11

PREDICTIVE METHODS USING PROTEIN SEQUENCES

Sharmila Banerjee-Basu

Genome Technology Branch
National Human Genome Research Institute
National Institutes of Health
Bethesda, Maryland

Andreas D. Baxevanis

Genome Technology Branch
National Human Genome Research Institute
National Institutes of Health
Bethesda, Maryland

The discussions of databases and information retrieval in earlier chapters of this book document the tremendous explosion in the amount of sequence information available in a variety of public databases. As we have already seen with nucleotide sequences, all protein sequences, whether determined directly or through the translation of an open reading frame in a nucleotide sequence, contain intrinsic information of value in determining their structure or function. Unfortunately, experiments aimed at extracting such information cannot keep pace with the rate at which raw sequence data are being produced. Techniques such as circular dichroism spectroscopy, optical rotatory dispersion, X-ray crystallography, and nuclear magnetic resonance are extremely powerful in determining structural features, but their execution requires many hours of highly skilled, technically demanding work. The gap in information becomes obvious in comparisons of the size of the protein sequence and structure databases; as of this writing, there were 87,143 protein entries (Release 39.0) in SWISS-PROT but only 12,624 structure entries (July, 2000) in PDB. Attempts to close the gap

center around theoretical approaches for structure and function prediction. These methods can provide insights as to the properties of a protein in the absence of biochemical data.

This chapter focuses on computational techniques that allow for biological discovery based on the protein sequence *itself* or on their comparison to protein families. Unlike nucleotide sequences, which are composed of four bases that are chemically rather similar (yet distinct), the alphabet of 20 amino acids found in proteins allows for much greater diversity of structure and function, primarily because the differences in the chemical makeup of these residues are more pronounced. Each residue can influence the overall physical properties of the protein because these amino acids are basic or acidic, hydrophobic or hydrophilic, and have straight chains, branched chains, or are aromatic. Thus, each residue has certain propensities to form structures of different types in the context of a protein domain. These properties, of course, are the basis for one of the central tenets of biochemistry: that *sequence specifies conformation* (Anfinsen et al., 1961).

The major precaution with respect to these or any other predictive techniques is that, regardless of the method, the results are *predictions*. Different methods, using different algorithms, may or may not produce different results, and it is important to understand *how* a particular predictive method works rather than just approaching the algorithm as a ''black box'': one method may be appropriate in a particular case but totally inappropriate in another. Even so, the potential for a powerful synergy exists: proper use of these techniques along with primary biochemical data can provide valuable insights into protein structure and function.

PROTEIN IDENTITY BASED ON COMPOSITION

The physical and chemical properties of each of the 20 amino acids are fairly well understood, and a number of useful computational tools have been developed for making predictions regarding the identification of unknown proteins based on these properties (and vice versa). Many of these tools are available through the ExPASy server at the Swiss Institute of Bioinformatics (Appel et al., 1994). The focus of the ExPASy tools is twofold: to assist in the analysis and identification of unknown proteins isolated through two-dimensional gel electrophoresis, as well as to predict basic physical properties of a known protein. These tools capitalize on the curated annotations in the SWISS-PROT database in making their predictions. Although calculations such as these are useful in electrophoretic analysis, they can be very valuable in any number of experimental areas, particularly in chromatographic and sedimentation studies. In this and the following section, tools in the ExPASy suite are identified, but the ensuing discussion also includes a number of useful programs made available by other groups. Internet resources related to these and other tools discussed in this chapter are listed at the end of the chapter.

AACompIdent and AACompSim (ExPASy)

Rather than using an amino acid sequence to search SWISS-PROT, AACompIdent uses the amino acid composition of an unknown protein to identify known proteins of the same composition (Wilkins et al., 1996). As inputs, the program requires the desired amino acid composition, the isoelectric point (pI) and molecular weight of

the protein (if known), the appropriate taxonomic class, and any special keywords. In addition, the user must select from one of six amino acid ''constellations,'' which influence how the analysis is performed; for example, certain constellations may combine residues like Asp/Asn (D/N) and Gln/Glu (Q/E) into Asx (B) and Glx (Z), or certain residues may be eliminated from the analysis altogether.

For each sequence in the database, the algorithm computes a score based on the difference in compositions between the sequence and the query composition. The results, returned by E-mail, are organized as three ranked lists:

- a list based on all proteins from the specified taxonomic class without taking pI or molecular weight into account;
- a list based on all proteins regardless of taxonomic class without taking pI or molecular weight into account; and
- a list based on the specified taxonomic class that does take pI and molecular weight into account.

Because the computed scores are a difference measure, a score of zero implies that there is exact correspondence between the query composition and that sequence entry.

AACompSim, a variant of AACompIdent, performs a similar type of analysis, but, rather than using an experimentally derived amino acid composition as the basis for searches, the sequence of a SWISS-PROT protein is used instead (Wilkins et al., 1996). A theoretical pI and molecular weight are computed before computation of the difference scores using Compute pI/MW (see below). It has been documented that amino acid composition across species boundaries is well conserved (Cordwell et al., 1995) and that, by considering amino acid composition, investigators can detect weak similarities between proteins whose sequence identity falls below 25% (Hobohm and Sander, 1995). Thus the consideration of composition in addition to the ability to perform ''traditional'' database searches may provide additional insight into the relationships between proteins.

PROPSEARCH

Along the same lines as AACompSim, PROPSEARCH uses the amino acid composition of a protein to detect weak relationships between proteins, and the authors have demonstrated that this technique can be used to easily discern members of the same protein family (Hobohm and Sander, 1995). However, this technique is more robust than AACompSim in that 144 different physical properties are used in performing the analysis, among which are molecular weight, the content of bulky residues, average hydrophobicity, and average charge. This collection of physical properties is called the query vector, and it is compared against the same type of vector precomputed for every sequence in the target databases (SWISS-PROT and PIR). Having this ''database of vectors'' calculated in advance vastly improves the processing time for a query.

The input to the PROPSEARCH Web server is just the query sequence, and an example of the program output is shown in Figure 11.1. Here, the sequence of human autoantigen NOR-90 was used as the input query. The results are ranked by a distance score, and this score represents the likelihood that the query sequence and new

Fragment search: OFF (POS1 and POS2 are begin and end of sequence)

Rank	ID	DIST	LEN2	POS1	POS2	pI	DE
1	>p1;s18193	0.00	727	1	727	5.33	autoantigen NOR-90 - human
2	ubf1_human	1.36	764	1	764	5.62	NUCLEOLAR TRANSCRIPTION FACTOR 1 (UPSTREAM BINDING FACTOR 1) (UBF-1)
3	ubf1_mouse	1.40	765	1	765	5.55	NUCLEOLAR TRANSCRIPTION FACTOR 1 (UPSTREAM BINDING FACTOR 1) (UBF-1).
4	ubf1_rat	1.57	764	1	764	5.61	NUCLEOLAR TRANSCRIPTION FACTOR 1 (UPSTREAM BINDING FACTOR 1) (UBF-1).
5	ubf1_xenla	3.95	677	1	677	5.79	NUCLEOLAR TRANSCRIPTION FACTOR 1 (UPSTREAM BINDING FACTOR-1) (UBF-1).
6	ubf2_xenla	4.18	701	1	701	6.05	NUCLEOLAR TRANSCRIPTION FACTOR 2 (UPSTREAM BINDING FACTOR-2) (UBF-2).
7	>p1;s57552	7.72	606	1	606	6.63	hypothetical protein YPR018w - yeast (Saccharomyces cerevisiae)
8	>p1;i50463	8.49	772	1	772	5.71	protein kinase - chicken
9	>p1;h54024	8.83	768	1	768	5.27	protein kinase (EC 2.7.1.37) cdc2-related PITSLRE alpha 2-3 - human
10	>p1;b54024	8.87	777	1	777	5.27	protein kinase (EC 2.7.1.37) cdc2-related PITSLRE alpha 2-2 - human
11	>p1;g54024	8.90	766	1	766	5.21	protein kinase (EC 2.7.1.37) cdc2-related PITSLRE beta 2-2 - human
12	>p1;a55817	9.00	783	1	783	5.19	cyclin-dependent kinase p130-PITSLRE - mouse
13	>p1;f54024	9.11	777	1	777	5.30	protein kinase (EC 2.7.1.37) cdc2-related PITSLRE beta 2-1 - human
14	>p1;e54024	9.11	779	1	779	5.42	protein kinase (EC 2.7.1.37) cdc2-related PITSLRE alpha 2-1 - human
15	yaa5_schpo	9.45	598	1	598	4.78	HYPOTHETICAL 69.5 KD PROTEIN C22G7.05 IN CHROMOSOME I.
16	>p1;s62449	9.45	598	1	598	4.78	hypothetical protein SPAC22G7.05 - fission yeast (Schizosaccharomyces pombe)
17	>f1;i58390	9.45	920	1	920	5.00	retinoblastoma binding protein 1 isoform I - human (fragment)
18	>p1;s63193	9.58	590	1	590	6.15	hypothetical protein YNL227c - yeast (Saccharomyces cerevisiae)
19	ynw7_yeast	9.58	590	1	590	6.15	HYPOTHETICAL 68.8 KD PROTEIN IN URE2-SSU72 INTERGENIC REGION.
20	>p1;s49634	9.74	899	1	899	4.79	hypothetical protein YML093w - yeast (Saccharomyces cerevisiae)
21	ymj3_yeast	9.74	899	1	899	4.79	HYPOTHETICAL 103.0 KD PROTEIN IN RAD10-PRS4 INTERGENIC REGION.
22	radi_human	9.76	583	1	583	6.33	RADIXIN.
23	radi_pig	9.81	583	1	583	6.21	RADIXIN (MOESIN B).
24	>f1;i78883	9.83	866	1	866	4.77	retinoblastoma binding protein 1 isoform II - human (fragment)
25	>p1;b42997	9.87	754	1	754	5.17	retinoblastoma-associated protein 2 - human
26	>p1;a57467	9.91	647	1	647	5.74	RalBP1 - rat

Figure 11.1. Results of a PROPSEARCH database query based on amino acid composition. The input sequence used was that of the human autoantigen NOR-90. Explanatory material and a histogram of distance scores against the entire target database have been removed for brevity. The columns in the table give the rank of the hit based on the distance score, the SWISS-PROT or PIR identifier, the distance score, the length of the overlap between the query and subject, the positions of the overlap (from POS1 to POS2), the calculated pI, and the definition line for the found sequence.

sequences found through PROPSEARCH belong to the same family, thereby implying common function in most cases. A distance score of 10 or below indicates that there is a better than 87% chance that there is similarity between the two proteins. A score below 8.7 increases the reliability to 94%, and a score below 7.5 increases the reliability to 99.6%. Examination of the results showed NOR-90 to be similar to a number of nucleolar transcription factors, protein kinases, a retinoblastoma-binding protein, the actin-binding protein radixin, and RalBP1, a putative GTPase target. None of these hits would necessarily be expected, since the functions of these proteins are dissimilar; however, a good number of these are DNA-binding proteins, opening the possibility that a very similar domain is being used in alternative functional contexts. At the very least, a BLASTP search would be necessary to both verify the results and identify critical residues.

MOWSE

The Molecular Weight Search (MOWSE) algorithm capitalizes on information obtained through mass spectrometric (MS) techniques (Pappin et al., 1993). With the use of both the molecular weights of intact proteins and those resulting from digestion of the same proteins with specific proteases, an unknown protein can be unambiguously identified given the results of several experimental determinations. This approach substantially cuts down on experimental time, since the unknown protein does not have to be sequenced in whole or in part.

The MOWSE Web front end requires the molecular weight of the starting sequence and the reagent used, as well as the resultant masses and composition of the peptides generated by the reagent. A tolerance value may be specified, indicating the error allowed in the accuracy of the determined fragment masses. Calculations are based on information contained in the OWL nonredundant protein sequence database (Akrigg et al., 1988). Scoring is based on how often a fragment molecular weight occurs in proteins within a given range of molecular weights, and the output is returned as a ranked list of the top 30 scores, with the OWL entry name, matching peptide sequences, and other statistical information. Simulation studies produced an accuracy rate of 99% using five or fewer input peptide weights.

PHYSICAL PROPERTIES BASED ON SEQUENCE

Compute pI/MW and ProtParam (ExPASy)

Compute pI/MW is a tool that calculates the isoelectric point and molecular weight of an input sequence. Determination of pI is based on p*K* values, as described in an earlier study on protein migration in denaturing conditions at neutral to acidic pH (Bjellqvist et al., 1993). Because of this, the authors caution that pI values determined for *basic* proteins may not be accurate. Molecular weights are calculated by the addition of the average isotopic mass of each amino acid in the sequence plus that of one water molecule. The sequence can be furnished by the user in FASTA format, or a SWISS-PROT identifier or accession number can be specified. If a sequence is furnished, the tool automatically computes the pI and molecular weight for the entire length of the sequence. If a SWISS-PROT identifier is given, the definition and organism lines of the entry are shown, and the user may specify a range of amino

acids so that the computation is done on a fragment rather than on the entire protein. ProtParam takes this process one step further. Based on the input sequence, ProtParam calculates the molecular weight, isoelectric point, overall amino acid composition, a theoretical extinction coefficient (Gill and von Hippel, 1989), aliphatic index (Ikai, 1980), the protein's grand average of hydrophobicity (GRAVY) value (Kyte and Doolittle, 1982), and other basic physicochemical parameters. Although this might seem to be a very simple program, one can begin to speculate about the cellular localization of the protein; for example, a basic protein with a high proportion of lysine and arginine residues may well be a DNA-binding protein.

PeptideMass (ExPASy)

Designed for use in peptide mapping experiments, PeptideMass determines the cleavage products of a protein after exposure to a given protease or chemical reagent (Wilkins et al., 1997). The enzymes and reagents available for cleavage through PeptideMass are trypsin, chymotrypsin, LysC, cyanogen bromide, ArgC, AspN, and GluC (bicarbonate or phosphate). Cysteines and methionines can be modified before the calculation of the molecular weight of the resultant peptides. By furnishing a SWISS-PROT identifier rather than pasting in a raw sequence, PeptideMass is able to use information within the SWISS-PROT annotation to improve the calculations, such as removing signal sequences or including known posttranslational modifications before cleavage. The results are returned in tabular format, giving a theoretical pI and molecular weight for the starting protein and then the mass, position, modified masses, information on variants from SWISS-PROT, and the sequence of the peptide fragments.

TGREASE

TGREASE calculates the hydrophobicity of a protein along its length (Kyte and Doolittle, 1982). Inherent in each of the 20 amino acids is its hydrophobicity: the relative propensity of the acid to bury itself in the core of a protein and away from surrounding water molecules. This tendency, coupled with steric and other considerations, influences how a protein ultimately folds into its final three-dimensional conformation. As such, TGREASE finds application in the determination of putative transmembrane sequences as well as the prediction of buried regions of globular proteins. TGREASE is part of the FASTA suite of programs available from the University of Virginia and runs as a stand-alone application that can be downloaded and run on either Macintosh or DOS-based computers.

The method relies on a hydropathy scale, in which each amino acid is assigned a score reflecting its relative hydrophobicity based on a number of physical characteristics (e.g., solubility, the free energy of transfer through a water-vapor phase transition, etc.). Amino acids with higher, positive scores are more hydrophobic; those with more negative scores are more hydrophilic. A moving average, or hydropathic index, is then calculated across the protein. The window length is adjustable, with a span of 7–11 residues recommended to minimize noise and maximize information content. The results are then plotted as hydropathic index versus residue number. The sequence for the human interleukin-8 receptor B was used to generate a TGREASE plot, as shown in Figure 11.2. Correspondence between the peaks and the actual location of the transmembrane segments, although not exact, is fairly good;

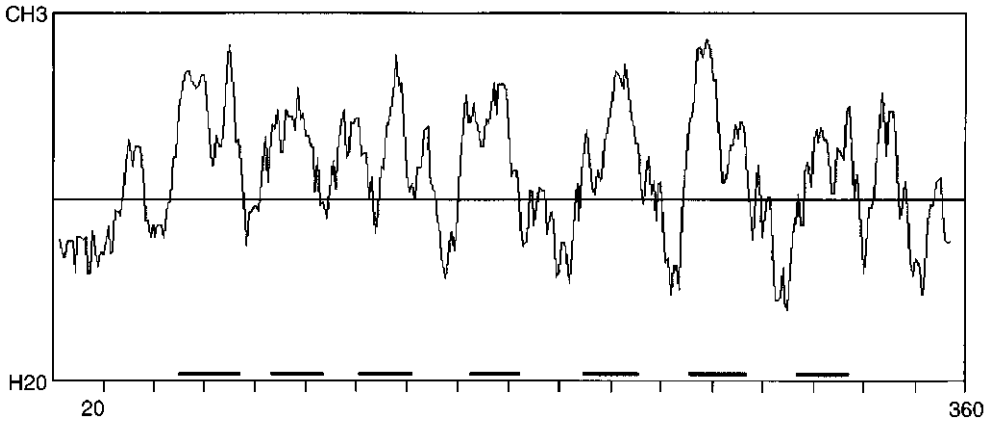

Figure 11.2. Results of a Kyte-Doolittle hydropathy determination using TGREASE. The input sequence was of the high affinity interleukin-8 receptor B from human. Default window lengths were used. The thick, horizontal bars across the bottom of the figure were added manually and represent the positions of the seven transmembrane regions of IL-8R-B, as given in the SWISS-PROT entry for this protein (P25025).

keep in mind that the method is predicting *all* hydrophobic regions, not just those located in transmembrane regions. The specific detection of transmembrane regions is discussed further below.

SAPS

The Statistical Analysis of Protein Sequences (SAPS) algorithm provides extensive statistical information for any given query sequence (Brendel et al., 1992). When a protein sequence is submitted via the SAPS Web interface, the server returns a large amount of physical and chemical information on the protein, based solely on what can be inferred from its sequence. The output begins with a composition analysis, with counts of amino acids by type. This is followed by a charge distribution analysis, including the locations of positively or negatively charged clusters, high-scoring charged and uncharged segments, and charge runs and patterns. The final sections present information on high-scoring hydrophobic and transmembrane segments, repetitive structures, and multiplets, as well as a periodicity analysis.

MOTIFS AND PATTERNS

In Chapter 8, the idea of direct sequence comparison was presented, where BLAST searches are performed to identify sequences in the public databases that are similar to a query sequence of interest. Often, this direct comparison may not yield any interesting results or may not yield any results at all. However, there may be very weak sequence determinants that are present that will allow the query sequence to be associated with a family of sequences. By the same token, a family of sequences can be used to identify new, distantly related members of the same protein family; an example of this is PSI-BLAST, discussed in Chapter 8.

Before discussing two of the methods that capitalize on such an approach, several terms have to be defined. The first is the concept of *profiles*. Profiles are, quite simply, a numerical representation of a multiple sequence alignment, much like the multiple sequence alignments derived from the methods discussed in Chapter 9. Imbedded within a multiple sequence alignment is intrinsic sequence information that represents the common characteristics of that particular collection of sequences, frequently a protein family. By using a profile, one is able to use these imbedded, common characteristics to find similarities between sequences with little or no absolute sequence identity, allowing for the identification and analysis of distantly related proteins. Profiles are constructed by taking a multiple sequence alignment representing a protein family and then asking a series of questions:

- What residues are seen at each position of the alignment?
- How often does a particular residue appear at each position of the alignment?
- Are there positions that show absolute conservation?
- Can gaps be introduced anywhere in the alignment?

Once those questions are answered, a *position-specific scoring table* (PSST) is constructed, and the numbers in the table now represent the multiple sequence alignment. The numbers within the PSST reflect the probability of any given amino acid occurring at each position. It also reflects the effect of a conservative or nonconservative substitution at each position in the alignment, much like a PAM or BLOSUM matrix does. This PSST can now be used for comparison against single sequences.

The second term requiring definition is *pattern* or *signature*. A signature also represents the common characteristics of a protein family (or a multiple sequence alignment) but does not contain any weighting information whatsoever—it simply provides a shorthand notation for what residues can be present at any given position. For example, the signature

```
[IV] - G - x - G - T -[LIVMF] - x(2) - [GS]
```

would be read as follows: the first position could contain *either* an isoleucine or a valine, the second position could contain only a glycine, and so on. An `x` means that any residue can appear at that position. The `x(2)` simply means that two positions can be occupied by any amino acid, the number just reflecting the length of the nonspecific run.

ProfileScan

Based on the classic Gribskov method of profile analysis (Gribskov et al., 1987, 1988), ProfileScan uses a method called pfscan to find similarities between a protein or nucleic acid query sequence and a profile library (Lüthy et al., 1994). In this case, three profile libraries are available for searching. First is PROSITE, an ExPASy database that catalogs biologically significant sites through the use of motif and sequence profiles and patterns (Hofmann, 1999). Second is Pfam, which is a collection of protein domain families that differ from most such collections in one important aspect: the initial alignment of the protein domains is done by hand, rather than by depending on automated methods. As such, Pfam contains slightly over 500 en-

tries, but the entries are potentially of higher quality. The third profile set is referred to as the Gribskov collection.

Searches against any of these collections can be done through the ProfileScan Web page, which simply requires either an input sequence in plain text format, or an identifier such as a SWISS-PROT ID. The user can select the sensitivity of the search, returning either significant matches only or all matches, including borderline cases. To illustrate the output format, the sequence of a human heat-shock-induced protein was submitted to the server for searching against PROSITE profiles only.

```
normalized raw      from -   to Profile | Description
355.9801 41556 pos.    6 - 612 PF00012 | HSP70 Heat shock hsp70 proteins
```

Although the actual PROSITE entry returned is no great surprise, the output contains scores that are worth understanding. The raw score is the actual score calculated from the scoring matrix used during the search. The more informative number is the normalized or *N*-score. The *N*-score formally represents the number of matches one would expect in a database of given size. Basically, the larger the *N*-score the lower the probability that the hit occurred by chance. In the example, the *N*-score of 355 translates to 1.94×10^{-349} expected chance matches when normalized against SWISS-PROT—an extremely low probability of this being a false positive. The `from` and `to` numbers simply show the positions of the overlap between the query and the matching profile.

BLOCKS

The BLOCKS database utilizes the concept of blocks to identify a family of proteins, rather than relying on the individual sequences themselves (Henikoff and Henikoff, 1996). The idea of a block is derived from the more familiar notion of a motif, which usually refers to a conserved stretch of amino acids that confer a specific function or structure to a protein. When these individual motifs from proteins in the same family are aligned without introducing gaps, the result is a block, with the term ''block'' referring to the alignment, not the individual sequences themselves. Obviously, an individual protein can contain one or more blocks, corresponding to each of its functional or structural motifs.

The BLOCKS database itself is derived from the entries in PROSITE. When a BLOCKS search is performed using a sequence of interest, the query sequence is aligned against all the blocks in the database at all possible positions. For each alignment, a score is calculated using a position-specific scoring matrix, and results of the best matches are returned to the user. Searches can be performed optionally against the PRINTS database, which includes information on more than 300 families that do not have corresponding entries in the BLOCKS database. To ensure complete coverage, it is recommended that both databases be searched.

BLOCKS searches can be performed using the BLOCKS Web site at the Fred Hutchinson Cancer Research Center in Seattle. The Web site is straightforward, allowing both sequence-based and keyword-based searches to be performed. If a DNA sequence is used as the input, users can specify which genetic code to use and which strand to search. Regardless of whether the query is performed via a sequence or via keywords, a successful search will return the relevant block. An example is shown in Figure 11.3. In this entry (for a nuclear hormone receptor called a steroid finger),

```
ID   NUCLEAR_RECEPTOR; BLOCK
AC   BL00031A; distance from previous block=(4,603)
DE   Nuclear hormones receptors DNA-binding region proteins.
BL   CCR;   width=33; seqs=177; 99.5%=1562; strength=1584
CSR1_CAEEL|Q17370   (  11)  CAVCDDIATGKHYSVASCNGCKTFFRRALVNNR  41

FTFB_DROME|Q05192   ( 376)  CPICGDKISGFHYGIFSCESCKGFFKRTVQNRK  16

HR96_DROME|Q24143   (   7)  CAVCGDKALGYNFNAVTCESCKAFFRRNALAKK  59

NER_HUMAN|P55055    (  87)  CRVCGDKASGFHYNVLSCEGCKGFFRRSVVRGG  12

NHR2_CAEEL|Q10902   ( 105)  CMVCGDNSTGYHYGVQSCEGCKGFFRRSVHKNI  16

ODR7_CAEEL|P41933   ( 331)  QVCLSTHANGLHFGARTCAACAAFFRRTISDDK  85

TLL_DROME|P18102    (  34)  CKVCRDHSSGKHYGIYACDGCAGFFKRSIRRSR  27

YKC8_CAEEL|P41999   (  18)  CLVCSDISTGYHYGVPSCNGCKTFFRRTIMKNQ  20

YQN7_CAEEL|Q09528   (  33)  CLICGEPSTGKHYGIVACLGCKTFFRRAVVQRQ  24

YRG4_CAEEL|Q09587   (  97)  HVCSSPTANTLHFGGRSCKACAAFFRRSVSMSM 100

7UP1_DROME|P16375   ( 200)  CVVCGDKSSGKHYGQFTCEGCKSFFKRSVRRNL   6
7UP2_DROME|P16376   ( 200)  CVVCGDKSSGKHYGQFTCEGCKSFFKRSVRRNL   6
ARP1_HUMAN|P24468   (  79)  CVVCGDKSSGKHYGQFTCEGCKSFFKRSVRRNL   6
ARP1_MOUSE|P43135   (  79)  CVVCGDKSSGKHYGQFTCEGCKSFFKRSVRRNL   6
COT1_MOUSE|Q60632   (  85)  CVVCGDKSSGKHYGQFTCEGCKSFFKRSVRRNL   6
COTF_HUMAN|P10589   (  86)  CVVCGDKSSGKHYGQFTCEGCKSFFKRSVRRNL   6
EAR2_HUMAN|P10588   (  56)  CVVCGDKSSGKHYGVFTCEGCKSFFKRTIRRNL   5
EAR2_MOUSE|P43136   (  57)  CVVCGDKSSGKHYGVFTCEGCKSFFKRTIRRNL   5
HR78_DROME|Q24142   (  52)  CLVCGDRASGRHYGAISCEGCKGFFKRSIRKQL   5
TR2_HUMAN|P13056    ( 113)  CVVCGDKASGRHYGAVTCEGCKGFFKRSIRKNL   5
TR4_HUMAN|P49116    ( 117)  CVVCGDKASGRHYGAVSCEGCKGFFKRSVRKNL   5
TR4_MOUSE|P49117    ( 117)  CVVCGDKASGRHYGAVSCEGCKGFFKRSVRKNL   5
TR4_RAT|P55094      ( 117)  CVVCGDKASGRHYGAVSCEGCKGFFKRSVRKNL   5
```

Figure 11.3. Structure of a typical BLOCKS entry. This is part of the entry for one block associated with steroid fingers. The structure of the entry is discussed in the text.

the header lines marked `ID`, `AC`, and `DE` give, in order, a short description of the family represented by this block, the BLOCKS database accession number, and a longer description of the family. The `BL` line gives information regarding the original sequence motif that was used to construct this particular block. The `width` and `seqs` parameters show how wide the block is, in residues, and how many sequences are in the block, respectively. Some information then follows regarding the statistical validity and the strength of the construct. Finally, a list of sequences is presented, showing only the part of the sequence corresponding to this particular motif. Each line begins with the SWISS-PROT accession number for the sequence, the number of the first residue shown based on the entire sequence, the sequence itself, and a position-based sequence weight. These values are scaled, with 100 representing the sequence that is most distant from the group. Notice that there are blank lines between some of the sequences; parts of the overall alignment are clustered, and, in each cluster, 80% of the sequence residues are identical.

CDD

Recently, NCBI introduced a new search service aimed at identifying conserved domains within a protein sequence. The source database for these searches is called the Conserved Domain Database or CDD. This is a secondary database, with entries

derived from both Pfam (described above) and SMART (Simple Modular Architecture Research Tool). SMART can be used to identify genetically mobile domains and analyze domain architectures and is discussed in greater detail within the context of comparative genomics in Chapter 15. The actual search is performed using reverse position-specific BLAST (RPS-BLAST), which uses the query sequence to search a database of precalculated PSSTs.

The CDD interface is simple, providing a box for the input sequence (alternatively, an accession number can be specified) and a pull-down menu for selecting the target database. If conserved domains are identified within the input sequence, a graphic is returned showing the position of each conserved domain, followed by the actual alignment of the query sequence to the target domain as generated by RPS-BLAST. In these alignments, the default view shows identical residues in red, whereas conservative substitutions are shown in blue; users can also select from a variety of representations, including the traditional BLAST-style alignment display. Hyperlinks are provided back to the source databases, providing more information on that particular domain. This ''CD Summary'' page gives the underlying source database information, references, the taxonomy spanned by this entry, and a sequence entry representative of the group. In the lower part of the page, the user can construct an alignment of sequences of interest from the group; alternatively, the user can allow the computer to select the top-ranked sequences or a subset of sequences that are most diverse within the group. If a three-dimensional structure corresponding to the CD is available, it can be viewed directly using Cn3D (see Chapter 5). Clicking on the CD link next to any of the entries on the CD Summary page will, in essence, start the whole process over again, using *that* sequence to perform a new RPS-BLAST search against CDD.

SECONDARY STRUCTURE AND FOLDING CLASSES

One of the first steps in the analysis of a newly discovered protein or gene product of unknown function is to perform a BLAST or other similar search against the public databases. However, such a search might not produce a match against a known protein; if there is a statistically significant hit, there may not be any information in the sequence record regarding the secondary structure of the protein, information that is very important in the rational design of biochemical experiments. In the absence of ''known'' information, there are methods available for predicting the ability of a sequence to form α-helices and β-strands. These methods rely on observations made from groups of proteins whose three-dimensional structure has been experimentally determined.

A brief review of secondary structure and folding classes is warranted before the techniques themselves are discussed. As already alluded to, a significant number of amino acids have hydrophobic side chains, whereas the main chain, or backbone, is hydrophilic. The required balance between these two seemingly opposing forces is accomplished through the formation of discrete secondary structural elements, first described by Linus Pauling and colleagues in 1951 (Pauling and Corey, 1951). An α-helix is a corkscrew-type structure with the main chain forming the backbone and the side chains of the amino acids projecting outward from the helix. The backbone is stabilized by the formation of hydrogen bonds between the CO group of each

amino acid and the NH group of the residue four positions C-terminal ($n + 4$), creating a tight, rodlike structure. Some residues form α-helices better than others; alanine, glutamine, leucine, and methionine are commonly found in α-helices, whereas proline, glycine, tyrosine, and serine usually are not. Proline is commonly thought of as a helix breaker because its bulky ring structure disrupts the formation of $n + 4$ hydrogen bonds.

In contrast, the β-strand is a much more extended structure. Rather than hydrogen bonds forming within the secondary structural unit itself, stabilization occurs through bonding with one or more *adjacent* β-strands. The overall structure formed through the interaction of these individual β-strands is known as a *β-pleated sheet*. These sheets can be parallel or antiparallel, depending on the orientation of the N- and C-terminal ends of each component β-strand. A variant of the β-sheet is the *β-turn*; in this structure the polypeptide chain makes a sharp, hairpin bend, producing an antiparallel β-sheet in the process.

In 1976, Levitt and Chothia proposed a classification system based on the order of secondary structural elements within a protein (Levitt and Chothia, 1976). Quite simply, an α-structure is made up primarily from α-helices, and a β-structure is made up of primarily β-strands. Myoglobin is the classic example of a protein composed entirely of α-helices, falling into the α class of structures (Takano, 1977). Plastocyanin is a good example of the β class, where the hydrogen-bonding pattern between eight β-strands form a compact, barrel-like structure (Guss and Freeman, 1983). The combination class, α/β, is made up of primarily β-strands alternating with α-helices. Flavodoxin is a good example of an α/β-protein; its β-strands form a central β-sheet, which is surrounded by α-helices (Burnett et al., 1974).

Predictive methods aimed at extracting secondary structural information from the linear primary sequence make extensive use of neural networks, traditionally used for analysis of patterns and trends. Basically, a neural network provides computational processes the ability to ''learn'' in an attempt to approximate human learning versus following instructions blindly in a sequential manner. Every neural network has an *input layer* and an *output layer*. In the case of secondary structure prediction, the input layer would be information from the sequence itself, and the output layer would be the probabilities of whether a particular residue could form a particular structure. Between the input and output layers would be one or more *hidden layers* where the actual ''learning'' would take place. This is accomplished by providing a training data set for the network. Here, an appropriate training set would be all sequences for which three-dimensional structures have been deduced. The network can process this information to look for what are possibly weak relationships between an amino acid sequence and the structures they can form in a particular context. A more complete discussion of neural networks as applied to secondary structure prediction can be found in Kneller et al. (1990).

nnpredict

The nnpredict algorithm uses a two-layer, feed-forward neural network to assign the predicted type for each residue (Kneller et al., 1990). In making the predictions, the server uses a FASTA format file with the sequence in either one-letter or three-letter code, as well as the folding class of the protein (α, β, or α/β). Residues are classified

as being within an α-helix (H), a β-strand (E), or neither (—). If no prediction can be made for a given residue, a question mark (?) is returned to indicate that an assignment cannot be made with confidence. If no information is available regarding the folding class, the prediction can be made without a folding class being specified; this is the default. For the best-case prediction, the accuracy rate of nnpredict is reported as being over 65%.

Sequences are submitted to nnpredict by either sending an E-mail message to *nnpredict@celeste.ucsf.edu* or by using the Web-based submission form. With the use of flavodoxin as an example, the format of the E-mail message would be as follows:

```
option: a/b
>flavodoxin - Anacystis nidulans
AKIGLFYGTQTGVTQTIAESIQQEFGGESIVDLNDIANADASDLNAYDYLIIGCPTWNVGELQSDWEGIY
DDLDSVNFQGKKVAYFGAGDQVGYSDNFQDAMGILEEKISSLGSQTVGYWPIEGYDFNESKAVRNNQFVG
LAIDEDNQPDLTKNRIKTWVSQLKSEFGL
```

The Option line specifies the folding class of the protein: `n` uses no folding class for the prediction, `a` specifies α, `b` specifies β, and `a/b` specifies α/β. Only one sequence may be submitted per E-mail message. The results returned by the server are shown in modified form in Figure 11.4.

PredictProtein

PredictProtein (Rost et al., 1994) uses a slightly different approach in making its predictions. First, the protein sequence is used as a query against SWISS-PROT to find similar sequences. When similar sequences are found, an algorithm called MaxHom is used to generate a profile-based multiple sequence alignment (Sander and Schneider, 1991). MaxHom uses an iterative method to construct the alignment: After the first search of SWISS-PROT, all found sequences are aligned against the query sequence and a profile is calculated for the alignment. The profile is then used to search SWISS-PROT again to locate new, matching sequences. The multiple alignment generated by MaxHom is subsequently fed into a neural network for prediction by one of a suite of methods collectively known as PHD (Rost, 1996). PHDsec, the method in this suite used for secondary structure prediction, not only assigns each residue to a secondary structure type, it provides statistics indicating the confidence of the prediction at each position in the sequence. The method produces an average accuracy of better than 72%; the best-case residue predictions have an accuracy rate of over 90%.

Sequences are submitted to PredictProtein either by sending an E-mail message or by using a Web front end. Several options are available for sequence submission; the query sequences can be submitted as single-letter amino acid code or by its SWISS-PROT identifier. In addition, a multiple sequence alignment in FASTA format or as a PIR alignment can also be submitted for secondary structure prediction.

The input message, sent to *predictprotein@embl-heidelberg.de*, takes the following form:

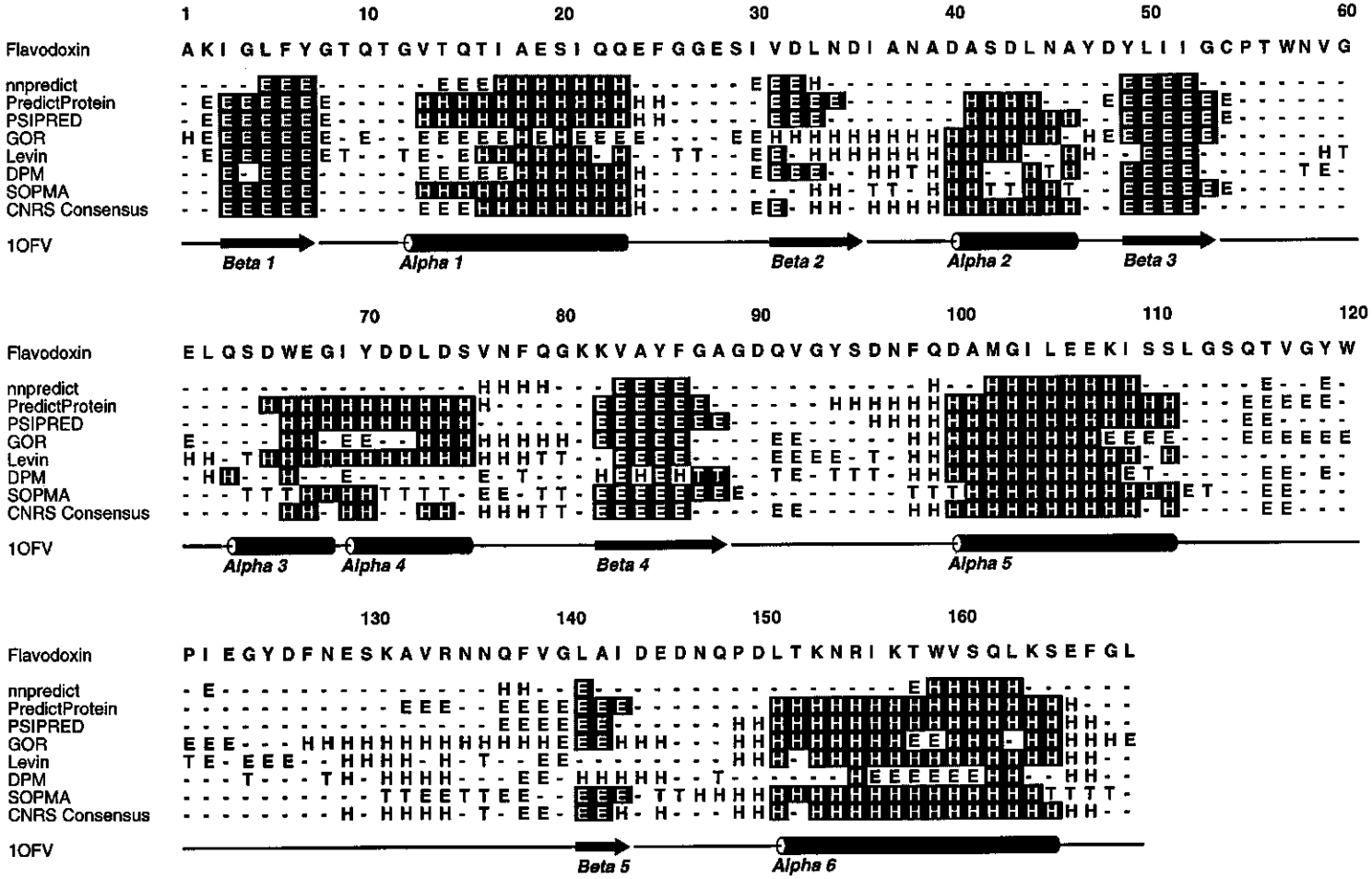

Figure 11.4. Comparison of secondary structure predictions by various methods. The sequence of flavodoxin was used as the query and is shown on the first line of the alignment. For each prediction, H denotes an α-helix, E a β-strand, and T a β-turn; all other positions are assumed to be random coil. Correctly assigned residues are shown in inverse type. The methods used are listed along the left side of the alignment and are described in the text. At the bottom of the figure is the secondary structure assignment given in the PDB file for flavodoxin (1OFV, Smith et al., 1983).

```
Joe Buzzcut
National Human Genome Research Institute, NIH
buzzcut@baldguys.org
do NOT align
# FASTA list     homeodomain proteins
>ANTP
---KRGRQTYTRYQTLELEKEFHFNRYLTRRRRIEIAHALSLTERQIKIWFQNRRMKWKK
>HDD
MDEKRPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKIKK
>DLX
-KIRKPRTIYSSLQLQALNHRFQQTQYLALPERAELAASLGLTQTQVKIWFQNKRSKFKK
>FTT
---RKRRVLFSQAQVYELERRFKQQKYLSAPEREHLASMIHLTPTQVKIWFQNHRYKMKR
>Pax6
--LQRNRTSFTQEQIEALEKEFERTHYPDVFARERLAAKIDLPEARIQVWFSNRRAKWRR
```

Above is an example of a FASTA-formatted multiple sequence alignment of homeodomain proteins submitted for secondary structure prediction. After the name, affiliation, and address lines, the # sign signals to the server that a sequence in one-letter code follows. The sequence format is essentially FASTA, except that blanks are not allowed. For this alignment, the phrase `do NOT align` before the line starting with # assures that the alignment will not be realigned. Nothing is allowed to follow the sequence. The output sent as an E-mail message is quite copious but contains a large amount of pertinent information. The results can also be retrieved from an ftp site by adding a qualifier `return no mail` in any line before the line starting with #. This might be a useful feature for those E-mail services that have difficulty handling very large output files. The format for the output file can be plain text or HTML files with or without PHD graphics.

The results of the MaxHom search are returned, complete with a multiple alignment that may be of use in further study, such as profile searches or phylogenetic studies. If the submitted sequence has a known homolog in PDB, the PDB identifiers are furnished. Information follows on the method itself and then the actual prediction will follow. In a recent release, the output can also be customized by specifying available options. Unlike nnpredict, PredictProtein returns a ''reliability index of prediction'' for each position ranging from 0 to 9, with 9 being the maximum confidence that a secondary structure assignment has been made correctly. The results returned by the server for this particular sequence, as compared with those obtained by other methods, are shown in modified form in Figure 11.4.

PREDATOR

The PREDATOR secondary structure prediction algorithm is based on recognition of potentially hydrogen-bonded residues in the amino acid sequence (Frishman and Argos, 1997). It uses database-derived statistics on residue-type occurrences in different classes of local hydrogen-bonded structures. The novel feature of this method is its reliance on local pairwise alignment of the sequence to be predicted between each related sequence. The input for this program can be a single sequence or a set of *unaligned*, related sequences. Sequences can be submitted to the PREDATOR server either by sending an E-mail message to *predator@embl-heidelberg.de* or by

using a Web front end. The input sequences can be either FASTA, MSF, or CLUSTAL format. The mean prediction accuracy of PREDATOR in three structural states is 68% for a single sequence and 75% for a set of related sequences.

PSIPRED

The PSIPRED method, developed at the University of Warwick, UK, uses the knowledge inferred from PSI-BLAST (Altschul et al., 1997; cf. Chapter 8) searches of the input sequence to perform predictions. PSIPRED uses two feedforward neural networks to perform the analysis on the profile obtained from PSI-BLAST. Sequences can be submitted through a simple Web front end, in either single-letter raw format or in FASTA format. The results from the PSIPRED prediction are returned as a text file in an E-mail message. In addition, a link is also provided in the E-mail message to a graphical representation of the secondary structure prediction, visualized using a Java application called PSIPREDview. In this representation, the positions of the helices and strands are schematically represented above the target sequence. The average prediction accuracy for PSIPRED in three structural states is 76.5%, which is higher than any of the other methods described here.

SOPMA

The Protein Sequence Analysis server at the Centre National de la Recherche Scientifique (CNRS) in Lyons, France, takes a unique approach in making secondary structure predictions: rather than using a single method, it uses *five*, the predictions from which are subsequently used to come up with a ''consensus prediction.'' The methods used are the Garnier–Gibrat–Robson (GOR) method (Garnier et al., 1996), the Levin homolog method (Levin et al., 1986), the double-prediction method (Deléage and Roux, 1987), the PHD method described above as part of PredictProtein, and the method of CNRS itself, called SOPMA (Geourjon and Déleage, 1995). Briefly, this self-optimized prediction method builds subdatabases of protein sequences with known secondary structures; each of the proteins in a subdatabase is then subjected to secondary structure prediction based on sequence similarity. The information from the subdatabases is then used to generate a prediction on the query sequence.

The method can be run by submitting just the sequence itself in single-letter format to *deleage@ibcp.fr*, using `SOPMA` as the subject of the mail message, or by using the SOPMA Web interface. The output from each of the component predictions, as well as the consensus, is shown in Figure 11.4.

Comparison of Methods

On the basis of Figure 11.4, it is immediately apparent that all the methods described above do a relatively good, but not perfect, job of predicting secondary structures. Where no other information is known, the best approach is to perform predictions using all the available algorithms and then to judge the validity of the predictions in comparison to one another. Flavodoxin was selected as the input query because it has a relatively intricate structure, falling into the α/β-folding class with its six α-helices and five β-sheets. Some assignments were consistently made by all methods; for example, all the methods detected $\beta 1$, $\beta 3$, $\beta 4$, and $\alpha 5$ fairly well. However, some

methods missed some elements altogether (e.g., nnpredict with α2, α3, and α4), and some predictions made no biological sense (e.g., the double-prediction method and β4, where helices, sheets, and turns alternate residue by residue). PredictProtein and PSIPRED, which both correctly found all the secondary structure elements and, in several places, identified structures of the correct length, *appear* to have made the best overall prediction. This is *not* to say that the other methods are not useful or not as good; undoubtedly, in some cases, another method would have emerged as having made a better prediction. This approach does not provide a fail-safe method of prediction, but it does reinforce the level of confidence resulting from these predictions.

A new Web-based server, Jpred, integrates six different structure prediction methods and returns a consensus prediction based on simple majority rule. The usefulness of this server is that it automatically generates the input and output requirements for all six prediction algorithms, which can be an important feature when handling large data sets. The input sequence for Jpred can be a single protein sequence in FASTA or PIR format, a set of unaligned sequences in PIR format, or a multiple sequence alignment in MSF or BLC format. In case of a single sequence, the server first generates a set of related sequences by searching the OWL database using the BLASTP algorithm. The sequence set is filtered using SCANPS and then pairwise-compared using AMPS. Finally, the sequence set is clustered using a 75% identity cutoff value to remove any bias in the sequence set, and the remaining sequences are aligned using CLUSTAL W. The Jpred server runs PHD (Rost and Sander, 1993), DSC (King and Sternberg, 1996), NNSSP (Salamov and Solovyev, 1995), PREDATOR (Frishman and Argos, 1997), ZPRED (Zvelebil et al., 1987), and MULPRED (Barton, 1988). The results from the Jpred server is returned as a text file in an E-mail message; a link is also provided to view the colored graphical representation in HTML or PostScript file format. The consensus prediction from the Jpred server has an accuracy of 72.9% in the three structural states.

SPECIALIZED STRUCTURES OR FEATURES

Just as the position of α-helices and β-sheets can be predicted with a relatively high degree of confidence, the presence of certain specialized structures or features, such as coiled coils and transmembrane regions, can be predicted. There are not as many methods for making such predictions as there are for secondary structures, primarily because the rules of folding that induce these structures are not completely understood. Despite this, when query sequences are searched against databases of known structures, the accuracy of prediction can be quite high.

Coiled Coils

The COILS algorithm runs a query sequence against a database of proteins known to have a coiled-coil structure (Lupas et al., 1991). The program also compares query sequences to a PDB subset containing globular sequences and on the basis of the differences in scoring between the PDB subset and the coiled coils database, determines the probability with which the input sequence can form a coiled coil. COILS can be downloaded for use with VAX/VMS or may more easily be used through a simple Web interface.

The program takes sequence data in GCG or FASTA format; one or more sequences can be submitted at once. In addition to the sequences, users may select one of two scoring matrices: MTK, based on the sequences of myosin, tropomyosin, and keratin, or MTIDK, based on myosin, tropomyosin, intermediate filaments types I–V, desmosomal proteins, and kinesins. The authors cite a trade-off between the scoring matrices, with MTK being better for detecting two-stranded structures and MTIDK being better for all other cases. Users may invoke an option that gives the same weight to the residues at the *a* and *d* positions of each coil (normally hydrophobic) as that given to the residues at the *b*, *c*, *e*, *f*, and *g* positions (normally hydrophilic). If the results of running COILS both weighted and unweighted are substantially different, it is likely that a false positive has been found. The authors caution that COILS is designed to detect solvent-exposed, left-handed coiled coils and that buried or right-handed coiled coils may not be detected. When a query is submitted to the Web server, a prediction graph showing the propensity toward the formation of a coiled coil along the length of the sequence is generated.

A slightly easier to interpret output comes from MacStripe, a Macintosh-based application that uses the Lupas COILS method to make its predictions (Knight, 1994). MacStripe takes an input file in FASTA, PIR, and other common file formats and, like COILS, produces a plot file containing a histogram of the probability of forming a coiled coil, along with bars showing the continuity of the heptad repeat pattern. The following portion of the statistics file generated by MacStripe uses the complete sequence of GCN4 as an example:

```
 89  89 L 5 a 0.760448 0.000047
 90  90 D 5 b 0.760448 0.000047
 91  91 D 5 c 0.760448 0.000047
 92  92 A 5 d 0.760448 0.000047
 93  93 V 5 e 0.760448 0.000047
 94  94 V 5 f 0.760448 0.000047
 95  95 E 5 g 0.760448 0.000047
 96  96 S 5 a 0.760448 0.000047
 97  97 F 5 b 0.760448 0.000047
 98  98 F 5 c 0.774300 0.000058
 99  99 S 5 d 0.812161 0.000101
100 100 S 5 e 0.812161 0.000101
101 101 S 5 f 0.812161 0.000101
102 102 T 5 g 0.812161 0.000101
```

The columns, from left to right, represent the residue number (shown twice), the amino acid, the heptad frame, the position of the residue within the heptad (`a-b-c-d-e-f-g`), the Lupas score, and the Lupas probability. In this case, from the fifth column, we can easily discern a heptad repeat pattern. Examination of the results for the entire GCN4 sequence shows that the heptad pattern is fairly well maintained but falls apart in certain areas. The statistics should not be ignored; however, the results are easier to interpret if the heptad pattern information is clearly presented. It is possible to get a similar type of output from COILS but not through the COILS Web server; instead, a C-based program must be installed on an appropriate Unix machine, a step that may be untenable for many users.

Transmembrane Regions

The Kyte-Doolittle TGREASE algorithm discussed above is very useful in detecting regions of high hydrophobicity, but, as such, it does not exclusively predict transmembrane regions because buried domains in soluble, globular proteins can also be primarily hydrophobic. We consider first a predictive method specifically for the prediction of transmembrane regions. This method, TMpred, relies on a database of transmembrane proteins called TMbase (Hofmann and Stoffel, 1993). TMbase, which is derived from SWISS-PROT, contains additional information on each sequence regarding the number of transmembrane domains they possess, the location of these domains, and the nature of the flanking sequences. TMpred uses this information in conjunction with several weight matrices in making its predictions.

The TMpred Web interface is very simple. The sequence, in one-letter code, is pasted into the query sequence box, and the user can specify the minimum and maximum lengths of the hydrophobic part of the transmembrane helix to be used in the analysis. The output has four sections: a list of possible transmembrane helices, a table of correspondences, suggested models for transmembrane topology, and a graphic representation of the same results. When the sequence of the G-protein-coupled receptor (P51684) served as the query, the following models were generated:

```
2 possible models considered, only significant TM-segments used

-----> STRONGLY preferred model: N-terminus outside
 7 strong transmembrane helices, total score : 14211
 # from  to length score orientation
 1   55  74 (20)    2707 o-i
 2   83 104 (22)    1914 i-o
 3  120 141 (22)    1451 o-i
 4  166 184 (19)    2170 i-o
 5  212 235 (24)    2530 o-i
 6  255 276 (22)    2140 i-o
 7  299 319 (21)    1299 o-i

-----> alternative model
 7 strong transmembrane helices, total score : 12079
 # from  to length score orientation
 1   47  69 (23)    2494 i-o
 2   84 104 (21)    1470 o-i
 3  123 141 (19)    1383 i-o
 4  166 185 (20)    1934 o-i
 5  219 236 (18)    2474 i-o
 6  252 274 (23)    1386 o-i
 7  303 319 (17)     938 i-o
```

Each of the proposed models indicates the starting and ending position of each segment, along with the relative orientation (inside-to-outside or outside-to-inside) of each segment. The authors appropriately caution that the models are based on the assumption that all transmembrane regions were found during the prediction. These models, then, should be considered in light of the raw data also generated by this method.

PHDtopology

One of the most useful methods for predicting transmembrane helices is PHDtopology, which is related to the PredictProtein secondary structure prediction method described above. Here, programs within the PHD suite are now used in an obviously different way to make a prediction on a membrane-bound rather than on a soluble protein. The method has reported accuracies that are nearly perfect: the accuracy of predicting a transmembrane helix is 92% and the accuracy for a loop is 96%, giving an overall two-state accuracy of 94.7%. One of the features of this program is that, in addition to predicting the putative transmembrane regions, it indicates the orientation of the loop regions with respect to the membrane.

As before, PHDtopology predictions can be made using either an E-mail server or a Web front end. If an E-mail server is used, the format is identical to that shown for PredictProtein above, except that the line `predict htm topology` must precede the line beginning with the pound sign. Regardless of submission method, results are returned by E-mail. An example of the output returned by PHDtopology is shown in Figure 11.5.

Signal Peptides

The Center for Biological Sequence Analysis at the Technical University of Denmark has developed SignalP, a powerful tool for the detection of signal peptides and their

```
Joe Buzzcut
National Human Genome Research Institute, NIH
buzzcut@nhgri.nih.gov
predict htm topology
# pendrin
MAAPGGRSEPPQLPEYSCSYMVSRPVYSELAFQQQHERRLQERKTLRESLAKCCSCSRKRAFGVLKTLVPILEWLPKYRV
KEWLLSDVISGVSTGLVATLQGMAYALLAAVPVGYGLYSAFFPILTYFIFGTSRHISVGPFPVVSLMVGSVVLSMAP...
```

```
                ....,....37...,....38...,....39...,....40...,....41...,....42
        AA      |YSLKYDYPLDGNQELIALGLGNIVCGVFRGFAGSTALSRSAVQESTGGKTQIAGLIGAII|
        PHD htm |                       HHHHHHHHHHHHHH              HHHHHHHHHH|
        Rel htm |3688999999999999986411046677776554312577788877776214677888888|
detail:         |                                                              |
        prH htm |3100000000000000001244578888888777653211100001111357888999999|
        prL htm |6899999999999999987554211111112223467888999988886421110000 0|
                                        .
                                        .
                                        .
        PHDThtm |iiiiiiiiiiiiiiiiiiTTTTTTTTTTTTTTTTTTooooooooooooooTTTTTTTT|
```

Figure 11.5. Partial output from a PHDtopology prediction. The input sequence is pendrin, which is responsible for Pendred syndrome (Everett et al., 1998). The row labeled `AA` shows a portion of the input sequence, and the row labeled `Rel htm` gives the reliability index of prediction at each position of the protein; values range from 0 to 9, with 9 representing the maximum possible confidence for the assignment at that position. The last line, labeled `PHDThm`, contains one of three letters: a `T` represents a transmembrane region, whereas an `i` or `o` represents the orientation of the loop with respect to the membrane (inside or outside).

cleavage sites (Nielsen et al., 1997). The algorithm is neural-network based, using separate sets of Gram-negative prokaryotic, Gram-positive prokaryotic, and eukaryotic sequences with known signal sequences as the training sets. SignalP predicts secretory signal peptides and not those that are involved in intracellular signal transduction.

Using the Web interface, the sequence of the human insulin-like growth factor IB precursor (somatomedin C, P05019), whose cleavage site is known, was submitted to SignalP for analysis. The eukaryotic training set was used in the prediction, and the results of the analysis are as follows:

```
****************** SignalP predictions ******************
Using networks trained on euk data
>IGF-IB length = 195
# pos aa C S Y
 .

 .
 .
 46 A 0.365 0.823 0.495
 47 T 0.450 0.654 0.577
 48 A 0.176 0.564 0.369
 49 G 0.925 0.205 0.855
 50 P 0.185 0.163 0.376
 .
 .
 .
< Is the sequence a signal peptide?
# Measure Position Value Cutoff Conclusion
 max. C 49 0.925 0.37 YES
 max. Y 49 0.855 0.34 YES
 max. S 37 0.973 0.88 YES
 mean S 1 - 48 0.550 0.48 YES
# Most likely cleavage site between pos. 48 and 49: ATA-GP
```

In the first part of the output, the column labeled `C` is a raw cleavage site score. The value of C is highest at the position C-terminal to the cleavage site. The column labeled `S` contains the signal peptide scores, which are high at all positions before the cleavage site and very low after the cleavage site. S is also low in the N-termini of nonsecretory proteins. Finally, the `Y` column gives the combined cleavage site score, a geometric average indicating when the C score is high and the point at which the S score shifts from high to low. The end of the output file asks the question, ''Is the sequence a signal peptide?'' On the basis of the statistics, the most likely cleavage site is deduced. On the basis of the SWISS-PROT entry for this protein, the mature chain begins at position 49, the same position predicted to be the most likely cleavage site by SignalP.

Nonglobular Regions

The use of the program SEG in the masking of low-complexity segments prior to database searches was discussed in Chapter 8. The same algorithm can also be used

```
                                  1-307  MAGAIASRMSFSSLKRKQPKTFTVRIVTMD
                                         AEMEFNCEMKWKGKDLFDLVCRTLGLRETW
                                         FFGLQYTIKDTVAWLKMDKKVLDHDVSKEE
                                         PVTFHFLAKFYPENAEEELVQEITQHLFFL
                                         QVKKQILDEKIYCPPEASVLLASYAVQAKY
                                         GDYDPSVHKRGFLAQEELLPKRVINLYQMT
                                         PEMWEERITAWYAEHRGRARDEAEMEYLKI
                                         AQDLEMYGVNYFAIRNKKGTELLLGVDALG
                                         LHIYDPENRLTPKISFPWNEIRNISYSDKE
                                         FTIKPLDKKIDVFKFNSSKLRVNKLILQLC
                                         IGNHDLF
mrrrkadslevqqmkaqareekarkqmerq    308-478
rlarekqmreeaertrdelerrllqmkeea
tmanealmrseetadllaekaqiteeeakl
laqkaaeaeqemqrikatairteeekrlme
qkvleaevlalkmaeeserrakeadqlkqd
         lqeareaerrakqkllleiatk
                                  479-496  PTYPPMNPIPAPLPPDIP
sfnligdslsfdfkdtdmkrlsmeiekekv    497-587
eymekskhlqeqlnelkteiealklkeret
aldilhnensdrggsskhntikkltlqsak
                            s
                                  588-595  RVAFFEEL
```

Figure 11.6. Predicted nonglobular regions for the protein product of the neurofibromatosis type 2 gene (L11353) as deduced by SEG. The nonglobular regions are shown in the left-hand column in lowercase. Numbers denote residue positions for each block.

to detect putative nonglobular regions of protein sequences by altering the trigger window length W, the trigger complexity K_1, and extension complexity K_2. When the command `seg sequence.txt 45 3.4 3.75` is received, SEG will use a longer window length than the default of 12, thereby detecting long, nonglobular domains. An example of using SEG to detect nonglobular regions is shown in Figure 11.6.

TERTIARY STRUCTURE

By far the most complex and technically demanding predictive method based on protein sequence data has to do with structure prediction. The importance of being able to adequately and accurately predict structure based on sequence is rooted in the knowledge that, whereas sequence may specify conformation, the same conformation may be specified by multiple sequences. The ideas that structure is conserved to a much greater extent than sequence and that there is a limited number of backbone motifs (Chothia and Lesk, 1986; Chothia, 1992) indicate that similarities between proteins may not necessarily be detected through traditional, sequence-based methods only. Deducing the relationship between sequence and structure is at the root of the ''protein-folding problem,'' and current research on the problem has been the focus of several reviews (Bryant and Altschul, 1995; Eisenhaber et al., 1995; Lemer et al., 1995).

The most robust of the structure prediction techniques is homology model building or ''threading'' (Bryant and Lawrence, 1993; Fetrow and Bryant, 1993; Jones and Thornton, 1996). The threading methods search for structures that have a similar

fold without apparent sequence similarity. This method takes a query sequence whose structure is not known and threads it through the coordinates of a target protein whose structure has been solved, either by X-ray crystallography or NMR imaging. The sequence is moved position by position through the structure, subject to some predetermined physical constraints; for example, the lengths of secondary structure elements and loop regions may be either fixed or varying within a given range. For each placement of sequence against structure, pairwise and hydrophobic interactions between nonlocal residues are determined. These thermodynamic calculations are used to determine the most energetically favorable and conformationally stable alignment of the query sequence against the target structure. Programs such as this are computationally intensive, requiring, at a minimum, a powerful UNIX workstation; they also require knowledge of specialized computer languages. The threading methods are useful when the sequence-based structure prediction methods fail to identify a suitable template structure.

Although techniques such as threading are obviously very powerful, their current requirements in terms of both hardware and expertise may prove to be obstacles to most biologists. In an attempt to lower the height of the barrier, easy-to-use programs have been developed to give the average biologist a good first approximation for comparative protein modeling. (Numerous commercial protein structure analysis tools, such as WHAT-IF and LOOK, provide advanced capabilities, but this discussion is limited to Web-based freeware.)

The use of SWISS-MODEL, a program that performs automated sequence-structure comparisons (Peitsch, 1996), is a two-step process. The First Approach mode is used to determine whether a sequence can be modeled at all; when a sequence is submitted, SWISS-MODEL compares it with the crystallographic database (ExPdb), and modeling is attempted only if there is a homolog in ExPdb to the query protein. The template structures are selected if there is at least 25% sequence identity in a region more than 20 residues long. If the first approach finds one or more appropriate entries in ExPdb, atomic models are built and energy minimization is performed to generate the best model. The atomic coordinates for the model as well as the structural alignments are returned as an E-mail message. Those results can be resubmitted to SWISS-MODEL using its Optimize mode, which allows for alteration of the proposed structure based on other knowledge, such as biochemical information. An example of the output from SWISS-MODEL is shown in Figure 11.7.

Another automated protein fold recognition method, developed at UCLA, incorporates predicted secondary structural information on the probe sequence in addition to sequence-based matches to assign a probable protein fold to the query sequence. In this method, correct assignment of the fold depends on the ranked scores generated for the probe sequence, based on its compatibility with each of the structures in a library of target three-dimensional structures. The inclusion of the predicted secondary structure in the analysis improves fold assignment by about 25%. The input for this method is a single protein sequence submitted through a Web front end. A Web page containing the results is returned to the user, and the results are physically stored on the UCLA server for future reference.

The second approach compares structures with structures, in the same light as the vector alignment search tool (VAST) discussed in Chapter 5 does. The DALI algorithm looks for similar contact patterns between two proteins, performs an optimization, and returns the best set of structure alignment solutions for those proteins (Holm and Sander, 1993). The method is flexible in that gaps may be of any length,

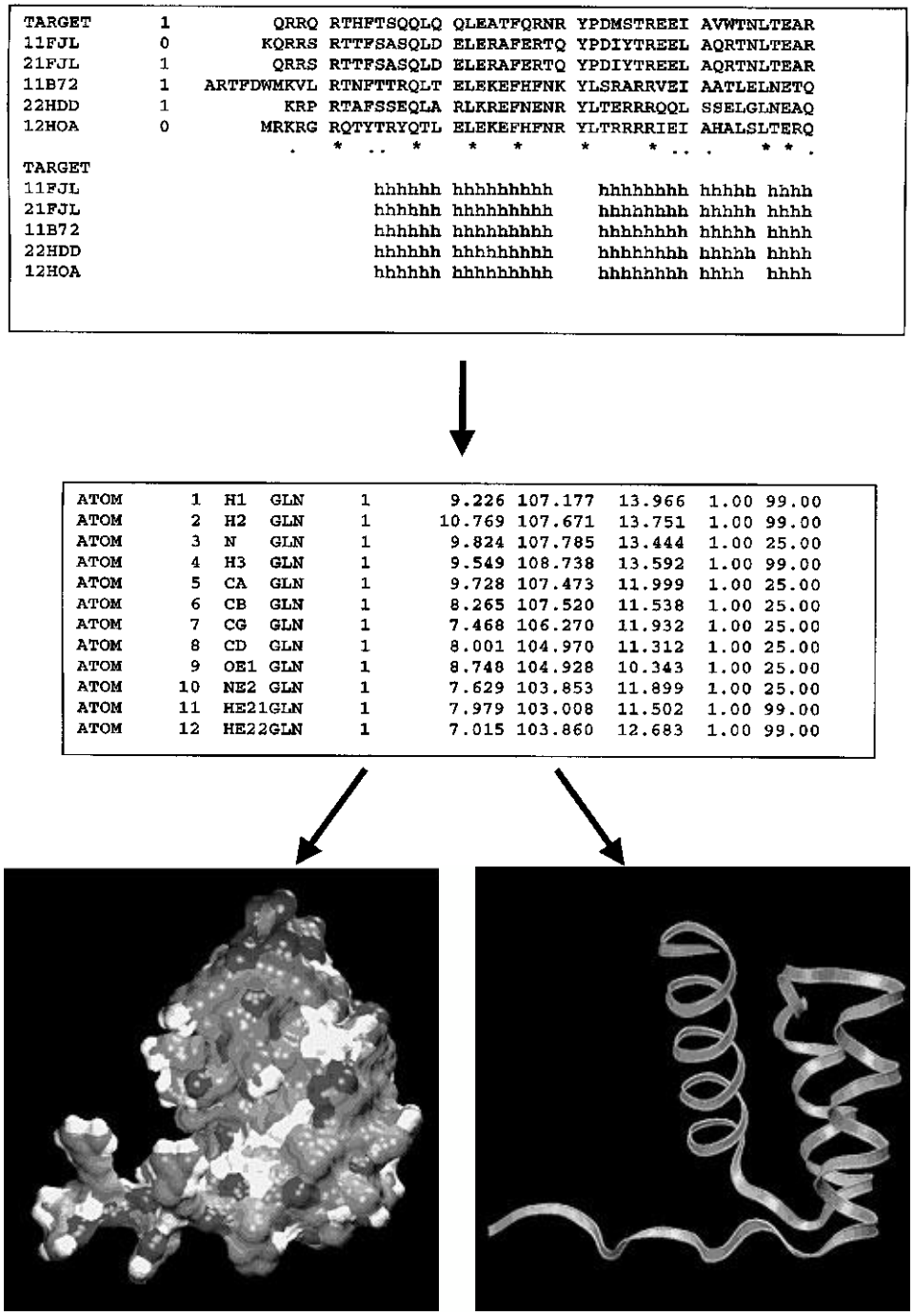

Figure 11.7. Molecular modeling using SWISS-MODEL. The input sequence for the structure prediction is the homeodomain region of human PITX2 protein. The output from SWISS-MODEL contains a text file containing a multiple sequence alignment, showing the alignment of the query against selected template structures from the Protein Data Bank (*top*). Also provided as part of the output is an atomic coordinate file for the target structure (*center*). In this example, the atomic coordinates of the target structure have been used to build a surface representation of the derived model using GRASP (*lower left*) and a ribbon representation of the derived model using RASMOL (*lower right*). (See color plate.)

and it allows for alternate connectivities between aligned segments, thereby facilitating identification of specific domains that are similar in two different proteins, even if the proteins as a whole are dissimilar. The DALI Web interface will perform the analysis on either two sets of coordinates already in PDB or by using a set of coordinates in PDB format submitted by the user. Alternatively, if both proteins of interest are present in PDB, their precomputed structural neighbors can be found by accessing the FSSP database of structurally aligned protein fold families (Holm and Sander, 1994), an ''all-against-all'' comparison of PDB entries.

The final method to be discussed here expands on the PHD secondary structure method discussed above. In the TOPITS method (Rost, 1995), a searchable database is created by translating the three-dimensional structure of proteins in PDB into one-dimensional ''strings'' of secondary structure. Then, the secondary structure and solvent accessibility of the query sequence is determined by the PHD method, with the results of this computation also being stored as a one-dimensional string. The query and target strings are then aligned by dynamic programming, to make the structure prediction. The results are returned as a ranked list, indicating the optimal alignment of the query sequence against the target structure, along with a probability estimate (*Z*-score) of the accuracy of the prediction.

The methods discussed here are fairly elementary, hence their speed in returning results and their ability to be adapted to a Web-style interface. Their level of performance is impressive in that they often can detect weak structural similarities between proteins. Although the protein-folding problem is nowhere near being solved, numerous protein folds can reliably be identified using intricate methods that are continuously being refined. Because different methods proved to have different strengths, it is always prudent to use a ''consensus approach,'' similar to the approach used in the secondary structure prediction examples given earlier. The timing of these computational developments is quite exciting, inasmuch as concurrence with the imminent completion of the Human Genome Project will give investigators a powerful handle for predicting structure-function relationships as putative gene products are identified.

INTERNET RESOURCES FOR TOPICS PRESENTED IN CHAPTER 11

PREDICTION OF PHYSICAL PROPERTIES

Compute pI/MW	*http://www.expasy.ch/tools/pi_tool.html*
MOWSE	*http://srs.hgmp.mrc.ac.uk/cgi-bin/mowse*
PeptideMass	*http://www.expasy.ch/tools/peptide-mass.html*
TGREASE	*ftp://ftp.virginia.edu/pub/fasta/*
SAPS	*http://www.isrec.isb-sib.ch/software/SAPS_form.html*

PREDICTION OF PROTEIN IDENTITY BASED ON COMPOSITION

AACompIdent	*http://www.expasy.ch/tools/aacomp/*
AACompSim	*http://www.expasy.ch/tools/aacsim/*
PROPSEARCH	*http://www.embl-heidelberg.de/prs.html*

MOTIFS AND PATTERNS

BLOCKS	*http://blocks.fhcrc.org*
Pfam	*http://www.sanger.ac.uk/Software/Pfam/*
PRINTS	*http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/PRINTS.html*

ProfileScan	*http://www.isrec.isb-sib.ch/software/PFSCAN_form.html*

PREDICTION OF SECONDARY STRUCTURE AND FOLDING CLASSES

nnpredict	*http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html*
PredictProtein	*http://www.embl-heidelberg.de/predictprotein/*
SOPMA	*http://pbil.ibcp.fr/*
Jpred	*http://jura.ebi.ac.uk:8888/*
PSIPRED	*http://insulin.brunel.ac.uk/psipred*
PREDATOR	*http://www.embl-heidelberg.de/predator/predator_info.html*

PREDICTION OF SPECIALIZED STRUCTURES OR FEATURES

COILS	*http://www.ch.embnet.org/software/COILS_form.html*
MacStripe	*http://www.york.ac.uk/depts/biol/units/coils/mstr2.html*
PHDtopology	*http://www.embl-heidelberg.de/predictprotein*
SignalP	*http://www.cbs.dtu.dk/services/SignalP/*
TMpred	*http://www.isrec.isb-sib.ch/ftp-server/tmpred/www/TMPRED_form.html*

STRUCTURE PREDICTION

DALI	*http://www2.ebi.ac.uk/dali/*
Bryant-Lawrence	*ftp://ncbi.nlm.nih.gov/pub/pkb/*
FSSP	*http://www2.ebi.ac.uk/dali/fssp/*
UCLA-DOE	*http://fold.doe-mbi.ucla.edu/Home*
SWISS-MODEL	*http://www.expasy.ch/swissmod/SWISS-MODEL.html*
TOPITS	*http://www.embl-heidelberg.de/predictprotein/*

PROBLEM SET

The sequence analyzed in the problem set in Chapter 10 yields at least one protein translation. Characterize this protein translation by answering the following questions.

1. Use ProtParam to determine the basic physicochemical properties of the unknown (leave the def line *out* when pasting the sequence into the query box).
- What is the molecular weight (in kilodaltons) and predicted isoelectric point (pI) for the protein?

2. Based on the pI and the distribution of charged residues, would this unknown possibly be involved in binding to DNA? Perform a BLASTP search on the unknown, using SWISS-PROT as the target database. Run BLASTP using pairwise as the Alignment View. *For each part of this question, consider the first protein in the hit list having a non-zero E-value.*
- What is the identity of this best, non-zero *E*-value hit, and what percent identity does the unknown share with this protein? For each alignment given, show the percent identity **and** the overall length of the alignment.
- Based on the BLASTP results *alone*, can any general observations be made regarding the putative function or cellular role of the unknown? *Do not just*

name the unknown—tell what you think the function of the unknown might be in the cell, based on ***all*** *of the significant hits in the BLASTP results.*

3. Does ProfileScan yield any additional information about the domain structure of this protein?
- What types of domains were found? How many of each of these domains are present in the unknown?
- Does the protein contain any low-complexity regions? If so, where?
- Following the PDOC links to the right of the found domains, can any conclusions be made as to the cellular localization of this protein?

4. Does this protein have a putative signal sequence, based on SignalP? If so, what residues comprise the signal sequence? Is the result obtained from SignalP consistent with the BLASTP results and any associated GenBank entries?

5. Submit the sequence of the unknown to PHDtopology. On the basis of the results, draw a schematic of the protein, showing
- the approximate location of any putative transmembrane helices and
- the orientation of the N- and C-termini with respect to the membrane.

REFERENCES

Akrigg, D., Bleasby, A. J., Dix, N. I. M., Findlay, J. B. C., North, A. C. T., Parry-Smith, D., Wootton, J. C., Blundell, T. I., Gardner, S. P., Hayes, F., Sternberg, M. J. E., Thornton, J. M., Tickle, I. J., and Murray-Rust, P. (1988). A protein sequence/structure database. *Nature* 335, 745–746.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25, 3389–3402.

Anfinsen, C. B., Haber, E., Sela, M., and White, F. H. (1961). The kinetics of the formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci. USA* 47, 1309–1314.

Appel, R. D., Bairoch, A., and Hochstrasser, D. F. (1994). A new generation of information retrieval tools for biologists: The example of the ExPASy WWW server. *Trends Biochem. Sci.* 19, 258–260.

Bjellqvist, B., Hughes, G., Pasquali, C., Paquet, N., Ravier, F., Sanchez, J.-C., Frutiger, S., and Hochstrasser, D. F. (1993). The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequence. *Electrophoresis* 14, 1023–1031.

Brendel, V., Bucher, P., Nourbakhsh, I., Blasidell, B. E., and Karlin, S. (1992). Methods and algorithms for statistical analysis of protein sequences. *Proc. Natl. Acad. Sci. USA* 89, 2002–2006.

Bryant, S. H., and Altschul, S. F. (1995). Statistics of sequence-structure threading. *Curr. Opin. Struct. Biol.* 5, 236–244.

Bryant, S. H., and Lawrence, C. E. (1993). An empirical energy function for threading protein sequence through the folding motif. *Proteins* 16, 92–112.

Burnett, R. M., Darling, G. D., Kendall, D. S., LeQuesne, M. E., Mayhew, S. G., Smith, W. W., and Ludwig, M. L. (1974). The structure of the oxidized form of clostridial flavodoxin at 1.9 Å resolution. *J. Biol. Chem.* 249, 4383–4392.

Chothia, C. (1992). One thousand families for the molecular biologist. *Nature* 357, 543–544.

Chothia, C., and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5, 823–826.

Cordwell, S. J., Wilkins, M. R., Cerpa-Poljak, A., Gooley, A. A., Duncan, M., Williams, K. L., and Humphery-Smith, I. (1995). Cross-species identification of proteins separated by two-dimensional electrophoresis using matrix-assisted laser desorption ionization/time-of-flight mass spectrometry and amino acid composition. *Electrophoresis* 16, 438–443.

Deléage, G., and Roux, B. (1987). An algorithm for protein secondary structure based on class prediction. *Protein Eng.* 1, 289–294.

Eisenhaber, F., Persson, B., and Argos, P. (1995). Protein structure prediction: Recognition of primary, secondary, and tertiary structural features from amino acid sequence. *Crit. Rev. Biochem. Mol. Biol.* 30, 1–94.

Fetrow, J. S., and Bryant, S. H. (1993). New programs for protein tertiary structure prediction. *Bio/Technology* 11, 479–484.

Frishman, D. and Argos, P. (1997) Seventy-five percent accuracy in protein secondary structure prediction. *Proteins* 27, 329–335.

Garnier, J., Gibrat, J.-F., and Robson, B. (1996). GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol*. 266, 540–553.

Geourjon, C., and Déleage, G. (1995). SOPMA: Significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *CABIOS* 11, 681–684.

Gill, S. C. and von Hippel, P. H. (1989) Calculation of protein extinction coefficients from amino acid sequence data. *Anal. Biochem*. 182, 319–326.

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

Gribskov, M., Homyak, M., Edenfield, J. and Eisenberg, D. (1988) Profile scanning for three-dimensional structural patterns in protein sequences. *Comput. Appl. Biosci.* 4, 61–66.

Guss, J. M., and Freeman, H. C. (1983). Structure of oxidized poplarplastocyanin at 1.6 Å resolution. *J. Mol. Biol.* 169, 521–563.

Henikoff, J. G. and Henikoff, S. (1996) Using substitution probabilities to improve position-specific scoring matrices. *Comput. Appl. Biosci.* 12, 135–43.

Hobohm, U., and Sander, C. (1995). A sequence property approach to searching protein databases. *J. Mol. Biol.* 251, 390–399.

Hofmann, K., and Stoffel, W. (1993). TMbase: A database of membrane-spanning protein segments. *Biol. Chem. Hoppe-Seyler* 347, 166.

Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.* 27, 215–219.

Holm, L., and Sander, C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* 233, 123–138.

Holm, L., and Sander, C. (1994). The FSSP database of structurally-aligned protein fold families. *Nucl. Acids Res.* 22, 3600–3609.

Ikai, A. (1980) Thermostability and aliphatic index of globular proteins. *J. Biochem. (Tokyo)* 88, 1895–1898.

Jones, D. T., and Thornton, J. M. (1996). Potential energy functions for threading. *Curr. Opin. Struct. Biol.* 6, 210–216.

King, R. D. and Sternberg, M. J. (1996) Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci.* 5, 2298–2310.

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

Knight, A. E. (1994). *The Diversity of Myosin-like Proteins* (Cambridge: Cambridge University Press).

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

Lemer, C. M., Rooman, M. J., and Wodak, S. J. (1995). Protein structure prediction by threading methods: Evaluation of current techniques. *Proteins* 23, 337–355.

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

Levitt, M., and Chothia, C. (1976). Structural patterns in globular proteins. *Nature* 261, 552–558.

Lupas, A., Van Dyke, M., and Stock, J. (1991). Predicting coiled coils from protein sequences. *Science* 252, 1162–1164.

Luthy, R., Xenarios, I. and Bucher, P. (1994) Improving the sensitivity of the sequence profile method. *Protein Sci.* 3, 139–146.

Mehta, P. K., Heringa, J., and Argos, P. (1995). A simple and fast approach to prediction of protein secondary structure from multiply aligned sequences with accuracy above 70%. *Protein Sci.* 4, 2517–2525.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* 10, 1–6.

Pappin, D. J. C., Hojrup, P., and Bleasby, A. J. (1993). Rapid identification of proteins by peptide-mass fingerprinting. *Curr. Biol.* 3, 327–332.

Pauling, L., and Corey, R. B. (1951). The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. USA* 37, 205–211.

Peitsch, M. C. (1996). ProMod and SWISS-MODEL: Internet-based tools for automated comparative protein modeling. Biochem. *Soc. Trans.* 24, 274–279.

Persson, B., and Argos, P. (1994). Prediction of transmembrane segments in proteins utilizing multiple sequence alignments. *J. Mol. Biol.* 237, 182–192.

Rost, B. (1995). TOPITS: Threading one-dimensional predictions into three-dimensional structures. In *Third International Conference on Intelligent Systems for Molecular Biology*, C. Rawlings, D. Clark, R. Altman, L. Hunter, T. Lengauer, and S. Wodak, Eds. (Cambridge: AAAI Press), p. 314–321.

Rost, B. (1996). PHD: Predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol*. 266, 525–539.

Rost, B. and Sander, C. (1993) Secondary structure prediction of all-helical proteins in two states. *Protein Eng.* 6, 831–836.

Rost, B., Sander, C., and Schneider, R. (1994). PHD: A mail server for protein secondary structure prediction. *CABIOS* 10, 53–60.

Salamov, A. A. and Solovyev, V. V. (1995) Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments *J. Mol. Biol.* 247, 11–15.

Sander, C., and Schneider, R. (1991). *Proteins* 9, 56–68.

Smith, W. W., Pattridge, K. A., Ludwig, M. L., Petsko, G. A., Tsernoglou, D., Tanaka, M., and Yasunobu, K. T. (1983). Structure of oxidized flavodoxin from *Anacystis nidulans. J. Mol. Biol.* 165, 737–755.

Takano, T. (1977). Structure of myoglobin refined at 2.0 Å. *J. Mol. Biol.* 110, 537–584.

Wilkins, M. R., Pasquali, C., Appel, R. D., Ou, K., Golaz, O., Sanchez, J.-C., Yan, J. X., Gooley, A. A., Hughes, G., Humphery-Smith, I., Williams, K. L., and Hochstrasser, D. F.

(1996). From proteins to proteomes: Large-scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Bio/Techniques* 14, 61–65.

Wilkins, M. R., Lindskog, I., Gasteiger, E., Bairoch, A., Sanchez, J.-C., Hochstrasser, D. F., and Appel, R. D. (1997). Detailed peptide characterization using PeptideMass, a World Wide Web-accessible tool. *Electrophoresis* 18, 403–408.

Zvelebil, M. J., Barton, G. J., Taylor, W. R. and Sternberg, M. J. (1987) Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* 195, 957–961.

12

EXPRESSED SEQUENCE TAGS (ESTs)

Tyra G. Wolfsberg

Genome Technology Branch
National Human Genome Research Institute
National Institutes of Health
Bethesda, Maryland

David Landsman

Computational Biology Branch
National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
Bethesda, Maryland

The benefits arising from the rapid generation of large numbers of low-quality cDNA sequences were not universally recognized when the concept was originally proposed in the late 1980s. Proponents of this approach argued that these cDNA sequences would allow for the quick discovery of hundreds or thousands of novel protein-coding genes. Their critics countered that cDNA sequencing would miss important regulatory elements that could be found only in the genomic DNA. In the end, the cDNA sequencing advocates appear to have won. Since the original description of 609 Expressed Sequence Tags (ESTs) in 1991 (Adams et al., 1991), the growth of ESTs in the public databases has been dramatic. The number of ESTs in GenBank surpassed the number of non-EST records in mid-1995; as of June 2000, the 4.6 million EST records comprised 62% of the sequences in GenBank. Although the original ESTs were of human origin, NCBI's EST database (dbEST) now contains ESTs from over 250 organisms, including mouse, rat, *Caenorhabditis elegans*, and *Drosophila melanogaster*. In addition, several commercial establishments maintain privately funded, in-house collections of ESTs. ESTs are now widely used throughout

the genomics and molecular biology communities for gene discovery, mapping, polymorphism analysis, expression studies, and gene prediction.

WHAT IS AN EST?

An overview of an EST sequencing project is shown in Figure 12.1. In brief, a cDNA library is constructed from a tissue or cell line of interest. Individual clones are picked from the library, and one sequence is generated from each end of the cDNA insert. Thus, each clone normally has a 5′ and 3′ EST associated with it. The sequences average ~400 bases in length. Because the ESTs are short, they generally represent only fragments of genes, not complete coding sequences. Many sequencing centers have automated the process of EST generation, producing ESTs at a rapid rate. For example, at the time of this writing, the Genome Sequencing Center at Washington University was producing over 20,000 ESTs per week.

The ESTs that have been submitted to the public sequence databases to date were created from thousands of different cDNA libraries representing over 250 organisms. The libraries may be from whole organs, such as human brain, liver, lung, or skeletal muscle, specialized tissues or cells, such as cerebral cortex or epidermal keratinocyte, or cultured cell lines such as liver HepG2 or gastric carcinoma. Some libraries have been constructed to compare transcripts from different developmental stages, such as fetal versus infant human brain or embryonic 7-day versus neonatal 10-day rat heart ventricle. Others are used to highlight gene expression differences between normal and transformed tissue, such as normal colonic epithelium and colorectal carcinoma cell line. The libraries are constructed by isolating mRNA from the tissue or cell line of interest. The mRNA is then reverse-transcribed into cDNA, usually with an oligo(dT) primer, so that one end of the cDNA insert derives from the polyA tail at the end of the mRNA. The other end of the cDNA is normally

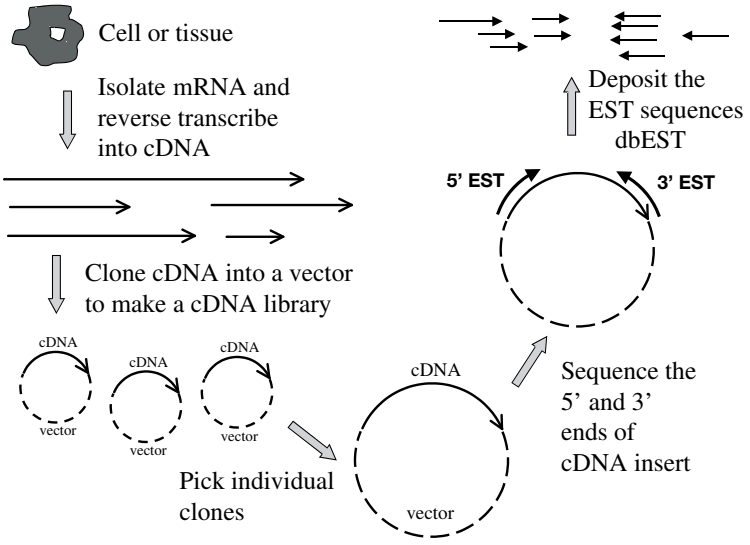

Figure 12.1. Overview of how ESTs are constructed.

within the coding sequence but may be in the 5′ untranslated region if the coding sequence is short. The resulting cDNA is cloned into a vector. In many libraries, the cDNA is cloned directionally. Some of the libraries are normalized to bring the frequency of occurrence of clones representing individual mRNA species into a narrow range (Bonaldo et al., 1996; Soares et al., 1994). Other libraries are constructed by a process of subtractive hybridization, in which a pool of mRNA sequences is removed from a library of interest, leaving behind sequences unique to that library (Bonaldo et al., 1996). For example, to construct a library for the study of bipolar disorder, researchers started with human frontal lobe cDNA from individuals with bipolar disorder, and subtracted out cDNA that hybridized to cDNA from mentally normal individuals (see *http://www.ncbi.nlm.nih.gov/dbEST/dbest_libs.html#lib1475*).

With the use of primers that hybridize to the vector sequence, the ends of the cDNA insert are sequenced. Automatic DNA sequencers generate most EST data. If the cDNA has been directionally cloned into the vector, the sequences can be classified as deriving from the 5′ or 3′ end of the clone. In most cases, both the 5′ and 3′ sequences are determined, but some EST projects have concentrated only on 5′ ESTs to maximize the amount of coding sequence determined. Because the sequence of each EST is generated only once, the sequences may (and often do) contain errors. Contaminating vector, mitochondrial, and bacterial sequences are routinely removed before the EST sequences are deposited into the public databases (Hillier et al., 1996). ESTs in the databases are identified by their clone number as well as their 5′ or 3′ orientation, if known.

The I.M.A.G.E. Consortium (Lennon et al., 1996) has picked individual clones from many of the libraries used for EST sequencing and arrayed them for easy distribution. These clones can be obtained royalty-free from I.M.A.G.E. Consortium distributors. As of the time of this writing, more than 3.8 million cDNA clones have been arrayed from 360 human and 108 mouse cDNA libraries; zebrafish and *Xenopus* clones have also been arrayed. I.M.A.G.E. Consortium sequences currently comprise more than half of the ESTs in GenBank. Most of the sequencing of I.M.A.G.E. clones is performed by the Genome Sequencing Center at Washington University/St. Louis. Merck sponsored human clone sequencing in 1995 and 1996; since then, the collaborative EST project has been sponsored by the National Cancer Institute as part of the Cancer Genome Anatomy Project. Sequencing by Washington University/St. Louis of mouse cDNAs is sponsored by the Howard Hughes Medical Institute. Sequence trace data from the ESTs sequenced by the Washington University/St. Louis projects are available online.

How to Access ESTs

ESTs are submitted to all three international sequence databases (GenBank, EMBL, and DDBJ), under the data-sharing agreement described in Chapter 2. Therefore, all ESTs can be accessed through all of these databases, regardless of where the sequence was originally submitted. The same ESTs are also available from the NCBI's dbEST, the database of Expressed Sequence Tags (Boguski et al., 1993). Instructions about how to submit EST sequences to GenBank are available online.

Like other sequences in GenBank, ESTs can be accessed through Entrez (see Chapter 7). Single ESTs are retrieved by accession or gi number. Advanced searches with multiple search terms can be limited to ESTs by selecting the `Properties` limit and entering `EST`. The two ESTs deriving from a particular I.M.A.G.E. clone

can be retrieved by searching for "IMAGE:clone_number" (e.g., "IMAGE: 743313"). The Entrez version of the EST with accession AW592465 is shown in Figure 12.2. Various identifiers for the EST, including the accession number and GenBank gi, are shown in the top block. The CLONE INFO section specifies the number of the clone (2934602) and whether this EST derives from the 5′ or 3′ end of the clone (here, 3′). The nucleotide sequence is shown next, along with a note supplied by the submitter about where the high-quality sequence stops. The COMMENTS block tells how to order the clone from the I.M.A.G.E. Consortium. The last few sections present other information supplied by the submitter, including details about the cDNA library. Although many ESTs (especially 5′ ESTs) can be translated into a partial or sometimes full-length protein sequence, coding sequence features are not provided. Other views of the data, including a FASTA-formatted DNA sequence, can be selected from a pull-down at the top of the Entrez entry (not shown).

EST sequences are also available for BLAST searching. Because ESTs are nucleotide sequences, they can be retrieved only by using BLAST programs that search nucleotide databases (BLASTN for a nucleotide sequence query, TBLASTN for a protein sequence query, and TBLASTX for a translated nucleotide sequence query). Because they make up such a high proportion of sequences in GenBank, ESTs are not included in the BLAST *nr* database. To search against ESTs, select the *dbest* database or, for a specific organism, the *mouse ests*, *human ests*, or *other ests* database. Note that ESTs *are* also included in the *month* database, which contains *all* new or revised sequences released in the last 30 days.

Limitations of EST Data

Although ESTs are an excellent source of sequence data, these data are not of as high a quality as sequences determined by conventional means. Because EST sequences are generated in a single pass, they have a higher error rate than sequences that are verified by multiple sequencing runs, on the order of 3% (Boguski et al., 1993). In contrast, the standard for the human genome project is an error rate of <0.01% (Collins et al., 1998). ESTs may contain substitutions, deletions, or insertions compared with the parent mRNA sequence. The region of an EST between positions 100 and 300 may be the most accurate part of the sequence (Hillier et al., 1996).

Hillier et al. (1996) have performed a comprehensive analysis of potential EST artifacts. They found that ESTs may contain bacterial, mitochondrial, or vector sequence contamination. Most EST cDNA libraries are oligo(dT) primed, and the 3′ EST derives from the 3′ untranslated region of the gene. However, Hillier et al. found that 1.5% of oligo(dT)-primed 3′ ESTs do not align with the known 3′ end of the mRNA. These ESTs either represent nonspecific priming or indicate alternative splicing. cDNA for some libraries is synthesized with random primers, so the location of the 3′ EST is unknown. Another potential problem comes from inverted clones in directionally cloned libraries, in which the 5′ and 3′ EST are mislabeled. cDNA inserts may be inverted because of failures in the directional cloning procedure, or simply because of human error. Hillier et al. found that 6.25% of ESTs that match a known mRNA align in an inverted orientation. Chimeric clones, in which the 5′ EST matches one mRNA and the 3′ EST another mRNA, may arise either during library construction or sample handling. Hillier et al. found a chimera frequency of

IDENTIFIERS

dbEST Id: **4025315**
EST name: hf43a02.x1
GenBank Acc: AW592465
GenBank gi: 7279647

CLONE INFO
Clone Id: IMAGE:2934602 (3')
Source: NCI
DNA type: cDNA

PRIMERS
Sequencing: -40UP from Gibco
PolyA Tail: Unknown

SEQUENCE

TTTTTTTTTAAATTGCCAAGTGATTTTACTTCAAGATGACATCAGAATTGCTAAAAGGTG
ATGTAACCGTCAGAGTGACTATTGATTATAACTCCCAGTAAGTGTCAACGTGATTTTCTC
CATTGTGTGGGCTTCCATTAGTATTTACTCATTAGGTTCAGTAGTTTTCATTATTTTCTC
TTCCATAAATTCTATTGCTTGTGAAAAGCCACCAAAGAGAAGTGAAACCAGAAAAAGGAT
GCAACGAGTAAATATTAAAAGTAGTGCTCAGTTTATATTCGCAAGTGTGCTGGCTGTAAT
ACGATATTGTTTGTCAGGTGGAGGGCCACTATCTATACTACCTCCTTTTCCTCAGTTCAC
ATGTTGGTGGTTGCCACCCATGCAGACAGTGACAATGTTTTTTGTTGTTACATACTCCTT
TGTAATTGCATGTGTTAAGAACACACTCAAAATGCAGGTCTTGATAAGAAGGCAATTGTG
TTTAAGACAGTAGCTGCCTGGGCCACAGGTTGCACCATCCACTGACCGCCCCATTTCTGG
CAAGTCTGGACCCTGGTGTGGCTAATAACCAAGGCATTTATT

Quality: High quality sequence stops at base: 356

Entry Created: Mar 22 2000
Last Updated: Mar 22 2000

COMMENTS

This clone is available royalty-free through LLNL ; contact the IMAGE Consortium (info@image.llnl.gov) for further information.

PUTATIVE ID Assigned by submitter
TR:Q60815 Q60815 ADAM 4 PROTEIN PRECURSOR ;

LIBRARY
Lib Name: Soares_NFL_T_GBC_S1
Organism: Homo sapiens
Organ: pooled
Lab host: DH10B
Vector: pT7T3D-Pac (Pharmacia) with a modified polylinker
R. Site 1: Not I
R. Site 2: Eco RI
Description: Equal amounts of plasmid DNA from three normalized libraries (fetal lung NbHL19W, testis NHT, and B-cell NCI_CGAP_GCB1) were mixed, and ss circles were made in vitro. Following HAP purification, this DNA was used as tracer in a subtractive hybridization reaction. The driver was PCR-amplified cDNAs from pools of 5,000 clones made from the same 3 libraries. The pools consisted of I.M.A.G.E. clones 297480-302087, 682632-687239, 726408-728711, and 729096-731399. Subtraction by Bento Soares and M. Fatima Bonaldo.

SUBMITTER
Name: Robert Strausberg, Ph.D.
Tel: (301) 496-1550
E-mail: Robert_Strausberg@nih.gov

CITATIONS
Title: National Cancer Institute, Cancer Genome Anatomy Project (CGAP), Tumor Gene Index
Authors: NCI-CGAP http://www.ncbi.nlm.nih.gov/ncicgap
Year: 1997
Status: Unpublished

MAP DATA

Figure 12.2. The Entrez view of an EST, accession AI273896.

1%, but a separate study estimated the frequency at 11% (Wolfsberg and Landsman, 1997).

EST CLUSTERING

As of mid-2000, GenBank contained just under 1.9 million human EST records. Although original estimates of the number of genes in the human genome hovered around the 100,000 mark, predictions made based on experimental data and presented at the 2000 Cold Spring Harbor Genome meeting have drastically reduced the estimate to below 50,000. In any event, it is clear, even without doing any sequence comparisons, that these ESTs cannot each represent a unique sequence. Even with the process of library normalization, abundant transcripts are represented more frequently in dbEST than rare ones. For example, dbEST contains more than 200 ESTs for human alpha-fetoprotein alone. A number of efforts are geared at simplifying this abundance of DNA sequences by grouping together records that likely derive from the same gene. Other resources, including those for mapping and gene discovery, can then make use of this condensed set of gene-based clusters, rather than the expansive and relatively unorganized collection of all ESTs and other mRNA sequences.

UniGene

The UniGene resource, developed at NCBI, clusters ESTs and other mRNA sequences, along with coding sequences (CDSs) annotated on genomic DNA, into subsets of related sequences (Boguski and Schuler, 1995; Wagner, L. et al., unpublished observations). In most cases, each cluster is made up of sequences produced by a single gene, including alternatively spliced transcripts (Fig. 12.3). However, some genes may be represented by more than one cluster. The clusters are organism specific and are currently available for human, mouse, rat, zebrafish, and cattle. They are built in several stages, using an automatic process based on special sequence comparison algorithms. First, the nucleotide sequences are searched for contaminants, such as mitochondrial, ribosomal, and vector sequence, repetitive elements, and low-complexity sequences. After a sequence is screened, it must contain at least 100 bases to be a candidate for entry into UniGene. mRNA and genomic DNA are clustered first into gene links. A second sequence comparison links ESTs to each other and to the gene links. At this stage, all clusters are ''anchored,'' and contain either a sequence with a polyadenylation site or two ESTs labeled as coming from the 3′ end of a clone. Clone-based edges are added by linking the 5′ and 3′ ESTs that derive from the same clone. In some cases, this linking may merge clusters identified at a previous stage. Finally, unanchored ESTs and gene clusters of size 1 (which may represent rare transcripts) are compared with other UniGene clusters at lower stringency. The UniGene build is updated weekly, and the sequences that make up a cluster may change. Thus, it is not safe to refer to a UniGene cluster by its cluster identifier; instead, one should use the GenBank accession numbers of the sequences in the cluster. A summary of the UniGene build procedure is shown in Figure 12.4a. Additional information about the UniGene build is available online.

As of July 2000, the human subset of UniGene contained 1.7 million sequences in 82,000 clusters; 98% of these clustered sequences were ESTs, and the remaining

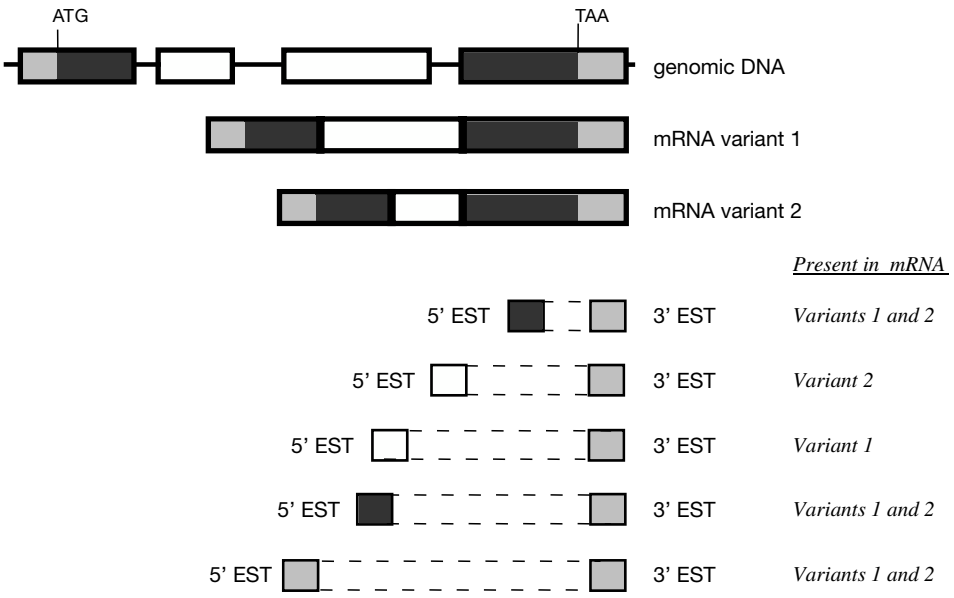

Figure 12.3. Sequences in a UniGene cluster. This cluster contains a genomic DNA sequence with an annotated coding sequence (CDS), two alternatively-spliced mRNA sequences, and 10 ESTs from five clones that derive from the mRNA sequences.

2% were from mRNAs or CDSs annotated on genomic DNA. These human clusters could represent fragments of up to ~82,000 unique human genes, implying that many human genes are now represented in a UniGene cluster. (This number is undoubtedly an overestimate of the number of genes in the human genome, as some genes may be represented by more than one cluster.) Only 1.4% of clusters totally lack ESTs, implying that most human genes are represented by at least one EST. Conversely, it appears that the majority of human genes have been identified *only* by ESTs; only 16% of clusters contain either an mRNA or a CDS annotated on a genomic DNA. Because fewer ESTs are available for mouse, rat, and zebrafish, the UniGene clusters are not as representative of the unique genes in the genome. Mouse UniGene contains 895,000 sequences in 88,000 clusters, and rat UniGene contains 170,000 sequences in 37,000 clusters.

A new UniGene resource, HomoloGene, includes curated and calculated orthologs and homologs for genes from human, mouse, rat, and zebrafish. Calculated orthologs and homologs are the result of nucleotide sequence comparisons between all UniGene clusters for each pair of organisms. Homologs are identified as the best match between a UniGene cluster in one organism and a cluster in a second organism. When two sequences in different organisms are best matches to one another (a reciprocal best match), the UniGene clusters corresponding to the pair of sequences are considered putative orthologs. A special symbol indicates that UniGene clusters in three or more organisms share a mutually consistent ortholog relationship. The calculated orthologs and homologs are considered putative, since they are based only on sequence comparisons. Curated orthologs are provided by the Mouse Genome Database (MGD) at the Jackson Laboratory and the Zebrafish Information Database (ZFIN) at the University of Oregon and can also be obtained from the scientific

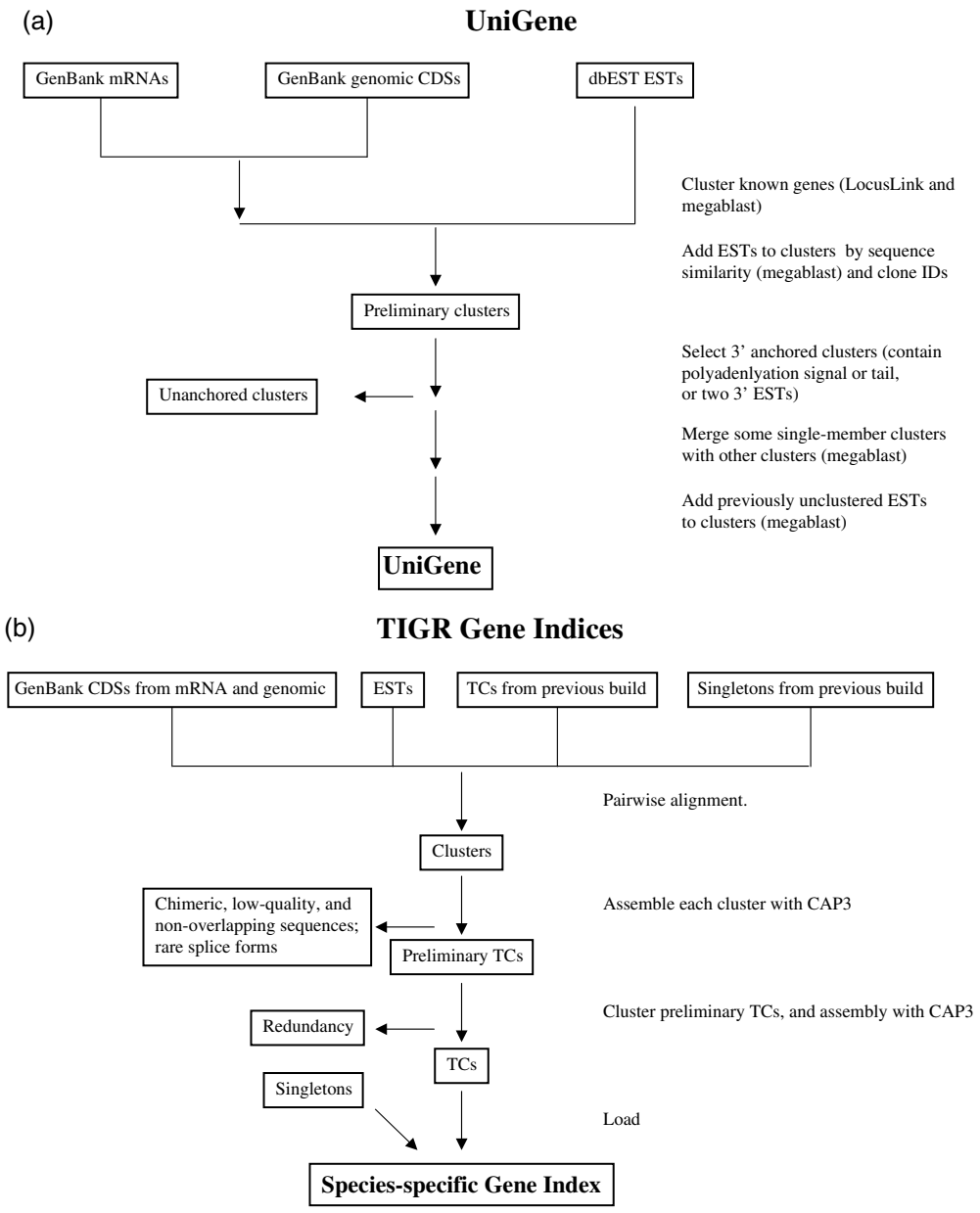

Figure 12.4. Schematics for clustering of ESTs. All three methods prescreen ESTs for contaminating sequence. (a) UniGene. Most sequence analysis is done with MegaBLAST (Zhang et al., 2000), a fast version of BLAST. The minimum alignment length is 70 nucleotides, and an alignment must extend over at least 70% of the alignable region in the first two steps or 55% of the alignable region in the last two steps. (b) TIGR Gene Indices. Sequences are clustered if they share a minimum of 95% identity over a 40 nucleotide region, with fewer than 20 nucleotides of mismatched sequence at either end. Sequences are assembled with CAP3 (Huang and Madan, 1999). (c) STACK. Sequences are clustered if they share 96% identity over 150 nucleotides. Clustering is done with d2_cluster (Burke et al., 1999) and aligned with PHRAP (Green, 1996).

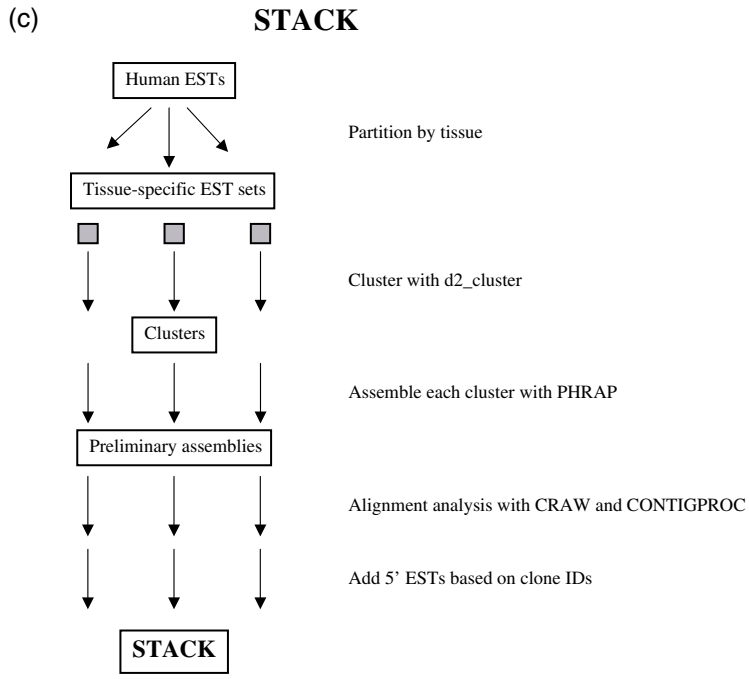

Figure 12.4. Continued

literature. Direct links to HomoloGene are provided for UniGene clusters that have a candidate ortholog or homolog.

Queries to UniGene are entered into a text box on any of the UniGene pages. Query terms can be, for example, the UniGene identifier, a gene name, a text term that is found somewhere in the UniGene record, or the accession number of an EST or gene sequence in the cluster. For example, the cluster entitled ''A disintegrin and metalloprotease domain 10'' that contains the sequence for human ADAM10 can be retrieved by entering `ADAM10`, `disintegrin`, `AF009615` (the GenBank accession number of ADAM10), or `H69859` (the GenBank accession number of an EST in the cluster). Enter multiple terms to get a list of entries containing all terms. To query a specific part of the UniGene record, use the @ symbol. For example, `@gene(symbol)` looks for genes with the name of the symbol enclosed in the parentheses, `@chr(num)` searches for entries that map to chromosome num, `@lib(id)` returns entries in a cDNA library identified by id, and `@pid(id)` selects entries associated with a GenBank protein identifier id.

The query results page contains a list of all UniGene clusters that match the query. Each cluster is identified by an identifier, a description, and a gene symbol, if available. Cluster identifiers are prefixed with Hs for *Homo sapiens*, Rn for *Rattus norvegicus*, Mm for *Mus musculus*, or Dn for *Danio rerio*. The descriptions of UniGene clusters are taken from LocusLink, if available, or from the title of a sequence in the cluster. The UniGene report page for each cluster links to data from other NCBI resources (Fig. 12.5). At the top of the page are links to LocusLink, which provides descriptive information about genetic loci (Pruitt et al., 2000), OMIM, a catalog of human genes and genetic disorders, and HomoloGene. Next are

Hs.172028 *Homo sapiens* **ADAM10**

A disintegrin and metalloprotease domain 10

SEE ALSO

LocusLink: 102
OMIM: 602192
HomoloGene: Hs.172028

SELECTED MODEL ORGANISM PROTEIN SIMILARITIES
organism, protein and percent identity and length of aligned region

H. sapiens :	PID:g2393947 - ADAM10	**100 % / 747 aa**
M. musculus :	PID:g2282608 - kuzbanian	**95 % / 747 aa**
R. norvegicus :	PIR:S52477 - disintegrin	**97 % / 543 aa**
D. melanogaster :	PID:g1531633 - kuzbanian	**46 % / 584 aa**
C. elegans :	PID:g3875352 - similar to Zinc-binding metalloprotease domain	**43 % / 715 aa**

MAPPING INFORMATION

Chromosome: 15
Cytogenetic Position: 15q22
Gene Map 98: Marker H69847 , Interval D15S117-D15S159
Gene Map 98: Marker stSG15641 , Interval D15S117-D15S159

EXPRESSION INFORMATION

cDNA sources: Brain, Breast, CNS, Foreskin, Heart, Kidney, Liver, Lung, Lymph, Muscle, Ovary, Pancreas, Pooled, Thyroid, Tonsil, Uterus, Whole embryo, breast_normal, colon, genitourinary tract, head_neck, kidney, lung
SAGE : Gene to Tag mapping

mRNA/GENE SEQUENCES (3)

AF009615	Homo sapiens ADAM10 (ADAM10) mRNA, complete cds	P
Z48579	H.sapiens mRNA for disintegrin-metalloprotease (partial)	P
NM_001110	Homo sapiens disintegrin and metalloprotease domain 10 (ADAM10) mRNA	

EST SEQUENCES (10 of 122)[Show all ESTs]

AI680390	cDNA clone IMAGE:2264325 Uterus	3' read 3.4 kb	C
AI188417	cDNA clone IMAGE:1723156	3' read 2.2 kb	A C
AA812209	cDNA clone IMAGE:1338085 Tonsil	2.0 kb	A C
AI302180	cDNA clone IMAGE:1902444 Kidney	3' read 1.9 kb	A C
AI742324	cDNA clone IMAGE:2368543 Pooled	3' read 1.8 kb	C
AI273896	cDNA clone IMAGE:1964283 Ovary	3' read 1.7 kb	C
AI472608	cDNA clone IMAGE:2149114 Pooled	3' read 1.7 kb	C
AI368402	cDNA clone IMAGE:2011372 Brain	3' read 1.6 kb	C
H69450	cDNA clone IMAGE:212359	5' read 1.5 kb	P
H69859	cDNA clone IMAGE:212359	3' read 1.5 kb	

Key to Symbols

P Has similarity to known **P**roteins (after translation)
A Contains a poly-**A**denylation signal
S Contains a mapped **S**equence-tagged site (STS)
C Clone source is a **C**GAP library

Figure 12.5. UniGene cluster for ADAM10. The contents of the cluster are subject to change as additional sequences are submitted to the public databases.

listed similarities between the translations of DNA sequences in the cluster and protein sequences from model organisms, including human, mouse, rat, fruit fly, and worm. The subsequent section describes relevant mapping information. It is followed by ''expression information,'' which lists the tissues from which the ESTs in the cluster have been created, along with links to the SAGE database (see below). Sequences making up the cluster are listed next, along with a link to download these sequences.

It is important to note that clusters that contain ESTs only (i.e., no mRNAs or annotated CDSs) will be missing some of these fields, such as LocusLink, OMIM, and mRNA/Gene links. UniGene titles for such clusters, such as ''EST, weakly similar to ORF2 contains a reverse transcriptase domain [H. sapiens],'' are derived from the title of a characterized protein with which the translated EST sequence aligns. The cluster title might be as simple as ''EST'' if the ESTs share no significant similarity with characterized proteins.

TIGR GENE INDICES

The TIGR Gene Indices represent another effort to consolidate EST and other annotated gene sequences (Quackenbush et al., 2000). A significant difference between the Gene Indices and UniGene is that the Gene Indices are assemblies of ESTs and other gene sequences rather than clusters (Figure 12.4b). The assemblies tend to represent one transcript, so alternatively spliced products are grouped separately. Furthermore, the process generates a single consensus sequence per assembly.

A Gene Index is maintained for 14 organisms, including human, mouse, rat, *Drosophila*, zebrafish, *Arabidopsis*, and several crop plants. Gene Indices are created from publicly available GenBank and dbEST sequences by clustering ESTs with the DNA sequences encoding the coding sequences annotated on DNA and mRNA sequences. The elements of a cluster are assembled with other EST sequences into tentative consensus sequences (TCs, or THCs for human). TCs are updated as sequence flow into the public databases. The TIGR databases, as of mid-2000, contain 85,000 THCs, 43,000 TCs from mouse, and 18,000 TCs from rat. These numbers are somewhat lower than the numbers of UniGene clusters, probably due to different methods used for clustering. The TIGR Gene Indices, like UniGene, can be queried with text searches. A BLAST interface, for BLASTN and TBLASTN, is also available. A related project at TIGR is to identify orthologous genes between human, mouse, and rat using the TCs. The TIGR Orthologous Gene Alignment (TOGA) database represents the ortholog sets.

STACK

The STACK resource at the South African National Bioinformatics Institute (SANBI) uses a third method, a combination of clustering and assembly, to group related ESTs into clusters (Burke et al., 1999; Miller et al., 1999). At this time, STACK clusters are available only for human ESTs (Fig. 12.4c). STACK clusters consolidate ESTs into a smaller number of groups than does UniGene. Unlike UniGene or the TIGR Gene Indices, ESTs in STACK are separated by tissue type before being clustered. BLAST queries can be performed on STACK clusters.

ESTs AND GENE DISCOVERY

ESTs have been widely used for gene discovery (Boguski et al., 1994). Because ESTs outnumber other nucleotide sequences in GenBank, researchers hunting for a novel gene are much more likely to find it in dbEST than in the rest of GenBank. ESTs are not included in the BLAST *nr* database, and sequence similarity searches for ESTs must target the est database. Gene discovery methods using ESTs include, for example, hunting for new members of gene families in the same species (paralogs), for functionally equivalent genes in other species (orthologs), or even for alternatively spliced forms of known genes.

Gene discovery using dbEST is very rapid, requiring only a few minutes for the BLAST search. For example, ADAM 10, also known as ''kuzbanian,'' is a well-studied gene with known orthologs in human, mouse, rat, cow, frog, pig, fruit fly, and worm. A TBLASTN search of *dbest* with the human protein sequence quickly reveals not only many of these known genes but also an additional likely ortholog in zebrafish. Discovering alternatively spliced transcripts among ESTs is more problematic. For one, it is difficult to determine if the new sequences are due to alternative splicing or to the presence of contaminating genomic DNA sequence in EST libraries (Wolfsberg and Landsman, 1997). An analysis of the TIGR Human Gene Index using a spliced alignment algorithm provided evidence that up to 35% of human genes may undergo alternative splicing and that the majority of these events occur in the 5′ untranslated regions (Mironov et al., 1999).

The uses of ESTs extend beyond mammals. For example, until recently, the public databases contained little sequence data from *Toxoplasma gondii*, a disease-causing protozoan parasite of human. A large scale project generated 7,000 5′ ESTs, representing ~4,000 unique sequences, from *T. gondii* (Ajioka et al., 1998). Comparisons between the ESTs and sequences in public databases identified potential functions for 500 novel *T. gondii* genes. Some ESTs are phylogenetically restricted to *T. gondii* and other members of the Apicomplexa phylum.

THE HUMAN GENE MAP

ESTs are also being used to create gene maps by the use of sequence-tagged sites (STSs), short stretches of unique sequence identified by polymerase chain reaction (PCR) assays. An international consortium agreed to coordinate a mapping effort for the human genome, using UniGene clusters to represent individual human genes. In the initial effort, the gene-based STSs were mapped relative to two radiation hybrid (RH) maps and one YAC panel; in subsequent work, the STSs have been mapped only to the two RH panels, the Genebridge4 (GB4) and Stanford G3. STSs were generated from the 3′ untranslated regions of UniGene clusters. GeneMap '96 reported the mapping of 16,000 gene-based STSs (Schuler et al., 1996), and GeneMap '98 nearly doubled that number to 30,000 (Deloukas et al., 1998). Thus, current maps detail the position of up to one-half of all human protein-coding genes. The gene map is updated as new STSs are mapped. GeneMap '98 is described in more detail in Chapter 6.

Information about the map location of individual ESTs is provided by UniGene. If an STS exists for an EST in the cluster, the map position of that STS is indicated in the record. This data may confirm what is already known about the location of a

mapped human disease gene, but, for the majority of cases involving yet unmapped ESTs, GeneMap '98 provides novel information. Because the STS markers correspond to individual genes, the project also shows the density of genes on each human chromosome. For example, chromosomes 1, 17, and 22 have a higher than expected gene density, and chromosomes 4, 13, 18, and X have a lower than expected density.

GENE PREDICTION IN GENOMIC DNA

Another use of ESTs is to predict or refine computational predictions of the location of genes in genomic DNA. With the appropriate use of sequence alignment parameters, up to 90% of genes annotated on human genomic DNA are also detected by ESTs (Bailey et al., 1998). ESTs can complement other algorithms used for gene prediction because they may do a better job at predicting alternative splicing and 3′ untranslated regions. A study is underway to reannotate the *C. elegans* genome using EST sequences (Kohara, Y., unpublished observations). With the use of the acembly program, 126,000 ESTs were aligned to 98 Mb of genomic DNA. The genes predicted by EST clones were compared with those predicted by the *C. elegans* genomic sequencing consortium, which were constructed using GeneFinder with hand editing (The *C. elegans* Sequencing Consortium, 1998). The following points are noteworthy:

1. In about half the cases, the computationally predicted genes were identical to the EST alignments; 25% of the genes were predicted with less accuracy, and the remaining 25% were predicted poorly. In some cases, 5′ sequences from ESTs showed that the gene predictions were either too long or too short.
2. Comparisons of the EST sequences to the genomic sequence confirm that the error rate of the worm genome sequence is less than 1 mistake per 10,000 nucleotides. Instances where many ESTs share identical sequence may indicate errors in genomic sequence. Alternatively, these differences could be sequence polymorphisms.
3. About 30% of the ESTs exist in alternatively-spliced forms. Many of these alternative splices are not annotated on the genomic sequence.
4. Computational methods may predict separate genes, whereas EST analysis shows that these segments are actually exons of a single gene. Conversely, the computational method may predict exons in cases that should be separate genes.

ESTs have also been used in genome sequencing projects to make estimates about gene expression along the chromosome. In the complete sequences of *Arabidopsis thaliana* chromosomes 2 (Lin et al., 1999) and 4 (Mayer et al., 1999), about one-third of the computationally predicted genes have an EST match. Histograms plotting the EST distribution along the chromosomes predict that some genes are highly expressed, at least within the tissues from which EST libraries were constructed. On chromosome 4, 75% of the matching ESTs aligned with only 6% of the genes, implying that these genes are transcribed at high rates.

ESTs AND SEQUENCE POLYMORPHISMS

Single nucleotide polymorphisms (SNPs) can help to associate sequence variations with heritable phenotypes, facilitate studies in population and evolutionary biology, and aid in positional cloning and physical mapping. On average, SNPs occur every 500–1,000 nucleotides in human DNA. Gene-associated SNPs are found in untranslated regions as well as coding sequences (cSNPs). Because ESTs are sequenced redundantly from libraries prepared from different individuals, they seem an ideal source of polymorphic data. Indeed, a number of recent studies demonstrate that analysis of aligned EST sequences can lead to SNP discovery (Buetow et al., 1999; Garg et al., 1999; Marth et al., 1999; Picoult-Newberg et al., 1999). All rely on alignment of EST sequences, identification of sequence differences, and a method to distinguish real polymorphisms from base calling (sequencing) errors and other artifacts. The public database for SNPs, dbSNP, is maintained at the NCBI (Sherry et al., 1999; Smigielski et al., 2000). dbSNP accepts submissions not only of single nucleotide polymorphisms but of other polymorphisms, such as short deletions and inserts, microsatellites, and polymorphic insertion elements like retrotransposons. As of mid-July 2000, dbSNP contains data from 600,000 SNPs. dbSNP is integrated with other NCBI resources such as GenBank, PubMed, genome sequences, and LocusLink.

ASSESSING LEVELS OF GENE EXPRESSION USING ESTs

Because ESTs are generated by random sequencing of clones from many different libraries, they appear, at least at first glance, to be a good source of data source for studies of gene expression levels. However, any conclusions about transcript levels must be made very carefully. Many libraries are normalized or generated by subtractive hybridization. Both of these processes change the relative representation of cDNAs. Normalization results in abundant messages being seen less frequently and rare messages more frequently, whereas subtractive hybridization removes entire pools of transcripts from the library. Although libraries made by these processes can provide very general ideas about which genes are expressed at higher levels, detailed analysis is not possible.

CGAP

A subset of the EST libraries was constructed for the purpose of gene expression profiling, and these libraries were not normalized or created by subtractive hybridization. Many of these libraries were constructed by the Cancer Genome Anatomy Project (CGAP), an NCI initiative that is working to decipher the molecular anatomy of the cancer cell (Wheeler et al., 2000). CGAP has developed libraries from normal, precancerous, and cancerous cell types. Comparing the genes expressed in these three tissue types can lead to predictions about the genes involved in cancer progression. ESTs from the CGAP project are submitted to dbEST and are available in UniGene. CGAP has developed online tools to compare computationally gene expression levels between libraries. Digital Differential Display (DDD) uses a statistical test to calculate the number of times sequences from different libraries are assigned to a par-

ticular UniGene cluster (Krizman et al., 1999). By selecting pools of libraries, users can compare gene expression levels between tissues (e.g., liver, lung, muscle, and spleen) or cancer stages (e.g., normal vs. premalignant vs. cancerous prostate tissue). The DDD results show the genes that are expressed at different levels in the selected pools (i.e., the UniGene clusters in which the number of ESTs from one set of libraries is significantly different from the number of ESTs from another set of libraries). The results are presented as an easily interpreted graphic similar to a Northern blot and also as text. A detailed explanation of how to perform a DDD experiment, including a worked example, is provided on the CGAP Web site. The CGAP xProfiler compares gene expression levels between two pools of libraries by listing the genes that are expressed either in both library pools or in one pool but not the other. The calculations are also based on the tissue distribution of ESTs in UniGene clusters.

SAGE

Serial analysis of gene expression (SAGE) is an experimental technique used for quantitative, high throughput gene expression analysis (Velculescu et al., 1995). SAGE involves the isolation of short unique sequence tags from a specific location within each transcript. These sequence tags are concatenated, cloned, and sequenced. The frequency of particular transcripts within the starting cell population is reflected by the number of times the associated sequence tag is encountered within the sequence population. The SAGEmap database is a repository for some of this SAGE data, and tools that allow gene expression analysis are also available on the SAGEmap Web site (Lal et al., 1999). The virtual Northern predicts the SAGE tag in a user-supplied mRNA sequence and calculates the distribution of the tag in the SAGE libraries, thus providing a virtual picture of the expression pattern of the mRNA. In the SAGE xProfiler, the user selects two pools of libraries, such as colon cancer versus normal colon. The tool calculates which SAGE tags are more abundant in one pool or the other. The SAGE tags are mapped to UniGene clusters to provide biological context for the results.

Microarrays

High-density oligonucleotide and cDNA microarrays are a relatively new technique being used to monitor gene expression on a genome-wide scale. The technique uses the same principles of nucleic acid hybridization as do Northern and Southern blots but on a much larger scale. Thousands of gene-specific probes are arrayed on a small matrix, such as a glass slide or microchip, and this matrix is probed with labeled nucleic acid synthesized from a tissue type, developmental stage, or other condition of interest. The expression profiles of thousands of genes under that condition can thus be assayed simultaneously. The array probes can be derived from oligonucleotides or cDNAs. In many cases, the probes for cDNA arrays are 3' ESTs (Duggan et al., 1999). For human expression analysis, UniGene clusters can be used as a source of additional information about the ESTs on the array. Microarray technologies and the bioinformatics challenges surrounding them are discussed in Chapter 16.

INTERNET RESOURCES FOR TOPICS PRESENTED IN CHAPTER 12

dbEST home page	*http://www.ncbi.nlm.nih.gov/dbEST/*
List of dbEST libraries	*http://www.ncbi.nlm.nih.gov/dbEST/libs_byorg.html*
dbEST summary by organism	*http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html*
How to submit ESTs to dbEST	*http://www.ncbi.nlm.nih.gov/dbEST/how_to_submit.html*
EST Projects at Washington University	*http://genome.wustl.edu/gsc/est/navest.pl*
The I.M.A.G.E. Consortium	*http://image.llnl.gov/*
UniGene	*http://www.ncbi.nlm.nih.gov/UniGene/*
The UniGene build procedure	*http://www.ncbi.nlm.nih.gov/UniGene/build.html*
UniGene query engine	*http://www.ncbi.nlm.nih.gov/UniGene/query.cgi*
HomoloGene	*http://www.ncbi.nlm.nih.gov/HomoloGene/*
STACK	*http://www.sanbi.ac.za/Dbases.html*
TIGR Gene Indices	*http://www.tigr.org/tdb/tgi.html*
TIGR Orthologous Gene Alignment database	*http://www.tigr.org/tdb/toga/toga.html*
GeneMap '98	*http://www.ncbi.nlm.nih.gov/genemap/*
dbSNP	*http://www.ncbi.nlm.nih.gov/SNP/*
Cancer Genome Anatomy Project (CGAP)	*http://www.ncbi.nlm.nih.gov/ncicgap/*
CGAP Digital Differential Display (DDD)	*http://www.ncbi.nlm.nih.gov/CGAP/info/ddd.cgi*
CGAP xProfiler	*http://www.ncbi.nlm.nih.gov/CGAP/hTGI/xprof/cgapxpsetup.cgi*
Serial Analysis of Gene Expression (SAGE)	*http://www.ncbi.nlm.nih.gov/SAGE/*
SAGE virtual Northern	*http://www.ncbi.nlm.nih.gov/SAGE/sagevn.cgi*
SAGE xProfiler	*http://www.ncbi.nlm.nih.gov/SAGE/sagexpsetup.cgi*

PROBLEM SET

You have been studying the histone deacetylase gene, RPD3, in the yeast *Saccharomyces cerevisiae*. You are moving to a lab that works on zebrafish, and you would like to continue your work on this gene. You wonder how difficult it will it be to clone the zebrafish ortholog of RPD3.

1. What is the GenBank accession number of the first listed RPD3 protein sequence from *Saccharomyces cerevisiae*?

2. Do the public sequence databases already contain any zebrafish proteins that are likely orthologs of RPD3?

a. What type of sequence comparison search should you perform?
b. To interpret the search results of your sequence comparison, you will need to know the scientific name for zebrafish. What is the scientific name?
c. Are there any zebrafish protein orthologs of yeast RPD3?

3. You remember that the EST database is an excellent source of sequence data.
a. What type of sequence comparison should you perform to find EST hits to the yeast protein sequence?
b. Are there any zebrafish EST hits to this yeast protein sequence?

4. Do the five top scoring ESTs belong to the same UniGene cluster?

5. What is the GenBank accession number of the human sequence that matches this UniGene cluster?

6. What cDNA clone does the top scoring EST hit come from?

7. Is this EST from the or 3′ end of the cDNA clone?

8. From which cDNA library was this clone sequenced?

9. Is the EST that comes from the opposite end of this cDNA clone also a member of this UniGene cluster?

10. Does this EST also align with the yeast RPD3 protein sequence? Why or why not?

11. Is the top-scoring zebrafish EST also present in the TIGR Zebrafish Gene Index?

12. Is the EST that comes from the opposite end of the cDNA clone also in this TIGR TC?

13. Are the sequences in the UniGene cluster and the TIGR TC basically the same?

14. How does the TIGR consensus sequence for the 5′ EST TC compare with that produced by UniGene?

15. Is the top-scoring zebrafish EST hit to the yeast RPD3 protein present in STACK?

16. Based on what you have learned, how would you get a cDNA clone of the zebrafish RPD3 gene?

REFERENCES

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., Kerlavage, A. R., McCombie, W. R., and Venter, J. C. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252: 1651–1656.

Ajioka, J. W., Boothroyd, J. C., Brunk, B. P., Hehl, A., Hillier, L., Manger, I. D., Marra, M., Overton, G. C., Roos, D. S., Wan, K. L., Waterston, R., and Sibley, L. D. (1998). Gene discovery by EST sequencing in *Toxoplasma gondii* reveals sequences restricted to the Apicomplexa. *Genome Res* 8: 18–28.

Bailey, L. C., Jr., Searls, D. B., and Overton, G. C. (1998). Analysis of EST-driven gene annotation in human genomic sequence. *Genome Res* 8: 362–376.

Boguski, M. S., Lowe, T. M., and Tolstoshev, C. M. (1993). dbEST—database for ''expressed sequence tags.'' *Nat Genet* 4: 332–333.

Boguski, M. S., and Schuler, G. D. (1995). ESTablishing a human transcript map. *Nat Genet* 10: 369–371.

Boguski, M. S., Tolstoshev, C. M., and Bassett, D. E., Jr. (1994). Gene discovery in dbEST. *Science* 265: 1993–1994.

Bonaldo, M. F., Lennon, G., and Soares, M. B. (1996). Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res* 6: 791–806.

Buetow, K. H., Edmonson, M. N., and Cassidy, A. B. (1999). Reliable identification of large numbers of candidate SNPs from public EST data. *Nat Genet* 21: 323–325.

Burke, J., Davison, D., and Hide, W. (1999). d2_cluster: A Validated Method for Clustering EST and Full-Length cDNA Sequences. *Genome Res* 9: 1135–1142.

Collins, F. S., Patrinos, A., Jordan, E., Chakravarti, A., Gesteland, R., and Walters, L. (1998). New goals for the U.S. Human Genome Project: 1998–2003. *Science* 282: 682–689.

Deloukas, P., Schuler, G. D., Gyapay, G., Beasley, E. M., Soderlund, C., Rodriguez-Tome, P., Hui, L., Matise, T. C., McKusick, K. B., Beckmann, J. S., Bentolila, S., Bihoreau, M., Birren, B. B., Browne, J., Butler, A., Castle, A. B., Chiannilkulchai, N., Clee, C., Day, P. J., Dehejia, A., Dibling, T., Drouot, N., Duprat, S., Fizames, C., Bentley, D. R., and et al. (1998). A physical map of 30,000 human genes. *Science* 282: 744–746.

Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P., and Trent, J. M. (1999). Expression profiling using cDNA microarrays. *Nat Genet* 21: 10–14.

Garg, K., Green, P., and Nickerson, D. A. (1999). Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. *Genome Res* 9: 1087–1092.

Green, P. (1996). http://www.genome.washington.edu/uwgc/analysistools/phrap.htm.

Hillier, L. D., Lennon, G., Becker, M., Bonaldo, M. F., Chiapelli, B., Chissoe, S., Dietrich, N., DuBuque, T., Favello, A., Gish, W., Hawkins, M., Hultman, M., Kucaba, T., Lacy, M., Le, M., Le, N., Mardis, E., Moore, B., Morris, M., Parsons, J., Prange, C., Rifkin, L., Rohlfing, T., Schellenberg, K., Marra, M., and et al. (1996). Generation and analysis of 280,000 human expressed sequence tags. *Genome Res* 6: 807–828.

Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Res* 9: 868–877.

Krizman, D. B., Wagner, L., Lash, A., Strausberg, R. L., and Emmert-Buck, M. R. (1999). The Cancer Genome Anatomy Project: EST Sequencing and the Genetics of Cancer Progression. *Neoplasia* 1: 101–106.

Lal, A., Lash, A. E., Altschul, S. F., Velculescu, V., Zhang, L., McLendon, R. E., Marra, M. A., Prange, C., Morin, P. J., Polyak, K., Papadopoulos, N., Vogelstein, B., Kinzler, K. W., Strausberg, R. L., and Riggins, G. J. (1999). A public database for gene expression in human cancers. *Cancer Res* 59: 5403–5407.

Lennon, G., Auffray, C., Polymeropoulos, M., and Soares, M. B. (1996). The I.M.A.G.E. Consortium: an integrated molecular analysis of genomes and their expression. *Genomics* 33: 151–152.

Lin, X., Kaul, S., Rounsley, S., Shea, T. P., Benito, M. I., Town, C. D., Fujii, C. Y., Mason, T., Bowman, C. L., Barnstead, M., Feldblyum, T. V., Buell, C. R., Ketchum, K. A., Lee, J., Ronning, C. M., Koo, H. L., Moffat, K. S., Cronin, L. A., Shen, M., Pai, G., Van Aken, S., Umayam, L., Tallon, L. J., Gill, J. E., and Venter, J. C. (1999). Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana. Nature* 402: 761–768.

Marth, G. T., Korf, I., Yandell, M. D., Yeh, R. T., Gu, Z., Zakeri, H., Stitziel, N. O., Hillier, L., Kwok, P. Y., and Gish, W. R. (1999). A general approach to single-nucleotide polymorphism discovery. *Nat Genet* 23: 452–456.

Mayer, K., Schuller, C., Wambutt, R., Murphy, G., Volckaert, G., Pohl, T., Dusterhoft, A., Stiekema, W., Entian, K. D., Terryn, N., Harris, B., Ansorge, W., Brandt, P., Grivell, L., Rieger, M., Weichselgartner, M., de Simone, V., Obermaier, B., Mache, R., Muller, M., Kreis, M., Delseny, M., Puigdomenech, P., Watson, M., and McCombie, W. R. (1999). Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana. Nature* 402: 769–777.

Miller, R. T., Christoffels, A. G., Gopalakrishnan, C., Burke, J., Ptitsyn, A. A., Broveak, T. R., and Hide, W. A. (1999). A comprehensive approach to clustering of expressed human gene sequence: The sequence tag alignment and consensus knowledge base. *Genome Res* 9: 1143–1155.

Mironov, A. A., Fickett, J. W., and Gelfand, M. S. (1999). Frequent alternative splicing of human genes. *Genome Res* 9: 1288–1293.

Picoult-Newberg, L., Ideker, T. E., Pohl, M. G., Taylor, S. L., Donaldson, M. A., Nickerson, D. A., and Boyce-Jacino, M. (1999). Mining SNPs from EST databases. *Genome Res* 9: 167–174.

Pruitt, K. D., Katz, K. S., Sicotte, H., and Maglott, D. R. (2000). Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet* 16: 44–47.

Quackenbush, J., Liang, F., Holt, I., Pertea, G., and Upton, J. (2000). The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res* 28: 141–145.

Schuler, G. D., Boguski, M. S., Stewart, E. A., Stein, L. D., Gyapay, G., Rice, K., White, R. E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., Bentolila, S., Birren, B. B., Butler, A., Castle, A. B., Chiannilkulchai, N., Chu, A., Clee, C., Cowles, S., Day, P. J., Dibling, T., Drouot, N., Dunham, I., Duprat, S., East, C., Hudson, T. J., and et al. (1996). A gene map of the human genome. *Science* 274: 540–546.

Sherry, S. T., Ward, M., and Sirotkin, K. (1999). dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res* 9: 677–679.

Smigielski, E. M., Sirotkin, K., Ward, M., and Sherry, S. T. (2000). dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* 28:352–355.

Soares, M. B., Bonaldo, M. F., Jelene, P., Su, L., Lawton, L., and Efstratiadis, A. (1994). Construction and characterization of a normalized cDNA library. *Proc Natl Acad Sci USA* 91: 9228–9232.

The *C. elegans* Sequencing Consortium. (1998). Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science* 282: 2012–2018.

Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. (1995). Serial analysis of gene expression. *Science* 270: 484–487.

Wheeler, D. L., Chappey, C., Lash, A. E., Leipe, D. D., Madden, T. L., Schuler, G. D., Tatusova, T. A., and Rapp, B. A. (2000). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 28: 10–14.

Wolfsberg, T. G., and Landsman, D. (1997). A comparison of expressed sequence tags (ESTs) to human genomic sequences. *Nucleic Acids Res* 25: 1626–1632.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. *J. Comput. Biol.* 7: 203–214.