Molecular Databases for Protein Sequence and Structure Studies

J. A. A. Sillince and M. Sillince, Springer-Verlag, 1991, 236 pages, $69; and

Patterns in Protein Sequence and Structure*

***W. R. Taylor, editor, Springer-Verlag, 1992, 262 pages, $89**

Reviewed by James V. White
TASC, 55 Walkers Brook Drive, Reading MA 01867

Here's the amino acid sequence I'm interested in. What does this protein do? What is its structure? Is this protein in any of the databases we use? Are there other databases I should consult? Maybe the sequence is very similar to some other sequences. Well, . . . probably just a few parts of my sequence are similar to parts of other sequences. If a structure and function are known for one of these partially matching proteins, then does my protein have these properties too? Wait, . . . how reliable are predictions of structure or function from sequence similarities. And what if there aren't any very similar sequences in the databases? How reliably can structure or function be predicted from the sequence alone? These questions are difficult to answer accurately. In particular, the big problem in structural biology of predicting the structures of sequenced proteins from their amino acid sequences is unsolved.

A protein's primary amino acid sequence, together with its environment, jointly determine its three-dimensional (3-D) structure. X-ray-diffraction and NMR studies have determined crystal structures and structures in solution, respectively, for only a small fraction of the sequenced proteins. For example, a recent release of PDB (the Brookhaven Protein Structure Databank) contains 1010 protein structures. Yet this number is less than 3% of the 26,706 sequences in release 23 of the SWISS-PROT protein sequence database. Furthermore, this percentage is expected to stay low, because x-ray and NMR studies are lengthy and expensive compared to sequencing. Therefore, the problem of predicting structures directly from sequences is of considerable interest to structural biologists.

Proteins of known structure usually fall into identifiable structural classes. Furthermore, many proteins are similar to other proteins that have already been sequenced. To investigate the 3-D structure and function of a protein, a researcher typically uses a sequence-comparison program to find all sequences in an online database that are similar to the new sequence. Typically, the program assembles and aligns a set of similar sequences using some reasonable definition of similarity and some accepted test of statistical significance. If the aligned proteins are homologous (i.e., if they are descendants of a common ancestral protein), they probably share a common folding structure and a common function. The researcher does the alignment in the hope that some sequence in the aligned set has a known structure, because then conclusions may be drawn about the structure and function of the new sequence, based on the homology. However, nonhomologous proteins may appear in the set, unless stringent conditions are placed on the allowed differences between the sequences. Moreover, when these stringent conditions are imposed, many homologous proteins may be missed by the available comparison algorithms. Thus, even though it is often extremely useful, determining structure and function "by homology" may not yield the desired results.

Some proteins have the same folding structure and function and, yet, have very different sequences. When such sequences are aligned with each other using an alignment algorithm that allows for gaps and insertions, less than 30% of the residues may match in the two sequences. Therefore, researchers have searched for residue patterns in protein sequences (expressed as regular expressions or side-chain properties) that indicate particular local folding structures. Examples are supersecondary motifs like the beta hair-pin or the helix-turn-helix motif of DNA binding proteins. To find such patterns and to determine the reliability of structure prediction algorithms based on them, requires the use of protein-sequence and protein-structure databases.

Which brings us to the book, *Molecular Databases for Protein Sequence and Structure Studies*, by John A. A. Sillince and Maria Sillince. The book covers a broader range of databases than its title implies: the range includes biochemistry, molecular biology, biotechnology, and genetics. The appendix lists information on 94 data bases.

The book represents the view points of a lecturer in information systems and an expert technical librarian, as opposed to those of a structural biologist, biochemist, or computer scientist. John Sillince is listed as a Lecturer in Management Information Systems at British Sheffield University, and Maria Sillince as Assistant Subject Librarian at Wolverhampton Polytechnic, UK. According to the Introduction, they wrote the book for beginning users of molecular data bases. However, this is not a "hands-on" book. It is not a compendium of insider tips from an expert user on how to get the most out of particular data bases with the least effort. Instead, the book pro-

0006-3495/93/06/1977/02 $2.00

vides ten chapters devoted to different aspects of molecular databases. For example, Chapter 4, "Methods for Computer Representation and Registration," discusses the problems of representing data on chemicals, not proteins. Chapter 5, "Database Searching in Biochemistry and Molecular Science," focuses on patent searching and substructure searching in chemical databases. DNA and protein data bases are discussed in Chapters 7 through 9. Chapters 6 and 10 discuss potential uses of expert systems for interactive database searching and for protein sequence and structure analysis. Chapter 10, "Case Study: Specification of Expert System for Protein Structure Prediction," summarizes the results of a questionnaire the authors sent to one hundred scientists involved in predicting protein structures from sequences.

One of the book's strengths is that it explains many deficiencies in the existing databases and points to potential ways of reducing these limitations. Therefore, this book will be useful for database designers who want to know about molecular databases from various users' points of view. The authors summarize a lot of material scattered in the literature. There are 256 references; the most recent is dated 1990, and their median year of publication is 1987. Virtually all of the 33 figures (line drawings) in the book are taken from cited references.

On the debit side, the skimpy two-page index may mislead beginning users. For example, the principal protein structure databank is PDB, which most structural biologists casually call the "Brookhaven databank." However, the book's index entry under "Brookhaven" references only one page, where PDB is listed in a parenthetical statement about computer networks. A naive reader who stopped at this point, would thereby miss the important references to PDB in Chapter 8. The book is a photo-offset reproduction of a typed manuscript, which gives it an outdated appearance in this day of electronic typesetting. Another symptom of how fast our technology is changing is that only two of the 94 databases listed in the appendix have email addresses listed. At a cost of $69, the book provides about 1000 words per dollar.

Patterns in Protein Sequence and Structure by William R. Taylor (editor), contains 15 papers by active researchers on the comparison and analysis of protein sequences and structures. The papers are based on presentations given at an EMBO workshop held at EMBL (Heidelberg) near the end of 1989. To help in the preparation of this book, the authors were sent transcripts of their oral presentations. This approach should shorten its preparation. Nevertheless, the book was not published until 1992. There are 442 references; the most recent is dated 1991, and their median year of publication is 1987. The book is attractively typeset, has 88 monochrome figures (mostly original), and also provides about 1000 words per dollar. There is no index.

A central concept unifies the book: specific patterns in amino acid sequences are correlated with specific types of folding environments or local structures in proteins. Such patterns may be defined as regular expressions of allowed amino acids. They may also be defined in terms of residue properties such as hydrophobicity, charge, polarity, etc.

The contents of the book fall into several broad areas: (*a*) approaches to sequence pattern matching, comparison, and alignment; (*b*) secondary structural motifs and their relations to some known tertiary structures; (*c*) approaches to describing, comparing, and aligning tertiary structures; and (*d*) structural motifs in more complex structures (e.g., enveloped and nonenveloped virus capsids and extracellular matrix proteins).

Although the book contains introductory material and sidesteps mathematical developments, it is not aimed at newcomers. The book emphasizes the structural and functional interpretation of protein sequence patterns from the view points of structural biologists. I think most active researchers in protein structure will enjoy consulting this book and will find the effort worthwhile. In my case, I found "The Helix-Turn-Helix Motif and the Cro Repressor" by W. F. Anderson particularly interesting. This contribution shows how important the global context of a structural motif can be in determining how a protein functions.

Does this book have a serious shortcoming? Yes, there aren't enough data quantifying the accuracy of pattern-based algorithms. The field of protein structure prediction has traditionally been data starved. There haven't been enough known structures to support the statistical analyses that researchers need to establish the sensitivity and specificity of pattern-based approaches on independent data sets. Perhaps as a result of this difficulty, the accepted (or tolerated?) standards for the objective evaluation of algorithms in this field are disappointing.