BIRS Workshop 04w5017, July 17 - 22, 2004 Modeling Protein Flexibility and Motions

Chief Organizer: Walter Whiteley (Mathematics and Statistics, York University)

March 8, 2006

Organizers: Michael Thorpe (Physics, Arizona State University),
Leslie Kuhn (Biochemistry, Michigan State University).

Overview of the Subject

Following from work on the genome, the focus is shifting to protein structure and function. Much of the function of a protein is determined by its 3-D structure and motions (often in complexes of several molecules). The structure of many new proteins is being determined by x-ray crystallography and by nuclear magnetic resonance techniques. One can then study both local flexibility (adapting shape to fit with other molecules) and larger motions. One can also study the impact of other contacts such as ligands (drugs), or binding into complexes of proteins, DNA etc. in changing the shape and flexibility. An important area of current research in biochemistry, computational geometry and in applied mathematics is the computer modeling of such behavior: which sections are rigid, under certain conditions; the possible motions; unfolding pathways; multiple configurations with different biological functions; and paths between these configurations.

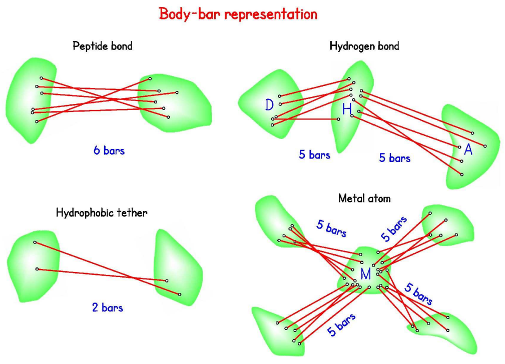

Figure 1: Showing the various elements in the body-bar representation of a folded protein structure that are used in many algorithms for rigidity including FIRST.

The mathematical theory of rigidity, and related techniques from geometric constraint theory (CAD, robotics), are one set of tools for such computer modeling. Applications of such techniques to protein flexibility have been expanding over the last few years, centered on the program FIRST. A short summary of the current state of the art for the combinatorics central to the rigidity methods and the robotics methods includes three factors:

1. the general problem of predicting whether a graph, build in 3-space as a bar and joint framework, will be rigid or flexible, for almost all realizations, is an long standing problem, going back at least to James Clerk Maxwell.

2. the general problem of predicting whether a graph built with vertices as rigid bodies, and edges as hinges, in 3-space, will be rigid or flexible for almost all choices of lines for the hinges, has a simple combinatorial solution and an efficient algorithm.

3. the general problem of frameworks extracted from covalent bonds of molecular structures (with fixed angles at the bonds) is conjectured to be covered by the algorithms of (2) (the Molecular Framework Conjecture) although it is a special class of frameworks under (1).

Algorithms have been implemented for certain models of proteins as frameworks within this mathematical theory. These models develop a graph based on the covalent bond network plus additional edges related to ionic bonding (salt bridges and hydrogen bonds, identified by proximity of these atoms in the 3-D structure) as well as graph edges for hydrophobic interactions, also identified by proximity of suitable heavy atoms in the 3-D structure. This graph yields a constraint matrix which will predict the first-order rigidity or flexibility of the corresponding model, and hopefully of the underlying molecules. However, for speed of computation (on works with up to 400,000 atoms), the rank is actually predicted from the combinatorics of the graph, using counting algorithms (often called the 'pebble game'). The accuracy of these combinatorial results to the rank of the underlying matrix would follow from the 'molecular conjectures' of Tay and Whiteley, and there is significant experimental evidence, as well as partial results to support this correctness. These algorithms are fast enough to be used as preliminary screening in areas such as ligands as drugs. Other partial results have been obtained, and interesting comparisons have been made with measured biological data. Recent work has scaled up from single proteins to complexes such viral coats and RNA protein complexes but much work remains. The program FIRST [Floppy Inclusions and Rigid Substructure Topography] was discussed at some length during the workshop, and is available on the net at flexweb.asu.edu. The use of this web site was demonstrated during the workshop.

This rigidity/constraint based work has been extended from first-order predictions (in the rank of the matrix), using rigidity decompositions, and Monte-Carlo steps, to simulate larger motions, including pathways between known conformations of the same molecule. This is embedded in the program ROCK [Rigidity Optimized Conformational Kinetics], which was also presented during the workshop, and is also available at flexweb.asu.edu

Gaussian Network Models (GNM) represent another combinatorial and computation method for integrating 'proximity' constraints into linear algebra and predictions of motions of large bio-molecular structures. These models build a simple "incidence' matrix for proximity of large atoms in the molecule, on a scale designed to ensure non-singular square matrices, then examine the dominant eigenvalues and vectors to predict significant overall motions. These methods were also presented at the workshop, along with some comparisons of predictions from GNM and FIRST and with known experimental measurements of forms of flexibility.

Recent work in computational has investigated the computational complexity of a variety of algorithms and questions around folding and unfolding chains, polygons and other simplified models which would relate to proteins. This includes the results in computational geometry (such as the Carpenter's rule problem that combined computational geometry with results in rigidity theory. Other work on linkages in 3-space confirms that the 3-D problem is significantly harder, but also indicates that some results can be obtained. Work in robotics has also studied the kinematics of larger scale structures subject to geometric constraints. In particular, the Probabilistic Road Map method from robotic motion planning has been applied by several groups to generate possible folding pathways for proteins. One version of this was presented at the workshop, along with a brief introduction to their on-line service that was being mounted as the workshop progressed, at parasol.tamu.edu/foldingserver/.

A number of computational biochemists and biophysicists have generated a broader range of algorithms for predicting and simulating the shape and motion of proteins. Many of these include minimizing energy functions - a process is can, at critical points of the functions, have relationships to rigidity theory. The most intensive of these methods are the Molecular Dynamics Simulations (MDS), which work with all atom models and energy functions for many interactions, to both simulate local motions, and to examine larger scale motions, up to the level of protein unfolding and folding.

Other Biochemists are working at a more detailed level of the local geometric configurations and choices made in the placement of small sections of the backbone, and side chains, in creating the initial protein data bank (pdb) models from X-ray crystallography and NMR data. The quality of this data can be crucial as input to various modeling methods (above) and some of the modeling and computational methods above can, in turn, contribute to the quality of the pdb data. The interplay of data and modeling is an important feature of the state of the art these days, and all communities share an interest in this interplay.

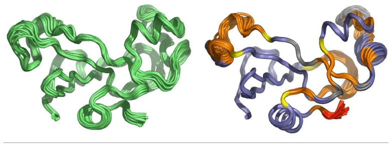

Figure 2: Showing the flexibility of the protein barnase. On the left is the structure as determined by NMR experiments. On the right additional conformers have been generated from the average X-ray structure using the programs FIRST and ROCK. [S. Menor, Ming Lei, M. Zavodszky and M.F. Thorpe, unpublished]

Each of these fields is in rapid evolution, due both to new theoretical results and to new experimental results that modify our assumptions and raise new questions. The work is increasingly interdisciplinary and the workshop reflected that reality.

Structure of the Workshop

There are a number of distinct communities working on computer (and mathematical) modeling of protein flexibility, rigidity, and folding, or on simplified and abstracted models with potential applications to these problems. This workshop brought together leading experts as well as current graduate students and post-docs from at least four of these communities: mathematicians working on the rigidity theory for structures (frameworks, molecular structures, tensegrity structures); computational geometers working on motions and paths of linkages, polygons, etc.; material scientists modeling rigidity in large molecular configurations; and biochemists modeling protein flexing, binding of molecules on proteins, detailed modeling of 3-D protein modeling and a variety of tools for predicting protein behavior.

The workshop gathered together members of these communities of researchers to:

1. summarize the state of the art (as this time) for modeling protein flexibility and motions using models such as frameworks, linkages, Gaussian network models, robotics kinematics, etc.;

2. describe unsolved critical problems about current and potential models (mathematical, computational and biochemical), helping to sort the potential significance of various problems and potential results;

3. provide some grounding of mathematical and computational modeling efforts in biochemical data, to explore the effective use of this knowledge within modeling programs, and offer some reality checks on the meaning of that data for predictions of flexibility and folding.

4. Participate in working sessions to explore ways to clarify, resolve or solve these problems and propose priority problems and approaches.

Full hour talks the first day provided a survey of all four areas, with an emphasis on posing questions, conjectures, and directions for work that would connect the represented audiences. This was followed by an evening 'problem session' of issues worthy of follow up. These problems were immediately posted to the Web Comptes Rendu site (see below) and integrated into the ongoing discussions.

In advance of the workshop, participants were encouraged to post relevant papers, presentations and unsolved problems on a web site (see below). About 25 participants loaded materials, and many participants were able to prepare for the exchange by downloading and reading the materials. This easy ability to share materials played an essential role in building sufficient common ground to support strong exchanges among participants from diverse backgrounds, who had never met prior to the workshop.

Some 'problems needing to be solved' were posed in advance, and least three of these were solved before the workshop was over. Other problems posed were discussed during the workshop and additional problems were posed and posted during the workshop. Solutions or other follow up commentary continue to be posted at this time. We anticipate further follow up materials will be posted to the site over the next few months. This valuable site linking material on flexibility of proteins has now been linked from the flexweb.asu.edu web site. It has also been linked from the home pages of several of the participants, to become a resource for the wider community of researchers. In this way, it offers a basic source of information for new people to this area, including graduate students just moving into these areas. As such, this resource represents a clear outcome from the workshop that will continue to assist the building of a larger community with common goals, and building comparisons of results towards shared standards.

The program was deliberately flexible. Participants brought along PowerPoint or other presentations that were adapted overnight to address questions raised, or new approaches and issues that were relevant. All talks generated extensive discussion both during and after - confirming that we had achieved the desired engagement of people in interdisciplinary conversations. We also scheduled, from the second day on, some time for focused conversations with leadership from one, or several participants, and guidance from one (or both) of the organizers. give time for organized and informal working groups. As proposed, we also offered on site, software, and web site demonstrations, and access on a demonstration basis to software.

From there, the program evolved with

1. themed sections (Biochemistry, mathematics, computer science, biophysics with comparisons of methods (see below).

2. some wide ranging discussions with experts leading off an everyone pitching in;

3. substantial unscheduled time (noon to 3:30 most days) for informal conversations;

4. some evening discussions and software / web site demonstrations, running up to 10:30 at night;

5. discussions of shared concerns, including one session on the community responses to issues of patents, university intellectual property rules, sharing of code, etc.

Overall of the program elements there was active, spirited and informative discussion. No talk passed without engaging in extensive conversation about the methods and the results, and some sessions became extended conversations focused on themes from the problem sessions, or from debates which arose during previous discussions.

The patience of all speakers contributed to an atmosphere of respect and debate through which our different priorities, approaches, solved and unsolved questions were compared, and sometimes contrasted. The discussions on into the night engaged people from distinct communities in more detailed sharing of approaches, resources, and 'gold standards' for evaluating the quality of conclusions.

Since even the people within a single community had not previously gathered in one spot, there were also exchanges among people with similar backgrounds, and these exchanges among mathematicians, some computer scientists and some biophysicists were further reinforced in the follow-up Calgary Workshop on Rigidity (see below).

Working within Diverse Communities

There were some spirited, good natured and insightful exchanges about the priorities and contributions of various communities to the conversations. Here is how one graduate student (in computer science) presented her observations:

1. The mathematicians like theorems, but don't really care if the theorems can be used to compute results.

2. The physicists like computing results, but don't really care if the theory behind the results is correct.

3. The computer scientists like to compute results and have the theory be correct, but the mathematicians don't like our proofs, and the physicists don't like our results.

Here are a couple of other noted quotes, illustrating the good-humored give and take:

"What we need is a good definition ... I am starting to think like a mathematician and the opposite was supposed to happen!" (A biophysicist)

"Mathematicians care about conjectures and things like that" (Post-doc).

Discussion and Outcomes

Everything said at the workshop supported the initial claim that flexibility of proteins (and other molecules) is an essential feature of the 3-D structure and its functioning. All of the speakers cast their work in terms of ways to explore this flexibility, to compare different measures and modeling methods and to predict flexibility and its impact on the interaction of molecules, both complexes of large molecules, and interactions of small drugs with large molecules.

Not surprisingly, everything said also supported the implicit theme that this modeling is hard. There were passing references to modeling protein folding (a very hard problem ab initio from the sequence) but this was not central to our discussions. The results, and the problems, addressed more modest goals, such as: - predicting, from a single 3-D structure, the regions of rigidity and flexibility within the molecular complex (e.g. FIRST); - predicting, from a single 3-D structure, the large scale patterns of the dominate modes of motion (e.g. Gaussian Network Models, ROCK) - searching for pathways between two known conformations of a molecule (e.g. ROCK PRM); - comparison of single molecule predictions with ensembles of structures generated by NMR methods, and increasingly generated by high quality X-ray crystallography methods (e.g. X-ray + ROCK = NMR), as illustrated in Figure 2. - ways of representing flexibility and motions of molecules, in ways that scale up and down for detail and overview; - providing solid mathematical foundations for the methods (above) and for comparisons among these methods. - improved models, computational techniques, and data base presentations, which will speed up work in many of these efforts; - incorporation of accurate biochemical information (e.g. rotomer data bases and improved Ramachandran plots) to improve the speed and accuracy of current algorithmic methods.

There was a good review of the successes (and limitations) of the current rigidity based algorithms. Discussion did confirm the advantages of switching from the former, 3-D bar and joint mathematical model to the molecular bar and hinge model. The advantages include: - a closer fit to kinematic models in use in robotics and in representations of molecular coordinates in terms of torsion angles; - simpler implementation of the combinatorial counting rules - more options for the value of constraints such as hydrophobic constraints, without resorting to 'pseudo-atom' insertions to trick to original algorithm - a more complete capture of 'stresses' and 'redundancy' in molecular models with small rings. - direct representation of possible 'collectively linked torsion angle changes' Other desirable features include: - equivalent matrix representations of first-order motions in the previous bar and joint models and the alternate models, - experimental equivalence with the prior algorithms for bar and joint models, - conjectures (with strong support) that the rigid region decompositions from the algorithms are identical; - close alignment with the mathematical conjectures and the broader mathematical model (body and hinge structures) for which the algorithms are known to be correct. An expository and research paper comparing the two models is being drafted by two of the mathematicians (Tay and Whiteley) to lay out the foundations of the molecular hinge model now being used, and to demonstrate the known equivalences and correspondences with the bar and joint model. This will support both those using these models for algorithmic work, and mathematicians investigating the conjectures about the algorithms for these models.

Extensive discussions, at Banff and the follow-up Calgary Workshop (see below) probed the mathematical complexities of the algorithms and the possible proofs of the molecular conjectures. A number of pieces were added to the puzzle, around connectivity and other features of 'rigid region decompositions' for general structures, and for molecular structures. As someone who has worked on these problems for over a decade, I was impressed by the new partial results, and new approaches which were discussed. Overall, the critical objective of engaging more mathematicians to work on the significant problems posed by the algorithms for modeling protein flexibility was achieved.

There were also discussions of how rigidity (FIRST) type results are being integrated into larger simulations, such as ROCK (flexweb.asu.edu/rock_index.html), and newer Parasol Folding Server (parasol.tamu.edu/foldingserver/). In each case, grouping the atoms of a 'rigid cluster' together as a single moving body reduces the complexity of the computations.

Ring Closure: A number of computations involve steps which perturb the torsion angles along a bonds, and then relaxing or filling in missing values to ensure the closure of loops formed either by other bonds, such as ionic bonds, or implicity by fixing the position of more distant atoms. This ring closure is a classical problem that forms a bottleneck in terms of simulations. There was a morning session on algorithms for ring closure, and their applications in programs such a ROCK, the Parasol server, and the Richardson's work on 'protein chiropraxis' (snapping pieces of the backbone to alternative positions to generate more relaxed configurations). This sharing of techniques again opened the possibility of further collaborations to improve current algorithms and other examples to benchmark proposed methods with.

We had a wide-ranging introduction to the programs, viewers, and data-bases at the Richardson Lab (kinemage.biochem.duke.edu/). This provoked a lot of discussion about the quality of different sources of data, what errors to suspect, and what could be done to anticipate, and perhaps correct, the errors in the data which might impact the performance of the algorithms. As follow up to this discussion, some existing software will build in use of the improved systems For example, following discussions on possible collaborations, FIRST will soon provide Reduce placement of hydrogen atoms as an alternative to WhatIf placement now recommended. Other software will look at using the 'penultimate rotamer library', and other refined Ramachandron Plots to control which torsion angles are permitted in the simulations.

In general, there was an underlying theme that algorithms for protein flexibility can improved by better incorporation of accurate biochemical information during the initial processing or during selection of steps for simulation. This discussion was only possible because we had people from the multiple communities debating and exchanging over the full five days of the workshop.

A second complementary theme is adding computation expertise to current biochemistry (and perhaps biophysics) algorithms to improve their performance. The discussion of ring closure (above), and another discussion of the use of singular value decomposition for 'collective motions' were two examples of this theme. No definitive conclusions were reached, but possibilities were explored, and comparisons generated for further reflection.

There were a number of new ideas and approaches that were sparked for individuals and for small groups. Here is one illustration of how this worked.

There were two presentations about the flexibility of icosahedral virus capsids, using different techniques (engineering model building and FIRST analysis). Looking at the illustrations, the question was raised - can we find algorithms to detect only the motions respecting certain symmetries which are subgroups of the symmetries of the molecule? (This question can also be asked for much smaller structures, such as dimers of two protein chains with half-turn symmetry.) After some exploration, discussion during a coffee break generated a proposal (from an engineering participant) for extracting the combinatorial counts for matrices from the irreducible representations of the symmetry group. From these counts, the corresponding combinatorial pebble games were proposed (from a mathematician), and some simple examples computed and analyzed to ensure the overall results fit the larger patterns in cases we understand. The tentative conclusion is that we do have a 'program' for adapting known methods to symmetric motions of symmetric molecules, and a research / writing project is underway to tie up the details and share the results.

Finally, the most difficult outcome to document, but achievement, is that each participant came way with a broader perspective on what questions are significant, what resources are available for pursing our old questions (and some new ones) and what people might offer that additional insight which make carry us over from initial ideas to promising methods and new results. All the individual feedback received by the organizers has confirmed that participants, from graduate students to senior faculty, made new connections, saw new possibilities, and developed new respect for the contributions, and the difficulties of each of the participating communities.

We ended with one consensus conclusion - we should do this again in the future!

Special Features around this Workshop

Travel Funds. In addition to the BIRS funding, and the associated MSRI travel support, the NIH grant of one of the organizers (Walter Whiteley) included some funding for a workshop in the summer of 2004. This funding was used to provide support for a number of additional graduate students and post-doctoral fellows, as well as a few more senior researchers who would otherwise not have been able to attend. This funding was also used to partially subsidize the related Workshop on Rigidity (see below) in Calgary. **Web Comptes Rendus**

biophysics.asu.edu/banf/list.php

This web site, hosted Arizona State University by one of the organizers, Michael Thorpe, permitted uploading of documents for sharing before, during and after the workshop. A number of people downloaded relevant articles and presentations to develop further background prior to the workshop. During the workshop, presentations, and related references were uploaded and accessed by people to further discussion. [The ready access to computers, wireless connections, and printers was a real assistance here.]

During the workshop, this web site became a 'value added' feature, as some presentations given in the morning became available to participants that evening. Overall, the easy access to the internet, via wireless and in-room connections, as well as the printers, was effectively mobilized to support the conversation, rather than distract from participation. When US visa problems forced a few cancellations, this web site also became a place to post their presentations, so these contributions could also be accessed.

University of Calgary Workshop on Rigidity July 22-24.

This Banff Workshop gathered an important segment of the mathematical (and computational) research community on rigidity in one place. While the interdisciplinary workshop was an important source of problems and an impetus for future work, we also wanted a few days where the full range of current mathematical questions of rigidity could be shared and discussed. With the cooperation of the Department of Mathematics at the University of Calgary, in particular of the Canada Research Chair in Geometry, Karoly Bezdek, Robert Connelly, and Walter Whiteley organized a follow-up Workshop on Rigidity on Friday and Saturday, July 23-24. Funds from the University of Calgary, and some travel / housing funds from the NIH subcontract of Walter Whiteley, we were able to cover costs for housing, lunches, and transportation from Banff to the Calgary Hotel.

As a surprise (to the Rigidity Workshop Organizers), people from the computational and biophysics communities also stayed on for this workshop and contributed greatly to the discussions. As a result, we were able to have some follow up talks addressing issues raised in Banff, and several focused discussions on core conceptual and computational issues around 'collective motions' and 'redundant constraints', as well as follow up tasks from Banff, such as extending algorithmic work to describe symmetric patterns of stress and motions for symmetric structures such as viral capsids.

These discussions consolidated mathematical and computational developments from the Banff workshop, and have been followed up by active electronic discussions, several initiatives among the participants to write papers (both expository and with new results) and follow up collaborations engaging people from the workshops along with other collaborators who were not able to attend.