CABIOS

Vol.10 no 2 1994
Pages 121 – 132

SEQSEE: a comprehensive program suite for protein sequence analysis

David S.Wishart, Robert F.Boyko, Leigh Willard, Frederic M.Richards[1] and Brian D.Sykes[2]

Abstract

SEQSEE (SEQuence SEEker) is a multi-purpose, menu-driven suite of programs designed to provide a fully integrated, state-of-the-art package for the analysis and display of protein sequences and protein databases. It is currently configured to run on most UNIX-based machines including Sun, SGI and NeXT workstations with conversion to other architectures (e.g. Vax or Cray) being a relatively simple task. SEQSEE is capable of performing nearly all of the analytical and comparative tasks found in most comprehensive commercially available software packages. These include sequence/database searching, sequence retrieval, sequence entry and editing, statistical sequence analysis, multiple sequence alignment, flexible pattern matching, and secondary structure prediction. SEQSEE also integrates a number of unique databases which allow it to perform many additional functions such as structure-based sequence alignments and homology-based secondary structure prediction. Additional enhancements to many previously published algorithms have substantially improved the performance of SEQSEE over that found for most other commercial products. The source code, the documentation and all of the required databases for SEQSEE are freely available and may be obtained by anonymous ftp.

Introduction

The past decade has seen an explosion in the use of computers in molecular biology. This is in no small part due to the tremendous success that 'gene cloners' have had at performing simple, yet important, computer experiments—experiments that have revealed many hundreds of unexpected relationships for thousands of newly sequenced gene products (Doolittle, 1987). The success shared by molecular biologists in these endeavors has induced others, including X-ray crystallographers, NMR spectroscopists, protein engineers and pharmaceutical chemists, to begin adapting the same molecular biology software to help answer important questions of their own. Indeed, both crystallographers and protein chemists alike are starting to write customized sequence analysis programs (incorporating multiple sequence alignments and advanced secondary structure prediction algorithms) to predict the tertiary structure of previously uncharacterized proteins (Schulz, 1988; Bazan, 1990; Niermann and Kirschner, 1990; Sali *et al.*, 1990). Others are pushing to develop more efficient methods for recognizing extremely remote protein sequence homologies (Gribskov *et al.*, 1987; Bowie *et al.*, 1991), while still others are modifying their software to identify automatically unique signature sequences and sequence motifs (Smith *et al.*, 1990). In addition to these programming developments, a host of new and very useful databases are beginning to appear in the literature, including protein secondary structure databanks (Kabsch and Sander, 1983; Kneller *et al.*, 1990), β-turn databases (Wilmont and Thornton, 1990), signature sequence databases (Bairoch, 1991) and NMR chemical shift databanks (Seavey *et al.*, 1991)—all of which are being integrated into biological software arsenals to permit even more specialized enquiries and analyses.

Quite clearly the development of sequence analysis packages and sequence/structure databases is entering a phase of very rapid expansion with many new and unforeseen applications being proposed for an increasingly diverse and substantially larger investigative population. Of course with this rapid expansion comes the usual problems of increased program specialization (i.e. single task programming), limited availability, hindered usability, restricted portability and reduced versatility—particularly for programs written for (and by) specialists. At the same time, this software expansion is leading to rapidly escalating costs for commercial software updates and increasingly longer lag-times between the period in which new algorithms are first described and then eventually incorporated into commercial packages. As a result, a rather problematic 'software stratification' is developing with the most powerful commercial software products becoming much more generalized, increasingly more expensive and much less up-to-date, while the freely available 'shareware' products are staying up-to-date but becoming far more specialized, much less versatile and, consequently, much less appealing.

Because most investigators would prefer not to be tied into integrating and updating dozens of separate 'shareware' programs supplied by a score of different programmers, many are turning to commercial vendors to supply them with fully integrated sequence software packages. Unfortunately, this prevents most investigators from being able to use (or program) the latest reported algorithms or to incorporate their own unique insights or specialized requirements into the programs they have

Protein Engineering Network of Centres of Excellence, Department of Biochemistry, University of Alberta, Edmonton, Alberta, Canada T6G 2S2 and [1]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, USA

[2]*To whom reprint requests should be sent*

on hand. Consequently, a great deal of new and important sequence information is probably being lost or misinterpreted because of this commercial software dependency. On the other hand, those choosing the 'shareware' route (in an effort to be as current as possible) find that a great deal of time is often wasted in converting dozens of different programs and a host of different formats to run on incompatible machines and incomprehensible compilers.

Clearly it would be of considerable help to the biochemical community to have an up-to-date, comprehensive, integrated protein sequence analysis package, similar to most commercial packages but freely available as 'shareware'. By making both the source code and databases available to the public at large, the problems of limited portability and restricted programmability could be simultaneously eliminated. Furthermore, such a package could serve as an excellent platform upon which other users could build, with either additional code or graphical enhancements to perform specialized tasks—thereby reducing the problems of excessive program diversification and software specialization that plague most 'shareware' products.

Because of these considerations we have undertaken to prepare just such a package. SEQSEE (SEQuence SEEker) is the result of nearly five years of development directed at preparing a freely distributable, easy-to-use, fully integrated, multi-purpose, menu-driven suite of programs for protein sequence analysis. SEQSEE has been designed with considerable flexibility in mind so as to permit the addition of new features and new algorithms when they are developed or as they are reported in the literature. SEQSEE contains many of the features for protein sequence analysis that are available on some of the most comprehensive commercially available products such as the Intelligenetics suite (Abarbanel *et al.*, 1984) and the GCG package (Devereux *et al.*, 1984). These features include rapid database searching, flexible pattern matching, multiple sequence alignment and sequence/database retrieval. SEQSEE also contains a large number of analytical and predictive programs which have been enhanced through the incorporation of several unique databases and a number of highly optimized algorithms. We believe that as a result of its comprehensive nature and complete source-code availability, SEQSEE offers a number of significant advantages over most commercial software packages and that it represents an important first step in bringing comprehensive shareware alternatives back into the field of protein sequence analysis.

System and methods

The current version of SEQSEE (v. 1.2) is written in standard C (consistent with both ANSI and Kernighan and Ritchie C). SEQSEE is compatible with most Unix-based machines including Sun 3, Sun 4 and Sun SPARCstations (operating under SunOS BSD), Silicon Graphics International workstations (using the IRIX operating system) and NeXT workstations (using the MACH operating system). Portability to other Unix-based architectures should be possible but has not yet been undertaken. Conversion of SEQSEE to run on Cray supercomputers requires only trivial modifications to integer size specifications (32 bits to 64 bits). Work is currently underway on converting SEQSEE to run on VAX, IBM and Macintosh computers.

In order to allow the widest possible distribution of SEQSEE, every effort has been made to keep the programs and I/O operations as machine-independent as possible. Consequently, SEQSEE does not offer any machine-specific graphics capability or machine-dependent windowing capacity. These enhancements (using X-Windows) may appear in later versions of the program.

Taken together, the various programs and subroutines in SEQSEE amount to >10 000 lines of extensively commented source code. If the accompanying databases, libraries, manuals and installation routines are included, the entire SEQSEE suite occupies some 7 Mbytes of memory. The protein sequence databanks (SwissProt, v. 23.0 and NBRF-PIR, v. 34.0) occupy an additional 115 Mbytes. Therefore, computers running the full suite of SEQSEE programs require at least 4 Mbytes of RAM (although 8–16 Mbytes is recommended) and at least 150 Mbytes of hard disk space to accommodate all of the programs and relevant databases. In addition to the minimum 4 Mbytes RAM requirement, SEQSEE also requires at least 16 Mbytes of 'swap-space' when running on most Unix-based machines.

Algorithms, programs and databases

A useful summary of most of SEQSEE's features is provided in Figure 1. As can be seen from this diagram, SEQSEE provides a very comprehensive set of sequence analysis tools based on a simple and consistent modular program design. Because SEQSEE is a menu-driven program, a more complete understanding of the package and its capabilities may be obtained if we look at the individual menu options and discuss these features in more detail. A sample of the SEQSEE menu is shown in Figure 2 and a brief description of each menu option is given below.

Help

The program SEQHELP contains an abridged version of the SEQSEE manual for on-line consultation. A menu is provided with a selection of various topics and accompanying descriptions. A brief tutorial is also provided.

Enter/Edit a Sequence

This menu option uses the program SEQED to allow entry, editing and storage of new (or old) sequence files. Sequences may be entered using the standard IUPAC single-letter amino acid code. Lower-case letters, upper-case letters or an arbitrary combination of both may be entered (i.e. sequence entry is case-independent). A 'sequence ruler' is presented at the top of each sequence entry line to permit quick identification of

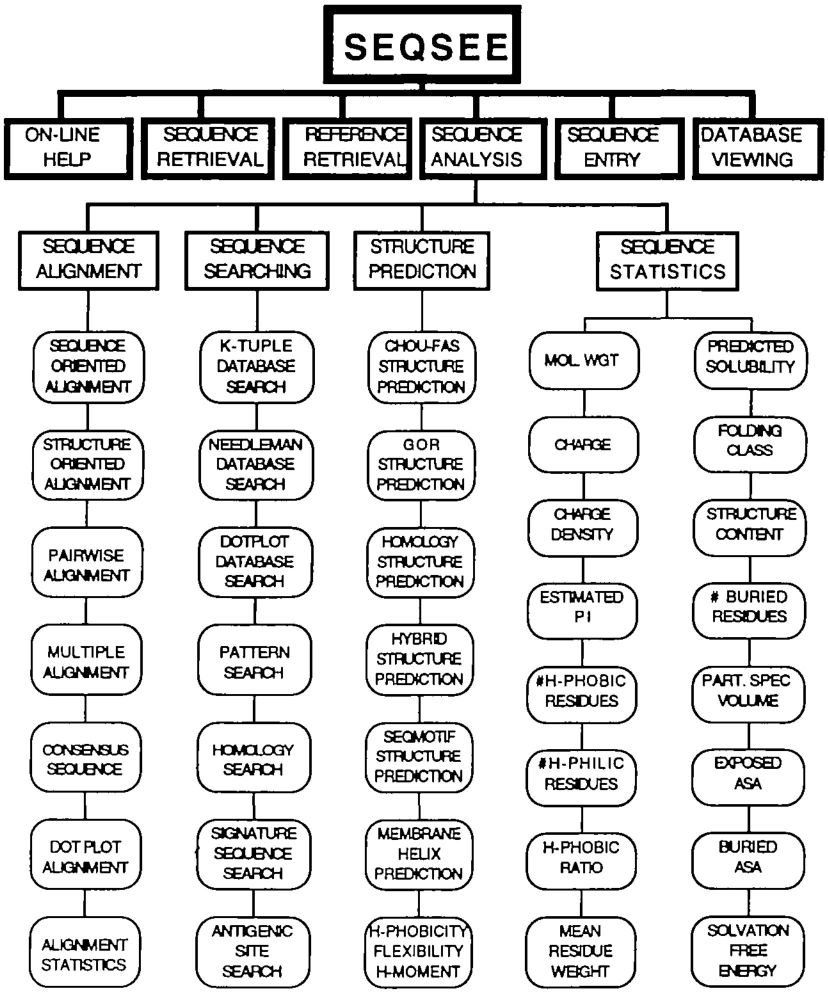


Fig. 1. A simplified overview of the program structure in SEQSEE. A modular programming architecture has been maintained throughout.

residue positions as they are typed. File editing conforms to the conventions of the 'vi' editor.

Retrieve Sequence from Database

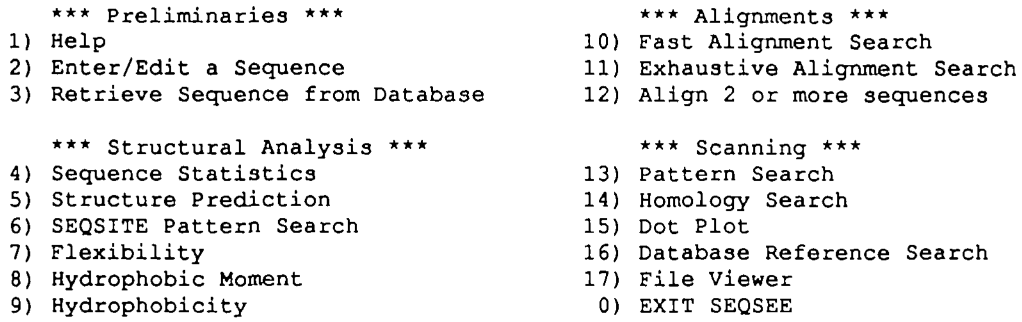
The program SEQRET allows the user to retrieve sequences or groups of sequences from either the PIR or SwissProt databases. Sequences may be retrieved either through their accession code or name (or portion thereof). Retrievals may be further 'fine-tuned' through the use of the conjunctive '&' symbol to select smaller subsets of sequences (i.e. **human & hemoglobin** selects for all human hemoglobins as opposed to all known hemoglobins). The '&' symbol is particularly useful for extracting sequences where a potential naming ambiguity may exist in a given sequence databank (not an infrequent occurrence).

Sequence Statistics

Uses the program STATS to conduct a statistical analysis of any given protein sequence. This program calculates and displays molecular weight, amino acid composition, average hydropathy (Kyte and Doolittle, 1982), total charge, predicted isoelectric point, expected quantity of exposed and interior surface area (Chothia, 1976; Richards, 1977; Miller *et al.*, 1987), expected packing volume (Richards, 1977; Janin, 1979), predicted specific volume (Zamayatnin, 1972), expected protein radius (Creighton, 1984), aggregation potential (Fisher, 1964), estimated solvation free energy of folding (Chiche *et al.*, 1990) and a host of other values that may be of structural or statistical interest. Some formulae for surface area, volume and solvation free energy have been adjusted slightly to account for recent remeasurements of these quantities (D.S.Wishart, L.Willard and B.D.Sykes, unpublished results).

Structure Prediction

Employs the program ALEXIS to identify the extent and location of potential membrane-spanning regions, the identification of short sequence folding motifs, the prediction of the protein folding class and the prediction of secondary structure



```
     *** Preliminaries ***                        *** Alignments ***
 1)  Help                                   10)  Fast Alignment Search
 2)  Enter/Edit a Sequence                  11)  Exhaustive Alignment Search
 3)  Retrieve Sequence from Database        12)  Align 2 or more sequences

     *** Structural Analysis ***                  *** Scanning ***
 4)  Sequence Statistics                    13)  Pattern Search
 5)  Structure Prediction                   14)  Homology Search
 6)  SEQSITE Pattern Search                 15)  Dot Plot
 7)  Flexibility                            16)  Database Reference Search
 8)  Hydrophobic Moment                     17)  File Viewer
 9)  Hydrophobicity                          0)  EXIT SEQSEE
```

Fig. 2. The SEQSEE menu. There are 18 menu options, divided into four broad categories. These categories include: preliminaries (i.e. sequence editing and retrieval); structural analysis, alignments; and sequence scanning.

using five different and well-tested methods. The membrane spanning predictor uses the central point maxima technique first described by Klein *et al.* (1985). Independent tests performed by Fasman and Gilbert (1990) have shown this to be the most accurate method for intramembrane helix identification. The prediction of protein structural class is based on a modification of the correlation-coefficient method proposed by Chou and Zhang (1992). On the other hand, because no single secondary structure prediction routine has been found to be significantly superior to any other, we have decided to incorporate five of the better performing approaches. They are:

1. The method of Chou and Fasman (1974, 1978) with modifications proposed by Williams *et al.* (1987). Accuracy = 61.5% (based on a three-sate prediction on the SEQBANK database of 267 solved protein structures).
2. The method of Garnier (1978) with modifications by Gibrat *et al.* (1987). Accuracy = 64.3%.
3. Homology-based secondary structure predictions based on the proposals of Levin *et al.* (1986), as well as Levin and Garnier (1988). Accuracy = 65.9% (based on a three-state prediction using a non-homologous database of 127 proteins).
4. Hydrophobic-moment secondary structure prediction based loosely on the Fourier analysis of hydrophobicity profiles described by Eisenberg *et al.* (1984). Accuracy = 64.5%.
5. Motif-based secondary structure prediction based on the methods first proposed by Rooman and Wodak (1988, 1990, 1991). Two predictions are performed, one based on literature-derived and the other on computer-derived sequence motifs. Accuracy >80% (where predicted).

Extensive optimization and reparameterization have been performed on all of the above algorithms. A more complete description of these enhancements will be forthcoming shortly (D.S.Wishart, R.F.Boyko and B.D.Sykes, manuscript in preparation).

SEQSITE Pattern Search

The program SEQSEARCH allows the user to search any given sequence for active sites, binding sites, signature sequences, phosphorylation sites and potential antigenic sites. A library of >1000 signature sequence patterns, 50 phosphorylation sites and 20 generalized antigenic regions can be scanned when this function is invoked. All sites are identified by residue location, matched template pattern and at least one current reference.

Flexibility

The program FLEQSEE predicts the flexibility of an input sequence on the basis of the Karplus algorithm (Karplus and Schulz, 1985). The procedure determines main-chain mobility by using smoothed averages of X-ray thermal B-factors taken from ~30 highly resolved crystal structures. A choice of both 'raw' and 'scaled' values is available.

Hydrophobic moment

The MOMENT program calculates the hydrophobic moment of a sequence using the Kyte–Doolittle (1982) scale of hydrophobicity and the Fourier analysis technique of Eisenberg *et al.* (1984). Calculations are performed over a preset sequence window using a range of values specific to helical periodicities (90–120°), exterior β-strand periodicities (160–180°) and interior β-strand periodicities (0°). A choice of both 'raw' and 'scaled' values is available.

Hydrophobicity

The program HYDRO calculates the smoothed hydrophobicity (over a window of predefined length) of any given sequence using a choice of several hydrophobicity scales. The user may select from the Eisenberg consensus scale (Eisenberg *et al.*, 1984), the Kyte–Doolittle scale (Kyte and Doolittle, 1982), the Cornette scale (Cornette *et al.*, 1987) or the Parker-HPLC scale (Parker *et al.*, 1986). The Hopp–Woods antigenicity scale (Hopp and Woods, 1981) is also available for antigenicity determination. A choice of both 'raw' and 'scaled' values is offered.

Fast Alignment Search

FAST_ALIGN is a *k*-tuple alignment algorithm intended for rapid database searching of input query sequences. It is able to search and align on both the SwissProt and PIR databases as well as selected user-defined databases. FAST__ALIGN is based loosely on the speed-up protocols incorporated into Pearson and Lipman's FASTA (1988) and Altschul *et al.*'s BLAST (1990) programs. The first stage of the FAST__ALIGN algorithm involves the generation of a table of similar 3-tuples from the query sequence, using a modified scoring matrix. In the second stage, a look-up table of these 3-tuples and their respective location is prepared. In the third stage a look-up table is prepared (on the fly) of 3-tuples for every sequence in the database. The two look-up tables (one from the query sequence, the other from a database sequence) are then compared and matches are identified. The result is a one-dimensional 'spectrogram' of homologies characterized by low-level noise (poor matches) and the occasional sharp peak (a string of matches). Database sequences with sufficiently high initial scores are then selected and rigorously aligned using the Needleman–Wunsch algorithm (1970) to assess the significance of the match. Several choices of scoring matrices, including the Unity matrix, the Dayhoff PAM 250 matrix (Dayhoff *et al.*, 1983), the McLachlan matrix (McLachlan, 1971) and the RBO matrix (see below) are available. The RBO matrix is the default scoring matrix.

Exhaustive Alignment Search

NW__ALIGN is a program which conducts an exhaustive pairwise alignment of any given query sequence to all other sequences in a given database. The algorithm is based on the Needleman–Wunsch (1970) dynamic programming approach for pairwise alignment. Alignments can be done against the PIR database, the SwissProt database, the SEQBANK database (see below) or any user-defined database of appropriate format. Alignment scores can be rigorously calculated on the basis of comparisons to randomized sequence alignments as recommended by Dayhoff *et al.* (1983). NW__ALIGN is intended to serve as a rigorous search and alignment tool for identifying extremely remote sequence homologies—homologies that are often missed by faster alignment algorithms. NW__ALIGN also incorporates another program called SB__ALIGN which is capable of performing structure-based alignments using the approach of Lesk *et al.* (1986). SB__ALIGN is only called when conducting alignments against the SEQBANK database.

Align 2 or more sequences

The program MULT__ALIGN uses a modification of the pairwise Needleman–Wunsch protocol (1970) to align two or more protein sequences. The method is closely related to the progressive alignment procedure, first described by Barton and Sternberg (1987), which permits rapid and accurate multiple alignments for up to several hundred proteins. A consensus sequence is also generated for each pair-wise or multiple alignment.

Pattern Search

PSEARCH allows pattern searches to be performed on either individual sequences or against large databanks, including the PIR, SwissProt and SEQBANK databases. Sequence matches are identified according to the following Pattern Query Language:

X	match exact residue specified, where X = any amino acid
!X	match any residue except X
*	wildcard character—matches any amino acid
[XYZ]	match X 'or' Y 'or' Z
X&Y	match residues X 'and' Y no matter what the separation distance
X{2,8}Y	match X and Y if separation is between 2 and 8 residues
$**X	match X if located 2 residues from N terminus

Combinations of all query types may be constructed to prepare sequence patterns of almost any description. PSEARCH is designed to allow the user to enter several patterns at once, either on a single line (using the '&' feature) or on separate lines.

Homology Search

The HSEARCH program searches either the PIR, SwissProt, SEQBANK or a compatible user-defined database to find the most homologous matches to any given sequence. Homologies are determined according to any one of the four user-defined scoring matrices described earlier. The default scoring matrix is the RBO matrix.

Dot Plot

The DOTPLOT program produces character representations of standard dot-matrix plots (Maizel and Lenk, 1981). The low resolution of most character-defined screens prevents the incorporation of a useful graphic representation of dot-plot results and consequently a simple character representation has been incorporated to overcome this problem. DOTPLOT may be used to identify internal repeats, to conduct pairwise alignments or to perform 'medium-speed' database searches and alignments.

Database Reference Search

The program REFSCAN is designed to allow the user to locate and retrieve specific sequence references from the PIR or SwissProt databases using either the accession code, the name (or portion thereof) or a bibliographic reference. This feature allows the user quickly to access important information pertaining to the function, structure or relationship of newly sequenced proteins to other proteins in the database.

File Viewer

The BROWSE program permits the user to edit or view a variety of files while still in the SEQSEE environment. Abbreviated versions of the SwissProt and PIR databases (which provide sequence name, source and accession code only) may be viewed directly with this command. Likewise, the complete SEQBANK database may also be displayed and scrolled through at leisure. BROWSE also permits the user to edit interactively the SEQSEE control file (seqsee.parms). This allows the user to customize SEQSEE program parameters in almost any manner desired. It is the SEQSEE control file which gives SEQSEE a degree of flexibility and versatility not often found in many sequence analysis packages.

Databases

In addition to the 17 analytical functions described above, >30 different databases, sequence libraries and scoring matrices have been prepared specifically for SEQSEE. Space limitations prevent us from describing all of them but there are at least four that deserve special mention. They are described in more detail below.

SEQBANK. This unique database contains a complete listing of the names, references, accession codes, sequences and secondary structures of peptides and proteins that have had their structures determined through X-ray crystallography or NMR spectroscopy as of January 1, 1991. A total of 267 protein sequences are contained in SEQBANK with no single entry being more than 52% homologous (as measured through sequence identity) to any other entry. SEQBANK is similar to the sequence/structure databases previously described by Chou and Fasman (1974), Levitt (1976) and Kabsch and Sander (1983) except that it is at least four times larger and at least 10 years more current. The secondary structures in SEQBANK represent consensus assignments obtained through careful compilation and cross-checking of close to 1000 source files based on author assignments, Kabsch–Sander assignments (1983), Levitt assignments (1976) and visual inspection. A more complete description of the database, its preparation and some of its more important features will be published shortly. SEQBANK is used by the programs SB_ALIGN, ALEXIS, BROWSE and PSEARCH.

SEQSITE. The SEQSITE database contains a library (including bibliographic entries) of >1000 sequence motifs and signature sequences which have been identified through extensive literature and computer searches. In addition to our own compilations, additional sequence motifs have been obtained from the PROSITE database (Bairoch, 1991), the compilations of Hodgman (1989), the sequence motif dictionary of Ogiwara *et al.* (1992) and the collection described by Doolittle (1987). All of the sequence motifs have been encoded using SEQSEE's own Pattern Query Language (see PSEARCH above) and all have had at least one current bibliographic reference assigned to them. For the sake of consistency we have attempted to adopt the nomenclature of Bairoch's more fully annotated PROSITE database in naming individual sequence motifs. SEQSITE is used in the program SEQSEARCH.

SEQMOTIF1 and SEQMOTIF2. The SEQMOTIF databases are two complementary databases containing sequence-related secondary structure patterns. SEQMOTIF1 contains ~150 of the longer and more complex sequence/structure patterns found in proteins of known structure such as the helix–loop–helix domain of calcium-binding proteins (Krestinger, 1980), the helix–turn–helix motif of DNA-binding proteins (Brennan and Matthew, 1989) and the ADP-nucleotide binding fold (Wierenga *et al.*, 1986). Many of these have been derived from extensive literature and crystallographic database searches. On the other hand, SEQMOTIF2 contains >300 much shorter and far simpler sequence strings which have been found to have a high propensity for certain secondary structures. These short structure-motifs were identified through computer searches of the SEQBANK database using criteria described previously by Rooman and Wodak (1988, 1990). Both SEQMOTIF databases are used in the ALEXIS program.

The RBO Weight Matrix. This weighting or distance matrix was originally developed to serve as an alternative to the Levin matrix (Levin *et al.*, 1986) for homology-based secondary structure prediction. The Levin matrix is a highly simplified, manually derived scoring matrix which uses only a small range of numbers (−1, 0, 1 and 2) inferred by studying amino acid substitutions found in a small (<60) database of proteins. In order to develop a more effective scoring scheme with a broader range of optimal values, we adjusted the original Levin matrix to allow its values to assume any positive integer value between 0 and 21. Beginning with this scaled matrix we used a computer-based Monte Carlo procedure to modify iteratively and selectively each of the 200 values in the initial scoring matrix by adding or subtracting integer values (with the requirement that no scoring matrix entry could be less than zero). Each modified matrix was then tested for its secondary structure prediction accuracy (using our version of the Levin homology algorithm) by comparing the predicted structures with the 267 known X-ray and NMR structures found in the SEQBANK database. This process was repeated exhaustively (for >100 CPU hours on a Sun workstation) until a stable, high-scoring matrix was found. The result of this exhaustive computer optimization is the RBO Weight Matrix, shown in Figure 3. This new matrix has been found to produce consistently superior results, compared to the Levin matrix, in homology-based secondary structure prediction of the SEQBANK database (65.9% versus 65.3%). Furthermore, the RBO matrix has also been found to perform at least as well as the Dayhoff PAM 250 matrix in identifying homologous sequences during FAST_ALIGN and NW_ALIGN database

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	10																			
C	2	17																		
D	2	2	10																	
E	2	2	6	10																
F	1	0	0	0	12															
G	2	2	2	2	0	10														
H	3	1	2	2	3	2	10													
I	2	2	0	0	6	0	1	10												
K	3	2	3	3	0	2	3	0	10											
L	2	2	0	0	6	0	0	4	0	10										
M	2	2	0	0	4	0	2	4	1	8	10									
N	2	0	6	4	0	3	3	0	7	0	1	10								
P	2	0	3	0	0	2	3	0	2	0	0	2	10							
Q	2	0	4	6	0	1	4	1	4	0	1	6	2	10						
R	2	2	2	2	0	2	4	0	7	0	1	2	2	4	10					
S	5	2	2	2	0	2	1	0	2	0	0	4	3	2	3	10				
T	3	2	2	2	0	2	2	2	2	2	2	2	2	2	2	4	10			
V	2	2	0	0	3	1	1	8	0	6	4	0	0	0	0	0	2	10		
W	0	0	0	0	5	0	1	0	0	1	1	0	0	1	3	1	0	0	16	
Y	0	1	0	1	9	0	2	3	0	2	3	0	1	1	0	1	0	3	5	12

Fig. 3. The RBO amino acid exchange matrix. This is the default matrix used in all of the SEQSEE's scanning, searching and homology routines.

scans (Upton *et al.*, 1992, 1993). Because of the general applicability of this 'derived' matrix we have adopted the RBO matrix as our default matrix in all of SEQSEE's alignment and homology routines.

Implementation

SEQSEE has been developed using a modular program design with each menu item representing a separate software module. Each of the modules is self-contained and operates independently of all other modules. Likewise, all database, libraries, sequence files, parameter files and program documents are similarly maintained in separate directories or modules. All menu options can be changed through a separate, centralized control file called 'seqsee.parms'. This control file, which contains >100 adjustable I/O, database and numeric parameters, can be altered either inside or outside the SEQSEE environment. By placing all adjustable parameters in a single file we have tried to make user-customization of SEQSEE a simple and straightforward task. Such program flexibility is not often available, nor is it easily performed in most other sequence analysis packages.

SEQSEE operates with a simple, user-friendly, menu-driven interface. The user is only required to choose menu options or to answer 'plain-English' questions by typing in a number, a filename or a sequence. Each menu or sub-menu provides clear directions and easily understood examples of how to answer questions or select menu items. In other words, questions are easily understood and answers are easily provided. In addition to making all of the menu and sub-menu options as clear as possible, we have also gone to considerable effort to make the interface as consistent and uniform as possible. A typical session with SEQSEE begins with the user selecting a main menu option by typing in its menu number and pressing the 'return' key. A response then appears asking the user to select one of several sub-menu options by typing in a number. If more information is needed, a second or a third sub-menu appears and so on. At almost every point the user is provided with the option to exit from the sub-menus and to return to the main menu. From the main menu, the user may also access the on-line help facility or any other main menu option. An extensive error-handling facility also permits the user to recover quickly from almost any kind of input error.

The results from every SEQSEE sequence analysis are automatically printed to the terminal screen (although the option does exits to have the results written to a file without being previewed). All output files have a file 'header' containing data on:

1. the name of the program module that was run;
2. the date;
3. the name of the protein sequence;
4. the length of the sequence;
5. the parameters used in the sequence analysis.

Whenever the output appears on the screen the user is automatically placed into an editor (a choice of 'vi' or 'emacs' is possible) to facilitate file editing, scanning and scrolling. Whenever the editor is exited, the user is asked if he or she wishes to save the file and, if it is to be saved, what the filename should be. After responding appropriately, the user is placed into the main menu again and can immediately begin another kind of sequence analysis. A sample SEQSEE session can be found in the Appendix.

Comparative performance

The developments and enhancements contained within SEQSEE have, in large part, been a response to the perceived disadvantages in other software packages. We strongly believe that the advantages of SEQSEE's user-friendliness, I/O consistency, program availability, interface flexibility and computer compatibility—relative to other packages—are just as important as SEQSEE's advantages in program comprehensiveness and prediction accuracy. These improvements in user-friendliness are particularly evident in the consistently short training period (~15 min) required by most new users. Indeed, the most difficult part of learning to use SEQSEE appears to be learning how to use the 'vi' or 'emacs' editor.

In Table I we compare many of the protein analysis features offered by SEQSEE with those offered by a widely used commercial package (the Intelligenetics suite). A quick inspection reveals that both packages are at least qualitatively comparable. However, SEQSEE does possess certain advantages in terms of the number of available secondary structure prediction schemes (five versus two), in alignment capabilities (sequence and structure-based alignments), in statistical analyses (volume and surface area calculations) and in database searching (exhaustive searches and FASTA-like searches). These performance advantages are, however, offset by the fact that SEQSEE does not yet offer the reverse translation or DNA database searching/translating facilities found in either the Intelligenetics or GCG packages.

In Table II we provide 'run-time' data on the two main types of database searching options found in SEQSEE. These tests were performed on a Silicon Graphics Crimson workstation, a Sun SPARCstation 2 and a Sun IPC workstation. The query sequence was that of the 108 residue protein known as *Escherichia coli* thioredoxin. The database selected for searching was v. 34.0 of the NBRF-PIR database (Intelligenetics format), which contains 44 890 sequences. It is clear that FAST__ALIGN performs at a speed comparable to FASTDB and that the overall search times are typically <2 min. We have also found that the search results are essentially identical for both FASTDB and FAST__ALIGN (data not shown). As might be expected, we generally find that the Needleman–Wunsch (exhaustive alignment) approach is a factor of 40–50 times slower than either FAST__ALIGN or FASTDB. These results are consistent with what has been previously noted by Pearson and Lipman (1988). However, the speed advantage of FAST__ALIGN and other heuristic *k*-tuple approaches (such as FASTA, FASTDB and BLAST) is generally offset by their reduced ability to detect remote sequence homologies—homologies that can often be detected through the Needleman–Wunsch protocol.

The sensitivity advantage offered by the Needleman–Wunsch algorithm has proven to be particularly effective in a number of studies (which will be discussed in the next section) and we believe this particular alignment option represents one of the great 'hidden' strengths of the SEQSEE package.

Table I. A detailed program comparison between SEQSEE and the protein analysis features of the Intelligenetics suite

	SEQSEE	Intelligenetics
Runs on Sun-yes	yes	yes
Runs on SGI	yes	
Runs on NeXT	yes	
Runs on VAX		yes
PIR/Swiss-Pro compatible	yes	yes
Sequence entry	yes	yes
Sequence editing	yes	yes
Sequence retrieval	yes	yes
Reference retrieval	yes	yes
FASTA-like search	yes	yes
Exhaustive search	yes	
Pattern search	yes	yes
Homology search	yes	
Motif search	yes	yes
Pairwise alignment	yes	yes
Multiple alignment	yes	yes
Structure alignment	yes	
Dotplot	yes	yes
Hydrophobicity	yes	yes
Antigenicity	yes	yes
Flexibility	yes	
Hydrophobic moment	yes	
Membrane prediction	yes	
Folding class prediction	yes	
Structure prediction	yes	yes
No. of structure prediction	5	2
Prediction accuracy	65%	55%
No. of sequence motifs	1500	750
No. of motif databases	5	2
Structure databases	yes	
Protease analysis		yes
Reverse translation		yes
Extensive statistics	yes	

Discussion

SEQSEE has been under development for nearly five years. Over the past year it has been undergoing extensive testing in more than a dozen different labs around North America. The current version (v. 1.2) has been modified in response to the suggestions and corrections provided by our test-site collaborators. Other versions will likely follow and we are hopeful that future users will provide programming suggestions to further the package's development and to enhance its general utility.

SEQSEE has already been used in a number of structural and database-oriented investigations. Early versions of the program, incorporating both the NW__ALIGN and GOR algorithms, were used in a blind test to predict the secondary and tertiary

Table II. Performance characteristics of SEQSEE's database alignment programs as measured in CPU seconds for three kinds of Unix machines, a sample scan-time for the Intelligenetics FASTDB alignment program is included as a benchmark[a]

Computer	Time (CPU s)		
	FASTDB[b]	FAST__ALIGN[c]	NW__ALIGN[d] (no jumbling)
IRIS Crimson (Elan) 50 MHz, 32 Mbytes RAM	–	36.0	1390
Sun SPARCstation 2 40 MHz, 16 Mbytes RAM	123.5	103.2	4950
Sun SPARCstation IPC 25 MHz, 24 Mbytes RAM	–	143.1	6830

[a] The query sequence was *E.coli* thioredoxin (108 residues). It was scanned against the NBRF-PIR database (v. 34.0) which contains 44 890 sequences (13.1 million residues).
[b] As implemented in the Intelligenetics sequence analysis package (IG suite).
[c] FAST__ALIGN is a novel *k*-tuple alignment algorithm written specifically for SEQSEE.
[d] NW__ALIGN is a rigorous alignment program written especially for SEQSEE. It is based on the Needleman–Wunsch dynamic programming algorithm (1970). Scores were determined on the basis of 'raw score/sequence length' with no randomization performed on the target sequence (i.e. no jumbling). If the jumbling option is used, scan times are ~10 times longer.

structures of three proteins (mandelate racemase, CBH I and haloalkane dehalogenase) in advance of the publication of their atomic structure (Wishart and Muir, 1990; Wishart, 1991). In Figure 4(A) we present a comparison between the recently solved X-ray structure of mandelate racemase (Neidhart *et al.*, 1991) and its predicted secondary structure (Wishart and Muir, 1990). The three-state prediction is found to be 77.5% correct. In Figure 4(B) we compare the predicted secondary structure with the observed NMR structure of the 36 residue C-terminal fragment of cellobiohydrolase I (Kraulis *et al.*, 1989). In this case, the prediction is found to be 75.0% correct. Furthermore, comparisons with the recently published X-ray structure of haloalkane dehalogenase (Franken *et al.*, 1991) indicate that our original prediction was 64.5% correct (data not shown). These results indicate that SEQSEE performs at a level comparable to some the better methods for secondary structure prediction (Schulz, 1988).

More recently, SEQSEE was used in the multiple alignment of a large number of species variants of CFTR proteins which had been isolated and sequenced at the University of Toronto. Evidently, other commercially available routines failed to produce a consistent alignment (A.Dulhanty, personal communication). The multiple alignments provided by SEQSEE have been used to ascertain the structural importance of a number of conserved residues in the human CFTR sequence and to suggest possible target residues for therapeutic intervention and prenatal diagnosis of cystic fibrosis.

A

```
      MSEVLITGLR TRAVNVPLAY PVHTAVGTVG TAPLVLIDLA TSAGVVGHSY  50
pred. cbbbbbbbbb bbbbcccccc cccccccccc cbbbbbbbbb bcccbbbbbb
obs.  cccbbbbbbb bbbbbbcccc cccccccccc cbbbbbbbbb bcccbbbbbb

      LFAYTPVALK SLKQLLDDMA AMIVNEPLAP VSLEAMLAKR FCLAGYTGLI 100
pred. bbbbcccccc ccchhhhhhh hhhhhhhhhh hhcchhhhhh hhhhhhhccc
obs.  bbbccccchhh hhhhhhhhhh hhhhccccch hhhhhhhhhh cccccccchh

      RMAAAGIDMA AWDALGKVHE TPLVKLLGAN ARPVQAYDSH SLDGVKLATE 150
pred. chhhhhhhhh hhhhhhcccc hhhhhhhhhh ccbbbbbbbc ccchhhhhhh
obs.  hhhhhhhhhh hhhhhhhccc chhhhhhhcc ccbbbbbbbb ccccccchhhh

      RAVTAAELGF RAVKTKIGYP ALDQDLAVVR SIRQAVGDDF GIMVDYNQSL 200
pred. hhhhhhhhcc cbbbbbbbcc ccchhhhhhh hhhhhccccb bbbbbbcccc
obs.  hhhhhhhhcc cbbbbbbccc chhhhhhhhh hhhhhhcccc bbbbbbcccc

      DVPAAIKRSQ ALQQEGVTWI EEPTLQHDYE GHQRIQSKLN VPVQMGENWL 250
pred. cchhhhhhhh hhhcccbbbb bcccchhhhhh hhhhhhhccc bbbbbbcccc
obs.  chhhhhhhhh hhhcccbbbb bbcccccchhh hhhhhhhccc cbbbbccccc

      GPEEMFKALS IGACRLAMPD AMKIGGVTGW IRASALAQQF GIPMSSHLFQ 300
pred. cchhhhhhhh hccbbbbbcc cccccchhhh hhhhhhhhhh hccbbbbbbb
obs.  chhhhhhhhh ccccbbbbcc cccccchhhh hhhhhhhhhh cbbbbbbbch

      EISAHLLAAT PTAHWLERLD LAGSVIEPTL TFEGGNAVIP DLPGVGIIWR 350
pred. cchhhhhhhh hhhhhhhhcc bbbbbbcccb bbbbcccccc cccccchhhh
obs.  hhhhhhhhhc ccbbbbbbcc cccbbbbccb bbbccbbbcc cccccccccc

      EKEIGKYLV                                               359
pred. hhhhhhhhc
obs.  hhhhhhccc
```

B

```
      TQSHYGQCGG IGYSGPTVCA SGTTCQVLNP YYSQCL 36
pred. ccbbbccbbb bbbcccbbbb cccbbbbbcc cbbbbb
obs.  cbbbcbbbbc ccccccbbbb cccbbbbccc cbbbbb
```

Fig. 4. (A) Comparison between the predicted (pred.) and observed (obs.) secondary structure of mandelate racemase (Neidhart *et al.*, 1991). (B) Comparison between the predicted (pred.) and observed (obs.) secondary structure of cellobiohydrolase I (Kraulis *et al.*, 1989) h, helix; b, β-strand; c, coil.

One of the most successful applications of SEQSEE has been in the identification of a number of previously uncharacterized gene products from the myxoma virus genome. Previous attempts to identify an abundant 37 kDa myxoma viral protein using commercially available sequence analysis software failed to detect any significant homology to known proteins. When the same sequence was run on SEQSEE using the exhaustive alignment option (NW_ALIGN), a statistically significant homology was detected to a family of γ-interferon receptors. Subsequent tests have revealed that this protein does indeed bind γ-interferon and that this activity may explain much of the myxoma's virus unusual virulence. Complete details can be found in the paper by Upton *et al.* (1992).

More recently, SEQSEE was used in the identification of a previously unrecognized family of uracil glycosylases belonging to the Shope fibroma, vaccinia and myxoma viruses (Upton *et al.*, 1993). As before, these homologies had escaped detection until SEQSEE was employed in the analysis.

Within our own protein engineering group, SEQSEE has been used to study numerous peptide and protein sequences, including anti-freeze proteins (Sonnichsen *et al.*, 1992), leucine zippers, interleukins, antibody fragments and desmopressin analogs. These studies have been directed at identifying evolutionary relatedness, searching for functional homologies, predicting secondary structure and characterizing re-engineered protein sequences. SEQSEE has also been used in a number of other unpublished studies both inside and outside our group. In general, we have found that SEQSEE has appealed to a broad range of specialists including molecular biologists, geneticists, X-ray crystallographers, NMR spectroscopists, protein engineers and peptide chemists. On account of its widespread appeal, SEQSEE has now been networked throughout the University of Alberta's Biochemistry, Medical Microbiology and Genetics departments. Negotiations for other network installations are in progress.

Future enhancements to SEQSEE include adding a DNA sequence analysis option, converting the interface to an X-Windows format, introducing a limited graphics capability and porting the package (or parts of the package) to smaller personal computers. The fact that the complete source code to SEQSEE is now available should facilitate additional enhancements should other members of the molecular biology computing community wish to take up the cause.

Conclusion

Advances in computer architecture coupled with the explosion in databases and database sizes are making high-speed/high-capacity workstations increasingly attractive to molecular biologists and protein chemists. As a result, it is becoming increasingly important for scientists to develop molecular biology software that is compatible with the new generation of workstations. We believe that SEQSEE represents one of the few comprehensive molecular biology software packages that is both compatible and easily operable on the new workstation designs. Furthermore, unlike any other package we are aware of, SEQSEE provides molecular biologists and protein chemists with a freely available, up-to-date program suite that is easily customized and openly accessible to modification and enhancement. We believe that these advantages in compatibility, accessibility and portability, along with SEQSEE's broad range of proven analytical capabilities, will make it a very useful addition to the field of computational molecular biology.

Availability

Copies of SEQSEE and its accompanying databases may be obtained through anonymous ftp by logging into our site at:

canopus.biochem.ualberta.ca (129.128.6.158)

and using your e-mail address as the password. All of the necessary installation routines, manuals, programs and databases are located in the /pub directory. Please note that all files must be uncompressed and untarred upon receipt. Specific installation information can be found in the /install directory. SEQSEE may also be obtained by sending a single 150 Mbyte Sun or Iris compatible ¼ in. cartridge tape to the Canadian address given above. Inquiries, suggestions and complaints about SEQSEE should be directed to seqsee@procyon.biochem.ualberta.ca.

Note

After the submission of this manuscript we learned that the University of Victoria has recently succeeded in installing SEQSEE's exhaustive alignment program (NW__ALIGN) on a massively parallel computer to facilitate hypersensitive sequence searching against the PIR and Swiss Prot databases (R.Olafson, personal communication). The resulting, highly parallelized program is now capable of doing rigorous Needleman–Wunsch alignments on a 100 residue protein against the entire PIR database in a little less than 4 min. This represents a speed-enhancement over the original serial algorithm of >30-fold. The package (SEQSEEMP) has been connected to an e-mail server to facilitate sequence queries in a manner similar to the BLAST and BLAZE servers. Interested users can find out more about how to use the SEQSEEMP server by sending mail to ibailey@galaxy.gov.bc.ca.

Acknowledgements

The authors wish to thank Chris Upton, Lorne Reid, Randy Read, Mark Dalton, Rod Olafson, Ray Oomen, Alastair Muir, Mike James, Bob Parker, Ann Dulhanty, Wendy Findlay, Frank Sonnichsen, Gordon Adamson, Jianjun Wang, Key-Sun Kim and Debby Waldman for their helpful suggestions and valuable advice. Financial support by the MRC Group in Protein Structure and Function (Canada), the Protein Engineering Networks of Centres of Excellence (Canada), the Alberta Heritage Fund for Medical Research and by the National Institutes of Health (NIH Grant GM-22778) is gratefully acknowledged.

References

Abarbanel,R.M., Wieneke,P.R., Mansfield,E., Jaffe,D.A. and Brutlag,D.L. (1984) Rapid searches for complex patterns in biological molecules. *Nucleic Acids Res.*, **12**, 263–280.

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bairoch,A. (1991) PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res.*, **19**, 2241–2245.

Barton,G.J. and Sternberg,M.J.E. (1987) A strategy for the rapid multiple alignment of protein sequences. *J. Mol. Biol.*, **198**, 327–337.

Bazan,J.F. (1990) Structural design and molecular evolution of a cytokine receptor superfamily. *Proc. Natl. Acad. Sci. USA*, **87**, 6934–6938.

Bowie,J.U., Luthy,R. and Eisenberg,D. (1991) A method of identifying protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.

Brennan,R.G. and Matthew,B.W. (1989) The helix–turn–helix DNA binding motif. *J. Biol. Chem.*, **264**, 1903–1906.

Chiche,L., Gregoret,L.M., Cohen,F.E. and Kollman,P.A. (1990) Protein model structure evaluation using the solvation free energy of folding. *Proc. Natl. Acad. Sci. USA*, **87**, 3240–3243.

Chothia,C. (1976) The nature of the accessible and buried surfaces in proteins. *J. Mol. Biol.*, **105**, 1–14.

Chou,K.-C. and Zhang,C.-T. (1992) A correlation-coefficient method to predicting protein-structural classes from amino acid compositions. *Eur. J. Biochem.*, **207**, 429–433.

Chou,P.Y. and Fasman,G.D (1974) Prediction of protein conformation *Biochemistry*, **13**, 222–245.

Chou,P.Y. and Fasman,G.D. (1978) Empirical predictions of protein conformation. *Annu. Rev. Biochem.*, **47**, 251–276.

Cornette,J.L., Cease,K.B., Margalit,H., Spouge,J.L., Berzofsky,J.A. and DeLisi,C. (1987) Hydrophobicity scales and computational techniques for detecting amphipathic structure in proteins. *J. Mol. Biol.*, **195**, 659–685.

Creighton,T.E. (1984) *Proteins: Structure and Molecular Properties*. W.H.Freeman, New York.

Dayhoff,M.O., Barker,W.C and Hunt,L.T. (1983) Establishing homologies in protein sequences. *Methods Enzymol.*, **91**, 524–545.

Devereux,J., Haeberli,P. and Smithies,O (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.*, **12**, 387–395.

Doolittle,R.F. (1987) *Of URFs and ORFs: A Primer of How to Analyze Derived Amino Acid Sequences*. University Science Books, CA.

Eisenberg,D., Weiss,R.M. and Terwilliger,R.C. (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. USA*, **81**, 140–144.

Fasman,G.D. and Gilbert,W.A. (1990) The prediction of transmembrane protein sequences and their conformation: an evaluation. *Trends Biochem. Sci.*, **15**, 89–92.

Fisher,H.F. (1964) A limiting law relating the size and shape of protein molecules to their compositions. *Proc. Natl. Acad. Sci. USA*, **51**, 1285–1290.

Franken,M., Rozeboon,H.J., Kalk,K.H. and Dijkstra,B.W. (1991) The crystal structure of haloalkane dehydrogenase: an enzyme to detoxify halogenated alkanes. *EMBO J.*, **10**, 1297–1302.

Garnier,J., Ogusthorpe,D.J. and Robson,B. (1978) Analysis of the accuracy and implementation of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.*, **120**, 97–120.

Gibrat,J.F., Garnier,J. and Robson,B. (1987) Further development of protein secondary structure prediction using information theory. *J Mol. Biol.*, **198**, 425–443

Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA*, **84**, 4355–4358.

Hodgman,T.C. (1989) The elucidation of protein function by sequence motif analysis. *Comput. Applic. Biosci.*, **5**, 1–13.

Hopp,T.P. and Woods,K.R. (1981) Prediction of protein antigenic determinants from amino acid sequence. *Proc. Natl. Acad. Sci. USA*, **78**, 3824–3828.

Janin,J. (1979) Surface and inside volumes in globular proteins. *Nature*, **277**, 491–492.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features *Biopolymers*, **22**, 2577–2638.

Karplus,P.A. and Schulz,G.E. (1985) Prediction of chain flexibility in proteins. *Naturewissenschaften*, **72**, 212–213.

Klein,P., Kanehisa,M. and DeLisi,C. (1985) The detection and classification of membrane-spanning proteins. *Biochim Biophys. Acta*, **815**, 468–476.

Kneller,D.G., Cohen,F.E. and Langridge,R. (1990) Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.*, **214**, 171–182.

Kraulis,P.J., Clore,G.M., Nilges,M., Jones,T.A., Petterson,G., Knowles,J. and Gronenborn,A.M. (1989) Determination of the three-dimensional structure of the C-terminal domain of cellobiohydrolase I from *Trichoderma reesei. Biochemistry*, **28**, 7241–7250.

Krestinger,R.H. (1980) Structure and evaluation of calcium-modulated proteins. *CRC Crit. Rev. Biochem.*, **8**, 119–174.

Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.

Lesk,A.M., Levitt,M. and Chothia,C (1986) Alignment of the amino acid sequences of distantly related proteins using variable gap penalties. *Prot. Engng*, **1**, 77–78.

Levin,J.M. and Garnier,J. (1988) Improvements in a secondary structure method based on a search for local sequence homologies and its use as a model building tool. *Biochim. Biophys. Acta*, **955**, 283–295.

Levin,J.M., Robson,B. and Garnier,J. (1986) An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Lett.*, **205**, 303–308.

Levitt,M. (1976) Automatic identification of secondary structure in globular proteins. *J. Mol. Biol.*, **114**, 181–293

Maizel,J.V. and Lenk,R.P. (1981) Enhanced graphic matrix analysis of nucleic acid and protein sequences. *Proc. Natl. Acad. Sci. USA*, **16**, 7665–7669.

McLachlan,A.D. (1971) Tests for comparing related amino-acid sequences. cytochrome c and cytochrome c551. *J. Mol. Biol.*, **61**, 409–423.

Miller,S., Janin,J., Lesk,A.M. and Chothia,C. (1987) Interior and surface of monomeric proteins *J. Mol. Biol.*, **196**, 641–656.

Needleman,S.B. and Wunsch,C D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453

Neidhart,D.J., Howell,P.L., Petsko,G.A., Powers,J M., Li,R., Kenyon,G.L. and Gerlt,J.A. (1991) Mechanism of the reaction catalyzed by mandelate racemase. *Biochemistry*, **30**, 9264–9273

Niermann,T. and Kirschner,K (1990) Improving the prediction of secondary structure of 'TIM-barrel' enzymes. *Prot. Engng*, **4**, 137–147.

Ogiwara,A., Uchiyama,I., Seto,Y. and Kanehisa,M. (1992) Construction of a dictionary of sequence motifs that characterize groups of related proteins. *Prot. Engng*, **5**, 479–488.

Parker,J.M.R., Guo,D. and Hodges,R.S. (1986) New hydrophobicity scale derived from HPLC peptide retention data. *Biochemistry*, **25**, 5425–5431.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.

Richards,F.M. (1977) Areas, volumes, packing and protein structure. *Annu Rev. Biophys. Bioengng*, **6**, 151–175.

Rooman,M.J. and Wodak,S.F. (1988) Identification of predictive sequence motifs limited by protein structure database size *Nature*, **335**, 45–49.

Rooman,M.J. and Wodak,S.J. (1991) Weak correlation between predictive power of individual sequence patterns and overall prediction accuracy in proteins. *Prot. Struct. Funct. Genet.*, **9**, 68–78.

Rooman,M.J., Rodriguez,J. and Wodak,S.J. (1990) Reactions between protein sequence and structure and their significance. *J. Mol. Biol.*, **213**, 337–350.

Sali,A., Overington,J.P., Johnson,M.S. and Blundell,T L (1990) From comparisons of protein sequences and structures to protein modelling and design. *Trends Biochem. Sci.*, **15**, 235–240.

Schulz,G.E. (1988) A critical evaluation of methods for prediction of protein secondary structures. *Annu. Rev. Biophys. Biophys. Chem.*, **17**, 1–22.

Seavey,B.R., Farr,E.A., Westler,W M. and Markely,J.L. (1991) A relational database for sequence-specific protein NMR data. *J. Biomol. NMR*, **1**, 217–236.

Smith,H.O., Annau,T.M. and Chandrasegaran,S. (1990) Finding sequence motifs in groups of functionally related proteins. *Proc. Natl. Acad. Sci. USA*, **87**, 826–830.

Sonnichsen,F.D., Sykes,B.D , Chao,H and Davies,P.L. (1992) The nonhelical structure of antifreeze protein type III. *Science*, **259**, 1154–1157.

Upton,C., Mossman,K. and McFadden,G. (1992) Encoding of a homolog of the IFN-γ receptor by myxoma virus. *Science*, **258**, 1369–1372.

Upton,C., Stuart,D T. and McFadden,G. (1993) Identification of a pox virus gene encoding a uracyl DNA glycosylase. *Proc. Natl. Acad. Sci. USA*, **90**, 4518–4522.

Wierenga,R.K., Terpstra,P. and Hol,W.G.J. (1986) Prediction of the occurrence of the ADP-binding $\beta-\alpha-\beta$ fold in proteins, using an amino acid sequence fingerprint. *J. Mol. Biol.*, **187**, 101–107.

Williams,R.W., Chang,A., Juretic,D. and Loughram,S. (1987) Secondary structure predictions and medium range interactions. *Biochim. Biophys. Acta*, **916**, 200–204.

Wilmont,C.M. and Thornton,J.M. (1990) β-Turns and their distortions: a proposed new nomenclature. *Prot. Engng*, **3**, 479–493.

Wishart,D.S (1991) Investigations into the denatured states of *E.coli* thioredoxin. Ph.D. thesis, Yale University, New Haven, CT.

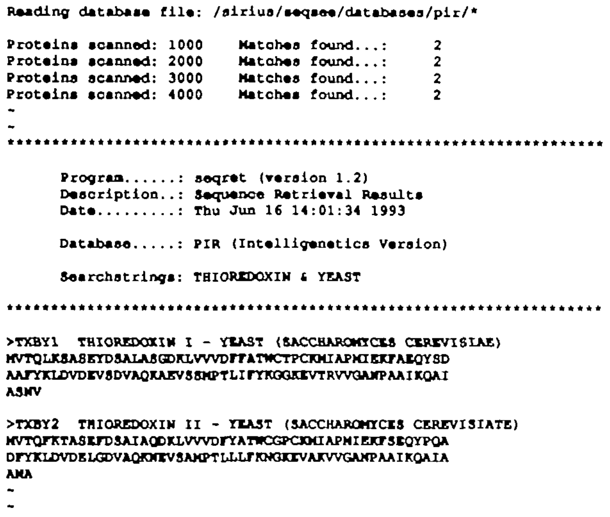
Wishart,D.S. and Muir,A.K. (1990) Protein structure prediction. In Villfranca,J.J. (ed.), *Current Research in Protein Chemistry: Techniques, Structure, Function.* Academic Press, San Diego, pp. 557–565.

Zamayatnin,A.A. (1972) Protein volume in solution. *Prog. Biophys. Mol. Biol.*, **24**, 107–123.

Appendix

The following is a sample SEQSEE session using the sequence retrieval option to select and retrieve sequences from the PIR (Intelligenetics format) database. The sequence name chosen for retrieval is yeast thioredoxin (a small, ubiquitous protein implicated in disulfide reduction and isomerization). Note that the user input is marked in bold face. Space limitations prevent us from presenting the full output for this particular session. The program is initiated by typing **seqsee**.

```
*****************************************************************
* Package...:                SEQSEE  Version 1.2 (c)           *
* Authors.. :       Robert Boyko / Leigh Willard / David Wishart *
*                          Fred Richards / Brian Sykes          *
* Location..:                University of Alberta              *
*              Protein Engineering Network of Centres of Excellence *
*****************************************************************

     *** Preliminaries ***                     *** Alignments ***
   1) Help                                  10) Fast Alignment Search
   2) Enter/Edit a Sequence                 11) Exhaustive Alignment Search
   3) Retrieve Sequence from Database       12) Align 2 or more sequences

     *** Structural Analysis ***               *** Scanning ***
   4) Sequence Statistics                   13) Pattern Search
   5) Structure Prediction                  14) Homology Search
   6) SEQSITE Pattern Search                15) Dot Plot
   7) Flexibility                           16) Database Reference Search
   8) Hydrophobic Moment                    17) File Viewer
   9) Hydrophobicity                         0) EXIT SEQSEE

   Enter the number of the desired function

>> 3                                              <--- <user input>


Seqret (Version 1.2)

How will you enter your search queries?

  1) Protein Name(s) entered from the keyboard
  2) Protein Name(s) taken from a file
  3) Protein Id(s) entered from the keyboard
  4) Protein Id(s) taken from a file
  5) Exit program

Enter a number (then press return)

>> 1                                              <--- <user input>


 Enter one search string per line.
 Use underscores instead of blanks (eg. CYSTIC_FIBROSIS).
 Use '&' symbol for conjugation (eg. FIBROSIS & CYSTIC).
 Type QUIT (then press return) when done.

 >> thioredoxin  &  yeast                         <--- <user input>

 >> quit                                          <--- <user input>
```



```
Reading database file: /sirius/seqsee/databases/pir/*

Proteins scanned: 1000    Matches found...:     2
Proteins scanned: 2000    Matches found...:     2
Proteins scanned: 3000    Matches found...:     2
Proteins scanned: 4000    Matches found...:     2
~
~
*******************************************************************

      Program......: seqret (version 1.2)
      Description..: Sequence Retrieval Results
      Date.........: Thu Jun 16 14:01:34 1993

      Database.....: PIR (Intelligenetics Version)

      Searchstrings: THIOREDOXIN & YEAST

*******************************************************************

>TXBY1  THIOREDOXIN I - YEAST (SACCHAROMYCES CEREVISIAE)
MVTQLKSASEYDSALASGDKLVVVDFFATWCTPCKMIAPMIEKFAEQYSD
AAFYKLDVDEVSDVAQKAEVSSMPTLIFYKGGKEVTRVVGANPAAIKQAI
ASNV

>TXBY2  THIOREDOXIN II - YEAST (SACCHAROMYCES CEREVISIATE)
MVTQFKTASEFDSAIAQDKLVVVDFYATWCGPCKMIAPMIEKFSEQYPQA
DFYKLDVDELGDVAQKNEVSAMPTLLLFKNGKEVAKVVGANPAAIKQAIA
ANA
~
~
```