

14 Tools for Protein Technologies

DAVID S. WISHART

Edmonton, Canada

1	Introduction	326
2	Protein Identification	326
2.1	Protein Identification from 2D Gels	327
2.2	Protein Identification from Mass Spectrometry	329
2.3	Protein Identification from Sequence Data	331
3	Protein Property Prediction	333
3.1	Predicting Bulk Properties (pI , Absorptivity, MW)	333
3.2	Predicting Active Sites	333
3.3	Predicting Modification Sites	335
3.4	Finding Protein Interactions and Pathways	335
3.5	Predicting Location or Localization	336
3.6	Predicting Stability, Globularity, and Shape	336
3.7	Predicting Protein Domains	337
3.8	Predicting Secondary Structure	337
3.9	Predicting 3D Folds (Threading)	338
3.10	Comprehensive Commercial Packages	339
4	References	342

1 Introduction

With the completion of the human genome now at hand, we will soon know the amino acid sequence of nearly every human protein. It is likely that within two more years, complete sets of protein sequences will be known for several important laboratory animals and key agricultural organisms (BRODER and VENTER, 2000; DUTT and LEE, 2000). The challenge over the coming decades will be to connect all those protein sequences with their respective actions and to translate that understanding into new approaches to manage or treat disease, to diagnose medical conditions, to monitor drug interactions, to improve crop yields, or to enhance the quality of our environment.

Key to translating this raw biological data to practical knowledge will be our ability to recognize or detect patterns that exist within these data (ATTWOOD, 2000). This is where bioinformatics comes in. Bioinformatics plays a vital role in all areas of proteomics (the study of proteins and their interactions) by providing the software tools that help sort, store, analyze, visualize, and extract important patterns from raw proteomic data. Computational tools such as correlational analysis, multiparametric fitting, dynamic programming, artificial intelligence, neural networks, and hidden Markov models are critical to teasing out many of the hidden patterns and relationships in sequence, 2D gel or mass spectrometric (MS) data. Complementing these tools are a growing array of queryable databases containing protein sequences, pre-calculated mass fragment data, 2D gel images, 3D structures, biochemical pathways, and functional sites that provide the critical “prior knowledge” necessary to extract additional information from unprocessed experimental data.

In previous chapters on protein and DNA technologies we have seen how the raw data on protein sequences, isoelectric points, and peptide mass fingerprints can be acquired. In this chapter we will see how these raw data can be transformed into useful biochemical knowledge. In particular, we will show how bioinformatics tools can be used to facilitate protein identification and characterization using 2D gel, MS, and unprocessed protein sequence

data. This chapter will be divided into two sections. The first section will be concerned with describing the software tools and algorithms that can facilitate protein identification from 2D gels, mass spectrometric, or protein sequence data. The second section will describe the bioinformatics tools and databases that may be used to predict the functions, locations, and properties of proteins once they have been identified. Particular emphasis will be placed on describing freely available Web tools or software packages that have been published in the scientific literature.

2 Protein Identification

Unfortunately for us, proteins do not come with name tags. What’s more, proteins like to hang out in crowds – usually with other proteins that look and behave almost identically. Indeed, the only way to uniquely identify a protein is to carefully separate it and painstakingly determine its sequence or precisely measure its mass. Consequently protein identification is an inherently difficult process that requires the close interplay between experimental and computational techniques (GEVAERT and VANDEKERCKHOVE, 2000). The experimental techniques provide the raw data while the computational techniques convert these raw data into a usable protein name or data bank accession number. These computational tools all rely on a common theme – i.e., they identify proteins by looking for close matches (in mass, in sequence, or in 2D gel position) to previously identified proteins. In this way protein identification is facilitated by making use of vast stores of previously accumulated knowledge about the 500,000 proteins already studied. In this section we will review three protein identification methods and their associated software tools:

- (1) identification by 2D gel spot position,
- (2) identification by mass spectrometry,
and
- (3) identification by sequence data.

2.1 Protein Identification from 2D Gels

As we have seen from Chapter 10, 2D gel electrophoresis allows for the precise and reproducible separation of up to 10,000 different proteins. The widespread use of 2D gels in functional genomics (i.e., proteomics) led to the development of some excellent software tools and a number of valuable 2D gel databases to facilitate protein identification and annotation. While most 2D gel analysis software is very image oriented, the fact that these packages can be used to measure physical properties (pI and MW) and identify proteins actually makes them a key part of the standard bioinformatics tool chest.

There are at least four major commercial programs for 2D gel analysis: Phoretix 2D, Investigator 2D, PDQuest, and Melanie 3 (Tab. 1). All four offer an impressive array of image manipulation facilities integrated into sophisticated graphical user interfaces. Three

are specific to Windows platforms (Phoretix 2D, Investigator 2D, Melanie 3) while Melanie II and PDQuest runs on both Windows and MacOS. Some commercial packages, such as Melanie (Medical ELeCtrophoresis ANalysis Interactive Expert system) began as academic projects and have been under development for many years (APPEL et al., 1988, 1997). Most of these packages make use of machine learning, heuristic clustering, artificial intelligence, and high-level image manipulation techniques to support some very complex 2D gel analyses. Several commercial packages are typically sold as part of larger equipment purchases (2D gel systems with robotic gel cutters) and are closely tied to the major proteomics or 2D gel vendors.

Essentially all commercial packages offer an array of automated or manual spot manipulations including: spot detection, spot editing, spot normalization, spot filtering, spot quantitation, and spot annotation. This allows users to compare, quantify and archive 2D gel spots quickly and accurately. A task which is particu-

Tab. 1. Protein Identification Tools – Web Links

Tool/Database	Web Address
Flicker (2D gels)	http://www.lecb.ncifcrf.gov/flicker/
Phoretix 2D	http://www.phoretix.com
PDQuest/Melanie II	http://www.proteomeworks.bio-rad.com/html/pdquest.html
Melanie 3	http://expasy.cbr.nrc.ca/melanie/
Investigator 2D	http://www.bioimage.com
2DWG (2D databases)	http://www.lecb.ncifcrf.gov/2dwgDB/
WebGel	http://www.lecb.ncifcrf.gov/webgel/
SWISS-2DPAGE	http://www.expasy.ch/
E. coli 2D database	http://pcsf.brcf.med.umich.edu/eco2dbase/
Yeast 2D database	http://www.ibgc.u-bordeaux2.fr/YPM/
PeptIdent (MS Fingerprint)	http://expasy.cbr.nrc.ca/tools/peptident.html
Profound (MS Fingerprint)	http://prowl.rockefeller.edu/cgi-bin/ProFound
Mowse (MS Fingerprint)	http://srs.hgmp.mrc.ac.uk/cgi-bin/mowse
Mascot (MS Fingerprint)	http://www.mascot.com
turboSEQUEST (MS/MS)	http://www.thermoquest.com/bioworks.html
PepSea (MS Fingerprint)	http://pepsea.protana.com/PA_PepSeaForm.html
BLAST	http://www.ncbi.nlm.nih.gov/BLAST/
PSI-BLAST	http://www.ncbi.nlm.nih.gov/blast/psiblast.cgi
Swiss-Prot Database	http://expasy.cbr.nrc.ca/sprot/
Owl Database	http://www.bioinf.man.ac.uk/dbbrowser/OWL/
PIR Database	http://www-nbrf.georgetown.edu/pirwww/pirhome.shtml
GenBank Database	http://www.ncbi.nlm.nih.gov
Protein Data Bank	http://www.rcsb.org

larly important in monitoring changes in protein expression from gel to gel or experiment to experiment. In addition to individual spot manipulation, whole gel manipulations such as rotating, overlaying, referencing, “synthesizing”, and averaging are typically supported in most commercial packages. This is done to facilitate inter-gel comparison and to calibrate gels to pI and molecular weight standards. Calibration is particularly important for 2D gels, if one wishes to extract accurate molecular weight or pI information for protein identification.

No matter how careful one is in casting or running a 2D gel, there is usually some inter-gel variability. Therefore, the ability to stretch or shrink certain gel regions (or even entire gels) is often necessary to permit direct comparisons. Techniques called spot matching, “landmarking”, and image “warping” are offered by most programs to allow this kind of forced matching. Once this kind of image transformation has been completed most commercial packages allow additional gels may be overlaid, subtracted, alternately flashed (flickered), or color contrasted to identify significant changes or significant spots.

In addition to the commercial 2D gel packages, a particularly nice freeware package known as Flicker (LIPKIN and LEMKIN, 1980; LEMKIN and THORNWALL, 1999) is also available for 2D gel image analysis and comparison (Tab. 1). While not quite as sophisticated as image manipulation, Flicker has recently been translated to Java, making it a platform independent package that runs on any Java-enabled Web browser. Flicker is quite useful for transforming (warping, rotating, etc.) and visualizing pairs of 2D gels so that the gel of interest can be easily compared to a pre-existing gel obtained over the Web. Its name comes from the fact that the program allows gel images to be flipped on and off (“flickered”) to facilitate visual comparison. When combined with other Web tools, such as WebGel and 2DWG (also developed by Lemkin), it is possible to create a very powerful suite for protein identification at essentially no cost.

Protein identification using 2D gels can be done any number of ways, be it through pI/MW measurements, Western blotting (if an antibody is known), or ^{32}P detection (if the

protein of interest is known to be phosphorylated). However, the best method for protein identification is through visual database comparisons to previously annotated gels (CELIS et al., 1998; LEMKIN and THORNWALL, 1999; HOOGLAND et al., 2000). Over the past 25 years, thousands of 2D gels have been run on cell extracts of many different organisms and human tissues. A large number of these gels have been analyzed and their protein spots identified through microsequencing or mass spectrometry. These carefully annotated gels have been deposited into more than 30 different “Federated” 2D gel databases (such as SWISS-2D PAGE) with the intention that others who may be studying similar systems could use these standardized, annotated gel images to overlay with their own gels and rapidly identify proteins of interest.

For instance, suppose you have decided to study *Saccharomyces cerevisiae* under anaerobic conditions. By running a 2D gel of the proteins expressed under anaerobic conditions, you may save literally months of effort by comparing this gel with the fully annotated *S. cerevisiae* 2D gel (grown in aerobic conditions) found at <http://www.ibgc.u-bordeaux2.fr/YPM/>. Using a software package like Flicker or more sophisticated commercial packages, it should be possible to visually transform the two gels, overlay them and identify nearly 400 yeast proteins or protein fragments in less than an hour. Quantification of the differences in expression might take only a few more hours. Indeed, the intent of these federated 2D gel databases is to avoid costly or repetitive efforts that only lead to the re-identification of previously mapped or previously known proteins. The utility of 2D gel databases is bound to grow as more gels are collected and as more spots are progressively identified in various labs around the world. Indeed, one might optimistically predict that sometime in the near future, mass spectrometry and microsequencing will no longer be needed to routinely identify protein spots as all detectable spots will have been annotated and archived in a set of Web-accessible 2D gel databases.

However, much still remains to be done before this vision could become a reality. 2D gel spot patterns are highly dependent on the methods used to isolate and prepare the initial

protein mixture (CELIS et al., 1998). Consequently, individuals wishing to do gel database comparisons must take into account such variables as the protein fraction that was isolated, how the sample was prepared, and how the gel was run. Even if sample preparation issues are eventually sorted out, continuing problems concerning 2D gel database maintenance and updates still persist. Indeed, most publicly available annotated 2D gels represent incomplete “best efforts” of a single graduate student rather than collective, sustained efforts arising from multiple laboratories. If the concept of 2D gel databases is going to succeed, it will need a well-funded central repository (like the NCBI or EBI) and open-minded funding agencies to support sustained gel annotation contributions from the whole scientific community.

In the future it is likely that other separation and display techniques such as, 2D HPLC, tandem capillary electrophoresis (see Chapter 11), and protein chips will gain greater prominence in functional proteomics. The resolution and separation reproducibility of these techniques suggests that similar database comparison methods (i.e., elution profile analysis) could eventually allow proteins to be identified without the need for MS or microsequencing analysis (LINK et al., 1999; YATES, 2000).

2.2 Protein Identification from Mass Spectrometry

Recent advances in mass spectrometry have led to a paradigm shift in the way peptides and proteins are identified (YATES, 2000). In particular, the introduction of “soft” ionization techniques (Electrospray and MALDI), coupled with substantial improvements in mass accuracy (5 ppm), resolution (MS/MS), and sensitivity (femtomoles) have made the rapid, high-throughput identification of peptides and proteins almost routine (DUTT and LEE, 2000). Key to making this paradigm shift possible has been the development of bioinformatics software that allows one to correlate biomolecular MS data directly with protein sequence databases. Two kinds of MS bioinformatics software exist:

- (1) software for identifying proteins from peptide mass fingerprints and
- (2) software for identifying peptides or proteins directly from uninterpreted tandem (MS/MS) mass spectra.

Peptide mass fingerprinting was developed in the early 1990s as a means to unambiguously identify proteins from proteolytic fragments (PAPPIN et al., 1993; YATES et al., 1993; MANN et al., 1993). Specifically, if a pure protein is digested with a protease that cuts at predictable locations (say trypsin), the result will be a peptide mixture containing a unique collection of between 10–50 different peptides, each with a different or characteristic mass. Running this mixture on a modern ESI or MALDI instrument will lead to an MS spectrum with dozens of peaks corresponding to the masses of each of these peptides. Because no two proteins are likely to share the same set of constituent peptides, this mixture is called a peptide mass fingerprint. By comparing the observed masses of the mixture with predicted peptide masses derived from all known protein sequences it is theoretically possible to identify the protein of interest (providing the protein has been previously sequenced). Specifically, in the course of performing a mass fingerprint search, database sequences are theoretically “cleaved” using known protease cutting rules, the resulting hypothetical peptide masses are calculated and the whole protein is ranked according to the number of exact (or near exact) cleavage fragment matches made to the observed set of peptide masses. The sequence with the highest number and quality of matches is usually selected as the most likely candidate.

There are nearly a dozen different types of mass fingerprinting software available for protein identification. While some are sold as commercial products, most are freely available over the Web (see Tab. 1 for a partial listing). Nearly all of the packages allow one to select a protein database (OWL, SWISS-PROT, or NCBI-nr), a source organism (to limit the search), a cleavage enzyme (trypsin is the most common), a cleavage tolerance (1 missed cleavage per peptide is usual), a mass tolerance (0.1 amu is typical), and a mass type (average or monoisotopic). Most of these values are pre-selected as defaults in the submission form

and do not normally need to be changed. All packages expect users to enter a list of masses (with at least 2 decimal point accuracy) read from the MS spectrum prior to launching the search. On most days a Web search result can be returned within 10–20 s.

When a peptide mass is read off from an MS spectrum, it is important to remember that mono-isotopic masses (the mass of the most abundant isotope for a given peptide) are only accurately readable from 500–3,000 m/z^{-1} . Peaks with mass-to-charge ratios above 3,000 generally correspond to an average mass and not to a mono-isotopic mass. Key to performing any successful peptide mass fingerprint search is to start with the most accurate masses possible. Internally calibrated mono-isotopic standards are essential. If one is very confident in the mass accuracy, then restricting the mass tolerance setting to less than 0.1 amu will generally improve the specificity of the search. Restricting the size of the database to search is also wise. In most cases the organism being studied is known, and so it is best to select only the portion of the protein database with protein sequences from the source organism (or very closely related organisms). It is not (yet) a good idea to search through translated EST databases as they have too many sequencing errors and contain only partial protein sequence information.

As a general rule one should try to use as many mass values as possible when performing an MS fingerprint analysis. An absolute minimum of 5, but more commonly 10 mass values should be entered to positively identify a protein. Typically the number of masses one enters should correspond to the molecular weight of the protein in kilodaltons (kDa). The need for so much mass data is primarily to compensate for the fact that experimental MS data is inherently “noisy”. Indeed, it is not uncommon to have up to half of all predicted peptide peaks absent from any given MS fingerprint spectrum along with any number of additional peaks arising from contaminating proteins. Consequently, one is usually quite content to get coverage (the fraction of predicted peptide masses closely matching with observed peptide masses) of only 40–50%. It is very rare to see a perfect match or 100% coverage.

Because of the experimental noise associated with MS data, the analysis of peptide fingerprint searches is not always easy. Some of the common complications include:

- (1) disappearance of key peaks due to non-specific ion suppression,
- (2) appearance of extra peaks from protease autolysis,
- (3) appearance of peaks from post-translational or artifactual chemical modification,
- (4) appearance of peaks from non-specific cleavage, or from contaminating proteases, and
- (5) appearance of peaks from contaminating impurities, contaminating homologs, or splice variants.

Because of these complications, the issue of how to score and rank peptide mass matches is actually quite critical to the performance and reliability of peptide mass fingerprint software. Most early fingerprinting programs used simple heuristic scoring schemes and arbitrary cut-offs to select candidate sequences. More recent programs such as Profound (ZHANG and CHAIT, 1995) use Bayesian statistics and posterior probabilities to rank database candidates. Some of the latest programs take at least some of the complications listed above into account and allow for secondary searches with so-called “orphan” masses. Mascot (PERKINS et al., 1999) is one program which uses a probabilistic model similar to an expectation or E-value to rank sequences. The use of probabilities allows a better estimation of significance (which guards against false positives). It also permits scores to be compared with other types of search algorithms (such as BLAST). Regardless of the advantages and disadvantages of individual programs, it is generally a good idea to run several different peptide mass fingerprinting programs and combine the results. This serves as a form of “signal averaging” and potentially reduces the occurrence of errors arising from algorithmic or database limitations in any single program.

Because peptide mass fingerprinting does not always work for unambiguous protein identification there has been increasing emphasis on using tandem mass spectrometers

equipped with collision induced dissociation (CID) cells to provide more precise and interpretable peptide data. SEQUEST (ENG et al., 1994; YATES et al., 1996) and Mascot (PERKINS et al., 1999) are two software packages that can be used to analyze tandem mass data of peptide fragments. Both programs take uninterpreted tandem mass spectral data (i.e., the actual spectrum), perform sequence database searches, and identify probable peptides or protein matches. Typically these programs work by first scanning the protein databases for potential matches to the precursor peptide ion, then ranking the candidate peptides on the basis of their predicted similarity (ion continuity, intensity, etc.) to the observed fragment ion masses. After this screening step a model MS/MS spectrum for each candidate peptide is generated and then compared, scored, and ranked with the observed MS/MS spectrum using correlational or probabilistic analysis. As with peptide mass fingerprinting, similar kinds of information (database, source organism, mass tolerance, cleavage specificity, etc.) must be provided before running the programs. The only difference is that instead of typing in a list of masses, the user is expected to provide a spectral filename containing the digitized MS/MS spectrum. Overall, the reported performance of both programs is quite impressive (YATES et al., 1996; PERKINS et al., 1999).

It is likely that protein identification via mass spectral analysis will continue to grow in popularity and in importance. The wide availability of easy-to-use, freely available peptide mass fingerprinting software has made the entire protein identification process very accessible. Furthermore, as more protein sequence data is deposited into sequence data banks around the world, the utility of these database-driven techniques is expected to grow accordingly. While sequence databases continue to grow, mass spectrometer technology is also progressing rapidly. With continuing improvements in mass resolution (i.e., Fourier Transform Cyclotron Mass Spectrometers with 1 ppm resolution are now available) it is likely that peptide mass fingerprinting will become less common as only a single tryptic peptide will be sufficient to positively identify a protein (GOODLETT et al., 2000).

2.3 Protein Identification from Sequence Data

The most precise and accurate way of unambiguously identifying a protein is through its sequence. Historically proteins were identified by direct sequencing using painstakingly difficult chemical or enzymatic methods (Edman degradation, proteolytic digests). All that changed with the development of DNA sequencing techniques which proved to be faster, cheaper, and more robust (SANGER et al., 1977). Now more than 99% of all protein sequences deposited in databases such as OWL (BLEASBY et al., 1994), PIR (BARKER et al., 2000), SWISS-PROT + trEMBL (BAIROCH and APWEILER, 2000, and GenBank (BENSON et al., 2000) are derived directly from DNA sequence data. While complete sequence data is normally obtained via DNA sequencing, improvements in mass spectrometry and chemical microsequencing now allow for routine sequencing of short (10–20 residue) peptides from subpicomole quantities of protein (SHEVCHENKO et al., 1997). With the availability of several different rapid sequencing methods (MS/MS, chemical microsequencers, DNA sequencers, ladder sequencing, etc.) and the growing number of protein sequences (>500,000) and sequence databases, there is now increasing pressure to develop and use specific bioinformatics tools to facilitate protein identification from partial or homologous sequence data.

Protein identification via sequence analysis can be performed either through exact substring matches or through local sequence similarity to a database of known protein sequences (ref). Exact matching of short peptide sequences to known protein sequences is ideal for identifying proteins from partial sequence data (obtained via Edman microsequencing or tandem MS). This type of text matching to sequence data is currently supported by the OWL, SWISS-PROT, and PIR Web servers (but not GenBank!). Given the current size of the databases and the number of residues they contain, it is usually wise to sequence 7 or 8 residues to prevent the occurrence of false positives. Alternately, if some information about the protein mass, predicted *pI*, or source or-

ganism is available, only 4 or 5 residues need to be determined to guarantee a unique match. Note that exact string matching will only identify a protein if it is already contained in a sequence database.

While exact string matching is useful for certain types of protein identification problems, by far the most common method for protein identification is through “fuzzy matching” via sequence similarity. Unlike exact string matching, sequence similarity is a robust technique which allows proteins to be identified even if there are sequencing errors in either the query or database sequence. Furthermore, sequence similarity allows one to potentially identify or ascribe a function to a protein even if it is not contained in the database. In particular, the identification of a similar sequence (>25% sequence identity to the query sequence) with a known function or name is usually sufficient to infer the function or name of an unknown protein (DOOLITTLE and BORK, 1993).

Sequence similarity is normally determined using database alignment algorithms wherein a query sequence is aligned and compared against all other sequences in a database. In many respects sequence alignment programs are just glorified spell checkers. Fundamentally there are two types of sequence alignment algorithms: dynamic programming methods (NEEDLEMAN and WUNSCH, 1970; SMITH and WATERMAN, 1981) and heuristic “fast” algorithms such as FASTA and BLAST (PEARSON and LIPMAN, 1988; ALTSCHUL et al., 1990). Both methods make use of amino acid substitution matrices such as PAM-250 and Blossum 62 (DAYHOFF et al., 1983; HENIKOFF and HENIKOFF, 1992) to score and assess pairwise sequence alignments. Dynamic programming methods are very slow N^2 type algorithms that are guaranteed to find the mathematically optimal alignment between any two sequences. On the other hand, heuristic methods such as FASTA and BLAST are much faster N -type algorithms that find short local alignments and attempt to string these local alignments into a longer global alignment. Heuristic algorithms make use of statistical models to rapidly assess the significance of any local alignments, making them particularly useful for biologists trying to understand the significance of their matches. Exact descriptions and detailed assess-

ments of these algorithms are beyond the scope of this chapter, but suffice it to say that BLAST and its successors such as FASTA3 (PEARSON, 2000); BLAST2 and PSI-BLAST (ALTSCHUL et al., 1997) have probably become the most commonly used “high-end” tools in all of biology.

BLAST-type searches are generally available for all major protein databases through a variety of mirror sites and Web servers (see Tab. 1). Most servers offer a range of databases which can be “BLASTed”. The largest and most complete database is GenBank’s non-redundant (nr) protein database, which is largely equivalent to the translated EMBL (TREMBL) database. The second largest, and one that is frequently used in mass fingerprinting, is the OWL database. This non-redundant database is updated every two months. The Swiss-Prot database is the most completely annotated protein database, but does not contain the quantity of sequence data found in OWL or GenBank. The PIR database, which was started in the 1960s, is actually the oldest protein sequence database and contains many protein sequences determined through direct chemical or MS methods (which are typically not in GenBank records). Most of these protein databases can be freely downloaded by academics, but industrial users must pay a fee.

The new version of BLAST (BLAST2) offers several improvements over the original BLAST program, particularly in its ability to create longer, near global alignments from preliminary local alignments. However, in terms of generating global alignments, FASTA3 is probably the best program to use. BLAST2 and FASTA3 are particularly good at identifying sequence matches sharing between 25% to 100% identity with the query sequence. PSI-BLAST (position-specific iterated BLAST), on the other hand, is exceptionally good at identifying matches in the so-called twilight zone of between 15–25% sequence identity. PSI-BLAST can also identify higher scoring similarities with the same accuracy as BLAST2. The trick to using PSI-BLAST is to repeatedly press the “Iterate” button until the program indicates that it has converged. Apparently many first-time users of PSI-BLAST fail to realize this by running the program only once and coming away with little more than a

regular BLAST2 output. Nevertheless, because of its near universal applicability, PSI-BLAST is probably the best all-round tool for protein identification from sequence analysis.

The stunning success that PSI-BLAST has had in “scraping the bottom of the barrel” in terms of its ability to identify sequence relationships is leading to increased efforts by bioinformaticians aimed at trying to develop methods to identify even more remote sequence similarities from database comparisons. This has led to the development of a number of techniques such as threading, neural network analysis, and Hidden Markov Modeling – all of which are aimed at extracting additional information hidden in the sequence databases. Many of these techniques are described in more detail in the following section.

3 Protein Property Prediction

Up to this point we have focused on how to identify a protein either from a spot on a gel, an MS fingerprint, or through DNA or protein sequencing. Once the identification problem has been solved, one is usually interested in finding out what this protein does and how/where it works. If a BLAST, PSI-BLAST, or PUBMED search turns up little in the way of useful information, it is still possible to employ a variety of bioinformatics tools to learn something directly from the protein’s sequence. Indeed, as we shall see in the following pages, protein property prediction methods can often allow one to make a very good guess as to the function, location, structure, shape, solubility, and binding partners of a novel protein long before one has even lifted a test-tube.

3.1 Predicting Bulk Properties (pI , Absorptivity, MW)

While the amino acid sequence of a protein largely defines its structure and function, a protein’s amino acid composition can also provide a great deal of information. Specifically,

amino acid composition can be used to predict a variety of bulk protein properties such as isoelectric point, UV absorptivity, molecular weight, radius of gyration, partial specific volume, solubility, and packing volume – all of which can be easily measured on commonly available instruments (gel electrophoresis systems, columns, mass spectrometers, UV spectrophotometers, amino acid analyzers, ultracentrifuges, etc.). Knowledge of these bulk properties can be particularly useful in cloning, expressing, isolating, purifying or characterizing any newly identified protein.

Many of these bulk properties can be calculated using simple formulas and commonly known parameters, some of which are presented in Tabs. 2 and 3. Typical ranges found in water-soluble globular proteins are also shown in Tab. 2. A large number of these calculations can also be performed with more comprehensive protein bioinformatics packages such as SEQSEE (WISHART et al., 1994, 2000) and ANTHEPROT (DELEAGE et al., 1988) as well as many commonly available commercial packages (GCG, LaserGene99, PepTool, VectorNTI).

3.2 Predicting Active Sites

As more and more protein sequences are being deposited into data banks, it is becoming increasingly obvious that certain amino acid residues remain highly conserved even among diverse members of protein families. These highly conserved sequence patterns are often called signature sequences and in many cases they define the active site of a protein. Because most signature patterns are relatively short (7–10 residues) this kind of sequence information is not easily detected from BLAST or FASTA searches. Consequently, it is always a good idea to scan against a signature sequences database (such as PROSITE) in an effort to detect additional information concerning a protein’s structure, function, or activity.

Active site or signature sequence databases come in two varieties:

- (1) pattern-based and
- (2) profile-based.