Sequences and topology
Editorial overview

Mark Gerstein* and Barry Honig†

Addresses
*Department of Molecular Biophysics and Biochemistry, 266 Whitney Avenue, Yale University, PO Box 208114, New Haven, CT 06520, USA;
e-mail: Mark.Gerstein@yale.edu
†Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street, New York, NY 10032, USA;
e-mail: bh6@columbia.edu

Current Opinion in Structural Biology 2001, 11:327–329

0959-440X/01/$ – see front matter

The reviews appearing in this section of *Current Opinion in Structural Biology* deal with many of the data sources currently used in bioinformatics — genome sequences, three-dimensional structures of proteins and expression data sets. They also include a wide variety of computational approaches — sequence and structure alignment and analysis, gene-expression clustering and biophysical analysis. Broadly, the section follows the molecular biological 'data flow' from raw genome sequences to detailed structural understanding and, in the process, touches on genome annotation, integration of expression information, fold assignments, structural alignment and the understanding of protein–protein interactions.

Bioinformatics is a new field and is still in the process of being defined. It is focused on analysis of genomic and, more recently, proteomic data. The range of tools that has been brought to bear on these data is enormous and, as reflected in the reviews in this section, has its sources in different disciplines: computer science, mathematics and statistics, physics, chemistry, biochemistry and biophysics. A remarkable aspect of the growth of bioinformatics has been the speed at which the various disciplines have been integrated into the research programs of individual laboratories. For example, biophysicists have learned and even developed dynamic programming methods, whereas computer scientists have made important contributions to the analysis of three-dimensional structure. A common language is emerging and some common goals exist, while others are in the process of being defined.

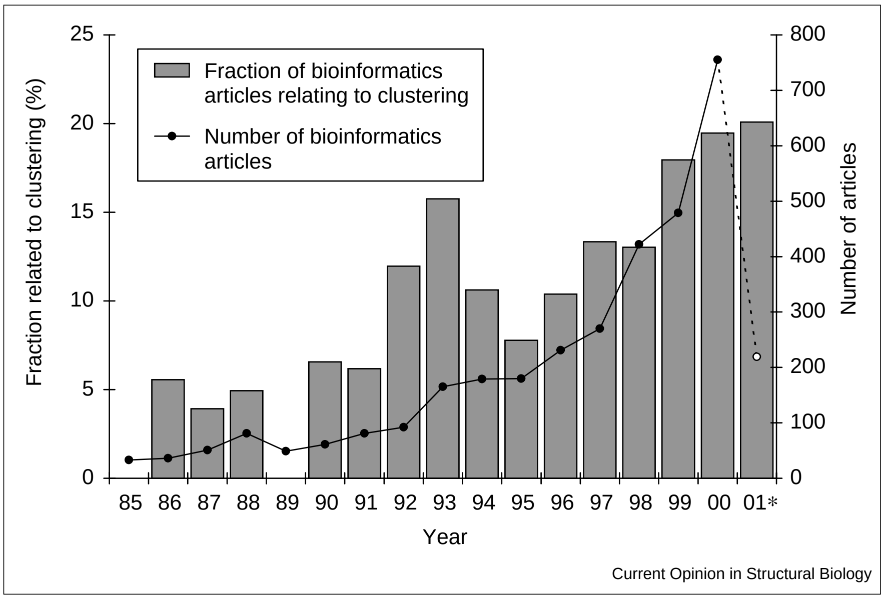
As is necessarily the case, different perspectives are evident in the articles in this section and in the literature that is reviewed. A common thread that runs through many of the reviews is the use of clustering to define biological 'parts' and then the use of these parts as frameworks for data integration and analysis. The clustering and classification of data is, of course, an essential element in the history of biology and the vast quantity of new data that has become available opens up a seemingly limitless set of possibilities for the derivation of new relationships and groupings. An alternate viewpoint emphasizes continuous aspects of biological data. We will return to the apparent 'conflict' between these two perspectives below. A second issue that arises from the emphasis on clustering involves the actual goals of the various analyses being carried out. On the one hand, the clustering of data and the derivation of new relationships is a valuable goal in its own right. On the other hand, it is not yet clear how to maximize the impact of bioinformatics on mainstream biology, which, for many years, has focused on the detailed characterization of individual systems in specific organisms. It is worth reading the reviews in this section with these questions and perspectives in mind.

Califano (pp 330–333) discusses recent trends in sequence analysis and describes how profile-based methods have, in most cases, replaced pairwise analysis methods. Profile methods of course rely either on the existence of multiple alignments or on the ability to generate them on-the-fly, and reflect major advances that have taken place in building consensus models for sequence families. A second theme discussed in this review is the combination and integration of different methods towards various goals. Examples include the identification of regulatory motifs in eukaryotes, protein structure prediction and the combined use of expression clustering and sequence data in promoter detection.

The review by Kriventseva, Biswas and Apweiler (pp 334–339) addresses the clustering and analysis of protein families in greater detail. They discuss the construction and use of protein family databases, and how a number of these are integrated in the new InterPro resource. The article also discusses clustering in structural databases and phylogenetic classification. The authors emphasize the challenge associated with the prediction of protein function and highlight the Gene Ontology consortium, whose goal is to produce a vocabulary for biological processes. It will be interesting to see the extent to which modern biology, and biologists, will be amenable to the construction of a controlled vocabulary.

Altman and Raychaudhuri (pp 340–347) discuss various applications of whole-genome expression analysis. A variety of clustering methods has been applied to microaray data and others have been developed. They describe how the clusters found from analyzing expression data can be used as a starting point for the prediction of regulatory elements, protein function, interactions and localization. In fact, their review artfully illustrates yet another application of clustering, that of clustering literature databases. They base their entire exposition on a clustering of expression analysis literature in terms of word counts.

Figure 1



Current Opinion in Structural Biology

The graph shows how the number of publications covering bioinformatics has increased over the past two decades and how the fraction of these publications devoted to 'clustering' has increased as well. These bibliometric statistics were obtained from the NCBI's PubMed database. For the first graph, showing the total number of bioinformatics publications (thin line), a query looking for either a 'computational biology' MeSH subheading or a small set of bioinformatics-only forums was used (i.e. *Bioinformatics* [previously *CABIOS*], the *Journal of Computational Biology* or the conference proceedings from ISMB or Pacific symposia). These journals and conference proceedings represent 'bioinformatics-only' forums, so one doesn't have the problem of the general increase in the number of papers in PubMed inflating the results. The results of the first query obviously understate the number of bioinformatics publications, but we believe do track some of the increase in the field. Note that the numbers for 2001 reflect that only about a third of the year had elapsed at the time of the search. For the second graph (gray columns), a query looking for the subset of matches to the first query that also contained words such as 'clustering' or the MeSH subheading 'cluster analysis' was used. The gray bars show the fraction of the second query in the first query per year. Notice how this also intelligently extrapolates a fraction for 2001.

The review by Gaasterland and Oprea (pp 377–381) discusses a fundamental problem in genomics, how to identify proteins in raw genome sequences. They provide a thoughtful discussion of the current state of the annotation of some of the larger eukaryotes, in particular, human and fly. They also discuss how the experimental evidence for verifying proposed new genes, for example, ESTs, cDNAs, microarrays and homology matches, integrates a number different data sources.

Koehl (pp 348–353) discusses various methods that have been developed to compare proteins in structural terms and highlights the difficulties in arriving at a quantitative measure of protein structure similarity. As discussed below, this is an area in which alternate viewpoints exist, in that structural classifications into discrete groupings are widely used, while there appears to be a continuous aspect to structural space as well. Structural alignments can, in many cases, produce sequence alignments that are superior to those obtained from pure sequence methods, and there is clearly much work to be done in integrating the two approaches, particularly as the amount of three-dimensional structural information continues to grow.

Teichmann, Murzin and Chothia (pp 354–363) focus on the information to be gained from large-scale structure determination, specifically structural genomics projects on model organisms. The review offers a number of interesting examples illustrating the fact that structure reveals a wealth of functional information and evolutionary relationships not available from sequence alone. The article also focuses on the use of different methods to deduce functional information. These range from the physical properties of active sites to the use of phylogeny to predict protein–protein interactions. The authors describe the assortment of methods that is being used to predict which proteins interact with one another, a problem that cannot be solved directly from structural genomics projects, which tend to focus on individual domains. An important lesson emphasized in the review involves the limitations, at least at present, of automatic methods. For example, the SCOP database relies on human decision about evolutionary relationships.

Protein–protein and protein–small molecule interactions are the focus of the review by Ma, Wolfson and Nussinov (pp 364–369), which also discusses how structure comparison can be used to study protein flexibility and plasticity. This article highlights the fact that the fold of the polypeptide chain is only one way of finding relationships between proteins; the nature of the protein surface, where interactions actually occur, is another. Far less attention has been paid to surface properties than to chain fold, but the situation is changing fairly rapidly.

Membrane proteins are, in some ways, fundamentally different than water-soluble proteins, because they exist, in part, in a nonpolar environment. Ubarretxena-Belandia and Engelman (pp 370–376) discuss some of the unique features of α-helical membrane proteins from this perspective, emphasizing, for example, the relative abundance of π-helical turns and the importance of interhelical hydrogen bonds. The α-helical membrane proteins illustrate how it is possible to generate considerable functional diversity in a single structural framework.

As mentioned above, many of the reviews in this section emphasize clustering, although the rationale and mathematical basis of the clustering process are often very different. One can get a broader sense of the importance of clustering the databases into parts than is indicated by the reviews in this section by directly analyzing the entirety of the recent bioinformatics literature, as catalogued by the National Center for Biological Information (NCBI) PubMed resource. Figure 1 shows how the prevalence of papers related to clustering and defining biological parts has increased markedly in the past few years (since 1996).

Of course, clustering and grouping data is a general mathematical issue confronted in a wide range of disciplines, from psychology to astronomy, and many generally applicable techniques, such as principal components analysis, have been developed. Some of the bioinformatics work, particularly that in expression analysis, makes explicit reference to previously developed approaches in computer science and statistics. In other areas, however, one sees more of a 'home-brew' effort, often reflecting the special structure of biological data. For instance, many of the methods in sequence clustering focus on the 'multidomain problem', which arises, for example, when domain A is inadvertently clustered with domain B because they both belong to multidomain protein AB.

In analyzing molecular biological information in terms of clusters, one confronts two very different perspectives. On one hand, there may be unique clusters suggested by the data, with clear divisions between each group. These, in turn, form natural foci for organizing databases into discrete parts and integrating information. On the other hand, clustering may be able to find groups in the data, but these may not be unique; that is, the division between two clusters may not involve crossing sharp boundaries, but rather might involve a continuous gradation. This suggests that the 'partslist' concept maybe of only limited utility.

The alternative viewpoints have emerged most clearly in the grouping of protein structures into folds. There are a number of well-established fold classifications, such as CATH, SCOP and FSSP, that divide protein structures into a very limited catalog of essential shapes [1–3]. This partslist has enabled it to serve as a basis for the integration of a rich amount of heterogeneous biological information (see, in particular, http://partslist.org [4]) and has encouraged speculation that, in total, there are only a very small number of folds (e.g. ~1000–5000). However, a number of recent papers have argued that 'fold space' is, in fact, continuous and that a description in terms of unique, discrete parts may not be possible [5,6]. It may be that both perspectives are valid and that they can be used either together, or individually, in different applications. An analogy that Janet Thornton has suggested involves the electromagnetic spectrum, which is clearly continuous, but which can be usefully divided, for many purposes, into discrete wavelength regions with somewhat nebulous boundaries between them.

A somewhat related issue arises in the definition of function. Classification schemes exist here as well and these will be expanded and integrated through efforts such as those of the Gene Ontology consortium. Nevertheless, there are limits to functional classification that are set, in part, by the limits of biological understanding. Classification can clearly help in the discovery of relationships between proteins and these are clearly extremely useful and important. On the other hand, there are unique aspects to each protein, or pathway, that will resist classification; for example, binding specificity has a discrete element by its very nature. Although the focus of much current bioinformatics research involves classification, it may be equally useful to develop tools that make it possible to bring a broad range of data and approaches to bear on the unique aspects of a particular problem, perhaps in the process allowing, in a dynamic sense, the discovery of new classifications and relationships.

References

1. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: CATH – a hierarchic classification of protein domain structures. *Structure* 1997, 5:1093-1108.

2. Murzin AG, Brenner SE, Hubbard T, Chothia C: SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995, 247:536-540.

3. Holm L, Sander C: Mapping the protein universe. *Science* 1996, 273:595-603.

4. Qian J, Stenger B, Wilson CA, Lin J, Jansen R, Teichmann SA, Park J, Krebs WG, Yu H, Alexandrov V *et al.*: PartsList: a web-based system for dynamically ranking protein folds based on disparate attributes, including whole-genome expression and interaction information. *Nucleic Acids Res* 2001, 29:1750-1764.

5. Yang AS, Honig B: An integrated approach to the analysis and modeling of protein sequences and structures. II. On the relationship between sequence and structural similarity for proteins that are not obviously related in sequence. *J Mol Biol* 2000, 301:679-689.

6. Shindyalov IN, Bourne PE: An alternative view of protein fold space. *Proteins* 2000, 38:247-260.