Description of Research in the Gerstein Lab

MB&B Department, Yale University

30 June 1999

Bioinformatics:

Large-scale Comparisons of Complete Genomes and Macromolecular Structures

It is expected that very soon after the turn of the century, the human genome and the genomes of a number of other organisms, comprising billions of basepairs, will be completely sequenced. By this time the number of known structures of protein domains, which provide the primary way to interpret gene sequences in physico-chemical terms, is expected to approach 100,000. Remarkably, all of this information is expected to fit onto a couple of second-generation compact disks.[1] Interpreting it will require new approaches.

Broadly, our goal is to prepare for and take advantage of this information deluge, by undertaking large-scale analyses of the expanding number of sequences and structures. It is hoped that these analyses will allow us to address a number of overall statistical questions about the properties and structures of proteins, their use in the cell, and their differential distribution in different organisms. More specifically, in the past few years, we have made several important contributions related to comparative genomics (analyzing the occurrence of protein folds in microbial genomes) and macromolecular motions (developing a database framework for classifying macromolecular motions). The ongoing research program in the lab extends and expands previous work as described below.

Our work is fundamentally data-driven and very different in conception from previous computational work related to macromolecules, which mostly concentrated on describing the physical process of protein folding or predicting the 3D structure of a protein given its amino-acid sequence.[2] Furthermore, our work, which loosely falls into the emerging field of bioinformatics, is interdisciplinary in character, combining questions drawn from biology and chemistry with quantitative approaches from computer science and statistics. Because of its interdisciplinary quality, it does not fit neatly into the funding structure for either medical or computer-science research.

1 Comparative Genomics

1A Surveys of Folds and Functions in Genomes. It is believed that there is a large, but finite number of protein folds (estimated to be about 1000).[3] As whole genomes are sequenced and more structures are determined, it will become possible to characterize all the folds used in a given organism -- statistically, in the sense of a population census. This will allow us to see whether certain folds and structural features are more common in certain organisms than in others. We have carried out a number of preliminary surveys related to these questions[4] and found that a number of folds, such as TIM-barrels, recur often in every (analyzed) genome, while other folds do not occur at all in certain genomes. We are currently expanding this work to relate folds to functions, looking at whether the combined distribution of folds and functions differs between organisms. Preliminary results show that preference of enzymes for alpha/beta folds appears to be fairly universal but metazoans have more non-enzymatic small folds than unicellular organisms. We eventually hope to include the analysis of biochemical pathways, looking at the presence or absence of whole metabolic systems in particular genomes. We are also simultaneously scaling up our database technology from handling microbial genomes to handling animal ones (~2000 to >20000 proteins).

1B Overall Statistical Characterization of Proteomes, Indication of Atypical Proteins

Our genome analyses have found many overall statistical differences between proteins from different phylogenetic groups -- e.g. longer and more all-beta proteins in eukaryotes than prokaryotes. We are currently extending them to compare proteins from different environments --- e.g. comparing the occurrence of salt-bridges in thermophiles vs mesophiles. We are also collaborating with L Regan to characterize biophysically the proteins in the *M. genitalium* identified as most atypical in the statistical surveys.

1C Comprehensive Analysis of Gene Expression and Regulation

We are developing ways to integrate analysis of expression and regulatory systems into our genome comparisons. This will involve using whole-genome expression data from GeneChip (cDNA microarray) and gene fusion experiments (to be partially done in collaboration with M Snyder). For instance, we want to find the most highly expressed folds in a genome. We are also doing surveys of the occurrence of homeodomain regulatory regions in the fly genome and trying to correlate these with gene expression data from the laboratory of M Biggin.

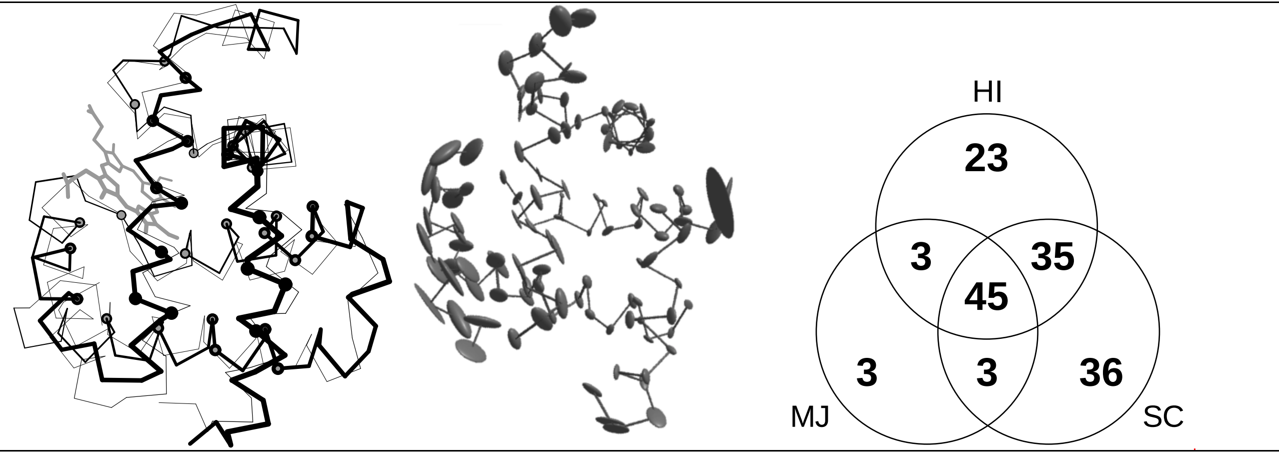
1D Building a Fold Library. It is possible to do the genomics work completely with sequence comparison.[5] However, actual 3D structures are necessary to more fully and precisely understand the elements involved (i.e domains or "modules"). This is not only because structure is the link between the gene and its chemical function but also because structure is much more conserved than sequence between different organisms.[6] In turn, to incorporate structures into the analysis of families, one needs some definition of what a fold is, a way of clustering together all structures with a given fold, and intelligent techniques for matching up sequences with unknown structure to those with known structure. We have completed work creating multiple alignments of protein structures and fusing these into consensus "fold templates".[7] A large number of templates, incorporating both sequences and structures, can be arranged into a library.[8] In the future we hope to expand such a library of fold templates into a useful biological resource, expressing each template in probabilistic terms for easy searching against other structures and sequences (using such approaches as Hidden-Markov Models or profiles[9]). Once completed, this library could become a sort of periodic table for biology, listing all the important molecular elements in an organized fashion.

2 Macromolecular Motions

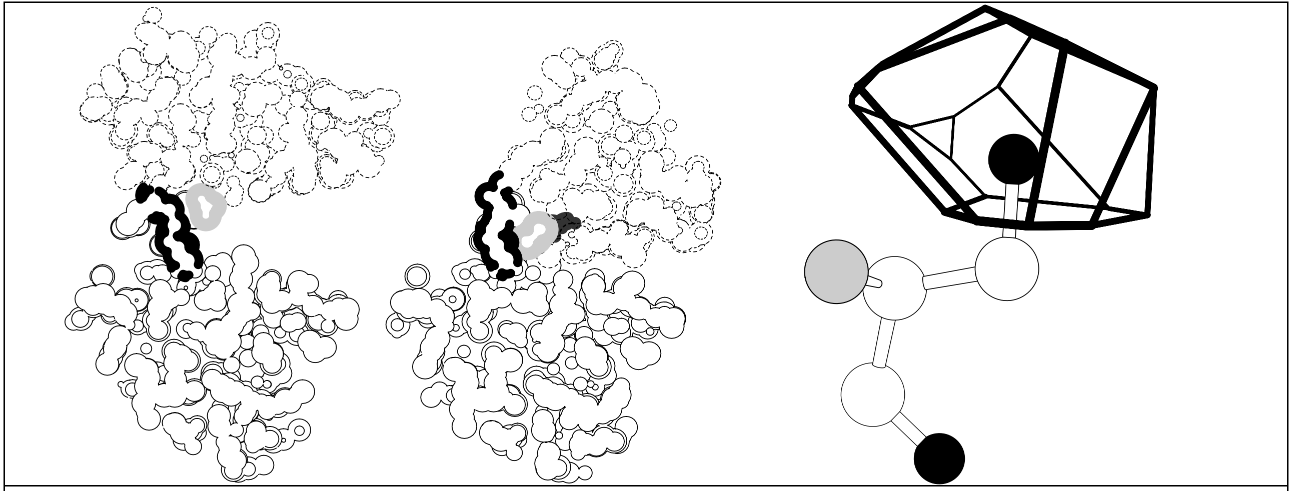
2A Motions Database. In addition to its use in genomes, one important aspect of the fold library will be its use in comprehensively surveying protein mobility and conformational variability. This occurs when two structures in the library share the same fold but still have substantial conformational differences, such as the disposition of an active site loop.

We are arranging all instances of conformational variability into a web-accessible database. Part of this project involves building a system for characterizing a protein motion in a highly standardized fashion.

2B Measurement of Packing. We have developed a simple scheme for classifying motions, concentrating on whether or not a well-packed interface between mobile elements is maintained throughout the motion.[10] This scheme works because protein interiors are packed exceedingly tightly, and this tight packing at most internal interfaces greatly constrains the way proteins can move.[11] Past research has involved measuring the packing efficiency at a few interfaces (e.g. interdomain, protein surface) using specialized geometric constructions, known as Voronoi polyhedra and Delauney Triangulations, in conjunction with limited amounts of molecular simulation.[12] We recently have developed a new parameter set for these calculations which we hope may be of general use. Future plans include developing software tools to automatically and rapidly scan the packing at a great variety of interfaces (e.g. helix-helix, interdomain, sheet-helix) and using these to enlarge the motions database and associated classification.



Analysis of Protein Families, Particularly in Genomes
LEFT shows a structural alignment of two globins. CENTER shows how a number of aligned globins can be fused into a “core structure template,” where the variability at each aligned position is represented with an “uncertainty ellipsoid.” A large number of these core structure templates could constitute a periodic table for molecular biology. RIGHT shows how this periodic table of folds could be applied to the analysis of genomes. In this subfigure, the number of common and unique folds in representatives genomes from each of the 3 major kingdoms (eukarya, bacteria, and archaea) are indicated using a Venn diagram.



Analysis of Protein Geometry, Especially as Relates to Motions
This figure shows how packing is related to protein motions. The domain motion in lactoferrin is depicted through the use of slices through the protein’s van der Waals envelope. Notice how most of the atoms in the sliced views are clearly tightly packed. The exceptions are the atoms in the hinge region, which are highlighted with a thick black line. RIGHT shows a Voronoi polyhedra, a specialized geometric construction used to precisely measure packing efficiency.

[1] The human genome is expected to be completed before 2005 [Schuler *et al.* (1996). *Science* **274**: 547-62; Pennisi (1999) *Science* **283**: 1822]. The growth in structures is extrapolated from current statistics available from the Protein Data Bank website (http://www.pdb.bnl.gov/statistics.html) or Orengo (1994, *Curr. Opin. Struc. Biol.* **4:** 429-440). Second generation compact disks (DVD technology) are expected to hold more than 8 Gb of uncompressed information.

[2] Levitt (1982). *Ann. Rev. Biophys. Bioeng.* **11:** 251-271. McCammon & Harvey (1987) *Dynamics of Proteins and Nucleic Acids*, Cambridge UP**.** Levitt *et al.* (1997). *Ann. Rev. Biochem*. **66**: 549. Karplus & Petsko (1990). *Nature* **347:** 631-639.

[3] Lander (1996). *Science* **274:** 536-539. Chothia (1992). *Nature*. **357:** 543-544.

[4] Gerstein. *J. Mol. Biol*. **274**: 562. Gerstein & Levitt. *PNAS* **94***:* 11911. M Gerstein (1998). *Proteins* **33**: 518-534. M Gerstein (1998). *Folding & Design* **3**: 497-512.

[5] Green *et al.* (1993). *Science* **259**: 1711-1716. Green (1994). *Cur. Opin. Struc. Biol*. **4**: 404-412.

[6] Chothia & Gerstein (1997). *Nature* **385:** 579-581. Chothia & Lesk (1986). *EMBO J.* **5:** 823-826. Gibrat *et al.* (1996). *Curr. Opin. Str. Biol.* **6:** 377-385.

[7] Gerstein & Altman (1995). *J. Mol. Biol.* **251:** 161-175; Gerstein and Altman (1995). *CABIOS*. **11:** 633-644; Gerstein and Levitt (1996). *Proc. Fourth Int. Conf. on Intell. Sys. Mol. Biol.* (Menlo Park, CA, AAAI Press)**,** 59-67. Suzuki *et al.* (1994). *Nuc. Acid. Res.* **22:** 3397-3405; Suzuki & Gerstein (1995). *Proteins* **23:** 525-535; Suzuki *et al*. (1995). *Prot. Eng.* **8:** 329-338. Gerstein *et al*. (1994). *J. Mol. Biol.* **236:** 1067-1078.

[8] Schmidt *et al*. (1996). *Prot. Sci*. **6**: 246-248. M Gerstein & M Levitt (1998). *Protein Science* **7**: 445-456. M Levitt & M Gerstein (1998). *Proc. Natl. Acad. Sci. USA* **95**: 5913-5920.See http://bioinfo.mbb.yale.edu/Align. Also, see Holm & Sander (1994). *Nuc. Acid Res.* **22:** 3600-3609. Holm & Sander (1996). *Science* **273:** 595-602. Holm & Sander (1997). *Structure* **5:** 165-171.

[9] Eddy (1996). *Curr. Opin. Struc. Biol.* **6:** 361-365. Krogh *et al*. (1994). *J. Mol. Biol.* **235:** 1501-1531. Bowie & Eisenberg (1993). *Curr Opin Struct Biol*. **3:** 437-444.

[10] Gerstein & Chothia (1991). *J. Mol. Biol.* **220:** 133. Gerstein *et al.* (1993). *J. Mol. Biol.* **234:** 357-372; Gerstein *et al.* (1994). *Biochemistry* **33:** 6739-6749. Database accessible from http://bioinfo.mbb.yale.edu/ProtMotDB.

[11] Harpaz *et al.* (1994). *Structure* **2:** 641-649. Gerstein & Krebs (1998), Nuc. Acids Res. **26**:4280

[12] Tsai *et al*. *Protein Science* **6**: 2606. Gerstein & Lynden-Bell (1993). *J. Phys. Chem.* **97:** 2991-2999. Gerstein & Lynden-Bell (1993). *J. Mol. Biol.* **230:** 641-650. Gerstein *et al.* (1995). *J. Mol. Biol.* **249:** 955-966; Gerstein & Chothia (1996). *Proc. Natl. Acad. Sci.* **93:** 10167-10172.