The Biology-Combinatorics Interface: Addressing new challenges in computational biology

David Bremner (University of New Brunswick)
Anne Condon (University of British Columbia)
Ken Dill (University of California at San Francisco)
Ron Elber (University of Texas at Austin)
Arvind Gupta (Simon Fraser University)
Ladislav Stacho (Simon Fraser University)

July 6, 2008–July 11, 2008

1 Overview of the Field

Math and biology have a long history together [**?**], beginning with the work of Mendel and Darwin. Today, combinatorics and discrete mathematics are key tools for genome sequencing alignment and assembly, Markov modeling of RNA sequences, gene expression haplotyping, phylogeny construction, and biomolecular statistical mechanics and protein structure modeling.

Now, on the horizon, are the new and bigger challenges posed by systems biology. Systems biologists are interested in models at various levels, from the microscopic (genes, protein structures, and signaling pathways) to the macroscopic (metabolic and genetic circuits, cells, organs, organisms, and populations). Mathematics is poised to help to understand emergent large-scale properties from properties of smaller subsystems, in ways that less quantitative modeling cannot. A key goal is ultimately a quantitative theory of biological cells that begins with the folding, structures and binding of proteins and the regulation of genes, and scales up to the properties of full metabolic and genetic networks, and to the evolutionary dynamical processes that lead to them.

2 Structure of the Workshop

The BIRS workshop on the *Biology-Combinatorics Interface* brought together bio-scientists, discrete mathematicians, and some researchers with a foot in both camps. Rather than look back at recent work in (discrete) mathematical biology, the goal was to look forward and to try to better understand the emerging challenges in systems biology, and how the tools of discrete mathematics can be brought to bear on these challenges. Because of this outlook, the workshop was organized into 3 phases. We began with a sequence of advanced tutorials to help explain the relevant systems biology. In the second phase we held a series of roundtable discussions to help better define mathematical problems. Finally, roughly the last two days of the workshop were dedicated to smaller working groups taking a more in-depth look at the problems developed in the first two phases.

3 Recent Developments and Open Problems

Recent developments are captured in the next section — Presentation Highlights. Open problems are presented in the next two sections — Presentation Highlights and Scientific Progress Made.

4 Presentation Highlights

Because of the format of our workshop, there were relatively few talks. It speaks to the quality of the presentations that the parallel workshop on Rigidity and Biology chose to attend almost all of our talks. We also attended several of their presentations; in fact the interaction between the two workshops could be characterized as one of the presentation related highlights.

4.1 Ken Dill

Ken Dill described some general unsolved problems at the interface of combinatorics and biology, described below.

Proteins & RNA molecules: predicting structures, stabilities, conformational changes, aggregation, rates, higher structures, loop modeling, interactions with ligands or biomolecules, mechanisms & motions. The key issue is that proteins are chain molecules with large numbers of degrees of freedom that must be searched by computer to find globally optimal states, but there are complex steric constraints (the chain cannot pass through itself – Hamilton walks, not random walks – and these problems are particularly acute for compact states, like the native states of proteins, which are the biologically important structures). An important problem is how to enumerate the relevant states efficiently by computer. Solving these problems is important for drug discovery in the pharmaceutical and biotech industries.

Zipping and assembly is a mechanism involving fast conformational searching for protein structures. Proteins and RNA molecules fold on funnel-shaped energy landscapes. These landscapes themselves help proteins search through their exponential search spaces with high efficiency.

Cells: characterize network topologies & traffic, flows, and robustness. Biochemical reactions inside biological cells are arranged as networks of molecules that take up substrates from some nodes and pass along products to other nodes. Problems in cell biophysics require understanding these networks and the flows of biochemicals along them. These flows are generally quite nonlinear, and there are many of them, coupled to each other. Mathematical methods are needed to better understand their global properties.

Evolution: How do networks arise? What are properties of fitness landscapes? Advantages of modularity, hierarchy? Cancer - a disease of evolution? How do pathogens become drug-resistant? Not only do we need to better understand the biochemical networks of biological cells, but it is also important to know how those networks evolved from simpler networks, as one species evolves from another. This is important for understanding the unity of life, but it is also of great practical importance. Cancer is a disease in which cells in the body evolve away from their canonical states and structures and become renegade. Cancer is called somatic cell evolution: it involves not the evolution of a species over long evolutionary time scales, but the evolution of a few cells in the body over the body's lifetime.

Nano- bio- tech: Make new structure: as scaffolds, drugs, delivery vehicles, biomaterials, test-tubes-in-a-cell, imaging agents. Biology provides powerful design principles for making functional machines on the nano size scale. However, to build machines on this scale requires overcoming Brownian and entropic forces. To better understand these forces requires better combinatorial techniques to understand the complexity of such systems – the entropies of protein chains, the diffusion of particles subject to constraints, the random navigation of particles through highly confined cell interior regions.

4.2 Paul Higgs

Paul Higgs spoke on the problem of evolution in a *pre-DNA World*. Several *RNA World conjectures* exist, the weaker being that RNA acted as a precursor to DNA and proteins for replication and catalysis. Models of growth by self-replication and by catalysis (for example in a "lipid world") can show heredity without using polymers for information storage, but the question of how evolution can occur remains. One hypothesis that is both biologically interesting and poses interesting questions of mathematical modelling and simulation is the notion of *vesicles* or autocatalytic set of molecules.

4.3 Ron Elber

One of the striking observations of modern structural biology is the relatively small number of experimentally-determined families of structures (about a thousand) that capture a significant portion of the empirically-known and much larger sequence space (several millions). The large capacity of protein structures (folds) to sequences motivates research along two directions. The first, which is perhaps more practical, is the modeling of protein structures from sequences based on similarity of the target sequence to another sequence for which the structure is known (homology). The second is an attempt to understand the "history" of sequence-structure relationships and analyze sequence evolution taking into account changes in the stability of protein structures upon mutations.

In the workshop we focused on the second item. To investigate the impact of stability on evolutionary processes we need to consider the options open to a mutating sequence. The protein mutant may be more or less stable than the native, it may unfold, or it may flip to a different structure.

The first question that was discussed concerned sequence capacity. Namely, how many sequences fit a particular fold and have energy lower than the energy of the native sequence? The capacity sets an upper bound on the evolvability of a protein fold. Besides stability there are other factors that impact mutability of a sequence. These factors include conservation of active site residues, of protein- protein interactions, of required flexibility, etc. We discussed randomized algorithms that sample sequences that fit a fold efficiently and allow for quantitative estimates of the capacity. It was found that the number of sequences accessible to a fold is exponential in the sequence length. The vast number of accessible sequences should be contrasted with the even (exponentially) larger number of possible sequences. Interestingly protein capacity correlates with protein mutation rates as observed empirically [**?**], supporting the idea that stability influences biological changes. Fold stability may provide a zero order model for molecular evolution.

Another intriguing view of protein evolution is the possibility that a protein sequence that folds into a particular family of structures flips into another family following a single amino acid change. Some experimental evidence is available that supports this idea [**?**] We discussed construction of a network of sequence of flow in which folds are nodes and weighted edges represent flow of sequences between structures. Hubs in the network are identified as structures rich in beta sheet.

4.4 Jie Liang

One of the most important topics in molecular biology is understanding the relationship between the sequence of amino acids in a protein and the resulting folded structure and function of the protein. Dr. Jie Liang presented recent work on using sequence (and hence evolutionary) information to understand geometric structure, and conversely, on extrapolating from known structure information to study evolutionary processes.

Proteins have many voids and pockets, which can be computed (identified and measured) through the aid of alpha complex and pocket algorithms. Identification of particular pockets as biologically important is challenging, since compact self-avoiding random walks exhibit similar behaviour to proteins. For enzymes, Dr. Liang explained that searching using sequence information, augmented by Markov model based processes had achieved success both in solving the resulting surface similarity problems and in distinguishing evolutionary factors from those related to physical folding stability.

A second issue related to protein evolution is the construction of protein fitness function important for protein design. The task is to develop a global fitness function that can discriminate correct protein sequence from unmatchable sequences for all known proteins structures simultaneously. Dr. Liang presented nonlinear potential function of Gaussian kernels. The resulting fitness landscape can be used to study protein evolution.

4.5 Christine Heitsch

An RNA molecule is a linear biochemical chain which folds into a three dimensional structure via a set of 2D base pairings known as a nested secondary structure. Reliably determining a secondary structure for large RNA molecules, such as the genomes of most viruses, is an important open problem in computational molecular biology. Dr. Heitsch discussed combinatorial results which yield insights into the interaction of local and global constraints in RNA secondary structures and suggest new directions in understanding the folding of RNA viral genomes.

4.6 Ján Maňuch, Ladislav Stacho, Arvind Gupta

The HP model for protein folding has been proposed in mid 80s as a simplification of complex interactions among protein molecules. The main problem is to understand the processes and determine the folds of proteins that minimize total free energy. To simplify the problem, the protein is laid out on a 2D lattice with each monomer occupying exactly one square and neighboring monomers occupy neighboring squares. The free energy is minimized when the maximum number of non-neighbor hydrophobic monomers are adjacent in the lattice. Even in its simplest version this problem is NP-complete.

In many applications such as drug design, we are actually interested in the inverse problem to protein folding: *protein design*. Again, most of the known research concentrates on the 2D and 3D designs in the HP model. Surprisingly, the complexity of this problem is not known, however it is generally believed that all interesting versions will be NP-complete. Recently research concentrated on designs based on domino like structures. Here the required design is composed of small base structures and it is done so that the resulting fold will have a minimum free energy. In 2D this method gives solutions that approximate required shapes arbitrary closely and in 3D such solutions are knows for a large classes of shapes. The main unsolved problem remains to prove the stability of the solutions, i.e. to prove that the folds will be unique in addition to their minimality. The stability of solutions is proved only for very basic shapes in both 2D and 3D cases.

Recently, an interesting refinement of the HP model, in which the cysteine and non-cysteine hydrophobic monomers are distinguished and SS-bridges which two cysteines can form are taken into account in the energy function. In 2D variant of this model with the additional assumption of two distinct SS-bridges types (that cannot interact between each other) it is possible to prove the stability of many solutions that can be used to approximate any structure. The main open question remains the stability of solutions when only one type of SS- bridges is used. In 3D variant stability still cannot be guaranteed and only results possible are that we get only few folds that are somehow structurally close to each other. The challenge remains to tweak these solutions so that stability is guaranteed.

4.7 Joint RNA Session

During the joint RNA session, participants from the Biology-Combinatorics Interface workshop and the Rigidity, Flexibility and Motion workshop met for a joint session on RNA structure. Jack Snoeyink described the computational challenges that arise in inferring RNA backbone conformations from NMR or X-ray crystallography data. The challenges arise because of the many possible dihedral angles per residue in the RNA sequence. Anne Condon described computational problems that arise in predicting RNA secondary structure. These include the high complexity of recognizing pseudoknotted structures, and the challenges in improving accuracy of both pseudoknot free and pseudoknotted structure predictions.

5 Scientific Progress Made

5.1 Evolutionary Capacity of Proteins

Inspired by the talk of Ron Elber, David Bremner and Jie Liang discussed how to construct simple exact models for the evolutionary capacity of known protein structures. By abstracting spatial information into a graph, and considering a simple hydrophobicity based energy model, exact and asymptotic bounds on the number of sequences that are a “good fit” to a given structure can be computed.

5.2 Energy barriers in RNA secondary structure

5.2.1 Topological Results

Christine Heitsch, Paul Higgs and David Bremner considered the barrier height problem for RNA molecules [**?**, **?**]. They investigated the existence of a transition path between any two RNA configurations where the energy of each move does not exceed some pre-determined threshold. Prior results [**?**] show that this is indeed possible in the most abstract situation. Preliminary results include that this extends to arbitrary balanced binary sequences of n 0's and n 1's; it turns out that every such sequence has a complete nested folding with n noncrossing $(0, 1)$ and $(1, 0)$ pairs, and that any two such foldings are connected by some minimal energy path of local moves. They are now considering the situations where the binary sequence is unbalanced and where only base pairings in the source or target structure are allowed. They will also investigate the problem of enumerating the number of possible foldings.

5.2.2 Computational Complexity

Problem worked on at the workshop: understanding the computational complexity of calculating energy barriers in RNA secondary structure pathways.

Context: RNA structures are determined largely by pairings of bases in the RNA molecules, with A's binding to T's and C's binding to G's. Given enough time, RNA molecules will fold to their minimum free energy (MFE) structures. There are many algorithms available that aim to predict MFE structures. However, the accuracy of such algorithms is still quite poor, suggesting that perhaps a molecule gets trapped in a locally optimal structure, from which there is a high energy barrier to the MFE structure. For this reason, it's interesting to calculate the energy barrier between two RNA secondary structures. Additionally, knowing the energy barrier can shed light on the pathway of structures formed by a molecule before it reaches it stable (MFE) structure.

Currently, the best methods for calculating the energy barrier between two RNA secondary structures are heuristic in nature, and it is an open problem whether the barrier can be calculated efficiently. To understand the complexity of calculating energy barriers, we cast the problem in a simple setting, and investigate whether the resulting simplified problem is in P or is NP-complete. We look at many problem variations.

Defining folding pathways: Let S be a secondary structure, and let —S— be the number of base pairs in secondary structure S. A secondary structure can be represented as an arc diagram, in which the bases of the molecule are arrayed along a horizontal line and arcs connect two paired bases. For this reason, we refer to base pairs as arcs. A structure is pseudoknotted if two base pairs cross.

A direct path P from secondary structure S to secondary structure S' is a sequence $S = S_0, S_1, \ldots S_t = S'$ of structures, each obtained from the previous one either by (i) removing an arc which is in S or (ii) adding an arc which is in S' and does not cross any arc of S which has not yet been removed.

The energy barrier of path P is

$$\max_{1 \leq k \leq t} |S| - |S_k|$$

The energy change of path P is $|S| - |S'|$.

One problem we considered is the Pseudoknot Free Folding Problem:

Given: S and S', which are two pseudoknot free secondary structures, and a positive integer k.

Question: Is there a direct path from S to S' with energy barrier at least k?

This problem does not allow intermediate arcs to be introduced along the path from S to S', by definition of direct path. A related problem, the Pseudoknotted Folding Problem is defined similarly, except that S and S' may be pseudoknotted. This problem is somewhat artificial, because the secondary structures S and S' may be pseudoknotted, yet when adding arcs of S', they are not allowed to cross those in S.

While we were unable to resolve the complexity of the pseudoknot free version of the problem, we did make progress during the workshop in showing that the pseudoknotted version is NP-hard.

5.3 Step self-assembly

Self-assembly is an autonomous process by which small simple parts assemble into larger and more complex objects. Self-assembly occurs in nature, for example, when atoms combine to form molecules, and molecules combine to form crystals. It has been suggested that intricate self-assembly schemes will ultimately be useful for circuit fabrication, nano-robotics, DNA computing, and amorphous computing. To study the process of self-assembly we use the Tile Assembly Model proposed by Rothemund and Winfree [**?**]. This model considers the assembly of square blocks called "tiles" and a set of glues called "binding domains". Each of the four sides of a tile can have a glue on it that determines interactions with neighbouring tiles. The process of self-assembly is initiated by a single seed tile and proceeds by attaching tiles one by one.

During the meeting we have discussed an extension of the basic tile model to step self-assembly as suggested by Ken Dill. In this extension, several sets of tiles are used. In each step one tile set is applies on the growing structure, which grows as long as possible. When the growth stops, the current tile set is washed away and a new tile set is applied on the structure. In this way, it is possible to build a square of size $2N + 1 \times 2N + 1$ in N steps using only 2 sets of tiles each having a constant number of tiles. This is true even at temperature 1 and when all neighboring tiles of the assembled square are connected by a glue. For comparison, using the original model to assemble $2N + 1 \times 2N + 1$ square in the same setting would require $(2N + 1)^2$ distinct tiles.

6 Outcome of the Meeting

The increasingly quantitative nature of biology is most readily evidenced in molecular genomics. To ensure rapid progress, matching the most appropriate mathematical techniques to specific problems is essential. This workshop was a major step in this direction. Outcomes can generally be grouped as follows:

1. Identify key problems amenable to combinatorial treatment. These range from understanding biomolecular molecular structures, build models for cellular processes such as signalling and network topologies, evolution and its role in disease and building nanotechnological structures using biomolecules.

2. New collaborations formed which should lead to rapid progress on a number of specific problems as outlined in Section 5.

3. Preliminary plans for continued and expanded collaborations. The main organizers are exploring possibilities to launch future meetings that bring together an expanded group of scientists.

A typical reaction from a participant was from mathematician Christine Heitsch

> The workshop was an exciting opportunity to interact with people interested in both discrete mathematics and molecular biology. I learned a lot about combinatorial models for protein structures, and had many stimulating conversations about RNA folding. I'm looking forward to future opportunities to interact with workshop participants, and hope that some of these connections will develop into research collaborations.

This reinforces the organizers' view that there are many interesting interactions to study between discrete and combinatorial mathematics and biology. One option we may explore is application to BIRS for *Research in Teams* programs to speed up progress on specific problems.

References

[1] Patrick A. Alexander, Yanan He, Yihong Chen, John Orban, and Philip N. Bryan The design and characterization of two proteins with 88% sequence identity but different structure and function *PNAS* 2007 104: 11963-11968].

[2] J.D. Bloom, D.A. Drummond, F.H. Arnold, and C.O. Wilke. Structural determinants of the rate of protein evolution in yeast. *Mol. Biol. Evol.* 23:1751-1761, 2006]

[3] C Flamm, I L Hofacker, S Maurer-Stroh, P F Stadler, and M Zehl. Design of multistable RNA molecules. *RNA*, 7(2):254 – 265, 2001.

[4] Christine E. Heitsch. A new metric on plane trees and RNA configurations. In revision.

[5] S R Morgan and Paul G Higgs. Barrier heights between ground states in a model of RNA secondary structure. *J Phys A (Math & Gen)*, 31:3153–3170, 1998.

[6] Margaret A. Palmer, Peter Arzberger, Joel E. Cohen, Alan Hastings, Robert D. Holt, Jennifer L. Morse, De Witt Sumners, and Zan Luthey-Schulten. Accelerating mathematical-biological linkages: Report of a joint nsf-nih workshop. Technical report, NIH, February 2003. `http://www.maa.org/mtc/NIH-feb03-report.pdf`.

[7] Paul W. K. Rothemund and Erik Winfree The Program-Size Complexity of Self-Assembled Squares. *STOC 2000* pp. 459–468