Background: Different Perspectives in Comparing Genomes and Structures

While there is a well-established practice of comparing protein structures -- often focusing on clustering structures into "fold families," comparing them in terms of simple geometric parameters (such as packing efficiency or inter-helical angles), or on understanding their motions [22,28,37,39,41] -- there has been little emphasis on comparing structures in terms of the organisms they come from. In marked contrast, recent work on genomes has (obviously) taken such an organism-comparative perspective, grouping sequences into families and seeing which families are present in which species. In particular, this sort of work has enabled the identification of particular sequences that are conserved over especially long time scales between very different organisms, such as vertebrates and bacteria [27,43]. Also, much effort has gone into functional genomics, assigning functions to genes, distinguishing between orthologs vs. paralogs, and, finally, relating genes to metabolic pathways [4,32,38].

The neglect so far of structure in genomics is unfortunate for a number of reasons. First of all, structural units (or domains) provide the logical way to subdivide proteins. Second, structure is conserved over much longer evolutionary times than sequence, allowing one to compare very distant organisms. Conservation of structure, moreover, is related in a more direct way to sequence divergence than that of function. Finally and most importantly, structure provides the connection between 1D genome sequences and functioning chemical entities. It, thus, provides an essential point of departure for those interested in designing drugs or other agents affecting proteins.

Our Objective: Bridging these Perspectives with Structural Genomics

Our objective is to bridge these two perspectives and bring a genome-comparative approach to protein structure analysis and a protein-structure angle to genome comparison. This work falls into a new subfield that has recently been dubbed "structural genomics."

A Census of Folds

More specifically, what we want to do is to build a library of folds organizing the universe of known protein structures and then to compare genomes in terms of their usage of folds from this master parts list -- in the sense of a large-scale "census" of structures. One interesting question addressed by such a census is to what degree certain folds occur only in certain regions of the "evolutionary tree." To put it in extreme terms, can one explain the obvious differences between yeast and E. coli in terms of their having different protein folds? Alternatively, it may be that most folds occur in every genome in the same way that the genetic code and many basic biochemical pathways (such as glycolysis) are almost universally shared. Thus far, it has been only possible to answer this question anecdotally. On the one hand, the immunoglobulin fold, which is usually closely associated with the vertebrate immune system, has been found in bacteria, where it carries out a different function [29]. On the other hand, the small DNA-binding fold known as the zinc finger so far appears to be confined only to eukaryotes [3]. Through our genome comparisons, we propose to address this question in a comprehensive fashion.

Identifying Folds Unique to Pathogens -- especially T. pallidum

We hope to concentrate on identifying folds unique to pathogenic organisms. We expect this work to be of clear medical relevance in the future, with the increasing prevalence of antibiotic resistant microbes, since finding folds unique to pathogens provides clear avenues for drug design.

Genomes of a number of pathogenic organisms have recently been sequenced, and already genes have been identified that *may* be unique to them (e.g. *H. pylori, B. burgdorferi, M. tuberculosis, T. pallidum* [7,10,11,45]). These would provide a logical place to begin studies. *T. pallidum*, the syphilis spirochete, in particular, presents an interesting structural problem [11]. This organism manages to evade immune system

detection to some degree. It has been suggested that its "stealth" characteristics may be due to its having a number of special proteins on its exterior. We would like to see whether these involve any unique folds. (We plan to collaborate on this specific subproject with two of the scientists who sequenced the *T. Pallidum* genome, G Weinstock and S Norris from the University of Texas.)

Our Approach to Comparative Genomics

To perform our structure census properly, we need to cluster together the known 3D structures into a library of folds and then match up genome sequences to folds in this library. We also need a way to characterize the sequences without structural homologues in, at least, rough structural terms. This is particularly important for membrane proteins. We have tackled many of these issues in the past, and our specific plans on how to proceed in the future logically follow from these experiences.

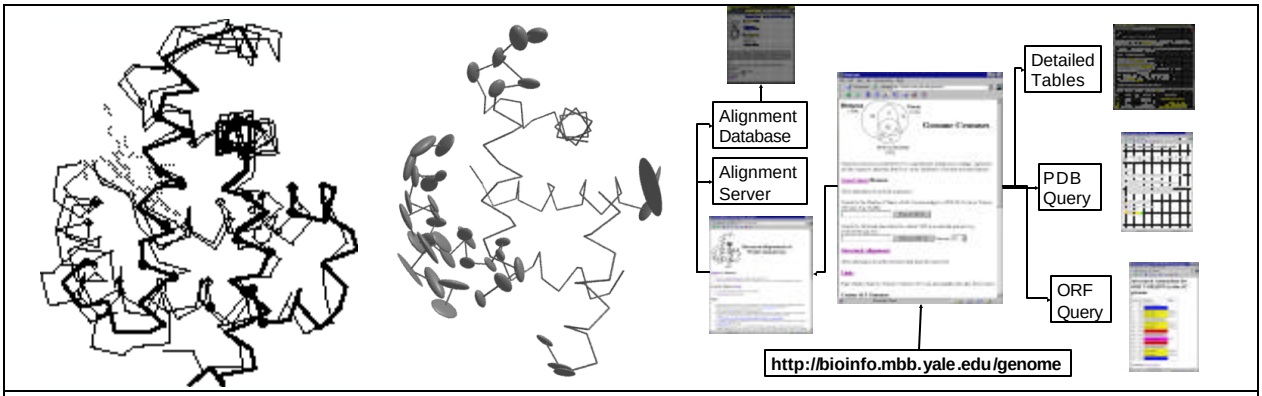


Fig. 1: Structure Alignment, Core Structures, and Database System. **LEFT** shows a structural alignment of 2 globins [25]. **MIDDLE** shows the mean positions and variability ellipsoids derived from an alignment of 8 globins [19]. Ellipsoids are only drawn around "high-variance" non-core atoms. **RIGHT** shows a schematic of the prototype structures-and-genomes database analysis system.

Step 1: Construct a Library of Folds and Match them against Genomes

A library of protein folds is expected to be an essential organizing principle in the huge but finite table of gene families, grouping together similar genes like the columns in the chemical periodic table [33]. Many groups worldwide are undertaking parallel efforts to build a fold library. We have developed preliminary versions of our library, based on a number of these other classifications (e.g. scop or FSSP [28,37]) [25,42]. In the future we hope to construct a self-contained fold library. This will require addressing three essential tasks:

(1) Alignment. We need a way of automatically comparing protein structures (see fig. 1). In the past, we developed a method to do pairwise alignments of protein structures using repeated application of dynamic programming [23,25]. This allowed structures to be aligned in a similar fashion to normal sequence alignment, in contrast to other structural alignment methods, which overlap distance matrices [28,44].

(2) P-value. We need to be able to assess the significance of a given 3D-comparison. This is often quite subtle and, in a sense, relates to the fundamental problem of what constitutes similarity in biology. We have recently developed an approach for evaluating significance based on how good a particular match is compared to one generated randomly (via a p-value) [34]. This is similar to the probabilistic schemes commonly used in sequence comparison -- e.g. in blast [30].

(3) Cores. Once all the structures in a family are aligned via statistically significant comparisons, we next want to know which regions are conserved and which are highly variable and to fuse all the conserved regions into a "core structure" template (see figure 1). We have developed a simple way to tackle this

problem through determining a mean and variance for an ensemble of multiply aligned structures and then picking the low variance atoms as "core" [18].

Sequence Comparison. Once the fold library has been built, there are a variety of ways to associate it with the genome sequences. The most straightforward approach is simply to compare each entry in the library directly against the genome sequences using traditional sequence matching programs, (e.g. blast or fasta [1,35]). Somewhat more sensitivity can be achieved through new approaches that indirectly link a query sequence to its match through a third, intermediate sequence or through some indirectly determined "property" of the sequences such as predicted secondary structure. We have recently developed some methods that accomplished this, and we hope to use them in conjunction with programs developed by others - e.g. PSI-blast [2,12,16,40]. Finally, it may be advantageous to fuse all the aligned sequences into some form of explicit consensus sequence template, such as a profile or Hidden Markov Model (HMM) [5,9], and then search with these.

Step 2: Prediction for Characterizing Sequences without a Structural Homologue

For the sequences without a clear structural homologue, we will try to characterize them in rough terms through a limited amount of structure prediction.

GOR. As we believe future improvements in secondary structure prediction will be limited, we plan to simply use an off-the-shelf method for this task, the well-established GOR program, which has an accuracy of 65% for single-sequence prediction and a somewhat higher value for multiple-sequence prediction [13].

TM-helices. In contrast, we believe that there is great room for improvement in TM-helix prediction. This is principally because of the rapidly increasing amount of structural data on membrane proteins -- e.g. the recent structures of cytochrome oxidase, potassium channel, and glycophorin A. In collaboration with J Beckwith at Harvard and D Engleman at Yale, we plan to assemble a set of TM-segments based on known structures as well as gene fusion experiments [36] and then use these to train statistical models (HMMs in particular) to recognize membrane proteins. Based on recent reports [6], we expect that membrane protein predication will be particularly useful for *T. Pallidum*.

We will use a "frequent-words" approach to assess the significance of differences in the number of predicted super-secondary structures [31].

Step 3: Results from Queries to an Integrated Database System

Our approach benefits greatly from comparing as many genomes as possible. The number of completely sequenced microbial genomes is currently >15 and rapidly increasing; we anticipate that within the five-year funding period there may be >100 genomes finished. Consequently, we plan to carry out genome annotation in as high-throughput and automated a fashion as possible, so we can rapidly integrate all the genomes into our analysis.

Database. Organizing all this data will require a sophisticated database system. We have recently received equipment grants from Informix and Intel allowing us to implement a sizeable, high-throughput system, and we have begun designing relational (and object-relational) schema to accommodate protein data [21,22]. A prototype version of our genome analysis system (configured for ~10 genomes) is available on the web (see Fig. 1). It contains 1522 tables that occupy a total of ~459 Mb. Extrapolation over the five-year funding period based on the increasing number and size of genomes plus the additional analysis we expect to do implies that our final database system will involve >25 Gb of data, a substantial scale up.

Statistics. Once the database is up and running, doing a structure census is simply a matter of executing a number of well-chosen queries cross-referencing folds and organisms. In particular, from querying the

database, we will construct Venn Diagrams, trees, and top-10 lists for the shared and most common folds in various organisms.

Biases. A most important issue in doing a large-scale survey is correcting for bias. Because of the preferences of investigators, some proteins are over-represented and others are under-represented in the databanks -- e.g. the PDB has an over-representation of globins from humans relative to those from plants. We have developed a weighting scheme that attempts to correct for this problem [26]. We have also developed resampling methods for assessing how representative the known structures are of the proteins in a complete genome [15]. These will be especially important for determining how applicable various prediction methods are to the genomes - since these methods are essentially extrapolations from the known structures.

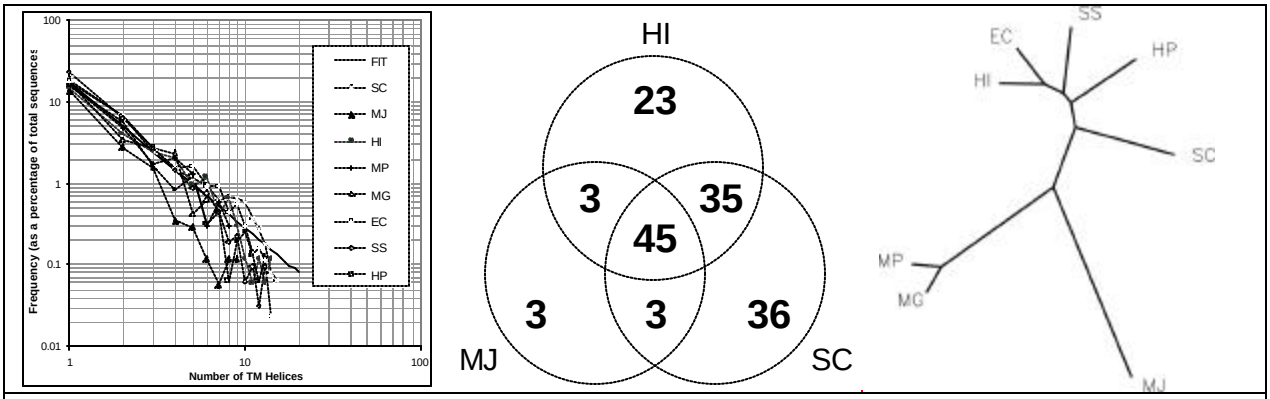
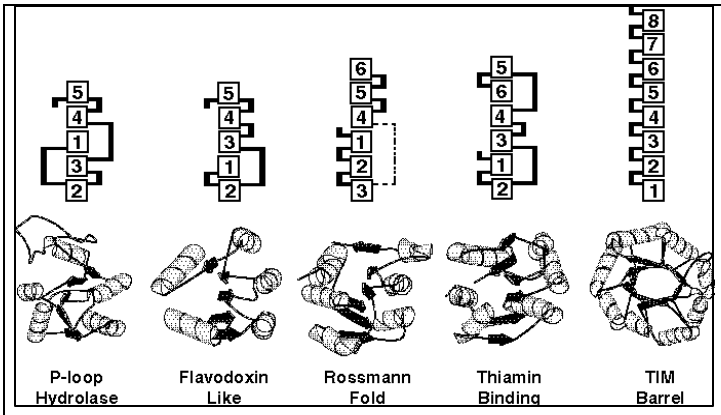


Fig. 2: Results of Initial Genome Analysis. **RIGHT**, Log-log graph showing the occurrence of membrane proteins with a given number of transmembrane (TM) helices in each of the eight genomes. The occurrence drops off in a similar fashion in all eight genomes, according to a Zipf-like law, and a fit to all eight is shown in the graph [20]. **MIDDLE,** Venn diagram showing how ~300 of the known folds are shared amongst 3 genomes [14]. **RIGHT**, Tree grouping 8 genomes together on the basis of shared folds [17]. **Genome names** are as follows: HI, *H. influenzae*; MG, *M. genitalium*; MJ, *M. jannaschii*; SS, *Synechocystis*; MP; *M. pneumoniae*; SC, *S. cerevisiae*; HP, *H. pylori*; EC, *E. coli*.

Preliminary Results: Structural Census of the First Genomes Sequenced

During the past year and a half, we have begun to assemble a prototype version of the database system and analyze the first 8 genomes sequenced [14,15,17,20,24]. Our initial results, shown in figures 2 and 3, illustrate what is possible. In particular, on the basis of shared folds, we were able to group these initial 8 genomes into a tree that is strikingly similar in topology to one based on conventional classifications. We also identified 45 ancient folds shared by the three kingdoms of life. The most common of these 45 had a remarkably similar architecture, consisting of repeated strand-helix-strand units joining adjacent strands. We were able to compare the most common folds in the yeast genome with expression level (using microarray data from Brown and colleagues [8]) and found clear differences between the most highly duplicated and most highly expressed structures.

Using structure-prediction, we found that the genomes had very similar secondary structure content even though their amino acid content differed widely. We also found that in each genome the occurrence of proteins with a given number of TM-helices falls off smoothly with increasing numbers of helices. This implies that there is no particular preference (i.e. local maximum) for proteins with 7 TM-helices and, thus, suggests that this heavily studied group of proteins is not exceptionally important in the context of microbial genomes.



Common Yeast Folds (scop)	Rep. Structure	Genome Duplication	Expression (aerobic)	Expression (anaerobic)
Protein kinases (cat. core)	1hcl	1	3	4
NTP Hydrolases with P-loop	1gky	2	1	2
Classic Zn finger	1ard	3	9	5
Ribonuclease H-like motif	2rn2	4	2	1
Rossmann Fold	1xel	5	4	3
Zn2/Cys6 DNA-binding dom.	125d	6	6	7
7-bladed beta-propeller	2bbk-H	7	8	16
TIM-barrel	1byb	8	5	6
like Ferrodoxin	1fxd	9	7	10
DNA-binding 3-helix bundle	1enh	10	30	36
…	…			
GroES-like	1lep-A	17	10	9
…	…			
like HSP70, Ct-dom.	1dkz-A	22	11	8

Fig. 3: Common Folds in the Genomes. LEFT, The "top five" folds common to all three kingdoms of life. Note how similar their super-secondary structure architecture is [14]. **RIGHT**, The top-10 folds in yeast ranked according to duplication and expression in two different conditions [17].

References

1. Altschul, S, *et al*. (1990). *J Mol Biol* **215**: 403.
2. Altschul, SF, *et al*. (1997). *Nuc Acids Res* **25**: 3389.
3. Berg, JM & Y Shi (1996). *Science* **217**: 1081.
4. Bork, P, *et al*. (1992). *Protein Science* **1**: 1677.
5. Bowie, JU, *et al*. (1991). *Science* **253**: 164.
6. Champion, CI, *et al*. (1997). *J Bacteriol* **179**: 1230.
7. Cole, ST, *et al*. (1998). *Nature* **393**: 537.
8. DeRisi, JL, *et al*. (1997). *Science* **278**: 680.
9. Eddy, SR (1996). *Curr Opin Struc Biol* **6**: 361.
10. Fraser, CM, *et al*. (1997). *Nature* **390**: 580.
11. Fraser, CM, *et al*. (1998). *Science* **281**: 375.
12. Fu, J, *et al*. (1998). *J Mol Biol* **280**: 317.
13. Garnier, J, *et al*. (1996). *Meth Enz* **266**: 540.
14. Gerstein, M (1997). *J Mol Biol* **274**: 562.
15. Gerstein, M (1998). *Folding & Design* (in press).
16. Gerstein, M (1998). *Bioinformatics* (in press).
17. Gerstein, M (1998). *Proteins* (in press).
18. Gerstein, M & R Altman (1995). *J Mol Biol* **251**: 161.
19. Gerstein, M & R Altman (1995). *CABIOS* **11**: 633.
20. Gerstein, M & H Hegyi (1998). *FEMS Microbiology Reviews* (in press).
21. Gerstein, M, *et al*. (1998). in *Rigidity theory and applications* (in press).
22. Gerstein, M & W Krebs (1998).*Nuc Acid Res* **26**:4280
23. Gerstein, M & M Levitt (1996). *ISMB* **4**: 59.
24. Gerstein, M & M Levitt (1997). *PNAS* **94**: 11911.
25. Gerstein, M & M Levitt (1998). *Prot Sci* **7**: 445.
26. Gerstein, M, *et al*. (1994). *J Mol Biol* **236**: 1067.
27. Green, P (1994). *Curr Opin Struc Biol* **4**: 404.
28. Holm, L & C Sander (1996). *Science* **273**: 595.
29. Holmgren, A & CI Branden (1989). *Nature* **342**: 248.
30. Karlin, S & SF Altschul (1993). *PNAS* **90**: 5873.
31. Karlin, S, *et al*. (1996). *Nuc Acid Res* **24**: 4263.
32. Karp, P, *et al*. (1996). *Nuc Acid Res* **24**: 32.
33. Lander, ES (1996). *Science* **274**: 536.
34. Levitt, M & M Gerstein (1998). *PNAS* **95**: 5913.
35. Lipman, DJ & W Pearson (1985). *Science* **227**: 1435.
36. Manoil, C & J Beckwith (1986). *Science* **233**: 1403.
37. Murzin, A, *et al*. (1995). *J Mol Biol* **247**: 536.
38. Mushegian, A & E Koonin (1996). *PNAS* **93**: 10268.
39. Orengo, CA, *et al*. (1994). *Nature* **372**: 631.
40. Park, J, *et al*. (1997). *J Mol Biol* **273**: 349.
41. Richards, FM (1985). *Meth Enz* **115**: 440.
42. Schmidt, R, *et al*. (1997). *Prot Sci* **6**: 246.
43. Tatusov, RL, *et al*. (1997). *Science* **278**: 631.
44. Taylor, WR & CA Orengo (1989). *J Mol Biol* **208**: 1.
45. Tomb, J-F, *et al*. (1997). *Nature* **388**: 539

Timeline

Specific Aims

Years → **1 2 3 4 5**

Build Prototype Fold Library
Apply Prototype Fold Library to ~10 Microbial Genomes
Refine Fold Library
Apply Existing Structure Prediction Methods
Develop New Methods of Structure Prediction
Apply New Structure Prediction Methods
Apply Everything on a Large-scale to >50 Microbial Genomes