

11 PROTEIN STRUCTURE PREDICTION

11.1 INTRODUCTION

Protein Structure Prediction (PSP) from a sequence is one of the high focus problems for researchers. This is a very useful application of bioinformatics as the experimental techniques like X-ray crystallography are time consuming. The fundamental issue is how can we predict the 3-D shape of a protein from its amino acid sequence. This chapter builds on the discussion on protein structure, classification and visualization discussed in Chapter 10. You will learn how to predict protein structure and function based on the amino acid sequence.

The Protein Folding Problem

According to the *Alfinsen's hypothesis*, the 3-D structure of a protein is determined solely by the amino-acid sequence information. The experimental support for this hypothesis was garnered as follows. Denaturants such as urea were added to the system of proteins that are folded in the native conformation. Denaturants destroy the tertiary structure so that the proteins are in the random coil state. After removal of the denaturants the proteins spontaneously fold back into their native conformation. This is an *in vitro* experiment where there is no cellular environment. The lack of any cellular environment and the capability of the protein to spontaneously fold back into its native conformation suggest that the information within the denatured sequence is enough for protein to fold itself. The results also suggest that the native conformation of the protein corresponds to the global-minimum state of the free energy.

The strong argument against the Alfinsen hypothesis is the *Levinthal's paradox*. Levinthal's paradox can be understood as follows. The 3-D structure of the main chain of a protein is determined by the dihedral angles ϕ and ψ (where ω is 180°). If only local interactions are considered, these dihedral angles have a few preferred values that correspond to the local minima of the torsion energy around each rotation bond. We may have to consider only about 10 conformations per each amino acid. However this implies that we have to examine as many as 10^N conformations for a protein with N amino acids. For a

protein with $N = 40$, there are 10^{40} possible conformations. Considering an average rotation frequency around each bond, one can assume that a protein can sample of the order of 10^{14} structures per second. Hence, it would take this protein about 10^{26} seconds $\equiv 10^{18}$ years to examine all the possible conformations. However, actually proteins fold into their native conformations on the time scale of milliseconds to minutes.

The computational difficulty of protein folding is classified as an NP-complete problem. If a problem is NP-complete, it means that a particular solution can be checked in polynomial time but to solve the whole problem requires an exponential time algorithm. A problem is in NP if it has a nondeterministic polynomial time solution. This means that the solution can be checked within polynomial time. As the exponential function in an NP-complete problem increases at a much more rapid rate than a polynomial, these problems are intractable.

There have been some thoughts on the resolution of Levinthal's paradox. These are summarized below:

1. The theoretical models used to prove hardness are not what nature is trying to optimize.
2. Evolution may have selected proteins which fold easily.
3. Proteins may well fold in locally, not globally optimal ways.

To summarize, it is difficult to predict structure from sequence. However, from the growing database of experimentally determined protein structures, some heuristics are emerging:

1. The number of unique protein folds is quite limited.
2. There are many proteins with the same fold, but no similarity of sequence.
3. 'Neutral' mutations altering the protein structure are likely.

11.2 PROTEIN IDENTIFICATION AND CHARACTERIZATION

Many of the tools for protein identification and characterization are available at ExPASy (<http://www.expasy.org/>). Some of these tools can be identified as unknown protein isolated through 2-D gel electrophoresis. Another set of these tools can help in predicting the physical properties of known proteins.

Some of the ExPASy tools and other tools are discussed as follows:

AACompIdent

AACompIdent (<http://us.expasy.org/tools/aacomp/>) is an important tool to identify a protein by its amino acid composition. It uses the amino acid composition of an unknown protein to identify known proteins of the same composition.

As the input to AACompIdent, you need to give the following information:

1. Amino acid composition of the protein to identify.
2. A name for this protein, so that you can recognize it later in the results.
3. The pI and Mw of that protein (if known).

4. The species or group of species for which you would like to perform the search. You may also just specify ALL for all SWISS-PROT/TrEMBL entries.
5. For scan in SWISS-PROT only: the keyword for which you would like to perform the search (example: ZINC-FINGER). This will produce the list of proteins matching this keyword. You may also just specify ALL for all SWISS-PROT entries.
6. Amino acid composition of a known protein, obtained in the same run as the amino acid composition of the unknown protein. This is for calibration. If you do not have a calibration protein, leave NULL.
7. The SWISS-PROT identifier (ID) of the calibration protein (example: ALBU_HUMAN).
8. Your e-mail address to get the search results mailed to you.

SWISS-PROT and TrEMBL are indexed into 6 constellations (groupings).

AACompSim (<http://us.expasy.org/tools/aacsim/>) is a variant of AACompIdent. It is used to compare the amino acid composition of a SWISS-PROT entry with all other entries.

TagIdent, PeptIdent and MultiIdent

TagIdent (<http://us.expasy.org/tools/tagident.html>) is a tool which allows the following:

1. The generation of a list of proteins close to a given pI and Mw.
2. The identification of proteins by matching a short sequence tag of up to 6 amino acids against proteins in the SWISS-PROT/TrEMBL databases close to a given pI and Mw.
3. The identification of proteins by their mass, if this mass has been determined by mass spectrometric techniques.

PeptIdent (<http://us.expasy.org/tools/peptident.html>) is used to identify proteins with peptide mass fingerprinting data, pI and Mw. Experimentally measured, user-specified peptide masses are compared with the theoretical peptides calculated for all proteins in SWISS-PROT, making extensive use of database annotations.

MultiIdent (<http://us.expasy.org/tools/multiident/>) is a tool that allows the identification of proteins using pI, MW, amino acid composition, sequence tag and peptide mass fingerprinting data. One or more species and a SWISS-PROT keyword can also be specified for the search.

PROPSEARCH

PROPSEARCH (<http://www.infobiosud.univmontp1.fr/SERVEUR/PROPSEARCHpropsearch.html>) is a tool to find the putative protein family if querying a new sequence has failed using alignment methods. PROPSEARCH uses the amino acid composition as the input. In addition, other properties like molecular weight, content of bulky residues, content of small residues, average hydrophobicity, average charge and the content of selected dipeptide-groups are calculated from the sequence as well. 144 such properties are weighed individually and are used as query vector. The weights have been trained on a set of protein families with known structures, using a genetic algorithm. Sequences in the database are transformed into vectors as

well, and the euclidian distance between the query and database sequences is calculated. Distances are rank ordered, and sequences with lowest distance are reported on top. Consider the following sequence (SwissProt Entry ID: Q969F8) as an example:

MHTVATSGPNASWGAPANASGCPGCGANASDGPVPSRAVDLWLVPLFFAAL
MLLGLVGNLSVIYVICRHKPMRTVTNFIYIANLAATDVTFLCCVPFTALLYPLPG
WVLGDFMCKFVNYIQQVSQATCATLTAMSVDRWYVTVFPLRALHRRTPRLAL
AVLSIWVGSAAVSAPVLALHRLSPGPRAVCSEAFPSRALERAFALYNLLALYLL
PLLATCACYAAMLRHLGRVAVRPAPADSALQGQVLAERAGAVRAKVSRLVAA
VLLFAACWGPIQLFLVLQALGPAGSWHPRSAAAYALKTWAHCMSYSNSALNP
LLYAFLGSHFRQAFFRRVCPAPRRPRRRPGSPDPAAPHAELHRLGSHAPARA
QKPGSSGLAARGLCVLGEDNAPL

Figure 11.1 is an excerpt from the PROPSEARCH output.

Rank	ID	DIST	LEN2	POS1	POS2	PI	DE
1	b3ar_felca	7.82	398	1	398	10.51	Beta-3 adrenergic receptor.
2	b3ar_macmu	8.00	418	1	418	9.40	Beta-3 adrenergic receptor.
3	b3ar_cavpo	8.20	351	1	351	10.64	Beta-3 adrenergic receptor (Fragment).
4	pi2r_human	8.24	386	1	386	8.37	Prostacyclin receptor (Prostanoid IP recepto
5	ur2r_human	8.26	389	1	389	11.42	Urotensin II receptor (UR-II-R).
6	b3ar_human	8.32	408	1	408	8.98	Beta-3 adrenergic receptor.
7	pi2r_mouse	8.40	415	1	415	8.16	Prostacyclin receptor (Prostanoid IP recepto
8	b3ar_caphi	8.49	405	1	405	11.01	Beta-3 adrenergic receptor.
9	b3ar_rat	8.52	400	1	400	10.40	Beta-3 adrenergic receptor.
10	b3ar_bovin	8.64	405	1	405	10.30	Beta-3 adrenergic receptor.
11	b3ar_sheep	8.65	405	1	405	10.62	Beta-3 adrenergic receptor.
12	ugat_human	8.70	396	1	396	10.81	UDP-galactose translocator (UDP-galactose tr
13	v2r_bovin	8.76	370	1	370	8.80	Vasopressin V2 receptor (Renal-type arginine
14	trab_rhien	8.76	387	1	387	10.92	Probable conjugal transfer protein traB.
15	gal2_human	8.87	387	1	387	10.05	Galanin receptor type 2 (GAL2-R) (GALR2).
16	b3ar_mouse	8.89	400	1	400	9.99	Beta-3 adrenergic receptor.
17	b3ar_canfa	9.02	405	1	405	11.00	Beta-3 adrenergic receptor.
18	ugat_mouse	9.09	390	1	390	10.71	UDP-galactose translocator (UDP-galactose tr
19	pi2r_bovin	9.26	385	1	385	8.85	Prostacyclin receptor (Prostanoid IP recepto
20	ta2r_human	9.27	369	1	369	10.89	Thromboxane A2 receptor (TXA2-R) (Prostanoid
21	v2r_human	9.29	371	1	371	9.62	Vasopressin V2 receptor (Renal-type arginine
22	p2y7_human	9.29	352	1	352	11.73	P2Y purinoceptor 7 (P2Y7) (Leukotriene B4 re
23	pe21_rat	9.40	405	1	405	12.07	Prostaglandin E2 receptor, EP1 subtype (Pros
24	pi2r_rat	9.44	416	1	416	8.04	Prostacyclin receptor (Prostanoid IP recepto
25	pe21_human	9.47	402	1	402	12.23	Prostaglandin E2 receptor, EP1 subtype (Pros
26	gal2_rat	9.53	372	1	372	10.21	Galanin receptor type 2 (GAL2-R) (GALR2).
27	ftsW_myce	9.60	465	1	465	10.63	Probable cell division protein ftsW.
28	atka_myctu	9.60	571	1	571	9.80	Potassium-transporting ATPase A chain (EC 3.
29	pucc_rhosu	9.62	454	1	454	11.39	Protein pucC.
30	pe21_mouse	9.68	405	1	405	12.14	Prostaglandin E2 receptor, EP1 subtype (Pros
31	a2ab_echte	9.72	375	1	375	10.13	Alpha-2B adrenergic receptor (Alpha-2B adren
32	rnfd_rhoa	9.72	358	1	358	8.65	Electron transport complex protein rnfd (Nit
33	secy_strsc	9.82	437	1	437	10.35	Preprotein translocase secY subunit.
34	polg_hcvh8	9.82	321	1	321	7.58	Genome polyprotein (Contains: Matrix protein
35

FIGURE 11.1 PROPSEARCH output.

A distance score ranks the results above. The first column give the rank, the second column gives the SWISSPROT or PIR id, then the distance score, followed by the length of the overlap between the query and the subject, the positions of overlap, the calculated pi and the definition line for the found sequence. A distance score of below 8.7 indicates a 94% chance of similarity between the two proteins.

SwissProt Entry ID Q969F8 represents Metastin. Metastin is an endogenous ligand for G-Protein Coupled Receptor hOT7T175. Metastin has a very high importance in cancer research. One endogenous mechanism for cell proliferation involves the product from the gene known as *KISS-1*, which has been shown to suppress metastasis of human melanomas and breast carcinomas. *KISS-1* encodes a 145-amino acid residue peptide, which is further processed to a final 54-amino acid peptide with C-terminal amidation. This final peptide is Metastin. Metastin is similar to a number of Beta-3 adrenergic receptors as indicated in Figure 11.1.

PepSea

PepSea (<http://195.41.108.38/PepSeaIntro.html>) is a tool for protein identification by peptide mapping or peptide sequencing. You can search the non-redundant protein sequence database by:

- ◆ A list of peptide masses
- ◆ A peptide sequence tag
- ◆ Sequence only

PepMAPPER, Mascot and PeptideSearch

These are various peptide mass fingerprinting tools.

PepMAPPER (<http://wolf.bms.umist.ac.uk/mapper/>) takes peptide mass as the key input as shown in the screen shot in Figure 11.2.

PepMAPPER (1)

Organism: Enzyme:

peptide parameters

Peptide Masses(m/z)

Masses are ☒ average ☐ monoisotopic

Charge ☒ Fixed state ☐ Unknown to

Missed Cleavages (max possible)

N-terminal amino acid

fixed peptide modifications

Acetylation (N-term.K)
 Biotinylated (N-term.K)
 Carbamidomethyl (C)
 Carbomyl (N-term)

protein parameters

Mass Range -

Isoelectric point to +/-

Error +/- ppm

Report top matches

[mapper1][mapper2][mapper3][mapper4][mapper5][Linkpage][Help Pages]

Mascot Search (http://www.matrixscience.com/cgi/index.pl?page=/search_form_select.html) can take the following as the input:

1. **Peptide mass fingerprint.** The experimental data are a list of peptide mass values from an enzymatic digest of a protein.
2. **Sequence query.** One or more peptide mass values associated with information such as partial or ambiguous sequence strings, amino acid composition information, MS/MS fragment ion masses, etc. This is a super-set of a sequence tag query.
3. **MS/MS ion search.** Identification based on raw MS/MS data from one or more peptides.

PeptideSearch (http://www.mann.embl-heidelberg.de/GroupPages/PageLink/peptide_searchpage.html) can be used for the following:

- ♦ List of peptide masses
- ♦ Peptide sequence tag—what is a sequence tag?
- ♦ Amino acid sequence

FindPept

FindPept (<http://ca.expasy.org/tools/findpept.html>) is an ExPASy tool. It can be used to identify peptides that result from unspecific cleavage of proteins from their experimental masses. FindPept takes into account artifactual chemical modifications, post-translational modifications (PTM) and protease autolytic cleavage. Experimentally measured peptide masses are compared with the theoretical peptides calculated from a specified SWISS-PROT entry or from a user-entered sequence.

Predicting Transmembrane Helices

Very little structural data is available for proteins that are not soluble in water. The major obstacle with these proteins is that they do not crystallise, and are hardly tractable by NMR spectroscopy. Consequently, for this class of proteins structure prediction methods are even more needed than for globular water-soluble proteins. Computational methods can be used to predict which proteins in a genome will be transmembrane proteins.

Transmembrane proteins show this property. It has been established that the hydrophobicity of a stretch of 20 residues is an excellent predictor of whether that sequence will be located within a membrane. However, for soluble proteins, not only does hydrophobicity correlate very poorly with helical nature but no other single predictor serves to reliably identify regions of secondary structure.

Membrane protein folding has been hypothesized as a two-stage model. In the first stage of this model, the insertion of hydrophobic helices into lipid bilayers generates independently stable transmembrane helices. These are thought to behave as autonomous domains that are unable to unfold or to leave the bilayer because of the high-energy penalties associated with breaking hydrogen bonds or exposing hydrophobic side chains to water. The second stage of the model consists of the lateral association of these helices. Interactions between the intra-membranous portions of these helices are supposed to be

responsible for the resulting tertiary and quaternary structures, although lipids, ligands, or extra-membranous loops can influence this process.

However, the prediction task is simplified because of environmental constraints on transmembrane proteins. The lipid bilayer of the membrane reduces the degrees of freedom to such an extent that 3-D structure formation becomes almost a 2-D problem.

TMAP (<http://www.mbb.ki.se/tmap/>) predicts transmembrane helices based on multiple sequence alignment. You can also give a single sequence as the input.

TMHMM (<http://www.cbs.dtu.dk/services/TMHMM/>) is used for prediction of transmembrane helices.

TMPRED (http://www.ch.embnet.org/software/TMPRED_form.html) is used to predict membrane-spanning regions and their orientation. The algorithm is based on the statistical analysis of TMbase, which is a database of naturally occurring transmembrane proteins. The prediction is made using a combination of several weight-matrices for scoring.

The output of TMPRED is in three parts. First is the listing of the possible transmembrane helices. The listing gives both inside-to-outside and outside-to-inside orientations. Only scores above 500 are considered significant. The second part is the table of correspondences. This shows that which of the inside \rightarrow outside helices correspond to which of the outside \rightarrow inside helices. A "+" symbol indicates a preference of this orientation and a "++" symbol indicates a strong preference of this orientation. The third part speculates on the suggested model for the transmembrane topology.

For Metastin (SWISS-PROT ID: Q969F8), there are two models suggested by TMPRED—one with 7 strong transmembrane helices and the second with 6. The output of TMPRED in a graphical form is shown in Figure 11.3.

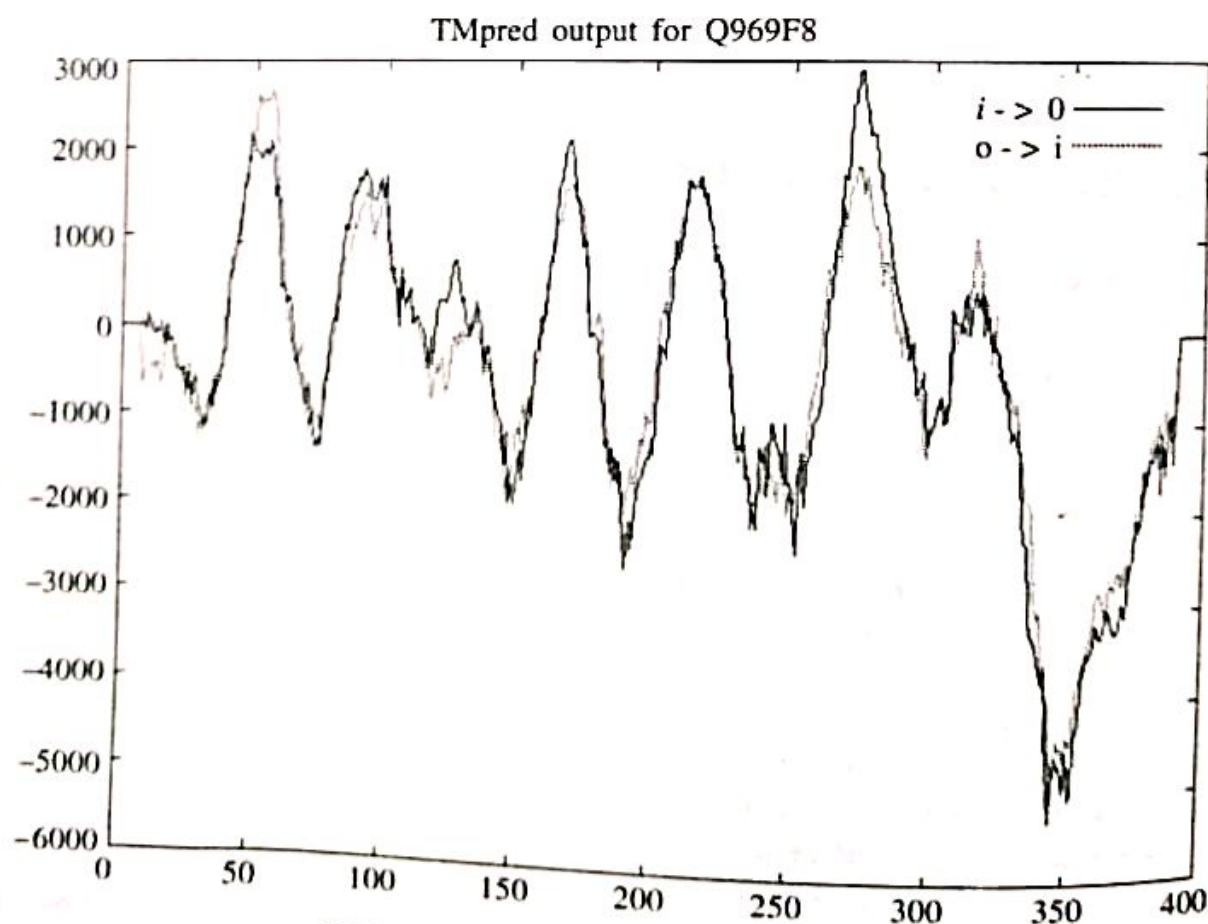


FIGURE 11.3 TMPRED output.

TopPred2 (<http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html>) can also be used for prediction of location and orientation of transmembrane helices.

PHDhtm (http://www.embl-heidelberg.de/predictprotein/Dtab/phd_htm.html) is a multiple alignment-based neural network system used to predict the locations of transmembrane helices. The shortcoming of the network system is that often too long helices are predicted. An empirical filter cuts these. The final prediction has an expected per-residue accuracy of about 95%. The number of false positives, i.e. transmembrane helices predicted in globular proteins, is about 2%.

An alternative to PHDhtm is DAS (<http://www.sbc.su.se/~miklos/DAS/>). The DAS server predicts transmembrane regions of a query sequence. The predictive power of DAS and PHDhtm is essentially the same while the single-sequence based methods perform slightly worse.

PHDhtm is refined by a dynamic programming-like algorithm—PHDtopolgy (http://www.embl-heidelberg.de/predictprotein/Dtab/phd_htm_top.html). This method resulted in correct predictions of all transmembrane helices for 89% of the 131 proteins used in a cross-validation test; more than 98% of the transmembrane helices were correctly predicted. The output of this method is used to predict topology, i.e. the orientation of the N-term with respect to the membrane. The expected accuracy of the topology prediction is >86%. Prediction accuracy is higher than average for eukaryotic proteins and lower than average for prokaryotes. PHDtopolgy is more accurate than all other methods tested on identical data sets.

11.3 PRIMARY STRUCTURE ANALYSIS AND PREDICTION

There are various tools for predicting the physical properties using the sequence information. Some of the major ones are discussed below:

Compute pI/Mw

Compute pI/Mw (http://ca.expasy.org/tools/pi_tool.html) is a tool that calculates the isoelectric point and molecular weight of an input sequence. The sequence can be input in the FASTA format, the output is the pI and molecular weight for the entire length of the sequence.

The pI/Mw for the protein sequence represented by SwissProt Entry ID Q969F8 (Metastin) is given below from Compute pI/Mw:

DE G protein-coupled receptor (Putative G protein-coupled receptor)
DE (G-protein-coupled receptor GPR54).
OS Homo sapiens (Human).

The computation has been carried out on the complete sequence

Molecular weight: 42610.02

Theoretical pI: 9.93

It is important to note the shortcoming of the theoretical calculation for computation of molecular weight. The molecular weight calculated based on the sequence cannot take into account post-translational modifications such as glycosylation, phosphorylation or other chemical modification of residues. It also cannot take into account information such as removal of a signal sequence or cleavage by a protease.

Peptide Mass (<http://www.expasy.org/tools/peptide-mass.html>) cleaves one or more protein sequences from the SWISS-PROT and/or TrEMBL databases or a user-entered protein sequence with a chosen enzyme and computes the masses of the generated peptides. Also returns theoretical isoelectric point and mass values for the proteins of interest.

If desired, PeptideMass can return the mass of peptides known to carry posttranslational modifications, and can highlight peptides whose masses may be affected by database conflicts, isoforms or splicing variants.

SAPS (http://www.isrec.isb-sib.ch/software/SAPS_form.html)

Statistical Analysis of Protein Sequences (SAPS) is a tool to evaluate a wide variety of protein sequence properties by using statistical criteria.

The output usually runs in several pages and is organized in the following sections:

- ◆ File name
- ◆ Sequence printout
- ◆ Compositional analysis
- ◆ Charge distributional analysis (charge clusters; high scoring (un)charged segments; charge runs and patterns)
- ◆ Distribution of other amino acid types (high scoring hydrophobic and transmembrane segments; cysteine spacings)
- ◆ Repetitive structures (in the amino acid alphabet and in a 11-letter reduced alphabet)
- ◆ Multiplets (counts, spacings, and clusters in the amino acid and charge alphabets)
- ◆ Periodicity analysis
- ◆ Spacing analysis

The output for ALBN_HUMAN in Swiss-prot notation is as follows (only the initial excerpts):

Protein 1 (File: wwwtmp/.SAPS.7177.445.seq)

SWISS-PROT ANNOTATION:

ID sp|P02768|ALBU_HUMAN

DE sp|P02768|ALBU_HUMAN (ALB)Serum albumin precursor.[Homo sapiens], 609 bases, D0E84FE4 checksum.

number of residues: 609; molecular weight: 69.4 kdal

1 MKWVTFISLL FLSSAYSRG VFRRDAHKSE VAHRFKDLGE ENFKALVLIAFAQYQQCPF
61 EDHVKL VNEV TEFAKTCVAD ESAENCDKSL HTLFGDKLCT VATLRETYGE MADCCAQEP
121 ERNECFLQHK DDNPNLPRLV RPEVDVMCTA FHDNEETFLK KYLYEIARRH PYFYAPELLF


```

181 FAKRYKAAFT ECCQAADKAA CLLPKLDELK DEGKASSAKQ RLKCAQLQKF GERAFKAWAV
241 ARLSQRFPKA EFAEVSKLVT DLTKVHTECC HGDLLCADD RADLAKYICE NQDSISSKLE
301 ECCEKPLEK SHCIAEVEND EMPADLPSLA ADFVESKDVC KNYAEAKDVF LGMELYEYAR
361 RHDPYSVLL LRLAKTYETT LEKCCAAADP HECYAKVFE FKPLVEEPQN LIKQNCLEFE
421 QLGEYKEQNA LLVRYTKKVP QVSTPTLVEV SRNLGKVGSK CCKHPEAKRM PCAEDYLSVV
481 LNQLCVLHEK TPVSDRVTKC CTESLVNRRP CFSALEVDET YVPKEFNAET FTFHADICTL
541 SEKERQIKKQ TALVELVKHK PKATKEQLKA VMDDFAAFVE KCCKADDDKET CFAEEGKKLV
601 AASQAALGL

```

ProtParam

ProtParam (<http://ca.expasy.org/tools/protparam.html>) is a tool, which allows the computation of various physical, and chemical parameters for a given protein stored in SWISS-PROT or TrEMBL or for a user entered sequence.

SAPS (Statistical Analysis of Protein Sequences)

SAPS (http://www.isrec.isb-sib.ch/software/SAPS_form.html) is a program that provides extensive statistical information for any given sequence. The output is organized in the following sections: file name, sequence printout, compositional analysis, charge distributional analysis (charge clusters; high scoring (un)charged segments; charge runs and patterns), distribution of other amino acid types (high scoring hydrophobic and transmembrane segments; cysteine spacings), repetitive structures (in the amino acid alphabet and in a 11-letter reduced alphabet), multiplets (counts, spacings, and clusters in the amino acid and charge alphabets), periodicity analysis, spacing analysis. The output is several pages long.

Predicting Protein Hydrophobicity

It has been hypothesized that if the segments of secondary structure could be accurately predicted, the 3-D structure could be predicted by simply trying different arrangements of the segments in space. One criterion for assessing each arrangement could be to use predictions of residue solvent accessibility. The principal goal is to predict the extent to which a residue of embedded in a protein structure is accessible to solvent. Solvent accessibility can be described in several ways. The simplest is a two-state description distinguishing between residues that are buried (relative solvent accessibility <16%) and exposed (relative solvent accessibility 16%). The classical method to predict accessibility is to assign either of the two states, buried or exposed, according to residue hydrophobicity. However, a neural network prediction using PHDacc (<http://www.embl-heidelberg.de/predictprotein/>) of accessibility has been shown to be superior to simple hydrophobicity analyses.

ProtScale (<http://ca.expasy.org/cgi-bin/protscale.pl>) can be used to calculate the hydro-phobicity. For example, the output for Q969F8 is as follows:

Using the Kyte & Doolittle scale, the individual values for the 20 amino acids are:

Ala: 1.800 Arg: -4.500 Asn: -3.500 Asp: -3.500 Cys: 2.500 Gln: -3.500
 Glu: -3.500 Gly: -0.400 His: -3.200 Ile: 4.500 Leu: 3.800 Lys: -3.900
 Met: 1.900 Phe: 2.800 Pro: -1.600 Ser: -0.800 Thr: -0.700 Trp: -0.900
 Tyr: -1.300 Val: 4.200 Asx: -3.500 Glx: -3.500 Xaa: -0.490

The ProtScale output in a graphical form is given in Figure 11.4.

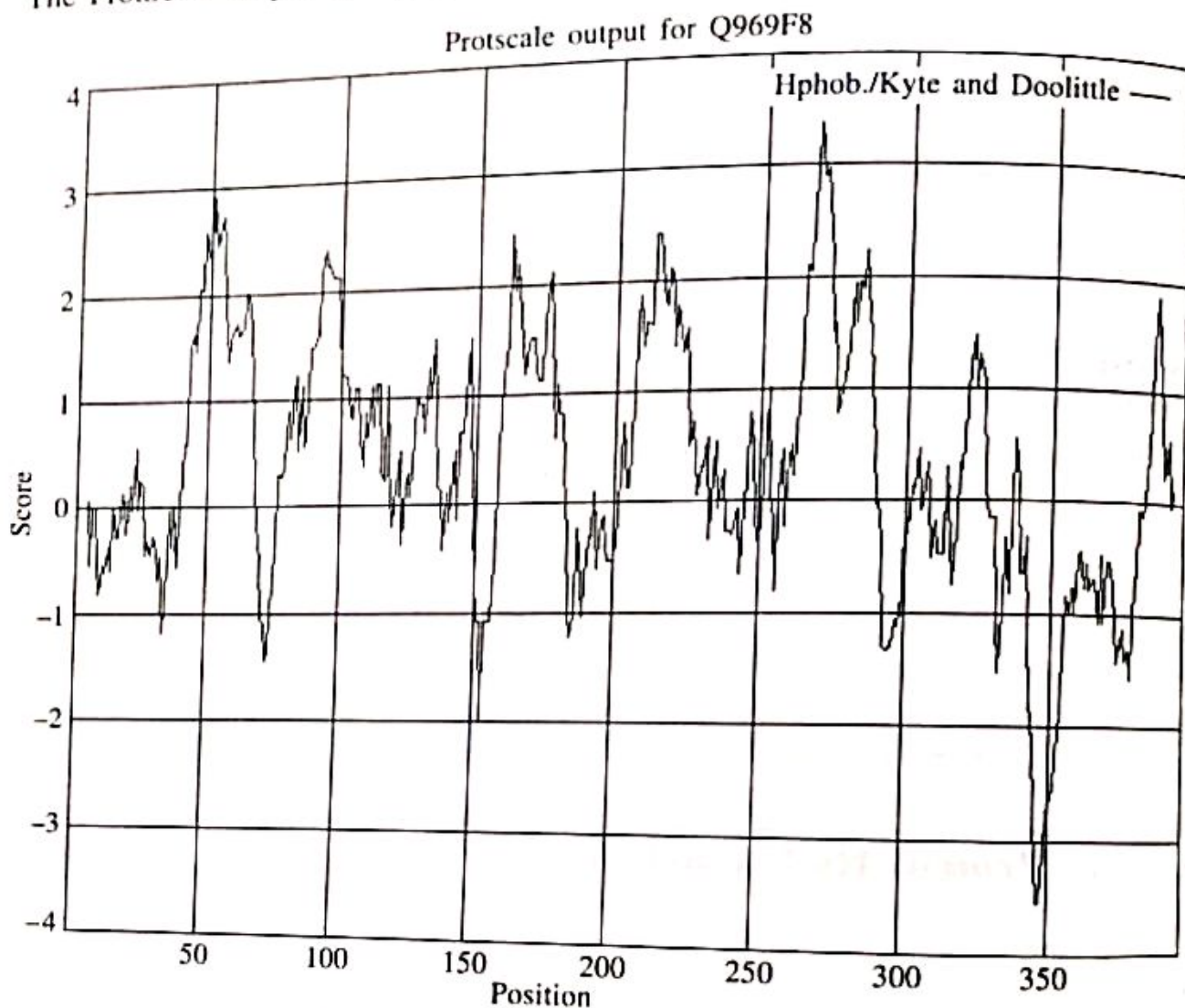


FIGURE 11.4 ProtScale output.

You can use drawhca (<http://smi.snv.jussieu.fr/hca/hca-form.html>) to draw an HCA (Hydrophobic Cluster Analysis) plot of a protein sequence.

PEST and PESTfind

Proteins with intracellular half-lives of less than two hours are found to contain regions rich in proline, glutamic acid, serine and threonine (P, E, S and T). These are called PEST regions and are generally flanked by clusters of positively charged amino acids.

PEST (<http://www.icnet.uk/LRITu/projects/pest/>) identifies possible PEST regions in a submitted probe using the Molecular fraction of the P, E, S and T components, and the hydrophobicity index of the region.

PESTfind (<http://embl.bee.univie.ac.at/embnet/tools/bio/PESTfind/>) is used to determine whether a protein contains a PEST region a computer program. The algorithm finds PEST sequences as hydrophilic stretches of amino acids greater than or equal to 12 residues in length. Such regions contain at least one P, one E or D and one S or T. They are flanked by lysine (K), arginine (R) or histidine (H) residues, but positively charged residues are disallowed within the PEST sequence.

Coils, Paircoil and Multicoil

Coils (http://www.ch.embnet.org/software/COILS_form.html) is a program that compares a sequence to a database of known parallel two-stranded coiled-coils and derives a similarity score. After comparing this score to the distribution of scores in globular and coiled-coil proteins, the program then calculates the probability that the sequence will adopt a coiled-coil conformation.

Paircoil (<http://nightingale.lcs.mit.edu/cgi-bin/score>) predicts the location of coiled-coil regions in amino acid sequences.

MultiCoil program (<http://nightingale.lcs.mit.edu/cgi-bin/multicoil>) predicts the location of coiled-coil regions in amino acid sequences and classifies the predictions as dimeric or trimeric. The method is based on the PairCoil algorithm.

4 SECONDARY STRUCTURE ANALYSIS AND PREDICTION

There are several protein secondary structure prediction methods and the most important of these methods are

- ✓ Chou-Fasman method
- ✓ GOR methods
- ♦ Nearest neighbour methods
- ♦ Hidden Markov models
- ♦ Neural networks
- ♦ Multiple alignments based self-optimization method

Chou-Fasman Method

The Chou-Fasman algorithm for the prediction of protein secondary structure is one of the most widely used predictive methods. The Chou-Fasman method of secondary structure prediction depends on assigning a set of prediction values to a residue and then applying an algorithm to the conformational parameters and positional frequencies.

The conformational parameters for each amino acid were calculated by considering the relative frequency of a given amino acid within a protein, its occurrence in a given type of secondary structure, and the fraction of residues occurring in that type of structure. These parameters are measures of a given amino acid's preference to be found in helix, sheet or coil.

$P(\alpha)$ $P(\beta)$ $P(\text{turn})$ are the preference parameters for the 20 amino acids for α -helix, β -strand and β -turn respectively.

The algorithm is as follows:

1. Assign all of the residues in the peptide the appropriate set of parameters.
2. Scan through the peptide and identify regions where 4 out of 6 contiguous residues have $P(\alpha\text{-helix}) > 1.00$. That region is declared an alpha-helix. Extend the helix in both directions until a set of four contiguous residues that have an average $P(\alpha\text{-helix}) < 1.00$ is reached. That is declared the end of the helix. If the segment defined by this procedure is longer than 5 residues and the average $P(\alpha\text{-helix}) > P(\beta\text{-sheet})$ for that segment, the segment can be assigned as a helix.
3. Repeat this procedure to locate all of the helical regions in the sequence.
4. Scan through the peptide and identify a region where 3 out of 5 of the residues have a value of $P(\beta\text{-sheet}) > 1.00$. That region is declared as a beta-sheet. Extend the sheet in both directions until a set of four contiguous residues that have an average $P(\beta\text{-sheet}) < 1.00$ is reached. That is declared the end of the beta-sheet. Any segment of the region located by this procedure is assigned as a beta-sheet if the average $P(\beta\text{-sheet}) > 1.05$ and the average $P(\beta\text{-sheet}) > P(\alpha\text{-helix})$ for that region.
5. Any region containing overlapping alpha-helical and beta-sheet assignments are taken to be helical if the average $P(\alpha\text{-helix}) > P(\beta\text{-sheet})$ for that region. It is a beta sheet if the average $P(\beta\text{-sheet}) > P(\alpha\text{-helix})$ for that region.
6. To identify a bend at residue number j , calculate the following value;

$$p(i) = f(j)f(j+1)f(j+2)f(j+3)$$

where the $f(j+1)$ value for the $j+1$ residue is used, the $f(j+2)$ value for the $j+2$ residue is used and the $f(j+3)$ value for the $j+3$ residue is used.

The main *helix forming residues* H are ala, glu, leu and met. The main *helix breaking residues* B are proline and glycine.

The main *beta sheet forming residues* H are ile, val, and tyr. The main *beta sheet breaking residues* B are pro, asp, and glu. Proline's unique structure in which the side chain is cyclically attached to the backbone gives it unique structural properties. It cannot assume the backbone dihedral angles typical of alpha and beta structures, nor can it form appropriate hydrogen bonds.

In the Chou and Fasman method, the central positions of the turn ($i+1$ and $i+2$ position) show strong preferences for pro (30%), ser (14%), lys, asp, arg, and thr (the latter four about 10% each) at the first position, and asn (19%), gly (19%), asp (18%), ser (13%), cys (12%) and tyr (11%) at the second position.

PeptideStructure (http://www.accelrys.com/products/gcg_wisconsin_package/Program_list.html) uses the original Chou-Fasman as well as a modification of the original method.

GOR (Garnier, Osguthorpe and Robson) Method

GOR is a method that assumes that amino acids up to 8 residues on each side influence the secondary structure of the central residue. This program is now in its fourth version. The accuracy of GOR when checked against a set of 267 proteins of known structure is 64%. This implies that 64% of the amino acids were correctly predicted as being helix, sheet or

coil. The algorithm uses a sliding window of 17 amino acids. All possible pairs of amino acids in this window are checked for their information content as to predicting the structure of the central amino acid by comparing them to a set of 266 other proteins of known structure. The method works better for helix than for sheet, because sheet is dependent on longer-range interactions between non-adjacent sequence fragments. GOR underpredicts the number of β strands and usually you can predict 36.5% of the β strands correctly.

GOR IV (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html) uses all possible pair frequencies within a window of 17 amino acid residues. One output gives the sequence and the predicted secondary structure in rows, H = helix, E = extended or β strand and C = coil. The other output gives the probability values for each secondary structure at each amino acid position.

Consider the following protein sequence (SWISS-PROT ID: Q969F8):

MHTVATSGPNASWGAPANASGCPGCGANASDGPVPSRAVDALVPLFFAAL
 MLLGLVGNLSVIYVICRHKPMRTVTNFYIANLAATDVTFLCCVPFTALLYPLPG
 WVLGDFMCKFVNYIQQVSVQATCATLTAMSVDRWYVTVFPLRALHRRTPRLAL
 AVLSIWVGSAAVSAPVLALHRLSPGPRAVCSEAFPSRALERAFALYNLLALYLL
 PLLATCACYAAMLRLGRVAVRPAPADSALQGQVLAERAGAVRAKVSRLVAA
 VLLFAACWGPIQLFLVLQALGPAGSWHPRSAAAYALKTWAHCMSYSNSALNP
 LYAFLGSHFRQAFRRVCPCAPRRPRRRPRRPGSPDPAAPHAELHRLGSHAPARA
 KPGSSGLAARGLCVLGEDNAPL

The first part of the GOR IV output is as in Figure 11.5.

```

      10      20      30      40      50      60      70
      |      |      |      |      |      |      |
MHTVATSGPNASWGAPANASGCPGCGANASDGPVPSRAVDALVPLFFAALMLLGLVGNLSVIYVICRHK
      hhhh hhhhhhhhhhhh eeeeeee
KFEPTVTNFYIANLAATDVTFLCCVPFTALLYPLPGWVLGDFMCKFVNYIQQVSVQATCATLTAMSVDR
      eee hhhh eeeee eee eeeee eeeee ee
WYVTVFPLRALHRRTPPLALAVLSIWVGSAAVSAPVLALHRLSPGPRAVCSEAFPSRALERAFALYNLL
      eeee hhhh hhhhhhhhhhhh hhhhhh hhhhhhhhhhhhhh
ALYLLPLLATCACYAAMLRLGRVAVRPAPADSALQGQVLAERAGAVRAKVSRLVAAVLLFAACWGPIQ
      hhhhhhhh hhhhhhhhhh eeee hhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhhh
LFLVLQALGPAGSWHPRSAAAYALKTWAHCMSYSNSALNPALLYAFLGSHFRQAFRRVCPCAPRRPRRPPR
      hhhhhhhh hhhhhhhh eeeee hhhhhh eeeeeeeeee
PGSPDPAAPHAELHRLGSHAPARAQKPGSSGLAARGLCVLGEDNAPL
      hhhhhh

```

Sequence length : 398

GOR4 :

Alpha helix	(Hh) :	138 15	34.67%
β helix	(Gg) :	0 15	0.00%
β helix	(Ii) :	0 15	0.00%
Beta bridge	(Eb) :	0 15	0.00%
Extended strand	(Ec) :	60 15	15.08%
Beta turn	(Tt) :	0 15	0.00%
Bend region	(B) :	0 15	0.00%
Pandom coil	(C) :	200 15	50.25%
Ambiguous states (7)	:	0 15	0.00%
Other states	:	0 15	0.00%



FIGURE 11.5 GOR IV output.

Nearest Neighbour Method

The Nearest neighbour method is based on the hypothesis that short homologous sequences of amino acids have the same secondary structure tendencies. A list of short sequence fragments is made by sliding a window of length n along a set of approximately 100–400 training sequences of known structure but minimal sequence similarity. For example in SIMPA96 (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_simpa96.html), one of the implementations of this method, n is 13 and there are 300 proteins. The secondary structure of the central amino acid in each training window is recorded and a sliding window of the same size is then selected from the query sequence.

The sequence in the window at each position of the query sequence is compared to each of the above training fragments and the 50 best matching fragments are identified. Scoring matrices, multiple sequence alignments, etc. may be used at this step. In SIMPA96 implementation, comparisons are made with the secondary structure assignments of Kabash and Sander from X-ray data and an empirically determined similarity matrix which assigns a sequence similarity score between any two sequences of 7 residues in length.

The frequencies of the known secondary structure of the middle amino acid in each of these matching fragments are then used to predict the secondary structure of the middle amino acid in the query window.

Figure 11.6 the output of SIMPA96 for the same protein sequence used with GOR4 earlier.

```

      10      20      30      40      50      60      70
MHTVATSGPNASWGAPANASGCGPGGANASDGPVSPRAVDAGLVPLFFAALMLLGLVGNLSVIYVICRH
EEEEe                hhhhhhhhhHHHHHhheee eeEEEEe
KPMRTVTNIFYIANLAATDVTFLLCCVPFTALLYPLPGWVLDGFMCKFVNTYIQWVQATCATLTAMSVDR
     eee     eEEEEe   eee hhhHHHHHHHHhLhhhh
WYTVTFPLRALHRTFRLALAVLSIWVGSAAVSAPVLALHPLSPGPATCSEAFPSALERAFALYNLL
eeee hhhHHHh hhhhhheeeeee hhhhhh hhhHHHHHHHHHHHH
ALYLLPLLATCACTAAMLPHLGRVAVRPAFADSAALQGVLAERAGAVFAKVSPLVAAVVLLFAACWGPIQ
HHHHHHHHHHHHHHHHHHHH eee hhhHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
LFLVLQALGPAGSVHPSPSYAAYALWTVAHCHMYENSALNPLLTAFLGSHFRQAFPPVCFPCAPPPPPPPPP
HHHHHHHh hhhhhHHHHHHhh hhhhhhh hhhHHHHHh
PGSPDPAAPHAEHLRGLGSHAPAPAAQKPGSSGLAARGLCVLGEDNAFL
     hhhhhh     eeee

```

Sequence length : 398

SIMPA96 :

Alpha helix	(Hh) :	160	13	40.10%
3 ₁₀ helix	(Gg) :	0	13	0.00%
Pi helix	(Ii) :	0	13	0.00%
Beta bridge	(Eb) :	0	13	0.00%
Extended strand	(Ee) :	44	13	11.03%
Beta turn	(Tt) :	0	13	0.00%
Send region	(Ss) :	0	13	0.00%
Random coil	(C) :	194	13	48.62%
Ambiguous states (?)	:	0	13	0.00%
Other states	:	1	13	0.25%



FIGURE 11.6 Output from SIMPA96. Compare it with the output of GOR4 for

NNSSP (<http://bioweb.pasteur.fr/seqanal/interfaces/nnssp-simple.html>) is another program that predicts the secondary structure combining the nearest neighbour and multiple sequence alignment approaches.

Hidden Markov Models (HMMs)

HMMs have been earlier discussed in Chapter 7. They can be used to predict the secondary structure of a protein of a given structural class (e.g. $\alpha + \beta$) as used in the structural classification databases. Each HMM is trained with the sequences of the proteins in that structural class. The models are used with a query sequence to predict both the class and the secondary structure of the protein. Pfam (<http://www.sanger.ac.uk/Software/Pfam/search.shtml>) uses the HMM approach.

Neural Networks

Most of the effective structure prediction models extract patterns from databases of known protein structures. Neural networks comprise a particular tool for pattern recognition and classification.

The simplest layered feed-forward neural network consists of a layer of input units and a layer of output units. Signals are transmitted from input to output layer (feed-forward) via the connections. In Figure 11.7, a simple neural network is shown. There are two input units (J's) and one output unit.

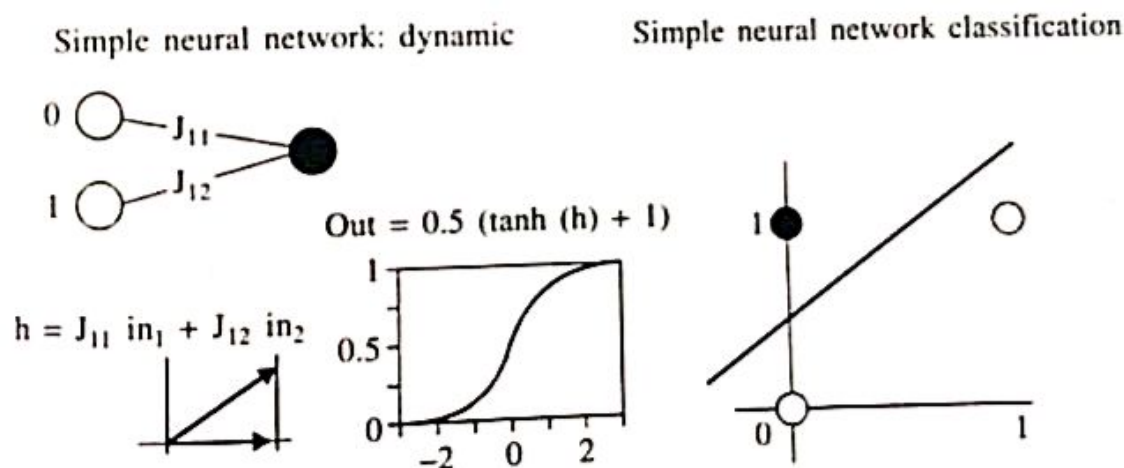


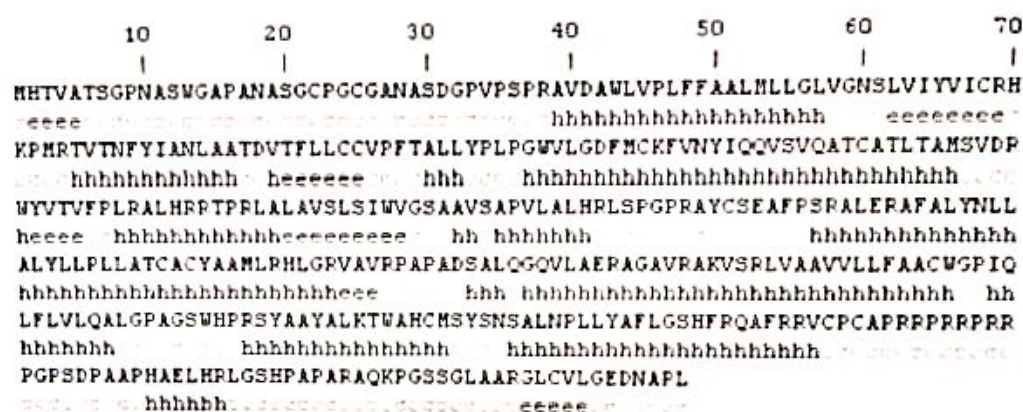
FIGURE 11.7 A simple neural network example.

The value of each input unit (example: 0 for unit 1; 1 for unit 2) is multiplied with the strength of the connection; the products sum to a local field (h) representing the signal that arrives at the output unit. The multiplication represents a projection of the input vector onto the vector of the connections. (2) The final output is determined by applying a sigmoid function (shown is the hyperbolic tangent) to the local field. The result is that the output is constrained to values between 0 and 1. On the right hand side the potential of such a network is illustrated: a line separates the open, and the dark circles.

Neural networks can be used for protein prediction. The protein sequence is translated into patterns by shifting a window of n adjacent residues (typical values of $n = 13-21$)

Training or learning a neural network implies changing the connections so that the error decreases for the given examples. A training set can comprises about 30,000 examples. If training is successful, the patterns are correctly classified. The network can succeed in extracting general rules by the classification of the training patterns. The generalization ability is checked by another set of test samples for which the mapping of sequence window to secondary structure is known as well. Sufficient testing is crucial and has to meet two requirements. First, any significant sequence similarity between test and training set has to be removed. Second, evaluations of expected prediction accuracy have to be based on a sufficient number of test proteins (>100).

HNN (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_nn.html) is Hierarchical Neural Network based program that gives a secondary structure prediction. The output of HNN for the above sequence (used for GORIV above) is shown in Figure 11.8.



Sequence length : 398

HNIN :

Alpha helix	(Hh) :	209	is	52.51%
3_{10} helix	(Gg) :	0	is	0.00%
Pi helix	(Ii) :	0	is	0.00%
Beta bridge	(Ebb) :	0	is	0.00%
Extended strand	(Ee) :	39	is	9.80%
Beta turn	(Tt) :	0	is	0.00%
Bend region	(Bb) :	0	is	0.00%
Random coil	(Cc) :	150	is	37.69%
Ambiguous states (?)	:	0	is	0.00%
Other states	:	0	is	0.00%

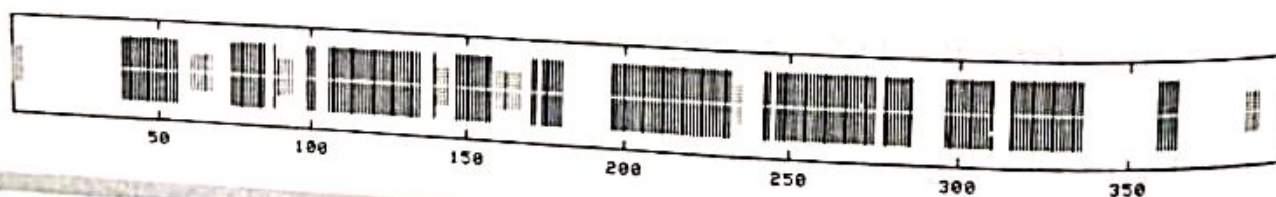


FIGURE 11.8 Output from HNN for the example protein sequence.

nnPredict (<http://www.empharm.ucsf.edu/%7Enomi/nnpredict.html>) predicts the secondary structure type for each residue in an amino acid sequence. The basis of the prediction is a two-layer, feed-forward neural network. The predicted type will be either: 'H',

Secondary structure prediction (H = helix, E = strand, - = no prediction):

-E-----H-NNNNNNNNNNNNNNNNNN--
 -EEEEEE-----E-NNENN--NNNNEEEE-----NNNNNNN-
 NNEEEE-EENNNNNN----EEEE-NNNNN----NNNNNNNEEEEE-----NNNN
 -E-----NNNNNNNNNNNNNNNNNNNN-NNNNNNNNNNNNNNNNNNNN-NNE-----

FIGURE 11.9 Output from nnPredict.

PSA (<http://bmerc-www.bu.edu/psa/request.htm>) is also a secondary structure prediction tool. It has 3 options for analysis: Monomeric-Soluble Type-1 analysis, Minimal Type-2 analysis, and WD-repeat WD-repeat analysis.

PSIPRED (<http://bioinf.cs.ucl.ac.uk/psipred/>) incorporates methods PSIPRED, GenTHREADER and MEMSAT 2 for predicting structural information about any given protein from its amino acid sequence alone. PSIPRED is a secondary structure prediction method, MEMSAT is a transmembrane topology prediction method and GenTHREADER is a new sequence profile based fold recognition method. PSIPRED carries out secondary structure prediction on a protein incorporating two feed-forward neural networks that perform an analysis on output obtained from PSI-BLAST. Version 2.0 of PSIPRED includes a new algorithm that averages the output from up to 4 separate neural networks in the prediction process to further increase prediction accuracy.

Multiple Alignments Based Self-Optimization Method

SOPMA (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopma.html) is a secondary structure prediction program (Self-Optimized Prediction Method) that uses multiple alignments. SOPMA correctly predicts 69.5% of amino acids for a three-state description of the secondary structure (alpha-helix, beta-sheet and coil) in a whole database containing 126 chains of non-homologous (less than 25% identity) proteins. Joint prediction with SOPMA and PHD correctly predicts 82.2% of residues for 74% of co-predicted amino acids. The output from SOPMA for the sequence used earlier is given in Figure 11.10.

11.5 MOTIFS, PROFILES, PATTERNS AND FINGERPRINTS SEARCH

Motifs extend the ideas of sequence/sequence comparison to use in sequence/motif or sequence/family comparisons. Use of motif-based information for comparisons is more useful as motifs are already associated with structural and functional information. Motif or family comparisons are also more sensitive because motifs represent a higher-level generalization of the features that are important for a given structural or functional feature.

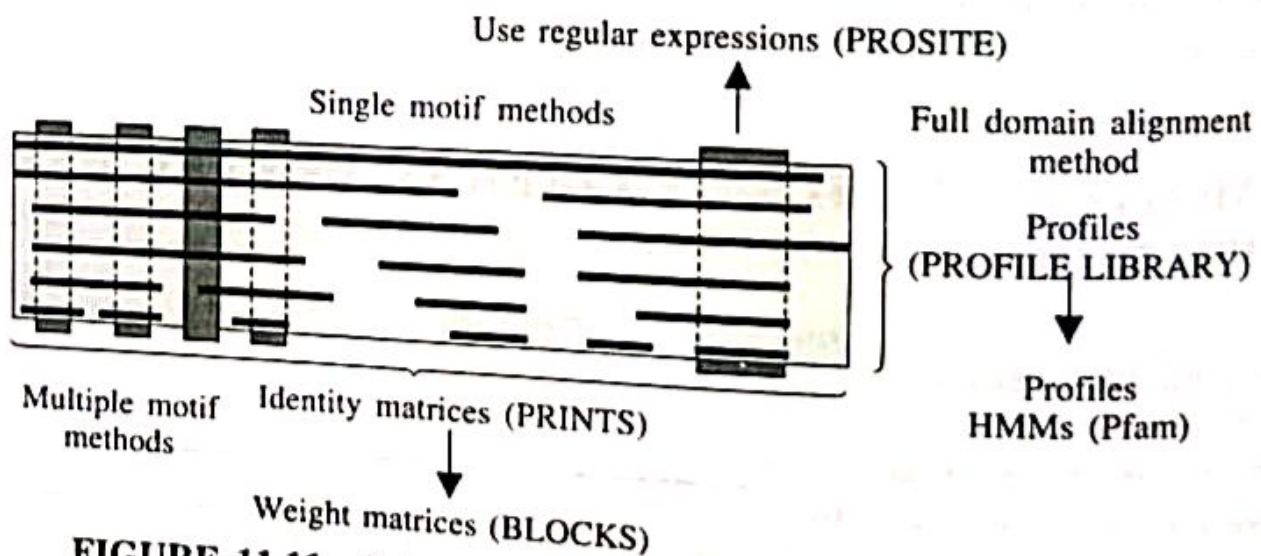
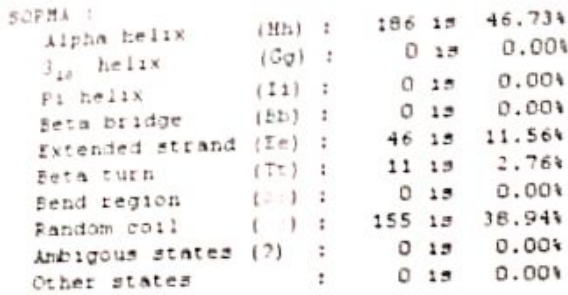


FIGURE 11.11 Schematic of pattern recognition methods.

Profiles

Profiles, as already discussed, are a numerical representation of a multiple sequence alignment. Within the multiple sequence alignment is the intrinsic sequence information that represents the common characteristics of that particular collection of sequences. Profiles help find the similarities between these sequences and help in identification and analysis of distant related proteins. Profiles are constructed by taking a multiple sequence alignment representing a protein family. A position-specific scoring table (PSSM) is constructed on the lines of PAM or BLOSUM.

Profilescan (<http://hits.isb-sib.ch/cgi-bin/PFSCAN>) uses a database of profiles to find structural and sequence motifs in protein sequences. Profilescan finds structural and sequence motifs in protein sequences. These motifs are represented as profiles in a library. ProfileScan aligns each profile motif to the sequence, and displays all alignments between the profile and sequence that have a normalized score above a set threshold.

The output for the Metastin sequence used above is given in Figure 11.12. PFSCAN has found two motifs in the sequence.


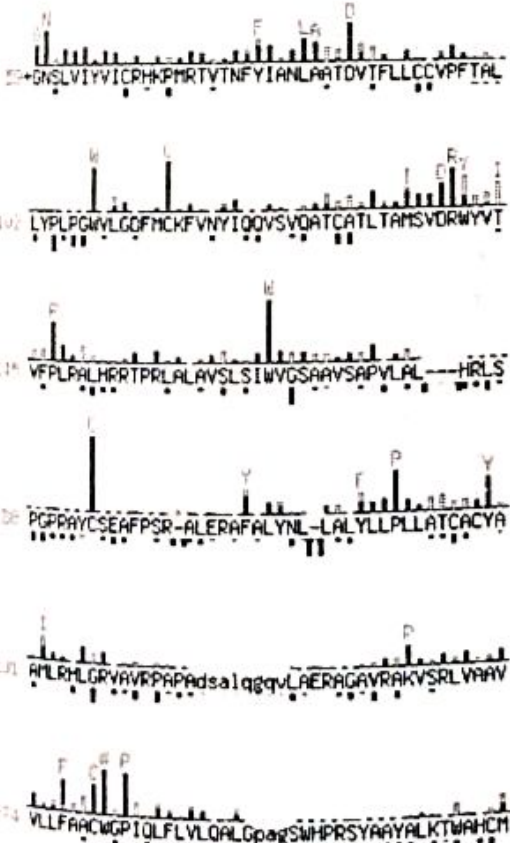
match detail	match score	motif information
 <p>335-352 RNVCPAPRRRRPPPG+</p>	<p>status: ? pos.: 335-352 raw-score = 3 N-score = 3.000 E-value = 2.1e+04</p>	<p>prf:NLS_BP Bipartite nuclear localization signal. PS50079 QDOC50079 InterPro</p>
 <p>59-323 GNSLVIVICPHMPRTVTNFIANLRAVDVTLCCVPFTAL LYPLPGVILGOFMCKFVNYIQOVSVQATCATLTAMSVORWYVT VFPLPALHRTPRALAYSLIIVGSAVSAVPLAL--HRLS PCPRAYCSEAFPSR-ALERAFALYNL-LALYLLPLLATCACYA RLRLHLGRVAVRPAPAdsalggqVLAERAGAVRAKYSRLVAAY VLLFAACWGPITQLFLVLOALGpagsWHPRSYRAYALKTWAHCH</p>	<p>status: ! pos.: 59-323 raw-score = 1843 N-score = 38.908 E-value = 2.1e-31</p>	<p>prf:G_PROTEIN_RECEP_F1_2 G-protein coupled receptors family 1 profile. PS50262 PDOC00210 InterPro</p>

FIGURE 11.12 PFSCAN output for Metastin.

Typical Profile searching goes through the following steps:

1. Assembly of a family of related sequences into a multiple sequence alignment with PileUp (<http://biobase.dk/gcgmanual/pileup.html>). LineUp (<http://biobase.dk/gcgmanual/lineup.html>) is a multiple sequence editor used to create multiple sequence alignments. Pretty (<http://biobase.dk/gcgmanual/pretty.html>) is used to display multiple sequence alignments.
2. Construction of a profile from the alignment with the program ProfileMake (<http://biobase.dk/gcgmanual/profilemake.html>).
3. Comparison of the profile to a database of sequences with ProfileSearch (<http://biobase.dk/gcgmanual/profilemake.html>).
4. Display of the optimal alignments between each sequence in the ProfileSearch output list and the group of aligned sequences (represented by the profile consensus) using ProfileSegments (<http://biobase.dk/gcgmanual/profilesegments.html>).
5. A single sequence can be searched with a library of different profiles using the ProfileScan program. ProfileGap (<http://biobase.dk/gcgmanual/profilegap.html>) can be used to make optimal alignments between one or more sequences and a group of aligned sequences represented as a profile.
6. Find structural and sequence motifs in protein sequences, using predetermined parameters to determine significance by using ProfileScan.
7. Compare one or more sequences to a database of profile HMMs (e.g.) the Pfam library, in order to identify known domains within the sequences using Hmmer-Pfam (<http://biobase.dk/gcgmanual/hmmerpfam.html>). HmmerIndex (<http://biobase.dk/gcgmanual/hmmerindex.html>) creates an index for a profile hidden Markov model database so that profile HMMs can be retrieved from the database with HmmerFetch (<http://biobase.dk/gcgmanual/hmmerfetch.html>).
8. Look for sequence motifs by searching through proteins for the patterns defined in the PROSITE Dictionary of Protein Sites and Patterns using Motifs (<http://biobase.dk/gcgmanual/motifs.html>).

All the above GCG programs are with Accelrys Inc. now. Accelrys is a wholly owned subsidiary of Pharmacia Inc. The program listing can be found at http://www.accelrys.com/products/gcg_wisconsin_package/program_list.html.

Frame-ProfileScan (http://www.isrec.isb-sib.ch/software/PFRAMESCAN_form.html) uses the frame-search capabilities of Pfsan to query the collection of Prosite profiles with a single DNA sequence. The six reading frames of the DNA query are inspected; coding frameshifts in the DNA sequence are supported.

Patterns

Patterns also represent the common characteristics of a protein family, but it does not contain any weighing information.

Pratt (<http://www.ebi.ac.uk/pratt/>) allows the user to search for patterns conserved in a set of protein sequences. The user can specify what kind of patterns should be searched for, and how many sequences should match a pattern to be reported—there are options for pattern conservation, restrictions, number of pattern symbols, flexible spacers, etc.

Aligned Segment Statistical Evaluation (ASSET) at http://bip.weizmann.ac.il/software/sg_software/asset.html, can locate patterns, combine related patterns and provide a measure of statistical significance of the patterns without any prior information that the patterns are actually present.

Motifs

Motifs are defined by a heterogeneous collection of predictors, which currently include regular expressions, generalized profiles and HMMs.

Hits (<http://hits.isb-sib.ch/>) is a database devoted to protein domains, also a collection of tools for the investigation of the relationships between protein sequences and motifs described on them.

The tools for querying and exploring the Hits database are as follows:

1. Query by protein produces a list of motifs present in one or several proteins.
2. Query by motif produces a list of proteins that contain one or several motifs.
3. "At least" query is another query by motif form that produces a list of proteins that share a minimal number of motifs.
4. Pattern search using a user-supplied regular expression to search protein databases.
5. Metamotif search looking for arrangements of motifs in protein databases.

The output from Hits for the following list of proteins is given in Figure 11.13.

Result

- Motif count: 3
- Motif names: prfCYS_RICH, pfamPEP_M12B_PROPEP, pfamREPROLYSIN, prfADAM_MEPRO, prfDISINTEGRIN_2, patZINC_PROTEASE, patDISINTEGRIN_1, pfamDISINTEGRIN

[more about these motifs](#)

- Match count: 43
- Matches:

sw:DISI_BOTCO	29	48	pat:DISINTEGRIN_1	-
sw:DISI_BOTCO	1	72	prf:DISINTEGRIN_2	17.279
sw:DISI_BOTCO	6	38	prf:CYS_RICH	9.433
sw:DISI_BOTCO	1	72	pfam:DISINTEGRIN	27.178
sw:DISI_BOTAT	29	48	pat:DISINTEGRIN_1	-
sw:DISI_BOTAT	1	71	prf:DISINTEGRIN_2	17.333
sw:DISI_BOTAT	6	38	prf:CYS_RICH	9.433
sw:DISI_BOTAT	1	71	pfam:DISINTEGRIN	26.229
sw:DISC_TRIFL	32	51	pat:DISINTEGRIN_1	-
sw:DISC_TRIFL	1	75	prf:DISINTEGRIN_2	15.177
sw:DISC_TRIFL	9	41	prf:CYS_RICH	9.433
sw:DISC_TRIFL	4	75	pfam:DISINTEGRIN	26.384
sw:DISI_AGKHA	29	48	pat:DISINTEGRIN_1	-
sw:DISI_AGKHA	1	71	prf:DISINTEGRIN_2	17.805
sw:DISI_AGKHA	6	38	prf:CYS_RICH	9.433
sw:DISI_AGKHA	1	71	pfam:DISINTEGRIN	26.104
sw:DISF_TRIFL	27	46	pat:DISINTEGRIN_1	-
sw:DISF_TRIFL	1	70	prf:DISINTEGRIN_2	15.675
sw:DISF_TRIFL	4	36	prf:CYS_RICH	9.433
sw:DISF_TRIFL	1	70	pfam:DISINTEGRIN	24.458
sw:DISI_CROVE	29	48	pat:DISINTEGRIN_1	-
sw:DISI_CROVE	1	72	prf:DISINTEGRIN_2	17.453
sw:DISI_CROVE	6	38	prf:CYS_RICH	9.433
sw:DISI_CROVE	1	72	pfam:DISINTEGRIN	27.739

FIGURE 11.13 Output from Hits.

sw:DISI_BOTCO, sw:DISI_BOTAT, sw:DISC_TRIFL, sw:DISI_AGKHA,
 sw:BOTR_BOTJA, sw:DISF_TRIFL, sw:DISI_CROVE, sw:DISI_TRIFL,
 sw:HRTE_CROAT, sw:DISA_ERIMA, sw:DISI_CROVL

MEME (<http://meme.sdsc.edu/meme/website/intro.html>) stands for Multiple EM for Motif elicitation. It is a tool for discovering motifs in a group of related protein sequences (it can take DNA sequences also as the input).

MEME represents motifs as position-dependent letter-probability matrices that are used to describe the probability of each possible letter at each position in the pattern. Individual MEME motifs do not contain gaps. Patterns with variable-length gaps are split by MEME into two or more separate motifs.

MEME takes as input a group of protein sequences (the training set) and outputs as many motifs as requested. MEME uses statistical modeling techniques to automatically choose the best width and description for each motif.

MEME works in tandem as a system with MAST (Motif Alignment and Search Tool). The MEME/MAST system allows you to discover motifs in groups of related protein sequences using MEME and then search sequence databases using motifs by utilizing MAST.

Meta-MEME (<http://metameme.sdsc.edu/>) combines motif models from MEME into a hidden Markov model framework for use in searching sequence databases. The input to Meta-MEME is a set of similar protein sequences, as well as a set of motif models discovered by MEME. Meta-MEME combines these models into a single, motif-based hidden Markov model and uses this model to produce a multiple alignment of the original set of sequences and to search a sequence database for homologs.

Gibbs Motif Sampler (<http://bayesweb.wadsworth.org/gibbs/gibbs.html>) allows you to identify motifs in protein sequences (or DNA sequences). The objective is to take a given set of amino acids (or nucleotide sequences) and determine common motif elements within them. One approach known as site sampling assumes that each sequence contains exactly one motif element for each motif type. The alternative Bernoulli motif sampler assumes that each sequence can contain zero or more motif elements of each motif type.

Blocks

Blocks are multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins. The Blocks Database (http://blocks.fhcrc.org/blocks/help/blocks_release.html) was constructed by the PROTOMAT system using the MOTIF algorithm.

A Blocks Search (http://blocks.fhcrc.org/blocks/blocks_search.html) for Metastin sequence used before gives the output as follows:

Size = 398 Amino Acids

Blocks Searched = 11182

Alignments Done = 4738033

Cutoff combined expected value for hits = 1

Cutoff block expected value for repeats/other = 1

Combined

Family	Strand	Blocks	E-value
IPB000276 Rhodopsin-like GPCR superfamily	1	4 of 4	2.5e-17
IPB002896 Herpesvirus glycoprotein D	1	1 of 6	0.18
IPB000444 Xanthine/uracil permeases family	1	1 of 2	0.21
IPB000832 G-protein coupled receptors family	1	2 of 7	0.82
IPB000542 Acyltransferase ChoActase/COT/C	1	1 of 10	0.91

Domain search using ProDom (<http://prodes.toulouse.inra.fr/prodom/2002.1/html/form.php?typeform=SPTR>) gives the results as shown in Figure 11.14.

ProDom

Home

Form

Contact

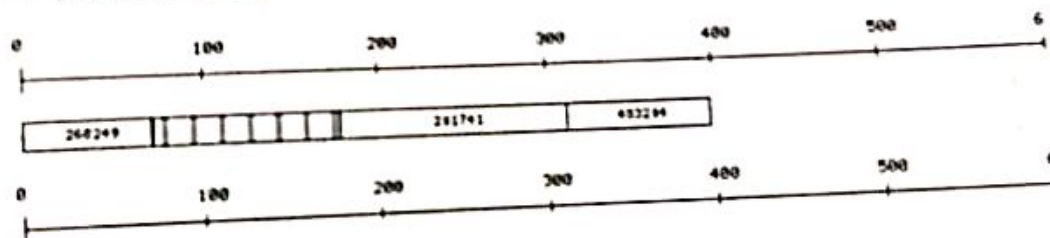
Site map

Release 2002.1

Graphical view of selected protein(s)

SEARCH

new window | close



new window | close

Domain arrangements 1 → 1 among 1 different arrangement for 1 protein

FIGURE 11.14 Output from ProDom.

Fingerprints

A fingerprint is a group of conserved motifs or elements that are used to characterize a particular protein family. Usually the motifs do not overlap, but are separated along a sequence, though they may be contiguous in 3-D space. Fingerprints can encode protein

folds and functionalities more flexibly and powerfully than can single motifs. Diagnostically, this is more powerful than using single motifs by virtue of the biological context afforded by matching motif neighbours.

PRINTS (<http://bioinf.man.ac.uk/dbbrowser/PRINTS/>) is a compendium of protein fingerprints. PRINTS is a companion to the BLOCKS, PROSITE, Pfam and ProDom databases.

Pscan (<http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/pscan.html>) finds matches between a query protein sequence and the motifs or elements in the PRINTS database. It reports various classes of matches:

1. Matches where all elements of a motif exist in the correct order
2. Matches where all elements exist but some are in the incorrect order
3. Matches where some elements match and are in the correct order
4. Miscellaneous matches.

Other Programs

InterPro (<http://www.ebi.ac.uk/interpro>) provides an integrated view of the commonly used signature databases, and has an intuitive interface for text- and sequence-based searches. InterPro is distributed by anonymous FTP and is accessible for interactive use via the EBI Web server, which can also be reached via each of the member databases. Where applicable, InterPro also has cross-references to the BLOCKS database.

InterPro release 5.3 (November 2002) was built from the following participating databases:

- ◆ Pfam 7.7
- ◆ PRINTS 33.0
- ◆ PROSITE 17.25
- ◆ ProDom 2001.3
- ◆ SMART 3.4
- ◆ TIGRFAMs 2.1
- ◆ Current SWISS-PROT + TrEMBL data.

It contains 6725 entries, representing 1453 domains, 5121 families, 136 repeats, and 15 post-translational modification sites. Overall, there are 2932939 InterPro hits from 850953 SWISS-PROT + TrEMBL protein sequences.

11.6 METHODS OF SEQUENCE-BASED PROTEIN PREDICTION

There are two fundamental approaches in using the sequence data for making protein structure prediction. One of the approaches uses pattern recognition methods. The pattern recognition approach is used to detect similarity between sequences. This gives indications to infer related structures & functions.

The other approach is to the sequence data without any template. This approach is called *ab initio* prediction. *Ab initio* approach is truly a prediction approach and is used to deduce structure and infer function directly from sequence.

In this section, you will learn the pattern recognition approaches for protein prediction. The use of pattern recognition approach is based on percentage identity, which is an important indicator of the level of evolutionary divergence and functional or structural similarity between compared sequences. Different alignment methods have different areas of optimum application.

Pair-wise alignment algorithms perform well at >50% identity. For <50%, but more than 30% identity, you can use consensus information from multiple alignments. For <30% identity, you can use the motifs or profile methods. At the lowest levels of identity (<20% identity), where alignments are no longer statistically significant, structure prediction algorithms like homology modeling or threading can be used.

Alignment and Database Search Methods

The most common tools for a database search are BLAST (PSI-BLAST, BLASTP) and FASTA. There are other tools also like MaxHom and SSEARCH.

MaxHom at PredictProtein server (<http://www.embl-heidelberg.de/predictprotein/predictprotein.html>) is a dynamic multiple sequence alignment programs that finds similar sequences in a database. MaxHom builds up a protein family (defined as all closely related proteins likely to have similar structures) in two steps:

1. In sweep 1, sequences are aligned consecutively to the search sequence by a standard dynamic programming method. After each sequence has been added a profile is compiled, and used to align the next sequence.
2. In sweep 2, after all sequences with significant homology have been picked from the BLASTP output, the profile is recompiled, and the dynamic programming algorithm starts once again to align consecutively the sequences, this time using the conservation profile as derived after completion of sweep 1.

You can use SSEARCH (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_ssearch.html) to search PDB.

You can use Consensus tool (<http://www.bork.embl-heidelberg.de:8080/Alignment/consensus.html>) to calculate the consensus for the CLUSTAL or MSF multiple alignment. You can also use Consensus Secondary Structure Prediction (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_seccons.html) that uses several protein prediction methods.

Homology or Comparative Modeling

The basic assumption of homology modeling is that the unknown structure and the homologous template protein of known structure have nearly identical backbone structure in the aligned regions. The basic action is to correctly place the side chains of U into the backbone of T

You can use MODELLER (<http://guitar.rockefeller.edu/modeller/modeller.html>) and SWISS-MODEL (<http://www.expasy.org/swissmod/SWISS-MODEL.html>) to do homology modelling.

MODELLER can be used for homology or comparative modeling of protein three-dimensional structures. You need to provide an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms. MODELLER implements comparative protein structure modeling by satisfaction of spatial restraints. MODELLER can also perform tasks like *de novo* modeling of loops in protein structures, optimization of various models of protein structure with respect to a flexibly defined objective function, multiple alignment of protein sequences and/or structures, clustering, searching of sequence databases, comparison of protein structures, etc.

SWISS-MODEL is a fully automated protein structure homology-modeling server accessible via the ExpASY web server, or from the program DeepView (Swiss Pdb-Viewer). SWISS-MODEL goes through the following five steps:

- ◆ Search for suitable templates
- ◆ Check sequence identity with target
- ◆ Create ProModII jobs
- ◆ Generate models with ProModII
- ◆ Energy minimization with Gromos96

Remote Homology Modeling (Threading)

Protein threading is the method of aligning a protein sequence (the target sequence) with a protein sequence whose structure is known (the template protein). The alignment of the two proteins is in such way that mapping residues of the target sequence onto a template sequence according to the alignment gives an accurate model of the backbone structure of the target. Threading is used for cases with <25% pair-wise sequence identity.

Threading algorithms use the database of known 3-D structures to classify new protein sequences and to predict their structures. Their premise is that detection of structural similarity by sequence-structure threading will recognize remote evolutionary relationships that are not detectable by sequence comparison alone. This premise is logical as protein evolution is known to strongly conserve the core structures of protein families. However, it is not clear how much improvement one should expect. The structures of remote homologs differ greatly in detail, with backbone root mean square residuals (RMS) only in the range of 2–3 Angstroms, and the conformational energy calculations used by threading algorithms may easily reject model structures with this level of error.

There are two basic algorithms for threading: Profile Method and Core threading model. A third model based on recursive dynamic programming has also been developed.

Profile method is also called the 1-D-3-D profiles method. In this method, the structure of the template sequence is aligned to a vector or a string of descriptors that describe the 3-D structure of the target sequence. This vector contains fingerprint of the structural environment of each residue inside the template protein. The main idea of this threading model is that the target protein reproduces the structural features of the template protein.

For each residue position in the structure the environment is described in terms of:

1. The local structure (α , β , or loop).
2. Solvent accessibility (3 states—buried, partially buried and exposed).
3. The degree of burial by polar or apolar atoms.

Hence, each position can be categorized into 18 environment classes. Each amino acid has a preference to a particular environment. For example Leu has a preference for being in a class with a high fraction of buried side chain area, whereas Asp has a very low preference for that position. A profile of the actual environments that any one of the amino acids resides in can be determined using PDB. A comparison can then be made between these environments and one of the 18 mentioned above. Computing a score for each environmental class and each amino acid can give the score table. You can then use a 2-D dynamic programming matrix to find the best score between the amino acid sequences of the unknown to the descriptors of the environmental classes of a target structure. The profile method fails when the structural profiles change from the template protein to the target.

The second model is called the core-threading model. It is based on the analysis of pair-wise interactions between structurally adjacent residues in the protein. It uses the branch-and bound method. The problem of this model is that it concentrates only on the core regions and overlooks the loop regions that are also very important for some proteins.

123-D+ (<http://123-D.ncifcrf.gov/123-D+.html>) is a program which combines sequence profiles, secondary structure prediction, and contact capacity potentials to thread a protein sequence through the set of structures.

LIBRA I (http://www.ddbj.nig.ac.jp/htmls/E-mail/libra/LIBRA_I.html), abbreviation for "Light Balance for Remote Analogous proteins" is another tool for threading.

TOPITS (<http://cubic.bioc.columbia.edu/predictprotein>) is a prediction-based threading program that finds remote structural homologues in the DSSP database.

Threader (<http://bioinf.cs.ucl.ac.uk/threader/threader.html>) is a program for predicting protein tertiary structure by recognizing the correct fold from a library of alternatives. However, if a fold similar to the native fold of the protein being predicted is not in the library, then this approach will fail.

The third threading model is a recursive dynamic programming (RDP) method for protein threading which can overcome the above problem. RDP is based on the divide-and-conquer paradigm and maps the target onto the template in a step-wise fashion. RDP has been implemented by ToPLign (<http://cartan.gmd.de/ToPLign.html>).

11.7 AB INITIO APPROACH FOR PROTEIN PREDICTION

In contrast to the above methods, the goal of *ab initio* prediction is to build a model for a given sequence without using a template. *Ab initio* prediction relies on the thermodynamic hypothesis of protein folding (Alfinsen hypothesis discussed in an earlier section). The *ab initio* prediction methods are based on the premise that the native structure of a protein sequence corresponds to its global free energy minimum state. Accordingly, the methods are generally formulated as optimizations.

The methods for *ab initio* prediction are of the following:

Molecular Dynamics (MD) Simulations

MD simulations are of proteins and protein-substrate complexes. MD methods provide a detailed and dynamic picture of the nature of inter-atomic interactions with regard to protein structure and function.

Monte Carlo (MC) Simulations

MC simulations are methods that do not use forces but rather compare energies via the use of Boltzmann probabilities.

Genetic Algorithms (GA) Simulations

GA methods try to improve on the sampling and the convergence of MC approaches.

Lattice Models

Lattice methods are based on using a crude/approximate fold representation (such as two residues per lattice point) and then exploring all or large amounts of conformational space given the crude representation.

The HMMSTR/I-sites/Rosetta Prediction Server (http://www.bioinfo.rpi.edu/~bystre_hmmstr/server.php) predicts the tertiary structure of proteins from the sequence. I-sites predicts local structure, expressed as backbone torsion angles, using a library of sequence-structure motifs. ROSETTA is a Monte Carlo Fragment Insertion protein folding program. HMMSTR, is HMM-based tool for local and secondary structure prediction, based on the I-sites Library.

Petra (<http://www-cryst.bioc.cam.ac.uk/cgi-bin/cgiwrap/charlotte/pet.cgi>) is an *ab initio* protein fragment prediction method.

11.8 METHODS OF 2-D STRUCTURE PREDICTION

Predicting Inter-residue Contacts

Some of the inter-residue contacts can be predicted—for example, helices and strands can be predicted based on hydrogen-bonding pattern between residues. The contacts predicted from secondary structure assignment are short-ranged, i.e. for between residues, nearby in sequence. For successful applications of distance geometry, you need to predict long-range contacts.

Analyses of correlated mutations are done to predict long-range inter-residue contacts. For example, PDGCON (http://www.pdg.cnb.uam.es:8081/pdg_contact_pred.html) is used to predict residue contacts based on correlated mutations derived from multiple alignments.

Other methods use statistics, mean-force potentials, or neural networks. For example, CORNET (<http://prion.biocomp.unibo.it/cornet.html>) is a neural network based predictor of residue contacts.

Predicting Inter-strand Contacts

Prediction of inter-residue contacts can be simplified by predicting the contacts between residues in adjacent strands. The method for predicting inter-strand contacts is based on potentials of mean-force. However, even if the locations of strands in the sequence are known exactly, the pseudo-potentials cannot predict the correct inter-strand contacts in most cases.

AGADIR (<http://www.embl-heidelberg.de/Services/serrano/agadir/agadir-start.html>) is an algorithm to predict the helical content of peptides.

There are a number of tools to predict the post-translational modification of proteins.

11.9 PROTEIN FUNCTION PREDICTION

Protein sequence determines protein structure determines protein function. We will first try to predict protein structure. Then use what we learned, both on the way to structure prediction, and from the predicted structure itself to predict function. Predicting protein function from sequence adds two additional problems in comparison to the unsolved task of structure prediction:

1. Function is not entirely determined by sequence; the environment is crucially important.
2. 'Protein function' is a rather intuitive but ill-defined term. Function is a complex phenomenon associated with many mutually overlapping levels: chemical, biochemical, cellular, physiological, organism mediated, and developmental.

These levels are related in complex ways, e.g. protein kinases can be related to different cellular functions (such as cell cycle), and to a chemical function (transferase) plus a complex control mechanism by interaction with other proteins.

Protein function prediction efforts generally involve attempts to predict biochemical function, cellular role predictions and subcellular location predictions.

HNB Network (http://dag.embl-heidelberg.de/hnb/cgi/show_overview_page.pl?MenuPath=%2Ftool_index) offers automated functional annotation of proteins. HNB Network has three tools for this purpose. Two of them, SMART and miniPEDANT are used to identify putative functional features/domains of the submitted protein, and the STRING tool is used to identify putative functional associations of the submitted protein to other proteins.

SMART is based on domain analysis. The presence of a domain in a protein of interest can be an indication of its function, since other proteins containing the same domain might have already been characterized experimentally. SMART is also used to annotate

transmembrane domains and other sequence features. miniPEDANT is used here to predict the secondary structure of the protein, to annotate charge clusters in the sequence and to identify matches to PROSITE patterns.

STRING is a tool to predict putative functional associations between proteins based on conserved genomic neighborhood—genes that tend to be close neighbours in several genomes (more often than expected by chance) are often transcriptionally coregulated and tend to be functionally associated.

Pfam can also be used similarly for protein function prediction. Pfam has groups of similar function proteins aligned and HMMs generated for each “cluster”. HMM can be generated for an unknown function protein and then compared to HMMs of known proteins for predicted function classification.

The above techniques for protein function prediction use the strategy of predicting the protein functions by classification into putative function groups. They usually fail to predict specific protein functions. Expert systems have been built for prediction at a protein level. One of such expert system is GeneQuiz (http://bric.postech.ac.kr/seminar/kjh/GeneQuiz_biowave/tsld001.htm).

11.10 PROTEIN PREDICTION FROM A DNA SEQUENCE

In the post-genome era, you may have the need to predict the protein structure from a DNA sequence. In such a case, you first need to translate the protein sequence from the DNA sequence. The summary of such tools is given in Table 11.1.

TABLE 11.1 Protein Sequence Prediction from the DNA Sequence

<i>Tool</i>	<i>Description</i>	<i>Address (URL)</i>
Translate	Translates a nucleotide sequence to a protein sequence	http://us.expasy.org/tools/dna.html
Backtranslation	Translates a protein sequence back to a nucleotide sequence	http://www.entelechon.com/eng/backtranslation.html
Transeq	Translates nucleic acid sequences to the corresponding peptide sequence	http://www.ebi.ac.uk/emboss/transeq/

SUMMARY

In this chapter, you have learnt protein structure prediction that is a very useful and important application in bioinformatics. If the amino acid sequence of a protein is known, one can predict the protein structure, its properties and functions, but the situation is compounded due to protein folding problem. A number of protein identification and

characterization tools are available. However, predicting the structure and functions of transmembrane helices, a special class of proteins that include GPCRs, is much needed, as they are important for therapeutic interactions. Although excellent tools and computational methods are available, none of the techniques is fool proof and the area remains a very exciting one for the researchers.

REVIEW QUESTIONS

1. What is Alfinsen's hypothesis?
2. Transmembrane proteins are important in drug discovery process. What are their properties and how can we generate their 3-D structures?
3. What properties you are likely to use for primary structure prediction?
4. Discuss the neural network method for analysis and prediction of secondary proteins?
5. What are the steps in profile searching?
6. What are fingerprints and what are their applications?
7. Explain the following:
 - (a) Motifs
 - (b) Patterns
 - (c) Profiles
8. How can you predict protein sequence from DNA sequence?
9. Predict the function of the following protein from *Methanobacterium thermoautotrophicum*.

>
 MYRITVIPGD GIGVEVMEAA LHVLAQALEIE FEFTHAEAGN ECFRRCGDTL
 PEETLKL VRK ADA TLFGA VT TVPGQKSAII TLRRELDLF A NLRPVKSLPG
 VPCLYPDLDF V~NTEDL YVGDEEYTPE GAVAKRIITR TASRRISQFA
 FQY AQKEGMQ KVT AVHK ANY LKKTDGIFRD FYKV ASEYPQ MEANDYYVDA
 TAMYLTQPQ EFQTIV_nNL FGDILSDEAA GLIGGLGLAP SANIGEKNAL.
 FEPVHGSAPQ IAGKNIANPT AMIL TITLML KHLNKKQEAQ KIEKALQKTL
 MRGIMTPDLG GT ASTMEMAE AIKEEIVKGE

Which functions have been described in this family of proteins? Which aspect of the protein function is conserved between the different functions? Which aspect is the least conserved? Find at least one sequence for each function that has experimental support.