit protein L1, regulates synthesis of L1 and L11
3985	255	198	50S ribosomal subunit protein L10
3986	255	174	50S ribosomal subunit protein L7/L12
3987	258	32	RNA polymerase, beta subunit
3988	262	78	RNA polymerase, beta prime subunit
Peak Number 2			
4170	473	18	enzyme in methyl-directed mismatch repair
4171	474	150	delta(2)-isopentenylpyrophosphate tRNA-adenosine transferase
4172	475	10	host factor I for bacteriophage Q beta replication, a growth-related protein

4173	476	10	GTP - binding subunit of protease specific for phage lambda cII repressor
4174	477	4	protease specific for phage lambda cII repressor
4175	478	10	protease specific for phage lambda cII repressor
4177	480	38	adenylosuccinate synthetase
4179	482	18	ribonuclease R
4180	484	172	orf, hypothetical protein
Peak Number 4			
23	737	174	30S ribosomal subunit protein S20
25	738	198	flavokinase and FAD synthetase
26	740	70	isoleucine tRNA synthetase
27	741	186	prolipoprotein signal peptidase (SPase II)
29	743	88	IspH protein
31	745	120	dihydrodipicolinate reductase
32	746	10	carbamoyl-phosphate synthetase, glutamine (small) subunit
33	748	8	carbamoyl-phosphate synthase large subunit
Peak Number 5			
72	796	30	3-isopropylmalate isomerase (dehydratase) subunit
74	799	10	2-isopropylmalate synthase
77	802	8	acetolactate synthase III, valine sensitive, large subunit
78	804	6	acetolactate synthase III, valine sensitive, small subunit
81	806	10	orf, hypothetical protein
82	806	204	putative apolipoprotein
84	808	190	septum formation; penicillin-binding protein 3; peptidoglycan synthetase
85	810	204	meso-diaminopimelate-adding enzyme
86	811	174	D-alanine:D-alanine-adding enzyme

87	812	88	phospho-N-acetylmuramoyl-pentapeptide transferase?
88	814	228	UDP-N-acetylmuramoylalanine-D-glutamate ligase
89	815	112	cell division; membrane protein involved in shape determination
90	816	234	UDP-N-acetylglucosamine:N-acetylmuramyl- (pentapeptide) pyrophosphoryl-undecaprenol N-acetylglucosamine transferase
91	817	186	L-alanine adding enzyme, UDP-N-acetyl-muramate:alanine ligase
92	819	4	D-alanine-D-alanine ligase B, affects cell division
93	819	14	cell division protein; ingrowth of wall at septum
94	821	96	ATP-binding cell division protein, septation process, complexes with FtsZ, associated with junctions of inner and outer membranes
95	822	32	cell division; forms circumferential ring; tubulin-like GTP-binding protein and GTPase
96	823	32	UDP-3-O-acyl N-acetylglucosamine deacetylase; lipid A biosynthesis
98	826	164	preprotein translocase; secretion protein
103	829	20	putative DNA repair protein
Peak Number 6			
166	901	14	2,3,4,5-tetrahydropyridine-2-carboxylate N-succinyltransferase
167	903	2	protein PII; uridylyltransferase acts on regulator of glnA
168	905	134	methionine aminopeptidase
169	906	44	30S ribosomal subunit protein S2
170	907	158	protein chain elongation factor EF-Ts
171	908	200	uridylate kinase

172	909	216	ribosome releasing factor
173	910	28	2-C-methyl-D-erythritol 4-phosphate synthase; 1-deoxy-D-xylulose 5-phosphate reductoisomerase
174	911	76	undecaprenyl pyrophosphate synthetase (di-trans,poly-cis-decaprenylcistransferase)
175	912	126	CDP-diglyceride synthetase
176	913	202	putative protease
177	915	26	putative outer membrane antigen
179	917	30	UDP-3-O-(3-hydroxymyristoyl)-glucosamine N-acyltransferase
180	918	54	(3R)-hydroxymyristol acyl carrier protein dehydratase
181	919	24	UDP-N-acetylglucosamine acetyltransferase; lipid A biosynthesis
182	920	42	tetraacyldisaccharide-1-P; lipid A biosynthesis, penultimate step
183	921	54	RNAse HII, degrades RNA of DNA-RNA hybrids
184	923	234	DNA polymerase III, alpha subunit
185	925	74	acetylCoA carboxylase, carboxytransferase component, alpha subunit
188	929	236	tRNA(Ile)-lysidine synthetase
194	934	140	proline tRNA synthetase
197	936	12	D-methionine transport protein (ABC superfamily, peri_bind)
198	937	12	D- and L-methionine transport protein (ABC superfamily, membrane)
199	938	10	D- and L-methionine transport protein (ABC superfamily, atp_bind)
Peak Number 7			
400	1134	16	positive and negative sensor protein for pho regulon
403	1139	2	maltodextrin glucosidase

405	1141	160	synthesis of queuine in tRNA; probably S-adenosylmethionine:tRNA ribosyltransferase-isomerase
406	1142	86	tRNA-guanine transglycosylase
407	1143	184	orf, hypothetical protein
408	1144	106	protein secretion; membrane protein, part of the channel
409	1145	14	protein secretion, membrane protein
413	1148	110	orf, hypothetical protein
414	1149	86	diaminohydroxyphosphoribosylaminopyrimidine deaminase; 5-amino-6-(5-phosphoribosylamino)uracil reductase
415	1150	48	riboflavin synthase, beta chain
416	1150	274	transcription termination; L factor
417	1151	26	thiamin-monophosphate kinase
418	1152	4	phosphatidylglycerophosphatase
421	1156	98	geranyltranstransferase (farnesyl-diphosphate synthase)
422	1156	100	exonuclease VII, small subunit
423	1157	2	sulfur transfer protein
428	1162	16	protoheme IX farnesyltransferase (haeme O biosynthesis)
431	1165	2	cytochrome o ubiquinol oxidase subunit I
435	1170	2	possible regulator of murein genes
436	1171	238	trigger factor; a molecular chaperone involved in cell division
437	1172	54	ATP-dependent proteolytic subunit of clpA-clpP serine protease, heat shock protein F21.5
438	1173	136	ATP-dependent specificity component of clpP serine protease, chaperone
439	1175	8	DNA-binding, ATP-dependent protease La; heat shock K-protein
440	1177	14	DNA-binding protein HU-beta, NS1 (HU-1)

441	1178	64	peptidyl-prolyl cis-trans isomerase D
442	1179	4	orf, hypothetical protein
Peak Number 8			
633	1380	50	a minor lipoprotein
634	1381	62	rod shape-determining membrane protein; sensitivity to radiation and drugs
635	1382	40	cell elongation, e phase; peptidoglycan synthetase; penicillin-binding protein 2
636	1384	74	orf, hypothetical protein
637	1384	238	orf, hypothetical protein
639	1385	84	NAMN adenylyltransferase
640	1386	6	DNA polymerase III, delta subunit
642	1389	110	leucine tRNA synthetase
Peak Number 9			
1090	1863	62	glycerolphosphate auxotrophy in plsB background
1091	1864	38	3-oxoacyl-[acyl-carrier-protein] synthase III; acetylCoA ACP transacylase
1092	1865	34	malonyl-CoA-[acyl-carrier-protein] transacylase
1093	1866	2	3-oxoacyl-[acyl-carrier-protein] reductase
1094	1867	80	acyl carrier protein
1095	1868	2	3-oxoacyl-[acyl-carrier-protein] synthase II
1097	1870	98	putative thymidylate kinase (EC 2.7.4.9)
1098	1871	40	thymidylate kinase
1100	1872	188	orf, hypothetical protein
1103	1877	104	orf, hypothetical protein
1107	1880	10	beta-N-acetylglucosaminidase
Peak Number 10			
1203	1972	168	putative GTP-binding protein

1204	1973	314	peptidyl-tRNA hydrolase
1207	1977	124	phosphoribosylpyrophosphate synthetase
1208	1978	186	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase
1209	1978	2	outer-membrane lipoprotein
1210	1979	4	glutamyl-tRNA reductase
1211	1981	228	peptide chain release factor RF-1
1212	1982	130	possible protoporphyrinogen oxidase
1215	1984	46	2-dehydro-3-deoxyphosphooctulonate aldolase
Peak Number 11			
1714	2127	150	phenylalanine tRNA synthetase, alpha-subunit
1716	2126	84	50S ribosomal subunit protein L20, and regulator
1717	2126	68	50S ribosomal subunit protein A
1718	2125	102	protein chain initiation factor IF-3
1719	2124	126	threonine tRNA synthetase
Peak Number 12			
1857	1984	14	High-affinity zinc uptake system periplasmic protein
1858	1983	2	High-affinity zinc uptake system ATP-binding protein
1859	1982	52	High-affinity zinc uptake system membrane protein
1860	1981	158	Holliday junction helicase subunit A; branch migration; repair
1861	1980	182	Holliday junction helicase subunit B; branch migration; repair
1862	1979	0	orf, hypothetical protein
1863	1979	90	Holliday junction nuclease; resolution of structures; repair
1864	1978	74	orf, hypothetical protein
1866	1976	158	aspartate tRNA synthetase

Peak Number 13			
2507	1294	96	GMP synthetase (glutamine-hydrolyzing)
2508	1292	46	IMP dehydrogenase
2509	1291	148	exonuclease VII, large subunit
2511	1289	180	putative GTP-binding factor
2514	1286	130	histidine tRNA synthetase
2515	1284	96	1-hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate synthase
2516	1283	4	putative membrane protein
2517	1282	98	orf, hypothetical protein
2518	1281	52	nucleoside diphosphate kinase
Peak Number 14			
2559	1228	20	tRNA-specific adenosine deaminase
2562	1226	10	orf, hypothetical protein
2563	1225	68	CoA:apo-[acyl-carrier-protein] pantetheinephosphotransferase
2564	1224	20	pyridoxine biosynthesis
2565	1224	4	protein interacts with RecR and possibly RecF proteins
2566	1223	156	GTP-binding protein
2567	1222	174	RNase III, ds RNA
2568	1221	218	leader peptidase (signal peptidase I)
2569	1220	236	GTP-binding elongation factor, may be inner membrane protein
2573	1216	8	RNA polymerase, sigma-E factor; heat shock and oxidative stress
Peak Number 15			
2593	1191	86	orf, hypothetical protein
2594	1190	158	23S rRNA pseudouridine synthase
2595	1189	48	orf, hypothetical protein
2599	1187	36	chorismate mutase-P and prephenate dehydratase

2606	1181	196	50S ribosomal subunit protein L19
2607	1181	248	tRNA methyltransferase; tRNA (guanine-7-)-methyltransferase
2608	1180	154	16S rRNA processing protein
2609	1180	70	30S ribosomal subunit protein S16
2610	1179	98	GTP-binding export factor binds to signal sequence, GTP and RNA
2614	1175	72	phage lambda replication; host DNA synthesis; heat shock protein; protein repair
2615	1174	80	orf, hypothetical protein
2616	1173	142	protein used in recombination and DNA repair
2619	1171	12	orf, hypothetical protein
2620	1171	280	small protein B
Peak Number 16			
2942	838	124	methionine adenosyltransferase 1 (AdoMet synthetase); methyl and propylamine donor, corepressor of met genes
2946	834	246	orf, hypothetical protein
2947	833	8	glutathione synthetase
2948	833	12	orf, hypothetical protein
2949	832	178	orf, hypothetical protein
2950	831	4	putative protein transport
2951	830	26	orf, hypothetical protein
2952	830	22	conserved hypothetical integral membrane protein
2954	829	44	putative ribosomal protein
2955	828	158	putative oxidase
2960	823	82	orf, hypothetical protein
Peak Number 17			
3064	716	182	putative O-sialoglycoprotein endopeptidase
3065	715	50	30S ribosomal subunit protein S21
3066	714	246	DNA biosynthesis; DNA primase

3067	712	126	RNA polymerase, sigma(70) factor; regulation of proteins induced at high temperatures
3071	709	2	orf, hypothetical protein
Peak Number 18			
3162	619	2	inducible ATP-independent RNA helicase
3163	617	0	lipoprotein
3164	616	104	polynucleotide phosphorylase; cytidylate kinase activity
3165	614	180	30S ribosomal subunit protein S15
3166	613	44	tRNA pseudouridine 5S synthase
3167	613	114	ribosome-binding factor A
3168	611	82	protein chain initiation factor IF-2
3169	609	288	transcription pausing; L factor
3170	608	176	orf, hypothetical protein
3172	606	22	argininosuccinate synthetase
3173	605	0	putative alkaline phosphatase I
3175	603	28	protein export - membrane protein
3176	602	8	phosphoglucosamine mutase
3177	601	16	7,8-dihydropteroate synthase
3178	600	72	degrades sigma32, integral membrane peptidase, cell division protein
3179	598	12	23 S rRNA methyltransferase
3180	598	6	orf, hypothetical protein
3181	597	208	transcription elongation factor: cleaves 3 nucleotide of paused mRNA
3182	596	0	D-alanyl-D-alanine carboxypeptidase, fraction B; penicillin-binding protein 4
3183	595	186	putative GTP-binding factor
3184	593	0	orf, hypothetical protein
3185	593	222	50S ribosomal subunit protein L27
3186	592	264	50S ribosomal subunit protein L21
3187	592	6	octaprenyl-diphosphate synthase

3188	591	0	regulatory factor of maltose metabolism; similar to Ner repressor protein of phage Mu
3189	590	202	first step in murein biosynthesis;UDP-N-glucosamine 1-carboxyvinyltransferase
3190	589	0	orf, hypothetical protein
3191	589	0	orf, hypothetical protein
3192	588	2	orf, hypothetical protein
3193	588	4	orf, hypothetical protein
3194	587	28	orf, hypothetical protein
3195	586	16	putative ATP-binding component of a transport system
3196	585	0	orf, hypothetical protein
3197	584	6	putative isomerase
3198	583	0	orf, hypothetical protein
3199	583	0	orf, hypothetical protein
3200	582	2	orf, hypothetical protein
3201	581	12	putative ATP-binding component of a transport system
3202	580	0	RNA polymerase, sigma(54 or 60) factor; nitrogen and fermentation regulation
3203	579	6	probable sigma-54 modulation protein
3204	579	0	phosphotransferase system enzyme IIA, regulates N metabolism
3205	578	22	orf, hypothetical protein
Peak Number 19			
3279	496	0	putative transferase
3280	496	0	orf, hypothetical protein
3281	495	0	dehydroshikimate reductase
3282	495	14	orf, hypothetical protein
3283	494	0	putative DNA topoisomerase
3284	494	0	orf, hypothetical protein
3287	492	152	peptide deformylase
3288	491	188	10-formyltetrahydrofolate:L-methionyl-tRNA(fMet) N-formyltransferase

3289	490	22	16S rRNA m5C967 methyltransferase
3290	489	0	transport of potassium
3291	488	0	mechanosensitive channel
3292	487	8	Zn(II)-responsive transcriptional regulator
3293	486	0	orf, hypothetical protein
3294	486	134	50S ribosomal subunit protein L17
3295	485	142	RNA polymerase, alpha subunit
3296	484	168	30S ribosomal subunit protein S4
3297	484	64	30S ribosomal subunit protein S11
3298	483	64	30S ribosomal subunit protein S13
3299	483	2	50S ribosomal subunit protein L36
3300	482	112	putative ATPase subunit of translocase
3301	481	198	50S ribosomal subunit protein L15
3302	481	70	50S ribosomal subunit protein L30
3303	481	94	30S ribosomal subunit protein S5
3304	480	166	50S ribosomal subunit protein L18
3305	480	160	50S ribosomal subunit protein L6
3306	479	96	30S ribosomal subunit protein S8, and regulator
3307	479	76	30S ribosomal subunit protein S14
3308	479	88	50S ribosomal subunit protein L5
3309	478	174	50S ribosomal subunit protein L24
3310	478	80	50S ribosomal subunit protein L14
3311	477	92	30S ribosomal subunit protein S17
3312	477	52	50S ribosomal subunit protein L29
3313	477	80	50S ribosomal subunit protein L16
3314	476	76	30S ribosomal subunit protein S3

3315	476	126	50S ribosomal subunit protein L22
3316	475	54	30S ribosomal subunit protein S19
3317	475	82	50S ribosomal subunit protein L2
3318	474	60	50S ribosomal subunit protein L23
3319	474	88	50S ribosomal subunit protein L4, regulates expression of S10 operon
3320	473	96	50S ribosomal subunit protein L3
3321	473	46	30S ribosomal subunit protein S10
3322	472	0	calcium-binding protein required for initiation of chromosome replication
3323	471	0	putative export protein A for general secretion pathway (GSP)
3324	470	0	putative export protein C for general secretion pathway (GSP)
3325	468	0	putative export protein D for general secretion pathway (GSP)
3326	467	2	putative export protein E for general secretion pathway (GSP); Type II traffic warden ATPase
3327	465	0	putative export protein F for general secretion pathway (GSP)
3328	465	0	putative export protein G for general secretion pathway (GSP); pilin-like
3329	464	0	putative export protein H for general secretion pathway (GSP)
3330	464	0	putative export protein I for general secretion pathway (GSP)
3331	463	0	putative export protein J for general secretion pathway (GSP)
3332	462	0	putative export protein K for general secretion pathway (GSP)
3333	461	0	putative export protein L for general secretion pathway (GSP)
3334	460	0	putative export protein M for general secretion pathway (GSP)

3335	460	0	bifunctional prepilin peptidase: leader peptidase; N-methyltransferase; part of general secretion pathway (GSP)
3336	459	0	bacterioferrin, an iron storage homoprotein
3337	459	0	regulatory or redox component complexing with Bfr, in iron storage and mobility
4473	493	0	orf, hypothetical protein
Peak Number 20			
3635	115	30	formamidopyrimidine DNA glycosylase
3636	114	52	50S ribosomal subunit protein L33
3637	114	100	50S ribosomal subunit protein L28
3638	114	26	DNA repair protein
3639	112	32	flavoprotein affecting synthesis of DNA and pantothenate metabolism
3640	112	18	deoxyuridinetriphosphatase
3641	111	0	putative transcriptional regulator
3642	110	4	orotate phosphoribosyltransferase
3643	110	0	RNase PH
3644	109	14	putative alpha helix protein
3645	108	0	DNA-damage-inducible protein
3646	107	0	orf, hypothetical protein
3647	105	0	putative enzyme
3648	104	142	guanylate kinase
3649	104	20	RNA polymerase, omega subunit
3650	102	90	(p)ppGpp synthetase II; also guanosine-3,5-bis pyrophosphate 3-pyrophosphohydrolase
3651	101	0	tRNA (Guanosine-2-O-)-methyltransferase
Peak Number 21			
3729	13	16	L-glutamine:D-fructose-6-phosphate aminotransferase

3730	11	96	N-acetyl glucosamine-1-phosphate uridyltransferase
3731	10	36	membrane-bound ATP synthase, F1 sector, epsilon-subunit
3732	9	50	membrane-bound ATP synthase, F1 sector, beta-subunit
3733	8	50	membrane-bound ATP synthase, F1 sector, gamma-subunit
3734	7	82	membrane-bound ATP synthase, F1 sector, alpha-subunit
3735	6	18	membrane-bound ATP synthase, F1 sector, delta-subunit
3736	5	38	membrane-bound ATP synthase, F0 sector, subunit b
3737	5	2	membrane-bound ATP synthase, F0 sector, subunit c
3738	4	66	membrane-bound ATP synthase, F0 sector, subunit a
3739	3	0	membrane-bound ATP synthase, dispensable protein, affects expression of atpB
3740	2	82	glucose-inhibited division; chromosome replication?
3741	1	42	glucose-inhibited division; chromosome replication?

Chapter 4

Conclusions and future directions

It now seems clear that structure, in biology once the province of x-ray crystallographers and electron microscopists is gradually entering the mainstream where it promises to yield a physical, not an abstract picture of cellular processes. It is likely that although three-dimensional space is now thought of as a confounding factor - an extra set of variables taking simple ordinary differential equations into the realm of less tractable partial differential equations, or an extra level of complexity in an already dizzying web of protein interactions, that it will instead prove to be a unifying and simplifying factor; the spatial integration of protein machines and functional modules is likely used by the cell, allowing it to control cross-talk between components or precisely control the positions at which cellular events take place. Thus an understanding of spatial organization may help us to rise in a natural way from the level of individual components to higher levels of organization. "From protein words, to the sentences and paragraphs of biology" [1].

It is interesting to note the history of structure in biology, starting from chemistry where the knowledge of the specific configuration of atoms, bond lengths, and angles gives an ability to predict chemical properties and reactivity. These structural properties prove often to be as important as the atoms themselves. We learn in organic chemistry that certain nucleophilic substitution reactions will occur by unimolecular as opposed to bimolecular reaction dynamics based on the steric hindrance of atoms bonded to the electrophile. And it is the specific configurations as well as the atoms in enzyme catalytic sites that allow them to lower the energy barrier of reactions. Indeed, it is in chemistry that we first learn of both the simplifying and the unifying power of understanding structure; the full quantum mechanical description of even simple organic molecules is not analytically solvable. Yet the simplifying properties of their structures along with simplified electronic properties make their behavior predictable.

In biology the investigation of structure has moved in two directions: up from simple molecules, from the first protein structure, myoglobin (~20KD)[2], solved in 1957 to more recent structures of massive machines like the ribosome (2500 KD) in 2000[3]; and down in microscopy from the resolution of simple cells, to subcellular organelles and to the <2nm resolution of 2-D electron crystallography [4]. Thus these two approaches to structure are converging such that they almost merge, for instance in the docking of the atomic structure of tubulin into electron density plots of electron tomograms. Even, three-dimensional reconstructions of entire bacterial cells do not seem so distant a dream.

As these processes converge, and as structure becomes more heavily integrated in biology it will be of great importance for structural biologists, grounded primarily in the atomic level structure of proteins, and molecular biologists and geneticists grounded in genes and genome sequences to learn each other's language. It is worth noting the role that theory may play. It was the theory of quantum mechanics that finally elucidated the rules behind chemical structures. Linus Pauling famously derived the alpha helix and beta sheet using paper models. Watson and Crick solved the structure of DNA building models from a simple x-ray diffraction pattern. And the physical theories of diffraction and the mathematics of Fourier transforms give us the atomic coordinates from crystal structures. It is likely that theory will play an important role in the generation of and the understanding of new structural data.

This thesis has centered around a set of theoretical approaches for unraveling one particular structure in the cell, the chromosome, and specifically, the bacterial chromosome. Chromosome structure both bacterial and eukaryotic represents an important link between the abstract logical representations of biological data as information and the physical structural ones. In chapter 2, I describe an extension of the framework of constraint based optimality, successful in other areas of systems biology, to the study of chromosome structure.

In chapter 3, I describe the first comparative genomics method of searching for actual constraints by finding patterns likely related to three-dimensional folding in genome sequences. I show that the pair-wise constraints from this search recapitulate known structural details about the chromosome and are strikingly correlated with genome-wide expression data suggesting that they are selected for proximity to the transcription machinery. Moreover, I show that specific periodicities in the positions of these pairs yield new insight into the likely fold of the chromosome

There are many extensions of the work in both chapters. While current experimental technologies are not of high enough resolution to see chromosome three-dimensional structure at the level of individual genes, signals of structure within the genome sequence may provide important insights. Even after high-resolution structures are available comparison between genomes will provide knowledge or at least suggestions of functional roles. In particular, if it can be rigorously shown that there is evolutionary selection for spatial relationships between points, this will provide strong evidence of functional links.

There are potentially immediate extensions of the methods of chapter 3 to eukaryotes. Since it is transcription that appears to be driving the signals we uncover and there is already ample evidence of transcriptional organization of the nucleus, we may expect to find similar signals in eukaryotic chromosomes. However, because eukaryotic nuclear organization seems flexible and perhaps even probabilistic in nature it will be important to understand what exactly these genomic signals mean before they can be applied for prediction [5]. Also, such pair-wise patterns are only one potential means of looking for structure in genome sequences. Other methods or more sophisticated adaptations of the methods we apply may generate cleaner structural signals.

The more theoretical methods of chapter 2, interestingly, are more applicable to concrete experimental data. It will, of course, be experiment that validates any of the predictions made in chapter 3 and experiments are the most obvious means to generate the large set of constraints necessary to solve the full structure of the chromosome even at a moderate level of resolution of 10kb (~8000 distance constraints for *C. crescentus*). I outline below two promising experimental methodologies to generate a large number of constraints for structure determination.

The first technology is an extension of the cross-linking methodology known as chromosome conformation capture or 3C [6]. The basic technology consists of measuring the frequency with which two chromosomal loci cross-link in vivo. This frequency is inversely related to the spatial distance between loci in the cell. Thus by measuring the frequency of pairing between all pairs of n loci, one can completely fill the matrix of distances between them and solve the structure.

Currently the method involves cross-linking, cutting the genome at known sites using a restriction enzyme, and joining the ends of the cross-linked material by ligation under dilute conditions where only cross-linked fragments should join. The sequences of the junctions between fragments are then quantitatively measured using polymerase chain reaction (PCR). PCR limits the number of pairs of points that can be examined, restricting an experiment to around 100-200 pairs of points or all pairs of 10-15 loci.

There are other means of measuring the amount of individual junctions which should require minimal technological innovation, and by virtue of their massive parallelism allow 400,000 to 1,000,000s of junctions to be measured (all pairs of 1000 to 2000 loci.) This would be over an order of magnitude above the resolution of fluorescence microscopy and give the relative positions of all points simultaneously where fluorescence microscopy can only follow a few..

The first technology is the custom microarray which can now contain over 400,000 features. Junctions could be labeled and bound, yielding in a single experiment the amounts of all 400,000 junctions. In an interesting parallel many of the considerations and algorithms described in Appendix 2 for the design of oligonucleotide probes for transcriptional analysis could be used to design oligonucleotide probes for 3C junctions. Some technical hurdles remain, such as the removal of fragments that ligate to themselves through intramolecular ligation which make up most of material after the ligation reaction and the 3C method has also not yet been optimized for bacteria. Both of these obstacles however, should be minor.

The second and perhaps even more promising technology for high-resolution 3C is polony sequencing.[7].Using massively parallel sequencing one can directly sequence millions of junctions, digitally counting the number of each. With current polony technology this would generate 1,000,000 or more reads from a single reaction. The dynamic range would be even broader than that of the microarray and the method could potentially count all junctions in an unbiased way, without having to select a subset to synthesize on an array or to amplify with PCR. The same technical hurdles as described for array 3C apply. Additionally, the increased dynamic range means that any highly abundant species could comprise most of the digital reads. However, such species could be selectively removed and the range of signal would give the advantage of even finer resolution on the distances.

Despite its many advantages 3C makes a measurement of chromosome structure in a population of cells. With cryoelectron tomography it is now possible to obtain three-dimensional high-resolution images of individual cells in almost their native state without sectioning. The technology is based on the ability to align images taken with a transmission electron microscope over a set of different tilt angles of the object. The two-dimensional diffraction pattern at each tilt makes up one slice of the complete three-dimensional Fourier transform and with enough tilts the original three-dimensional image can be reconstructed. Because the information in many of the tilts is partially redundant, each image can be taken at a relatively low dose of electrons which prevents damage to the structures, but also has the consequence of high level of noise. By averaging over all of the aligned images the noise can be removed and a full three-dimensional structure at < 6nm resolution can be obtained.

In order to distinguish the chromosome in an electron microscopic image the DNA must be labeled with an electron dense material so that it is highly diffractive to incident electrons. To track both the structure and the sequence simultaneously one would like both a distinguishable nonspecific DNA label and a sequence specific label. This would allow the curve of the chromosome to be tracked and specific sequences pinpointed in the same image. One method for sequence specifically labeling DNA is to use small ~1 nm gold clusters, by coupling them to labeled nucleotides. One can incorporate these labeled clusters at specific sequences, and introduce sequence specific numbers of clusters at each of these incorporation sites. The most difficult hurdle is how to introduce any such label without grossly altering the structure of the cell or the chromosome. Various means might provide a solution including cross-linking before labeling, or performing the entire process in polyacrylamide gels which could capture the chromosome structure in a three-dimensional net of polymer. Though the hurdles here are more significant than for 3C, it is likely that cryoelectron tomography may provide the most detailed views of individual chromosomes in the near future. In fact, labeling of individual loci using the FLASH technique has already been performed (Shapiro personal communication).

The combined set of cross-linking and electron microscopic technologies promise to deliver high resolution structures of the chromosome in the near future. I emphasize that ultimately the genome sequence and its structure may be viewed as one entity, in the same way that the actual coordinates of amino acids on a protein fold yield an understanding of catalysis or binding, the positions of genes and regulatory regions, heterochromatin, and supercoils in space will tell us about factories, and assembly and regulation in the cell. Indeed it seems that the structure of an entire cell may not be so far off. Then, the relation of the information stored in the genome and its realization in proteins and membranes and RNA will be encompassed in an integrated, physical whole.

References

1. Sali, A., et al., *From words to literature in structural proteomics.* Nature, 2003. **422**(6928): p. 216-25.
2. Kendrew, J.C., et al., *A three-dimensional model of the myoglobin molecule obtained by x-ray analysis.* Nature, 1958. **181**(4610): p. 662-6.
3. Ban, N., et al., *The complete atomic structure of the large ribosomal subunit at 2.4 A resolution.* Science, 2000. **289**(5481): p. 905-20.
4. Baumeister, W., *From proteomic inventory to architecture.* FEBS Lett, 2005. **579**(4): p. 933-7.
5. Misteli, T., *Concepts in nuclear architecture.* Bioessays, 2005. **27**(5): p. 477-87.
6. Dekker, J., et al., *Capturing chromosome conformation.* Science, 2002. **295**(5558): p. 1306-11.
7. Shendure, J., et al., *Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome.* Science, 2005.

Appendix 1

On the complete determination of biological systems

(previously published as Selinger DW, Wright MA, and Church GM
Trends in Biotechnology 2003 Jun;21(6):251-4)

Abstract

The nascent field of systems biology ambitiously proposes to integrate information from large-scale biology projects to create computational models which are, in some sense, complete. However, the details of what would constitute a complete systems level description of an organism are far from clear. To provide a framework for this difficult question, it is useful to define a model as a set of rules that maps a set of inputs, e.g. descriptions of the cell's environment, to a set of outputs, e.g. the concentrations of all its RNAs, proteins, etc. We show how the properties of a model affect the required experimental sampling and we estimate the number of experiments needed to "complete" a particular model. Based on these estimates, we suggest that the complete determination of a biological system is a concrete, achievable goal.

Scientific investigation has long been a technology-limited endeavor: from Aristotle's passive observations, to Galileo's experimental probings, to our own elaborately contrived and controlled micro-dissections of nature. New technologies, in the form of systematic, quantitative, large-scale experiments with machine-readable outputs have recently unleashed a torrent of data onto the biological community, resulting in abundant speculation about the future of post genomic biology.

With new tools, naturally come new goals. Classical molecular methods forced us to focus our gaze on small numbers of molecules at a time, so we laboriously built up descriptions in human language, pictures, and the occasional video clip. The overarching goal of biology, if there was one, was to compile a large number of systems that are interesting (those that define a general rule, break one, or appeal to us as idiosyncratic human beings) or applicable (those that contribute to the engineering, reverse-engineering, or modification of a system). The defining feature of this "compilation strategy" is that it is more a process than a goal. It specifies no endpoint other than continual accumulation.

Completion in biology

Long the goal of physicists searching for a "theory of everything", completion has now become a pervasive idea in biology, raising the question of where it rightfully applies and whether it constitutes a new sort of goal for biological inquiry. The proliferation of the "-ome" suffix attests to widespread acceptance that biology is rife with things to be completed, whether it's the genome, the transcriptome, or the proteome. Already genome projects and large scale experiments have yielded important advances in medicine, biotechnology and basic research. Systems level descriptions promise predictions for cell, organ, and organismal behavior.

There seem to be two distinct levels of completion. The first, and conceptually simpler of the two, is "parts list completion." Completion at this level is defined as the fraction of observed to total predicted parts. This is well underway, and consists of the various "ome" projects. The second, more ambitious and less well-defined level of completion, is at the level of "systems biology", the study of how the parts work together to form a functioning biological system[1,2]. There is no clear correspondence between these two levels of completion. A nearly complete parts list could lead to an inaccurate description of the system if the missing parts were crucial for system function. For example, a model may be wildly inaccurate due to the omission of a single essential gene.

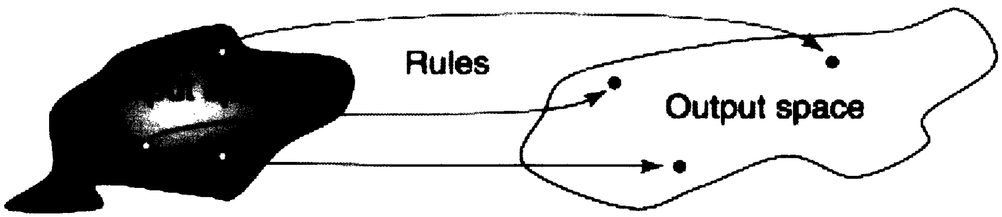
But how can we know when a systems level description is complete? Whereas crystallographers can state an R_{free} to describe the extent of agreement between a structural model and the data from which it was derived, biologists still lack a coherent framework for deciding the extent to which a model of a biological system is consistent with experimental data. Such a framework would be useful for setting systems biology goals, assessing progress, and identifying areas in need of further investigation.

A model for modeling

We can think of a systems level description as a formal mathematical construct, or model. Thus, consideration of the properties of a model is necessary to understand in what sense one might be considered complete. A model can be defined as a set of rules which maps a set of inputs (Fig. 1, blue area), e.g. descriptions of the cell's environment, to a set of outputs (Fig. 1, yellow area), e.g. the concentration of all of its RNAs. Large-scale experimental sampling of input-output pairs (Fig. 1, yellow-red dots), such as condition-transcriptome pairs, may be used to derive these rules[3].

In order to specify a particular model we must make certain decisions about its basic properties (Table 1). Firstly, we must decide on the inputs and outputs. This choice will depend, for example, on whether we are interested in predicting transcriptomes from temperature and pH, or in predicting successive molecular states. Secondly, we must decide on the range of values the inputs can assume. Finally, we must decide on what level of accuracy and precision we require in our predictions. If we are predicting relative RNA levels, do we need predictions such as, "upregulation by a factor of 3.3 ±0.1" or would a predicted factor of 3 ±1 allow us to reach the same biological conclusions? Once we have made these three decisions, we must choose a rule type that will allow us to realize the model, i.e. one that will allow us to map our chosen input space to our chosen output space with the desired level of accuracy and precision.

From these basic model properties we can determine how many measurements, at least to the order of magnitude, it would take to populate the space of all possible inputs



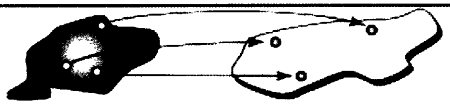
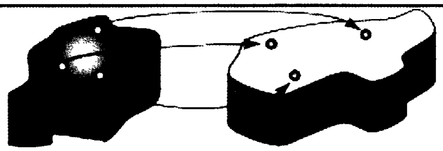
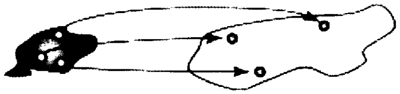
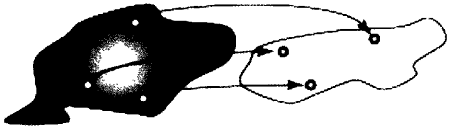
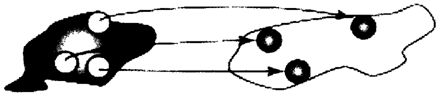
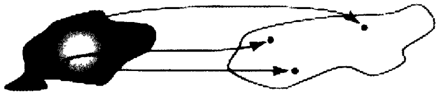
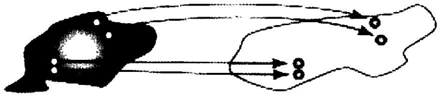
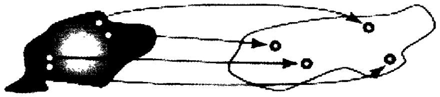
Input		Output
External environment	→	Transcriptome, proteome, etc.
External environment	→	Molecular fluxes, growthrate
Cell type	→	Transcriptome, proteome, etc.
Cell state at time *t*	→	Cell state at time *t* + 1

Figure 1. A general schema for modeling as an exercise in mapping input space (blue area), e.g. all possible environments in which a cell can live, to output space (yellow area), e.g. all possible cellular responses. The yellow-red dot pairs represent measured input-output pairs, which, in large numbers, can be used to derive rules (arrows) to predict outputs for novel inputs. Examples of possible inputs and outputs are given below.

(e.g. conditions) with enough measured outputs (e.g. transcriptomes, proteomes, etc.) to make prediction feasible. In other words, how many measurements are needed to adequately sample input space to allow the rule parameters to be determined. A similar issue has been addressed in the field of supervised learning by the theory PAC (Probably Approximately Correct)[4] which gives the probability that a given number of measurements will generate rules of a given accuracy. Of course, once a model has been generated, its accuracy must then be verified by additional experiments that were not used to infer the rules.

We can readily determine how the properties of a model affect the number of measurements required to derive its rules (Table 1). A larger number of inputs and outputs requires more individual measurements per input-output pair, i.e. sample (Row I). A larger range of input values may require a greater number of samples (Row II). A higher desired accuracy generally will require more samples and increased accuracy for the individual measurements (Row III). Finally, a more complex rule type will likely require more samples (Row IV). If nearby points in input space do not map to nearby points in output space then we must sample the space more densely. It should be noted that, because the output space is simply a function of the input space, we can focus exclusively on the properties of the input space and rule type when considering the required experimental sampling density.

Table 1.

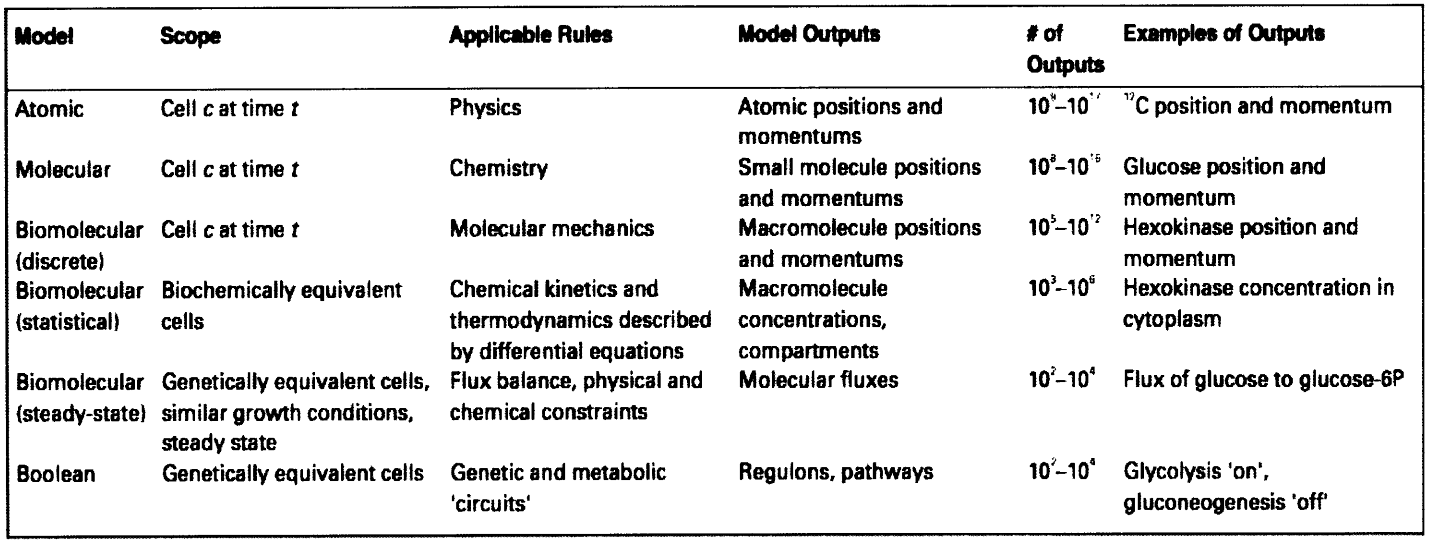
Property	Minimizing (*maximizing*) case	Minimizing case	Maximizing case
I. Number of inputs and outputs	Low (*high*) level of model detail, less (*more*) comprehensive model		
II. Range of inputs	Can live in few (*many*) environments		
III. Accuracy and precision	Predictions useful at low (*only at high*) level of accuracy		
IV. Rule type	Similar inputs give similar (*different*) outputs, requires simple (*complex*) rule types.		

Model Types

The choice of a model type is a critical part of any completion effort as it determines the type of rules which need to be discovered and the number and type of measurements which need to be made. There are a host of issues, discussed in several recent reviews[5-8], which must be considered when planning a modeling strategy. Table 2 gives examples of model types organized by the level of detail of their predictions i.e. outputs. On one end of the spectrum, we can imagine atomic level, or even subatomic level descriptions of a complete cell. While large-scale measurements at this level are not forthcoming in the foreseeable future these model types set an upper bound on detail. Towards the lower end of the detail spectrum we have boolean models, which we can build from logical statements such as, "if the *lac* repressor is bound to the operator then the *lac* operon is off."

As we move from more to less detailed models we make certain trade-offs. The more detailed models make fewer assumptions, and are therefore potentially more accurate for the systems they describe. On the other hand, they tend to be more problematic with regard to computability and measurement, and are therefore difficult to apply to large systems. Furthermore, computational predictions at too high a level of detail may not be useful for human understanding of the biological phenomena under study. As we enhance our ability to make large numbers of measurements, we may be able to generate enough input-output pairs to allow the complete determination of more and more detailed model types.

Table 2.



Model	Scope	Applicable Rules	Model Outputs	# of Outputs	Examples of Outputs
Atomic	Cell *c* at time *t*	Physics	Atomic positions and momentums	10^{9}–10^{17}	^{12}C position and momentum
Molecular	Cell *c* at time *t*	Chemistry	Small molecule positions and momentums	10^{8}–10^{15}	Glucose position and momentum
Biomolecular (discrete)	Cell *c* at time *t*	Molecular mechanics	Macromolecule positions and momentums	10^{5}–10^{12}	Hexokinase position and momentum
Biomolecular (statistical)	Biochemically equivalent cells	Chemical kinetics and thermodynamics described by differential equations	Macromolecule concentrations, compartments	10^{3}–10^{6}	Hexokinase concentration in cytoplasm
Biomolecular (steady-state)	Genetically equivalent cells, similar growth conditions, steady state	Flux balance, physical and chemical constraints	Molecular fluxes	10^{2}–10^{4}	Flux of glucose to glucose-6P
Boolean	Genetically equivalent cells	Genetic and metabolic 'circuits'	Regulons, pathways	10^{2}–10^{4}	Glycolysis 'on', gluconeogenesis 'off'

Table 2. Examples of hypothetical levels at which a systems biology project can be completed, listed from most detailed (top) to least (bottom). The number of outputs is estimated for single cells. We may soon be able to collect complete datasets for some classes of biomolecules at the level of macromolecular concentrations. Several recent reviews discuss biological systems modeling in more detail5-8. The details of these calculations can be found at http://arep.med.harvard.edu/completion.

Practical Application

Now let's consider specific examples of projects we might wish to complete. A useful model for many biological purposes is one in which the resulting expression level of each gene can be predicted using the input levels of all of the genes. Such a model would predict the effects of overexpression, genetic knockouts, or even various environmental stimuli, provided that the effects of those stimuli on individual genes are known. In fact, historically, much of genetic research has been devoted to finding small parts of such a model. Specifically, we consider a discrete transcriptional network model which maps all N genes as inputs to all N genes as outputs, where the genes can take on three levels of expression (low, medium, and high) and each gene has at most K direct regulators (Table 3). We consider this model for organisms that span a wide range of complexity: *Mycoplasma pneumoniae*, *Escherichia coli*, and *Homo sapiens*.

A very simple cell, like *M. pneumoniae*, is an example of a cell with a minimal number of genes (low N) which also seems to lack any transcriptional regulation (low K) and lives in an exquisitely controlled environment within its human host [9]. At an intermediate level is *E. coli* which can live in many environments and consequently has more genes and requires more genetic regulation (intermediate N and K). At the upper extreme are humans which have a large number of genes and highly complex regulatory mechanisms. Additionally, as multicellular organisms, humans have abundant intercellular communication and a large number of cell types, each potentially with its own set of transcriptional states.

Table 3.

Organism	N	K	Estimated number of microarrays	
			Lower bound	Upper bound
M. pneumoniae	688	1	10	80
E. coli	4,288	3	50	40 000
H. sapiens	50 000	4	100	700 000

Table 3. Upper and lower bounds on the number of measurements to complete discrete transcriptional network models for various organisms, calculated according to Krupa10. N represents the number of nodes (genes in this example). K represents the maximum number of regulatory connections per node. Measurements are assumed to be microarrays where the expression level of each gene is categorized as high, medium, or low (ξ = 3). The lower bound (information-theoretic) is given by . The upper bound is given by , where the measurements fail to determine the model with probability 1/C. Here we set 1/C equal to 0.01. It is important to note that the upper bound estimate increases exponentially with K, making it the dominant parameter.

In Table 3 we use formulae given by Krupa[10] to estimate the upper and lower bounds for the number of microarray experiments needed to complete the discrete transcriptional network model described above. The lower bound is related to the amount of information needed to specify the network structure and mapping functions. The lower bound assumes that each microarray experiment is maximally informative and independent from previous measurements. It also assumes perfectly efficient use of experimental measurements to determine model parameters. These assumptions make it likely that this estimate is far below the actual number of measurements needed. The upper bound reflects the number of random experiments needed to complete the model with a 99% probability of success, and is probably a more realistic estimate. It is important to note that the upper bound estimate grows rapidly (exponentially) with the maximum number of regulatory connections (K) per gene. Fortunately, genes tend to have a low number of regulatory inputs, making an estimate based on a low K reasonable. It is also encouraging to note that the upper bound estimate grows very slowly (logarithmically) with the number of genes (N), making determination feasible even for very large genetic networks.

The upper bound of 80 experiments for *M. pneumoniae* is already feasible with current technology. While 40,000 microarrays for *E. coli* and 700,000 for human may seem daunting, we should keep in mind that the initial version of the human genome required approximately 30-40 million sequencing reads[11,12], a number that was not practical with the technology available when the project was first proposed.

Other methods have been described for inferring rules directly from large-scale datasets and for estimating the number of measurements necessary for a given level of accuracy[3,13,14]. Additionally, current microbial models based on flux balance analysis have shown considerable progress towards a complete description of metabolism, with mappings from culture conditions and genotype (input) to growth phenotype (output) that reach accuracies greater than 90% (106/116)[15]. Models of this type have even been shown to be predictive of the biological evolution of metabolic fluxes[16]. Additional refinements promise to further increase their accuracy[17].

Conclusion

With the advent of large-scale projects, synthesis has become an important goal in biology. Completion of a large number of genome projects, and the pursued completion of other "ome" projects, raises the question of what it might mean to complete a systems biology project and what might be gained from such an effort. We have found it useful to consider this question within a framework for modeling and show how the number of experiments necessary can be related to the model properties. Furthermore, we present an example of a discrete transcriptional network model and estimate the number of experiments necessary for its completion. When viewed through the framework of modeling, the complete determination of a biological system becomes a concrete, achievable goal.

Acknowledgements

We thank Patrik D'haeseleer, Daniel Segrè, Jeremy Edwards, and many members of the Church lab for helpful discussions.

References

1 Kitano, H. (2001) Systems Biology: Toward System-level Understanding of Biological Systems. In *Foundations of Systems Biology* (Kitano, H., ed.), pp. 1-36, The MIT Press

2 Kitano, H. (2002) Looking beyond the details: a rise in system-oriented approaches in genetics and molecular biology. *Curr Genet* 41 (1), 1-10

3 D'haeseleer, P. et al. (2000) Genetic network inference: from co-expression clustering to reverse engineering. *Bioinformatics* 16 (8), 707-726

4 Valiant, L.G. (1984) A theory of the learnable. *Comm. ACM* 27 (11), 1134-1142

5 Palsson, B. (2000) The challenges of in silico biology. *Nat Biotechnol* 18 (11), 1147-1150

6 Kitano, H. (2002) Systems biology: a brief overview. *Science* 295 (5560), 1662-1664

7 Palsson, B. (2002) In silico biology through "omics". *Nat Biotechnol* 20 (7), 649-650

8 Kitano, H. (2002) Computational systems biology. *Nature* 420 (6912), 206-210

9 Razin, S. et al. (1998) Molecular biology and pathogenicity of mycoplasmas. *Microbiol Mol Biol Rev* 62 (4), 1094-1156

10 Krupa, B. (2002) On the Number of Experiments Required to Find the Causal Structure of Complex Systems. *J Theor Biol* 219 (2), 257-267

11 Lander, E.S. et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409 (6822), 860-921

12 Venter, J.C. et al. (2001) The sequence of the human genome. *Science* 291 (5507), 1304-1351

13 Ideker, T. et al. (2002) Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18 Suppl 1, S233-240

14 Akutsu, T. et al. (1999) Identification of genetic networks from a small number of gene expression patterns under the Boolean network model. *Pac Symp Biocomput*, 17-28

15 Covert, M.W. and Palsson, B.O. (2002) Transcriptional regulation in constraints-based metabolic models of *Escherichia coli. J Biol Chem* 277 (31), 28058-28064

16 Ibarra, R.U. et al. (2002) *Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* 420 (6912), 186-189

17 Segrè, D. et al. (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A* 99 (23), 15112-15117

Appendix 2

An open-source oligomicroarray standard for human and mouse

(Previously published as Wright MA and Church GM
Nature Biotechnology 20, 1082 - 1083 (2002))

DNA microarrays have transformed biology, allowing the expression levels of thousands of genes to be measured simultaneously [1, 2.] However, the variety of chip designs and the lack of information about the probes used severely limits the use of the data. As a way of addressing these problems, we have designed a probe standard, a set of 121,000 70-base oligonucleotides for the mouse and human transcriptomes, together with a probe selection algorithm. We make the probe sequences and the code for the algorithm freely available on our web site, http://arep.med.harvard.edu/probes.htm.

Open-source standards of this sort have many advantages. With sequence and design information available, probe quality can be closely monitored and the community can suggest improvements. Probe sequences can be directly linked to microarray measurements, allowing detailed troubleshooting and analysis of biologically meaningful fine structure [3, 4]. By annotating the probe sequences, the probes and design methods can be optimized by iteration of experiment and probe design.

Our algorithm terminates once it has found probes that satisfy a set of conditions for sensitivity, specificity, and uniformity, rather than searching for optimal probes as, for 70-mers, many equally specific and sensitive probes may be chosen for the same target[5]. Using this strategy, our algorithm can select probes for human and mouse where algorithms that demand optimality are unsuccessful[6.]

The probes are 70-mers, selected from the representative transcripts of the UniGene database (http://www.ncbi.nlm.nih.gov/UniGene; August 3, 2001, build #138). We identify them by the GenBank identifier and the UniGene cluster of their target so that they can be tracked through future UniGene updates. Our criteria are a BLAST [7] threshold for specificity [5, 8], secondary structure prediction [9], and sequence complexity calculations for sensitivity (J. Deris, personal communication), and a melting temperature window (see http://www.operon.com/oligos/webfaq.php#calculations) for uniformity. We also select probes from within 1,000 bases of the 3' end of the transcript to reflect the 3' bias generated by poly(A) priming during the cDNA synthesis step of many microarray protocols. Although there are alternative ways of predicting specificity [6, 8] and melting temperature [10], none of these alternatives has been shown to be significantly more valuable for 70-mer selection. More details of our methods are available at http://arep.med.harvard.edu/probes.htm.

Table 1 summarizes data for our probe sets and compares human probes selected using our design criteria with those selected by Operon Technologies (Alameda, CA; http://www.operon.com/arrays/arraysets.php). The comparison was carried out between probes designed for the same UniGene cluster by calculating the score of each probe for each of the criteria used in our algorithm. Our human probe set is much larger than Operon's (65,062 versus 13,975) and contains no significant BLAST hits (bit score > 50), whereas Operon's contains 1,496.

Using our initial criteria, we were able to choose probes from 56,037 mouse and 65,062 human UniGene clusters [11]. Our algorithm did not find probes that met these criteria

Table 1. Summary statistics for human and mouse probe sets and comparison with human probe set from Operon Technologies*

Shaded Rows: Our Oligos/ Unshaded Rows: Operon Oligos

	Human				Mouse			
	(N=65,062)[1]				(N=56,037)			
	Min	Max	Average	Stdev	Min	Max	Average	Stdev
Tm[2]	73	83	76.2	2.9	70	78	74	2.7
	71	83	77.1	2.4	-	-	-	-
LZ[3]	21	39	29.3	1.2	26	31	29.2	0.9
	10	41	29.2	1.2	-	-	-	-
RNA[4]	.1	33.8	10.2	4.6	0.1	32.6	7.8	3.7
	.1	36.5	13.8	4.7	-	-	-	-
Distance[5]	1	1000	334.7	242.3	1	1000	219	197.7
	1	4297	244.8	179.1	-	-	-	-
BLAST[6]	0	0	0	0	0	0	0	0
	0	70	14.5	25.9	-	-	-	-

1. Number of transcripts from which probes were selected 2. Melting temperature calculated by the formula $T_m = 81.5 + 16.6*\log[Na+] +41*(G+C)/length - 500/length$ with [Na+] taken to be [.1] see http://www.operon.com/oligos/webfaq.php#calculations. We note that other methods of Tm calculation exist, notably [11] but that for 70-mers the difference between the Tms calculated using these methods is negligible 3. Sequence complexity score calculated by comparing the length of the string (in characters) before and after gzip. 4. RNAfold score -(ΔG) calculated by the Vienna group's RNAfold algorithn 6. Distance of 3' most base of the probe from the 3' end of the transcript. 7. Maximum number of identities in a BLAST alignment (with bit score over 50) with a transcript in a different cluster of unigene.

from among the remaining one-third of the UniGene clusters. We are currently working to choose probes from these remaining clusters by making the criteria less stringent, specifically the 3' distance restriction. It should be noted that some UniGene clusters may contain multiple splice variants. In these cases, the probes target a particular exon or splice junction in the representative transcript and are linked to this transcript by their position in the sequence and GenBank identifier.

In the future, as knowledge of the human and mouse transcriptomes increases, the list of probes will include each exon and splice junction of each transcript from which subsets could be chosen for particular applications. The list presented here is currently the largest freely available list of probes for which the selection criteria are rigorous and publicly disclosed. Synthesis and testing of the probes is currently underway. They are intended as a community resource, and input about the probes or the design methods is welcome. This is a starting point from which to develop an optimized and comprehensive transcriptional oligonucleotide standard. Indeed, our designs are currently being tested by the Programs in Genomics Applications initiative of the National Heart, Lung, and Blood Institute, a large-scale collaborative functional genomics effort that involves groups from across the United States. In choosing our criteria and developing our algorithm, we are grateful to Joe DeRisi and to Operon Technologies, as well as to Affymetrix (Santa Clara, CA), which has released sequence information for its Escherichia coli chip.

References

1. Schena, M. et al. Science 270, 467–470 (1995)

2. Gress, T.M. et al. Mamm. Genome 3, 609–619 (1992)

3. Badarinarayana, V. et al. Nat. Biotechnol. 19, 1060–1065 (2001)

4. Selinger, D.W. et al. Nat. Biotechnol. 18, 1262–1268 (2000)

5. Hughes, T.R. et al. Nat. Biotechnol. 19, 342–347 (2001)

6. Li, F. & Stormo, G.D. Bioinformatics 17, 1067–1076 (2001)

7. Altschul, S.F. et al. J. Mol. Biol. 215, 403–410 (1990)

8. Kane, M.D. et al. Nucleic Acids Res. 28, 4552–4557 (2000)

9. Wuchty, S. et al. Biopolymers. 49, 145–165 (1999)

10. SantaLucia, J., Jr., Allawi, H.T. & Seneviratne, P.A. Biochemistry 35, 3555–3562 (1996)

11. Boguski, M.S. & Schuler, G.D. Nat. Genet. 10, 369–371 (1995)

Supplemental Material on Oligonucleotide Design Considerations

Length

Shorter probe sequences are more specific to their targets since a mismatch will alter the stability of a shorter duplex by a greater percentage than it will alter the stability of a larger duplex. On the contrary, longer sequences form more stable duplexes and are consequently more sensitive; they will bind reliably at lower target concentrations. 50 to 70 base probes represent an optimum of sensitivity and specificity for transcriptional profiling [5].

Uniformity

Since all probe hybridizations occur simultaneously on a chip, the entire set of probe/target duplexes must have similar stabilities. Otherwise, at a given hybridization temperature some probes will form duplexes that are too stable (leading to cross-hybridization with close matches) and other duplexes will be unstable and unable to effectively bind their target.

The melting temperature or Tm, the temperature at which equal numbers of single and double stranded DNA forms are present at equilibrium, is a reliable measure of duplex stability for 50 –70mer probes [6]. We define a range of acceptable Tm for probes to enforce uniform stabilities across the probe set.

Sensitivity

Probes should be capable of detecting a wide range of target concentrations and therefore capable of forming stable probe/target duplexes at both low and high concentrations of the target. By choosing 50-70mer probes within a restricted Tm range close to the temperature of the experiment, we can ensure that most probes have high sensitivity.

Intrastrand basepairing can impair sensitivity by causing probes to preferentially pair with themselves, instead of target. For each potential probe we calculate the most stable secondary structure[7] and attempt to choose probes with low propensity for stable secondary structure formation.

Specificity

Under hybridization conditions probe/non-target DNA duplexes should be unstable and the probe target duplex should be stable. To calculate the potential contribution of probe/non-target duplexes we must consider all possible hybridizations between each probe in the probe set with each DNA in the target pool.

If the target species are of average length M nucleotides with probe sequences of length L nucleotides, there are a total of

$$\binom{M+L}{M}$$

hybridizations (global alignments) per probe sequence pair. There are M-L+1 potential probes per target and N-1 non target sequences to calculate alignments. The thermal stabilities of each of the hybridized duplexes will depend on the specific base pairing, stacking interactions, loops, and mismatches of its configuration. Although it is possible to calculate these stabilities to high precision using measured thermodynamic parameters for each type of interaction, for problems of this scale these calculations are prohibitively expensive. We make approximations that allow us to use sequence alignment methods to calculate potential probe/DNA duplexes. These approximations amount to assuming averages values for each type of interaction and searching with word matches to seed longer alignments. We use the megaBLAST [8] algorithm to perform all alignments and use the bit score to approximate the energy of duplex formation. We then use a threshold for the maximum bit score of a probe alignment with non-target DNA.

Low sequence complexity can also lower specificity. Low complexity (simple repeat) sequences are prevalent in many transcriptomes and also prevalent within the pool of cDNAs. Probes with low complexity regions are therefore likely to form partially stable duplexes with many non-targets and, since the ratio of cross-hybridization to true signal can be approximated by

$$\frac{\sum_{j \geq i} e^{(-\beta \Delta G_{ij})}}{e^{(-\beta \Delta G_{ii})}}$$

(where ΔG_{ij} represents the energy of duplex formation for sequence i with sequence j,) the sum of these many partially stable duplexes can make a large contribution to the signal. We calculate the complexity of potential probes by compression (high compressibility indicates low complexity) and attempt to choose probes that are of high complexity.

The Algorithm

Our algorithm, based on the above criteria, is written in perl, uses BLAST to evaluate non-target hybridizations, calls the RNAfold program[7] to calculate secondary structure, and gzip to calculate sequence complexity based on compressibility. The probes it selects meet a set of quality criteria defined by thresholds. It is divided into two programs and requires a set of parameters to be specified by the user.

Array.pl

Construct an array containing the information from BLAST alignments for each sequence and select at most n sequences of Lmax > length > Lmin with BLAST alignments of BLAST bit score < t

Score.pl

Score each of the initial n probes per sequence chosen by array.pl for Tm, ΔG_{ss} and C. Select at most m that have scores in the required ranges.

Outline of the Algorithm

The steps of the algorithm are listed below

1. Align target sequences against database of repeats and do not consider any aligned subsequences for potential probes

2. Align target sequences against database of all sequences

3. Parse to remove alignments with sequences in target's sequence cluster

4. Parse to generate an alignment array

5. Starting at 3' end, select first n probes with BLAST scores below t. If a sequence is accepted into the set of n, then jump j bases toward the 5' end. If not, move i bases towards the 5' end. Stop examining sequences if Dmax is reached

6. Calculate Tm, C, and ΔG_{ss} for the n probes from step 5

7. Select first m probes with scores that lie in the required ranges for each of the criteria

Score.pl outputs the probes and the scores in a text file.

First Selection

Table 2: Parameter Values used in the Selection

	Human	*Mouse*
Number of Initial Probes	7	7
Number of Final Probes	1	1
BLAST Threshold	50	50
Tm Window	72.5 – 83.0	70-80
RNAfold Maximum	Best	Best
Compressibility Maximum	35	35
Length Window	70 – 70	60-70
Maximum 3' Distance	1000	1000
Increment size	1	1
Jump size	39	39

Results

The algorithm chose a total of 121,000 probes from the human and mouse UniGene sets. Using stringent criteria, its success rate per cluster was about 2/3. Some statistics are summarized below.

Table 3: Selection Statistics

	Input	*Output*
Human	96,826	65,062
Mouse	85,923	56,037

References

1. Gress, T.M., et al., *Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues.* Mamm Genome, 1992. **3**(11): p. 609-19.
2. Selinger, D.W., et al., *RNA expression analysis using a 30 base pair resolution Escherichia coli genome array.* Nat Biotechnol, 2000. **18**(12): p. 1262-8.
3. Schena, M., et al., *Quantitative monitoring of gene expression patterns with a complementary DNA microarray.* Science, 1995. **270**(5235): p. 467-70.
4. Li, F. and G.D. Stormo, *Selection of optimal DNA oligos for gene expression arrays.* Bioinformatics, 2001. **17**(11): p. 1067-76.
5. Hughes, T.R., et al., *Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer.* Nat Biotechnol, 2001. **19**(4): p. 342-7.
6. SantaLucia, J., Jr., H.T. Allawi, and P.A. Seneviratne, *Improved nearest-neighbor parameters for predicting DNA duplex stability.* Biochemistry, 1996. **35**(11): p. 3555-62.
7. Wuchty, S., et al., *Complete suboptimal folding of RNA and the stability of secondary structures.* Biopolymers, 1999. **49**(2): p. 145-65.
8. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.
9. Boguski, M.S. and G.D. Schuler, *ESTablishing a human transcript map.* Nat Genet, 1995. **10**(4): p. 369-71.
10. Dandekar, T., et al., *Re-annotating the Mycoplasma pneumoniae genome sequence: adding value, function and reading frames.* Nucleic Acids Res, 2000. **28**(17): p. 3278-88.

MATTHEW A. WRIGHT

George M. Church Laboratory
232 New Research Building
Harvard Medical School
77 Avenue Louis Pasteur, Boston MA 02115

Phone: 617 432 5917
E-mail: maw@mit.edu

EDUCATION

Massachusetts Institute of Technology (Ph.D. expected September 2005)
Theoretical Chemistry with coursework in Chemistry, Physics, Biophysics, Applied Mathematics, Computer Science, and Biology
Thesis Topic: Approaches to Determining the Three-Dimensional Structure and Dynamics of Bacterial Chromosomes

University of Southern Maine, B.A., B.M. *Summa Cum Laude* 1999
Majors in Chemistry and Music Performance with coursework in Chemistry, Physics, Mathematics, and Music
Thesis Topic: Synthesis of Functionalized Group 14 and 15 Metallocyclopentadienes

RESEARCH EXPERIENCE

Harvard Medical School - George M. Church 2000-2005
Developed first methods to detect evolutionary signals of large-scale 3-D chromosome structure in genome sequences. Developed novel methods for analyzing the relationship between gene function and 3-D chromosome structure using constraint based optimization and the mathematics of distance geometry. Developed algorithms for designing oligonucleotide microarray probes.

Caltech – Rudolf A. Marcus 1999-2000
Studied the rate of long-range electron transfer in DNA in collaboration with J. Barton

University of Southern Maine – Henry J. Tracy 1998-1999
Developed a general reaction scheme for the synthesis of functionalized metallocyclopentadienes.

TEACHING EXPERIENCE

Teaching Fellow, Harvard University – Genomics and Computational Biology	Fall	2001
Teaching Assistant, Massachusetts Institute of Technology - General Chemistry	Fall	2000
Teaching Assistant, Caltech – Synthesis and Analysis of Organic and Inorganic Compounds	Spring	1999

PUBLICATIONS

Wright MA, Kharchenko P, Church GM, Segrè D
Optimal Three-Dimensional Functional Organization of Bacterial Chromosomes
(Manuscript in Preparation)

Segrè D, Zucker J, Katz J, Lin X, D'Haeseleer P, Rindone WP, Kharchenko P, Nguyen D, **Wright MA**, Church GM.
From annotated genomes to metabolic flux models and kinetic parameter fitting.
OMICS. 2003 Fall;7(3):301-16.

Selinger DW, **Wright MA**, and Church GM
On the complete determination of biological systems
Trends in Biotechnology 2003 Jun;21(6):251-4

Wright MA Segrè D, Church GM
4-D Modeling of Bacterial Chromosome Structure
International Conference on Systems Biology, 2002
(Extended Abstract)

Wright MA, Church GM
An open-source oligomicroarray standard for human and mouse.
Nature Biotechnology 2002 Nov;20(11):1082-3

DISCLAIMER OF QUALITY

Due to the condition of the original material, there are unavoidable flaws in this reproduction. We have made every effort possible to provide you with the best copy available. If you are dissatisfied with this product and find it unusable, please contact Document Services as soon as possible.

Thank you.

Some pages in the original document contain color pictures or graphics that will not scan or reproduce well.