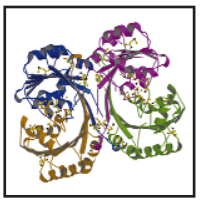
Integrative database analysis in structural genomics

Mark Gerstein

An important aspect of structural genomics is connecting coordinate data with whole-genome information related to phylogenetic occurrence, protein function, gene expression, and protein–protein interactions. Integrative database analysis allows one to survey the 'finite parts list' of protein folds from many perspectives, highlighting certain folds and structural features that stand out in particular ways.



Individual bits of genomic data need to be put in a context to be meaningful. For instance, the isolated fact that yeast gene YBR191w is expressed at a level of 65 copies per cell in microarray experiments is, by itself, meaningless. However, if one can connect this measurement to those of other genes and an overall functional classification, one can determine that this gene codes for a ribosomal protein and that ribosomal proteins have among the highest levels of expression in yeast. The same logic applies to structure. Coordinates by themselves just specify shape and are not necessarily of intrinsic biological value, unless they can be related to other information. In the past, for 'single-molecule' experiments, formal integration was unnecessary; one got the whole picture through reading the literature. However, this is impossible for all ~18,000 proteins in the worm. Thus, integrative database analysis is essential in structural genomics. Specifically, it allows one to think broadly about structure in terms of the distribution of properties of many molecules in a genome, rather than about the individual details of a particular one, and to highlight certain folds and features that stand out against this distribution. Furthermore, it potentially gives one an unbiased view of the full universe of macromolecular structure.

Database integration is of great value for companies producing propriety genome-scale datasets, as their data become more valuable when packaged with other genomic information. In particular, a number of companies offer integrated views of the human genome. Currently, these focus more on genetic rather than structural features, such as allowing one to find all the domain homologies in genes with splice variants.

Integrated database surveys (or censuses) are useful in both prospective and retrospective senses. In the former, one uses genomic information to pick targets for large-scale structure determination efforts. In the latter, one does data mining on the results of many structure determinations, trying to glean interesting statistics about a large population of structures. As illustrated in Figs 1 and 2, the main sources of information to inter-relate with structures are fold and function classifications, patterns of phylogenetic occurrence, expression data and protein–protein interactions.

Finite parts list, fold classifications and assigment

A key idea in structural genomics is that of a finite list of protein 'parts', a 'lego-kit' from which all proteins can be assembled. Parts can be defined as sequence modules, in terms of families of homologous sequences (for example, from PFAM, PROTOMAP, CDD, and COGs[1–3]) and associated structures. Alternatively, they can be defined as folds purely based on similarity of three-dimensional structure, with one fold combining a number of sequence modules. The fact that the number of folds is considerably smaller than that of modules provides a valuable simplification in interpreting complex genomic information (although there is the complication that folds can unite analogous rather than distantly homologous sequences).

There are a number of different classifications of folds, derived from manual or automatic structure comparison (for example, SCOP, CATH, FSSP[4–6]). For structural genomics, these are essential for putting individual structures into proper context in fold-space and measuring the scale of the structure data bank and its rate of increase. By one measure there are ~550 known folds (SCOP 1.50) out of an estimated total of only 1,000–10,000[7,8].

To directly cross-reference folds against genomes one needs sensitive procedures for sequence comparison with the sequences corresponding to known structures. There are a variety of techniques for this, ranging from standard and reliable pairwise comparison (such as fasta and blast[9,10]), to multiple-sequence comparison (PSI-blast and variants[11,12]), to more sensitive, though more speculative, threading methods[13,14]. One important issue in these calculations is the degree that they are biased by the incomplete nature of the structure data bank and the varying sensitivity of some comparison programs, especially the profile-based ones, which find disproportionately more homologs for certain families[15].

Phylogenetic occurrence information

If one carefully tracks the species of each sequence assigned a fold, one can use structural genomics to address certain evolutionary questions[16,17]. Are specific folds associated with particular phylogenetic groups — that is, are there metazoan-only folds? To what degree are folds shared between related organisms and does this degree of sharing parallel measures of relatedness derived from the traditional evolutionary trees? Initial analyses indicate that the sharing of folds does indeed parallel the tradi-

Department of Molecular Biophysics & Biochemistry, 266 Whitney Avenue, Yale University. PO Box 208114, New Haven, Connecticut 06520, USA.
email: Mark.Gerstein@yale.edu

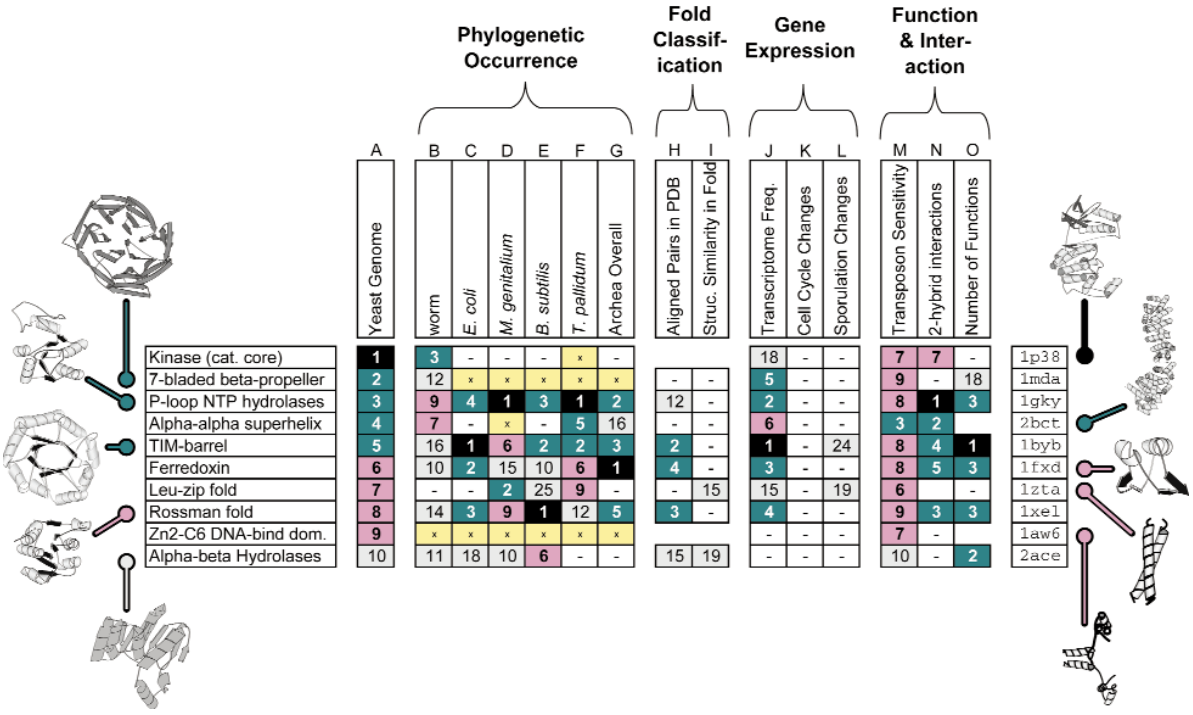


Fig. 1 An example of structural genomics data integration is shown for yeast. The figure shows the ten most common folds in the yeast genome and the rankings these have when they have been arranged according to measures other than level of genome duplication. It gives an overview of the degree to which the common parts in yeast occur in other genomes, are prevalent in the structure data bank, have many functions and interactions, and are highly expressed. In general, ranking is useful for bringing together many disparate properties of folds into a common numerical framework. In the table, the numbers and color coding indicate the rank, with black for the top rank, followed by teal for ranks between 2 and 5, pink for ranks between 6 and 9, and white for ranks between 10 and 25. Ranks >25 are just indicated by a white box containing a "-". A known "not present" (zero value) is shown as a yellow box with an "x", whereas a fold with insufficient information to be ranked according to a particular attribute is indicated by a completely empty square. Note that the precise values for the rankings are contingent on the evolving contents of various data banks. Thus, over time as more structures are determined, one should expect statistics such as the most common folds in a particular genome to change somewhat. Specific discussion of the ranking attributes in each column follows. The columns headed "phylogenetic occurrence" (B–G) show the rankings of each fold in a number of other representative genomes. (These are based on previously described PSI-blast sequence comparison of genomic sequences against the PDB[11,19,25].) The columns headed "fold classification" show some typical ways of ranking folds based on their prevalence in the structure databank (from structural alignments of proteins in ref. 28). Column H shows a rough ranking in terms of frequency in the PDB. Comparing it to columns B–G gives one a rough measure of the 'biases' in the PDB as compared to the natural occurrence of folds. Column I shows how each fold ranks against others in the data bank in terms of the overall structural similarity of the representatives of the fold. The columns headed "gene expression" show rankings of folds from weighting them either by their mRNA population (column J)[34] using data from ref. 35 or in terms of the degree to which they change in expression during a gene expression time course (either "cell cycle" or "sporulation", columns K or L). The columns headed "function & interaction" show how further functional genomics information can be integrated. Column M shows the sensitivity of the common yeast folds (that is, of all ORFs containing that fold) to an inserted transposon when yeast is grown in a specific condition[33], and column N ranks these folds in terms of the number of interactions in the two-hybrid experiment[29]. Finally, column O gives the number of functions found for the fold in a survey of the whole data bank (as defined in ref. 25). Further fold rankings are available from *bioinfo.mbb.yale.edu/partslist*.

tional tree[18]. Furthermore, one can look at the prevalence of particular folds in various organisms. Initial surveys show that the frequency of folds differs considerably among organisms but there are a few folds, such as those of TIM-barrels and P-loop hydrolases, that are common in all genomes studied[19] (Fig. 1).

While these analyses are useful retrospectively, the phylogenetic distribution of folds and sequence families is also useful prospectively in target selection (see the article by Brenner). One can choose to focus on folds and families unique to an organism or those shared among many organisms — that is, atypical or typical proteins. Straightforward, sequence-based clustering of proteins can readily identify large, shared families that represent typical proteins[1–3]. Alternatively, folds and families unique to pathogenic organisms may provide good drug targets. While speculative, this idea is partially borne out by the recent structure of OspA, a protein that has a fold unique to the pathogen *B. burgdorferi* and also functions as the antigen for a vaccine against it[20].

Functional classification and protein interactions

Integrated structural genomics analysis must include functional classification. However, there is currently no 'universal' classification, covering all functions in all organisms, that could be applied uniformly to all structures. Most of the existing schemes (such as GO, MIPS, GenProtEC, Enzyme, and COG[2,21–24]) focus on all functions in specific organisms or specific functions (such as enzyme reactions) across many different organisms. Furthermore, classifications may mean different things when they refer to function, conflating biochemical mechanism, cellular role, and phenotypic manifestation (for example 'is-hydrolase' *versus* 'in-glycolytic-pathway' *versus* 'cancer-causing'). Finally, many proteins have multiple functions and some functions require multiple proteins.

One of the greatest potential retrospective uses of structural genomics is making more precise the annotation of function. Certain folds are related to specific biochemical functions, and, broadly, certain classes of folds tend to be associated with certain

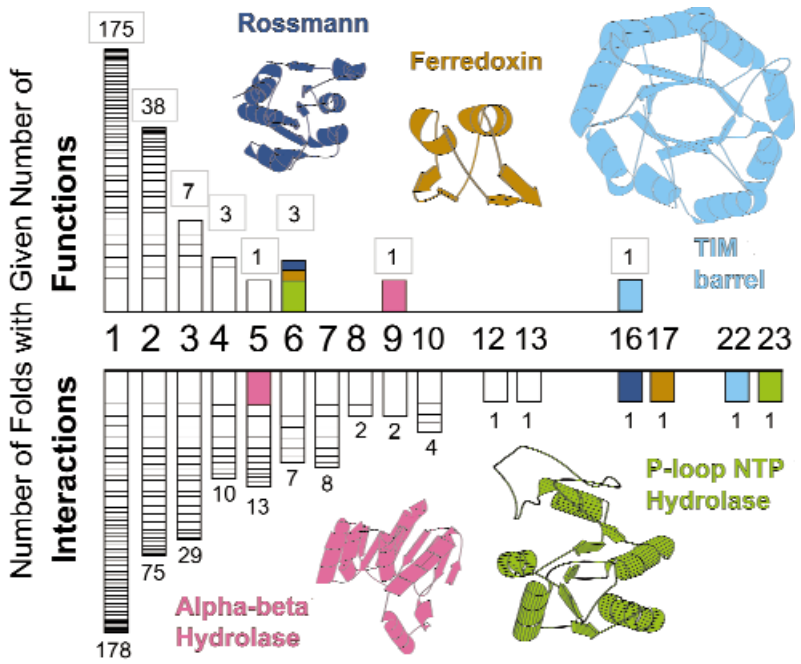


Fig. 2 Results of an integrated database analysis on the relationship between fold, function and interactions and its implications for structural genomics. The top part of the figure shows a histogram of the number of folds with a given number of functions, from previous tabulations[21,25]. Only a few folds, which are highlighted, have many functions, whereas most have only one or two functions. This has implications for data mining and function prediction based on structural genomics. In particular, it implies that if a structure with an unknown function is solved, one may be able to confidently infer function if is it is not one of the few multi-functional folds. The bottom part of the figure shows a histogram, similarly formatted to the one at top, with the number folds with a given number of interactions, where the number of interactions for a given fold is the number of other folds it interacts with in the PDB. This highlights how folds with many functions also interact with many other folds. The axes on both histograms are not drawn to scale, but exact number of functions (boxed) and interactions are listed above each bar. (For clarity, immune system folds were omitted from the histogram.) Data interactions are from Ref. 40. See *bioinfo.mbb.yale.edu/partslist* for further details.

classes of functions (for example, α/β folds with enzymes)[25,26]. Moreover, the concept of 'fold' while not perfect, is more precise than that of 'function', and there is a clearly defined relationship between the degree of similarity in sequence and the corresponding degree of similarity in structure, while the analogous relationship for function is less well understood[27,28].

One can take these ideas further and, prospectively, try to predict function given just structure (see the article by Thornton and colleagues). This is in essence a speculative extrapolation from the known fold-function relationships in the database. The existence of folds that have many functions confounds this to some degree. However, there are actually only a few multipurpose scaffolds, with most folds only having a single function, suggesting that function prediction may be realistic for a subset of proteins (Fig. 2)[21,25]. (This situation has a direct analog in day-to-day experience, where given the shape of a mechanical part one can usually, but not always, guess what it does.)

Protein function is often closely related to protein–protein interactions. The structure data bank itself and some whole-genome experiments (particularly the yeast two-hybrid[29]) now allow one to survey interactions on a large-scale and relate them to structure. Broadly, one sees patterns, such as folds that interact with many other different folds having many functions (Fig. 2). One of the most interesting questions suggested by such comprehensive data is the prediction of the entire protein–protein interaction map for an organism given all the structures in its genome. That is, can one correctly dock the structures in an organism's parts list to predict their associations?

Expression data and related functional information

An exciting new source of information is whole-genome expression data, which gives the level of expression of a particular gene in the context of all the genes in the genome (reviewed in refs 30,31). Two-dimensional gel experiments provide analogous information on cellular protein abundance[32], and for select organisms there is further related genomic information, such as the essentiality of a given gene and the subcellular localization of its protein product[23,33]. Overall these functional genomics data sets are by far the largest source of information in genomics; for yeast, they now dwarf the information in the sequence alone. Combining expression information with genome fold assignments allows one to see whether highly expressed or highly abundant proteins share particular features, which might, for instance, better stabilize them[34]. Expression time courses may also be useful for detecting and studying proteins in large complexes as well as proteins that strongly interact[35], as these often show concerted changes in expression. Finally, expression information will be useful prospectively in target selection, for highlighting proteins that may be more readily expressed and purified.

Technical issues: interconnecting databases

The most important issue in integrative database analysis and data mining is determining scientifically relevant questions to address and interesting statistics to compute. One cannot understand how to design, build, and interrelate genomic and structural databases in the abstract without a sense of the types of questions that integration can address. Furthermore, beyond conventional relational databases, robust file systems, and standard statistical techniques, there are few generic tools and approaches.

That said, one of the major practical issues confronting structural genomics today is bringing together, in the computer, many different data sets. This process differs depending on the overall architecture of the information: whether it is stored in a single centralized repository or in a federation of different resources. The former has the advantages of efficiency and uniformity and is the solution adopted by the major archival databases, GenBank and the PDB (see the article by Berman and colleagues). It clearly works well for bulk data in standardized formats, such as coordinates and sequences. However, much of the information generated by functional and structural genomics projects will be more heterogeneous, such as large-scale data sets reporting crystallizability or the binding of metabolites to protein arrays. Furthermore, it will be collected in many locations, reflecting the distributed character of biological research. It is impossible for all this information to be

kept in a single repository in a single format; rather, it will be stored in distributed resources. This federated structure has the advantage that it can harness many people in the genome-annotation effort. Moreover, it is similar in spirit to the open-source software movement, which gave rise to the popular linux operating system.

Given the federated structure of genomic information, one has the problem of database interoperability[36]. Currently, the most common interface involves reports on a single protein 'joined' together by web hyperlinks. This provides a simple and effective way of traversing multiple information sources for a single protein. However, it is ineffective for genome-scale queries. There are a variety of technologies (such as CORBA and SRS[37]) for addressing this, and a number of novel approaches for creating virtual meta-databases through which one can perform queries across many information sources. Nevertheless, at present the solution often adopted is transferring structured data files. Ideally these come in standard formats (such as XML and ASN.1) with metadata describing their contents. For effective use, all these approaches require more standardized nomenclature than we currently have, and there are a number of proposals for creating ontologies and controlled vocabularies for biological function and structure[24,38]. Specifying a 'version history' on information is also essential; in reporting the results of a database survey reproducibly one needs a way of referring to particular 'frozen' snapshots of a number of continually growing data bases.

The major information resource in science is the literature. This is often not discussed in the way data bases are, but it should be[39]. Papers are the way sequences and structures have traditionally been 'annotated'. With the advent of on-line journals and the way they can be queried in an integrated fashion (*via* PubMed), there may be little distinction between future data bases and journals, or between curators and editors.

Structure as the 'final' annotation for the genome

Structural information can and should be tightly integrated with genomic information. Now that the human genome has been sequenced, attention is turning to annotation. Considering a long-time horizon, one can see that there will be essentially an infinite amount of resources for annotating the human genome. Given this, what would one want as the 'final' annotation? Structure will undoubtedly be vital. It connects genomics with chemistry, which is invaluable for pharmaceuticals. Moreover, structural domains provide a natural way of specifying a basic unit in annotation, as the definition of modules purely in terms of conserved sequence motifs is not nearly as unambiguous and rigorous. Finally, the definition of protein fold, while not perfect, is more precise than that of function, providing a valuable reference point in annotation.

Acknowledgments

Thanks are given to M. Schultz, N. Luscombe, D. Greenbaum, J. Junker, P. Bertone, W. Krebs, P. Miller, and K. Cheung for carefully reading the draft; to S. Teichmann and J. Park for helping with protein–protein interaction numbers for the figures; and to the NIH and Keck foundation for financial support.

1. Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A. & Durbin, R. *Nucleic Acids Res.* **26**, 320-322 (1998). http://www.sanger.ac.uk/Pfam
2. Tatusov, R.L., Koonin, E.V. & Lipman, D.J. *Science* **278**, 631-637 (1997). http://www.ncbi.nlm.nih.gov/COG
3. Yona, G., Linial, N. & Linial, M. *Nucleic Acids Res* **28**, 49-55 (2000). http://protomap.stanford.edu
4. Holm, L. & Sander, C. *Nucleic Acids Res.* **22**, 3600-3609 (1994). http://www.ebi.ac.uk/dali/fssp
5. Murzin, A., Brenner, S.E., Hubbard, T. & Chothia, C. *J. Mol. Biol.* **247**, 536-540 (1995). http://scop.mrc-lmb.cam.ac.uk/scop
6. Orengo, C.A. *et al. Structure* **5**, 1093-1108 (1997). http://www.biochem.ucl.ac.uk/bsm/cath
7. Brenner, S.E., Chothia, C. & Hubbard, T.J. *Curr. Opin. Struct. Biol.* **7**, 369-376 (1997).
8. Wolf, Y.I., Grishin, N.V. & Koonin, E.V. *J. Mol. Biol.* **299**, 897-905 (2000).
9. Altschul, S., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. *J. Mol. Biol.* **215**, 403-410 (1990). http://www.ncbi.nlm.nih.gov/BLAST
10. Pearson, W.R. *J. Mol. Biol.* **276**, 71-84 (1998). http://fasta.bioch.virginia.edu
11. Altschul, S.F. *et al. Nucleic Acids Res.* **25**, 3389-3402 (1997). http://www.ncbi.nlm.nih.gov/BLAST
12. Kelley, L.A., MacCallum, R.M. & Sternberg, M.J. *J. Mol. Biol.* **299**, 523-44 (2000). http://www.bmm.icnet.uk/~3dpssm
13. Fischer, D. & Eisenberg, D. *Curr. Opin. Struct. Biol.* **9**, 208-211 (1999). http://www.doe-mbi.ucla.edu/people/frsvr/preds/MG/MG.html
14. Jones, D.T. *J. Mol. Biol.* **287**, 797–815 (1999). http://insulin.brunel.ac.uk/threader/threader.html
15. Gerstein, M. *Folding & Design* **3**, 497-512 (1998). http://bioinfo.mbb.yale.edu/genecensus
16. Gerstein, M. *J. Mol. Biol.* **274**, 562-576 (1997). http://bioinfo.mbb.yale.edu/genome/browser
17. Wolf, Y.I., Brenner, S.E., Bash, P.A. & Koonin, E.V. *Genome Res.* **9**, 17-26 (1999). ftp://ftp.ncbi.nlm.nih.gov/pub/koonin/FOLDS/index.html
18. Lin, J. & Gerstein, M. *Genome Res.* **10**, 808-818 (2000). http://bioinfo.mbb.yale.edu/genome/tree
19. Gerstein, M. *Proteins* **33**, 518-534 (1998). http://bioinfo.mbb.yale.edu/partslist
20. Li, H., Dunn, J.J., Luft, B.J. & Lawson, C.L. *Proc. Natl. Acad. Sci. USA* **94**, 3584-3589 (1997).
21. Thornton, J.M., Orengo, C.A., Todd, A.E. & Pearl, F.M. J. Mol. Biol. **293**, 333-342 (1999). http://www.biochem.ucl.ac.uk/bsm/cathwheels
22. Karp, P.D., *et al. Nucleic Acids Res.* **28**, 56-59 (2000). http://ecocyc.DoubleTwist.com/ecocyc
23. Mewes, H.W. *et al. Nucleic Acids Res.* **27**, 44-48 (1999). http://www.mips.biochem.mpg.de
24. Ashburner, M. *et al. Nature Genet.* **25**, 25-29 (2000). http://geneontology.org
25. Hegyi, H. & Gerstein, M. *J. Mol. Biol.* **288**, 147-164 (1999). http://bioinfo.mbb.yale.edu/genome/foldfunc
26. Martin, A.C. *et al. Structure* **6**, 875-884 (1998). http://www.biochem.ucl.ac.uk/bsm/cathwheels
27. Chothia, C. & Lesk, A.M. *EMBO J.* **5**, 823-826 (1986).
28. Wilson, C.A., Kreychman, J. & Gerstein, M. *J. Mol. Biol.* **297**, 233-249 (2000). http://bioinfo.mbb.yale.edu/partslist/scop
29. Uetz, P. *et al. Nature* **403**, 623-627 (2000). http://depts.washington.edu/sfields/projects/YPLM
30. Lipshutz, R.J., Fodor, S.P., Gingeras, T.R. & Lockhart, D.J. *Nature Genet* **21**, 20-24 (1999).
31. Brown, P.O. & Botstein, D. *Nature Genet.* **21**, 33-37 (1999). http://genome-www4.stanford.edu/MicroArray/SMD
32. Gygi, S.P., Rochon, Y., Franza, B.R. & Aebersold, R. *Mol Cell. Biol.* **19**, 1720-1730 (1999).
33. Ross-Macdonald, P. *et al. Nature* **402**, 413-418 (1999). http://www.yale.edu/snyder
34. Jansen, R. & Gerstein, M. *Nucleic Acids Res.* **28**, 1481-1488 (2000). http://bioinfo.mbb.yale.edu/genome/expression
35. Holstege, F.C. *et al. Cell* **95**, 717-28 (1998). http://web.wi.mit.edu/young/pub/regulation.html
36. Frishman, D., Heumann, K., Lesk, A. & Mewes, H.W. *Bioinformatics* **14**, 551-61 (1998).
37. Etzold, T., Ulyanov, A. & Argos, P. *Methods Enzymol.* **266**, 114-128 (1996). http://srs6.ebi.ac.uk
38. Westbrook, J.D. & Bourne, P.E. *Bioinformatics* **16**, 159-168 (2000). http://ndbserver.rutgers.edu/mmcif
39. Gerstein, M. *Bioinformatics* **15**, 429-431 (1999).
40. Park, J., Lappe, M., & Teichman, F. *Trends Genet.* **in the press** (2000).