Review of the *SciProtein* Molecular Modeling Program[1]

Peter R. Bergethon[†]

Department of Biochemistry, Boston University School of Medicine, 80 East Concord Street, Boston, Massachusetts 02118, and Symmetry Research, 20 Whitney Drive, Sherborn, Massachusetts 01770

Received July 8, 1997

Since the seminal work of C. Anfinsen, it has been accepted that the primary amino acid sequence of a protein directs its folding into secondary and tertiary functional form. Though elegant in its simplicity, the understanding of how this primary sequence actually directs the three-dimensional folding of a complicated structure, with literally an astronomical number of available conformations, remains a substantial practical problem and one of fundamental interest to theoretical computational chemists. With the growing knowledge of primary sequence, mostly from the success of the human genome project, this question of discovering a protein folding code that is linked to the pattern of DNA/RNA protein production is an important area of active research. However, most biological scientist are unwilling to wait for the full deciphering of this physical/chemical encryption scheme. To this end, numerous techniques have been developed in computational chemistry laboratories to assist at least with a partial prediction of structure based on primary code.

There is an ever-increasing array of commercial products designed to assist with the manipulation and exploration of primary sequence; most of these relate primary and secondary sequence structure. *SciProtein*, by SciVision (Lexington, MA), has recently brought out Version 2 of their program that engages this problem. *SciProtein* is described as a "comprehensive tool kit for proteins and peptides" and is an integrated package structured around homology search and structural prediction engines. The homology engine is capable of searching the Brookhaven database using the FASTP Algorithm and allows search of both primary and secondary structure finding elements or complete proteins with similar primary and secondary structure to a peptide under investigation. The prediction engine is based on the work of R. J. Gilbert and uses a prediction scheme utilizing a digital encoding algorithm in which a set of training proteins is used to generate a database. Based on the principle that the secondary structure is influenced by the local already formed secondary structural domains, the database is drawn from a set of known protein structures related to the function or milieu of the unknown protein. Thus secondary structure prediction proceeds based on standard statistical methods (hydropathy, etc.) that are modified and guided in a heuristic manner by the empirical database. Using properly chosen training sets, the prediction scheme can be somewhat superior to the more commonly used statistical methodologies.

SciProtein is integrated with two of the more popular PC-based programs, HyperChem and Alchemy 2000. This integration allows *SciProtein* to be used as a module in each of these programs and thus pass structures back and forth for interpretation. Thus the power of both the molecular modeling program as well as the protein sequence prediction system are mutually enhanced. Within the *SciProtein* program itself are adequate search and find engines and the ability to produce database and property maps necessary to do the common bookkeeping for the more fundamental homology and prediction searching. At the outset the combination of these internal integrated modules in *SciProtein* provide a powerful tool for the analysis, manipulation, and prediction of primary structure which is necessary in the day to day benchwork of a biochemist working in peptidyl based chemistry.

SciProtein is a Windows based system which arrives on a pair of 3 1/2" disks and is easily loaded using Windows 3.1 or Windows 95 environments. There are another seven disks which provide SciVision's specially compressed Brookhaven protein database. The database provides over 3800 structures which are then available for further analysis in the *SciProtein* program. These structures can be read directly into another program capable of reading SciVision's *.seq file format (such as Alchemy 2000) or can be translated into Brookhaven PDB format after import into Alchemy or HyperChem. *SciProtein* is easy to load, and the instruction manual walks a new user through the installation procedure with ease. Included is appropriate information that tells you how to establish the paths both from *SciProtein* to Alchemy 2000 and HyperChem and vice versa. It took me no more than a few minutes to load these programs, and they ran flawlessly. The longer aspect of installation is the loading of the Brookhaven protein database which requires the unzipping of the loaded files. Depending on the speed of your hard drive this can take a variable amount of time. I loaded the program on two separate PC based molecular modeling systems in use in our laboratory. One is a Cyrix based 166 MHz machine with a SCSI hard drive, and another is an Intel Pentium 90 MHz based machine with a standard ISA hard drive. In both cases the loading and the running of the program was easily accomplished within half an hour. The program runs smoothly on both machines, with no difficulty and no fussing with configuration or auto.exec files. This is a delight even for the experienced computer user and is a pleasant experience in the loading of scientific software which frequently needs constant configurational manipulation to get the programs to run.

The manual provides a brief but excellent tutorial that, if followed, certainly eases sophisticated use of the program within a day or two. However, for those who are resistant to reading instruction manuals, the program itself has help menus and is standard enough in terms of a graphic user interface that the user can fumble their way around the program with some success from the outset. In general, however, for the proper use of databases a particular sophisticated knowledge of how to build a training set is required. The wise user will take advantage of the manual to ease this process. In fact, this very aspect of the use of

[†] Please contact the author at Symmetry Research.

the program is fundamental to the successful professional use of *SciProtein*. Like so many graphically based systems, it is easy to quickly generate beautiful pictures without knowing the inner workings of the system. Thus a program like SciProtein can quickly generate beautiful pictures, especially when linked to a modeling system like HyperChem or Alchemy 2000. However, the prediction and homology search engines which are the real fundamental value of this program require a certain care ane excecution of scientific knowledge if useful predictions and meaningful data are to be generated. It is not a program for playing, although it is a program with which it is easy enough to play.

The program's power is substantial, and it is useful for studying a variety of problems. A shorter peptide can be searched and a secondary structure generated. However, the true test comes in evaluating a large protein and examining the entire Brookhaven database. I undertook this on my Cyrix machine, and in a homology search I examined all 3800 proteins provided in the Brookhaven database. A complete homology search, searching for secondary and primary structure motif, was accomplished in no more than 25 min. This is a substantial statistical search, and I was pleased that it could be accomplished in a single session sitting at the desk. This speaks well to the implementation of the algorithms of the system and to its file and data handling manipulation. While this experiment is a little artificial, it is comforting to know that with adequate memory and a fast enough machine such a motif search can be successfully performed. The sequence editor of the program is easy to use and allows for quick mutation of a protein as well as bookmarking and highlighting of certain sequence structures or individual amino acids of interest. The major frustration I have with the editor is that it shows a limited number of lines on the screen and a sequence of no more than about 250 amino acids can be displayed at once. However, the occasional need for such a complete display is probably rare enough that this is not a major inconvenience. The one other complaint I have with the program is that there is no direct printing implemented in the program, although it is possible to provide output by cutting and pasting to the clipboard or by an Excel DDE link. This is frustrating when you simply want to print out a hard copy of the problem that you are working on for your notebook. I am sure that in the next version this relatively small gripe will be fixed.

Overall *SciProtein* is a great value for the money and provides the day to day bench worker with a fast, efficient, user-friendly system for the examination, manipulation, and effective secondary structure prediction for a wide variety of biological proteins under common investigation today. I have now been using the program in both its first and second versions for the past several years and have found it to be a dependable and useful tool for exploration of protein structure. Though it requires a certain level of *a priori* sophistication to ask the correct questions, its developmental evolution and easy implementation with other programs suggests that it will continue to be a core tool, useful in the coming years as the ultimate goal of discovering the protein-folding encryption scheme is finally pried from nature's hands.

REFERENCES AND NOTES

(1) Gilbert, R. J. Protein Structure Prediction from Predicted Residue Properties Utilizing a Digital Encoding Algorithm. J. *Mol. Graphics* **1992**, *10*, 112−119.

CI970345+

S0095-2338(97)00345-4