Basic Protein Sequence Analysis

Protein sequence analysis can be performed from many perspectives and with a vast array of bioinformatics methods. This unit will focus on those analyses that can be performed using only a computer (desktop workstation or laptop), Internet access, and a laboratory notebook to record results. All the resources needed are accessible via publicly available Web servers and databases, and require little or no computational expertise. These tools enable biologists to predict the structure of a protein, the presence of functional motifs or domains, cellular localization (e.g., membrane-bound, secreted, chloroplast, or nuclear), and post-translational modifications. Integration of all these data enables the biologist to make a more informed prediction of the likely molecular functions of proteins of interest.

A plethora of tools are available for these tasks, and in choosing preferred methods for each of the separate protocols, the authors have been guided primarily by two criteria: user-friendliness and utility. Some tools are extremely powerful, but inappropriate for the computational novice. Other tools may be easy to use, but relatively ineffectual. We have kept working biologists in mind, and asked ourselves which tools our biologist colleagues would find easiest to use and understand. With these criteria in mind, we selected one or two protocols for each task. Identifying structural domains using different types of servers are presented (Basic Protocol 1 and Alternate Protocols 1 and 2), as are procedures for predicting transmembrane regions and subcellular localization (Basic Protocol 2 and Alternate Protocol 3) and key functional residues and motifs (Basic Protocol 3). Each procedure also includes a brief discussion of some of the Web servers and databases that we feel deserve attention. Support Protocols for each method present a strategy for the integration of the results from the different procedures (Support Protocols 1, 2, 3, 4, 5, 6, and 7).

IDENTIFYING STRUCTURAL AND FUNCTIONAL DOMAINS USING INTEGRATED META-SERVERS

BASIC PROTOCOL 1

Below is a protocol for the use of integrated meta-servers. Meta-servers include prediction results from several independent prediction tools to produce a consensus prediction. This helps a biologist avoid false positive predictions. Our primary protocol for this task involves the use of the SMART (Simple Modular Architecture Research Tool) server provided by the European Molecular Biology Laboratory (EMBL) at Heidelberg (Schultz et al., 1998; Letunic et al., 2004). This server integrates structure and function prediction tools with pre-calculated analyses (including domain structure prediction, intron/exon boundary detection and ortholog identification) for proteins in the Swiss-Prot and TrEMBL databases (Boeckmann et al., 2003) and all proteomes analyzed by the Ensembl group (Hubbard et al., 2002). The extensive results and appealing interface make a biologist's work much easier, and this is our reason for recommending the use of this server.

Materials

Input data

The sequence to be analyzed should be in FASTA format, or may alternatively be submitted as raw sequence (the single-letter representation of the protein sequence without the definition line). An example FASTA format is shown in Figure 19.5.1. In cases where the sequence of interest is in the Swiss-Prot or

Contributed by Nandini Krishnamurthy and Kimmen V. Sjölander
Current Protocols in Molecular Biology (2005) 19.5.1-19.5.24

```
>sp|P27037|AVR2_HUMAN Activin receptor type II precursor (ACTR-II)
MGAAAKLAFAVFLISCSSGAILGRSETQECLFFNANWEKDRTNQTGVEPCYGDKDKRRHCFATWKNISGS
IEIVKQGCWLDDINCYDRTDCVEKKDSPEVYFCCCEGNMCNEKFSYFPEMEVTQPTSNPVTPKPPYYNIL
LYSLVPLMLIAGIVICAFWVYRHHKMAYPPVLVPTQDPGPPPPSPLLGLKPLQLLEVKARGRFGCVWKAQ
LLNEYVAVKIFPIQDKQSWQNEYEVYSLPGMKHENILQFIGAEKRGTSVDVDLWLITAFHEKGSLSDFLK
ANVVSWNELCHIAETMARGLAYLHEDIPGLKDGHKPAISHRDIKSKNVLLKNNLTACIADFGLALKFEAG
KSAGDTHGQVGTRRYMAPEVLEGAINFQRDAFLRIDMYAMGLVLWELASRCTAADGPVDEYMLPFEEEIG
QHPSLEDMQEVVVHKKKRPVLRDYWQKHAGMAMLCETIEECWDHDAEARLSAGCVGERITQMQRLTNIIT
TEDIVTVVTMVTNVDFPPKESSL
```

Figure 19.5.1 FASTA format of AVR2_HUMAN protein sequence. The Swiss-Prot accession number is P27037 and the Swiss-Prot ID is AVR2_HUMAN.

TrEMBL databases, the accession number is normally sufficient to retrieve precalculated results. Otherwise, the protein sequence is required as described in the protocol below.

1. Point the browser at the SMART Web site (*http://smart.embl-heidelberg.de/*). On the SMART main page that appears, click either SMART MODE: NORMAL or SMART MODE: GENOMIC. In NORMAL SMART, the database contains Swiss-Prot, SP-TrEMBL and stable Ensembl proteomes. In GENOMIC SMART, only the proteomes of completely sequenced genomes are used. We recommend the use of NORMAL SMART. In the page that now appears, paste the sequence of interest in FASTA format in the box provided. If the sequence is from Swiss-Prot, TREMBL, or the EMBL database, one could alternatively submit the Accession number or Entry name in the box marked Sequence ID or ACC (Fig. 19.5.2).

2. Select additional analyses. At this point, one can either click the Sequence SMART button or select additional analyses. In this example (shown in Fig. 19.5.3), we have selected all available analyses.

   a. Outlier homologs.

      This option will search sequence databases derived from proteins of solved structures and return the ID and E-value (see Table 19.5.1 for definition) of the top hit and the database searched. The search is done on PDB (Protein Data Bank, a database of solved structures) and position-specific iterated BLAST (PSI-BLAST) profiles derived from SCOP (Structural Classification Of Proteins; Table 19.5.1), a system of classifying proteins based on structural and functional similarity (Murzin et al., 1995).

   b. PFAM domains.

      This option searches for the presence of PFAM domains (Bateman et al., 2002) in the query sequence. Since SMART (as of November, 2004) has 667 HMMs, and PFAM (August, 2004) has 7503 hidden Marker models (HMMs), searching PFAM HMMs increases the likelihood of a match.

   c. Signal peptides.

      This gives information on whether the query sequence contains a signal peptide, using SigCleave program.

   d. Internal repeats.

      This option shows any repeated motifs found in the query.

3. Select the Sequence SMART button. This will retrieve precalculated results (when available) or search the sequence against the library of HMMs available.

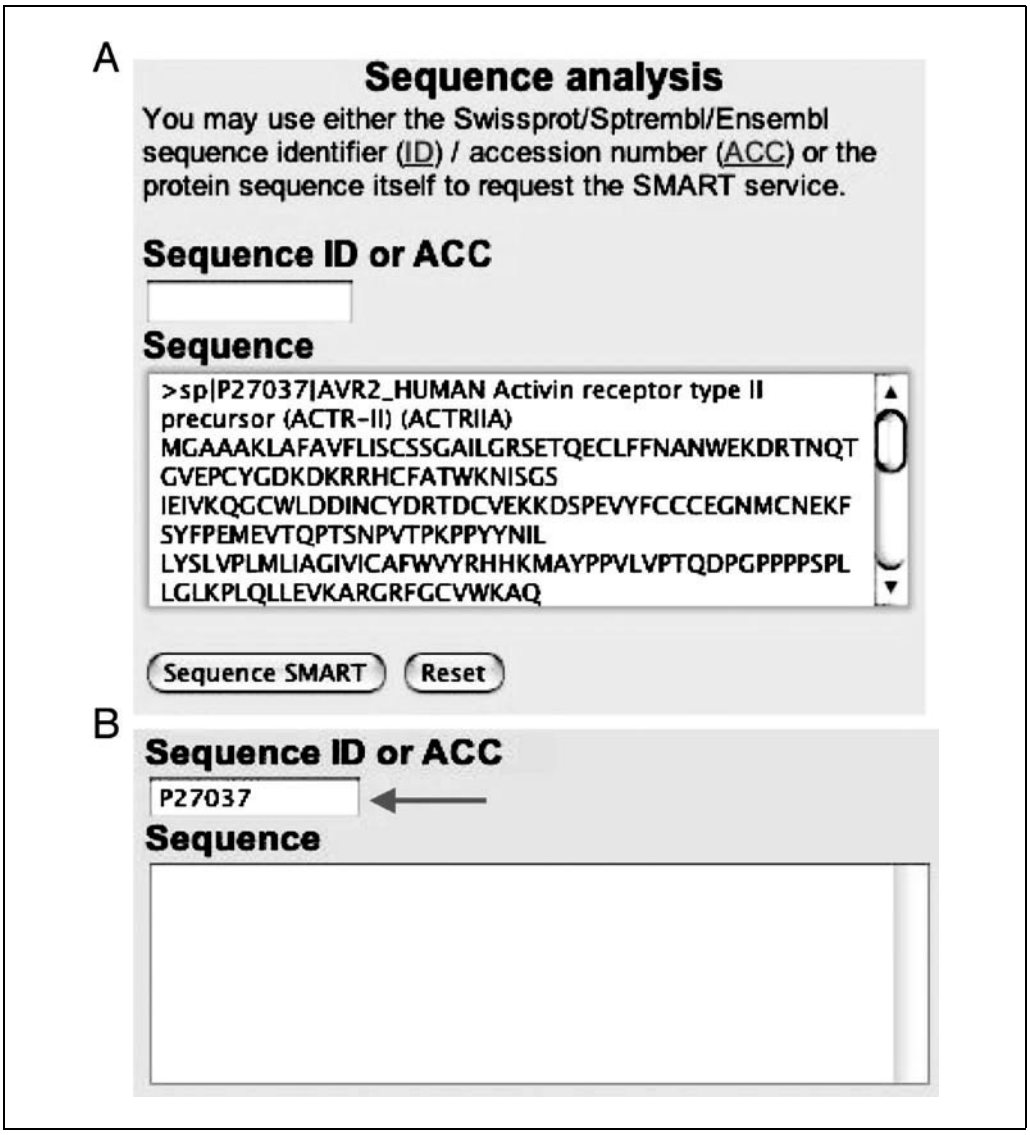


Figure 19.5.2 SMART sequence submission form. (**A**) Submission of sequence in FASTA format. (**B**) Submission using the Swiss-Prot accession number in the box marked with arrow.

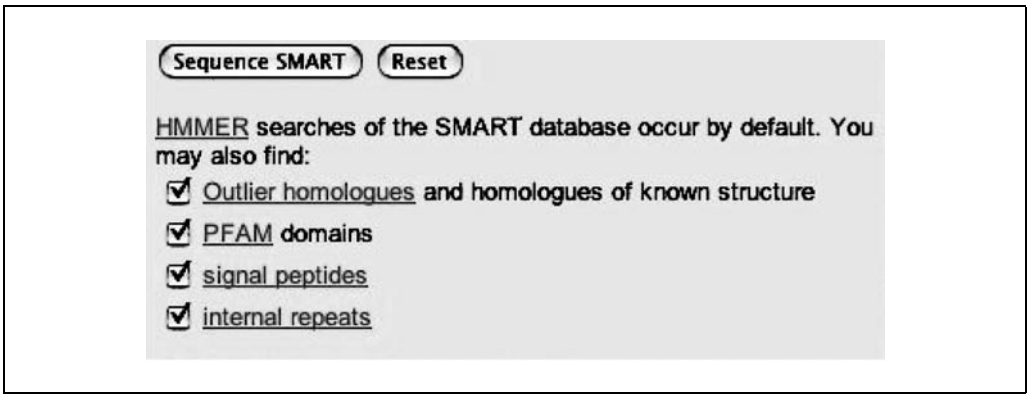


Figure 19.5.3 Sequence submission options in SMART.

Table 19.5.1 Definitions of Acronyms and Abbreviations Used in this Unit

Acronym/ abbreviation	Full name	Description
3D-PSSM	3-D Position Specific Scoring Matrix	The 3D-PSSM is a protein fold recognition server that uses 1-D and 3-D sequence profiles coupled with secondary structure and solvation potential information
BLAST	Basic Local Alignment Search Tool	BLAST is a sequence comparison algorithm optimized for speed, used to search sequence databases for optimal local alignments to a query. Variants of BLAST enable searches of protein databases using nucleotide queries and vice versa.
CDART	Conserved Domain Architecture Retrieval Tool	For a given protein query sequence, CDART displays the predicted functional/structural domains that make up the protein and lists proteins with similar domain architectures
EMBL	European Molecular Biology Laboratory	EMBL is an international research organization with its main laboratory in Heidelberg, Germany, and four outstations in Hinxton, U.K. (the European Bioinformatics Institute, EBI), Grenoble, France, Hamburg, Germany, and Monterotondo, Italy.
E-value	Expectation value	For a given score S and a database size, the E-value is the number of hits with scores equivalent to or better than S that are expected to occur in a database of that size by chance. The lower the E value, the more significant the score.
HMM	Hidden Markov Model	HMMs for proteins are statistical models (akin to profiles and PSSMs) that represent the amino acid preferences and insertion/deletion likelihoods at each position. They are used to detect remote homologs and to generate alignments of new sequences to a family.
MSA	Multiple Sequence Alignment	A multiple sequence alignment refers to an alignment of more than two sequences; pairwise alignment refers to an alignment of two sequences. Gap characters (typically, dots or dashes) are inserted between amino acids so that all sequences have the same length, and can be arranged in a matrix to display the amino acids conserved across the family. MSAs are used as input to construction of PSSMs and HMMs, to estimate phylogenetic trees, and for numerous other tasks.
NCBI	National Center for Biotechnology Information	The NCBI is the main resource for computational biologists in the U.S., and includes a number of important tools (e.g., BLAST) and databases. Databases include the CDD (Conserved Domain Database), GenBank (a comprehensive sequence database), PubMed (for retrieving papers in biomedical literature), OMIM (Online Mendelian Inheritance in Man), and a host of other important resources.

continued

Table 19.5.1 Definitions of Acronyms and Abbreviations Used in this Unit, *continued*

Acronym/ abbreviation	Full name	Description
PFAM	Protein Families	PFAM is a large collection of multiple sequence alignments and hidden Markov models representing many protein domains and families. It has a web interface for identification of domains in a given protein, and includes hyperlinks to relevant literature and functional/structural information for many domains.
PSI-BLAST	Position-Specific Iterated BLAST	An iterative search using the BLAST algorithm and profile construction, enabling identification of remote homologs
PSIPRED		PSIPRED is a simple and reliable secondary-structure prediction method, incorporating two feed-forward neural networks, which perform an analysis on output obtained from PSI-BLAST to predict whether a position is likely to be found in an α helix, β strand, or coil.
PSSM	Position-specific scoring matrix	A PSSM is a statistical representation of the amino acids in a multiple sequence alignment, and is used to provide a log-odds score for each amino acid at each position in the alignment. PSSMs, HMMs, and profiles are related.
RasMol	Raster Display of Molecules.	Molecular visualization freeware to visualize 3-D structure of proteins
RPSBLAST	Reverse PSI-BLAST	RPS-BLAST (Reverse PSI-BLAST) searches a query sequence against a database of profiles (i.e., the reverse of PSI-BLAST that searches a profile against a database of sequences).
SAWTED	Structure Assignment With Text Description	SAWTED compares annotations of the query and closely related sequences with annotations associated with database hits having weak scores. This is included to improve the ability to differentiate between true homologs and spurious matches (i.e., SAWTED enables sequences with weak scores but similar annotations to be given more credence in homology search).
SCOP	Structure Classification of Proteins	The SCOP database classifies protein structural domains into a hierarchy (class, fold, superfamily and family), based on functional and structural data. Proteins in the same SCOP superfamily are expected to be related by divergent evolution from a common ancestral protein and to share a similar (albeit potentially high-level) function. Proteins in different SCOP folds are believed to not be homologous.
SMART	Simple Modular Architecture Research Tool	The SMART Web server allows rapid identification of the domain architecture of a given protein, and includes precomputed results for many sequences

Figure 19.5.4 SMART results for AVR2_HUMAN. Integrated results from several servers show an N-terminal signal peptide, two PFAM domains, a transmembrane domain, and a low-complexity region. This black and white facsimile of the figure is intended only as a placeholder; for full-color version of figure go to *http://www.interscience.wiley.com/c_p/colorfigures.htm*.

4. Examine predicted domains and motifs. The features represented in the example (Fig. 19.5.4) are:
   a. Signal peptide: the small red box at the N-terminus.
   b. Two PFAM domains: Activin receptor and Pkinase.
   c. Low-complexity region: pink box at C-terminus.
   d. Transmembrane domain: Blue vertical box.
5. Place the mouse cursor over predicted domains and unidentified regions to see their start and end amino acid positions.
6. Click on the domains to get a detailed description of the domain.
7. Scroll down to see the E-value of the predicted matches.
8. Analyze results (see Support Protocol 1).

SUPPORT PROTOCOL 1

GUIDELINES FOR UNDERSTANDING RESULTS OF ANALYSES FROM INTEGRATED META-SERVERS

The domains in SMART (see Basic Protocol 1) are represented by seed alignments (alignments used to build the HMMs). Homologs are gathered by different iterative methods and included in a multiple sequence alignment (MSA) after estimating the statistical significance of sequence similarities and examining experimental biological information, if available. To reduce redundancy in a family of sequences representing a domain, a phylogenetic tree is constructed using the MSA of all the sequences in the family, and, from every branch that is less than a distance of 0.2 on the tree, only a single sequence is chosen. This corresponds to about 80% sequence identity. These sequences are then used in the seed alignment from which the profiles and HMMs representing a domain are constructed (Schulz et al., 1998).

Interpreting the Significance of a Match to a Profile

Most similarity search methods employ E-values (Expect values) to report the statistical significance of a match. How an E-value is calculated varies with each method. Generally, the E-value represents the number of matches expected to be found merely by chance. For instance if an E-value is 10, then 10 matches are expected to be found by chance alone. In general, an E-value <0.001 is a good indicator of the credibility of predicted homology, and the likelihood of a spurious result increases as the E-value approaches 1. A significant E-value implies similarity in structure; the functional specificity may have changed. To confirm functional similarity, it is critical to check for agreement at key functional residues (particularly active site residues in the case of an enzyme), and for a high percent identity (as a conservative rule of thumb, we recommend requiring

a minimum of 40%). If the profile or HMM represents a structural domain, the query should align over a significant fraction of the profile/HMM. As a basic rule of thumb, we recommend requiring a minimum of 75% coverage.

Note that the SMART output does not provide this information (i.e., percent identity, coverage of the profile and agreement with key positions); these criteria will need to be confirmed separately.

Factors Affecting E-Value

E-values are affected by the length of the profile or HMM, with longer profile/HMMs producing more significant E-values for related sequences and shorter profile/HMMs producing weaker E-values. Weak E-values are particularly common when the profile or HMM is based on very short motifs (e.g., the PFAM Leucine Rich Repeat HMMs, having between 20 and 25 residues). Because a near-exact match to a short motif can occur by chance alone, E-values will be correspondingly high (i.e., approaching or exceeding 1) even for closely related sequences. For this reason, many HMM and profile libraries employ empirically determined score cutoffs instead of E-values (see, e.g., the PFAM gathering threshold); these tend to be more relaxed for shorter profile/HMMs than for longer.

Another factor that affects E-values is the size of the database searched. The probability of finding a match by chance increases with the size of the database searched. The upshot of this is that a match between a given query and a hit will have different E-values depending on the database searched. For example, if the E-value of a match between query X and hit Y is 0.01 on searching the National Center for Biotechnology Information (NCBI) nonredundant database (2,182,528 sequences as of November, 2004), then it will be smaller (e.g., 0.001), and hence more significant, on searching Swiss-Prot (159,078 sequences as of November, 2004) using the same search method. The database size becomes an issue especially when E-values are not extremely significant.

IDENTIFYING STRUCTURAL AND FUNCTIONAL DOMAINS USING THE NCBI CD-SEARCH

ALTERNATE PROTOCOL 1

The Conserved Domain Database (CDD; Marchler-Bauer et al., 2003) at NCBI is a database of Position Specific Scoring Matrices (PSSMs) generated from alignments imported from PFAM and SMART and automated alignments from COGs (Cluster of Orthologous Groups; Tatusov, 2001) CD-Search is a tool that uses RPS-BLAST (Reverse PSI-BLAST) to search a sequence against a database of profiles, namely the CDD (Marchler-Bauer et al., 2004). This is the reverse of PSI-BLAST, which searches a profile against a database of sequences.

Materials

See Basic Protocol 1.

1. Point the browser at the CDD (*http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml*).
2. Input a sequence in FASTA format (or, alternatively, the NCBI accession or GenBank accession number) in the box provided (Fig. 19.5.5).
3. Set the Search Database from the pull-down menu of that name. The default option is CDD, but users can also search SMART, PFAM, and COGs.
4. Set other parameters using the “click here for advanced options” link, if desired.
5. Click Submit.

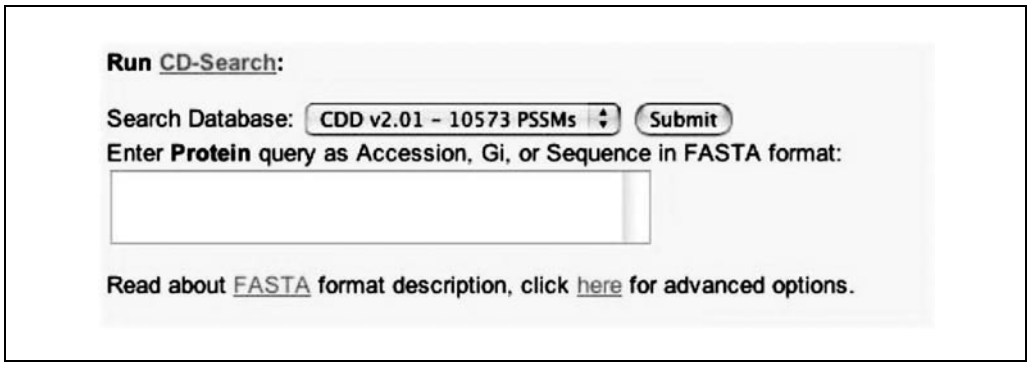


Figure 19.5.5 CD-Search sequence submission form.

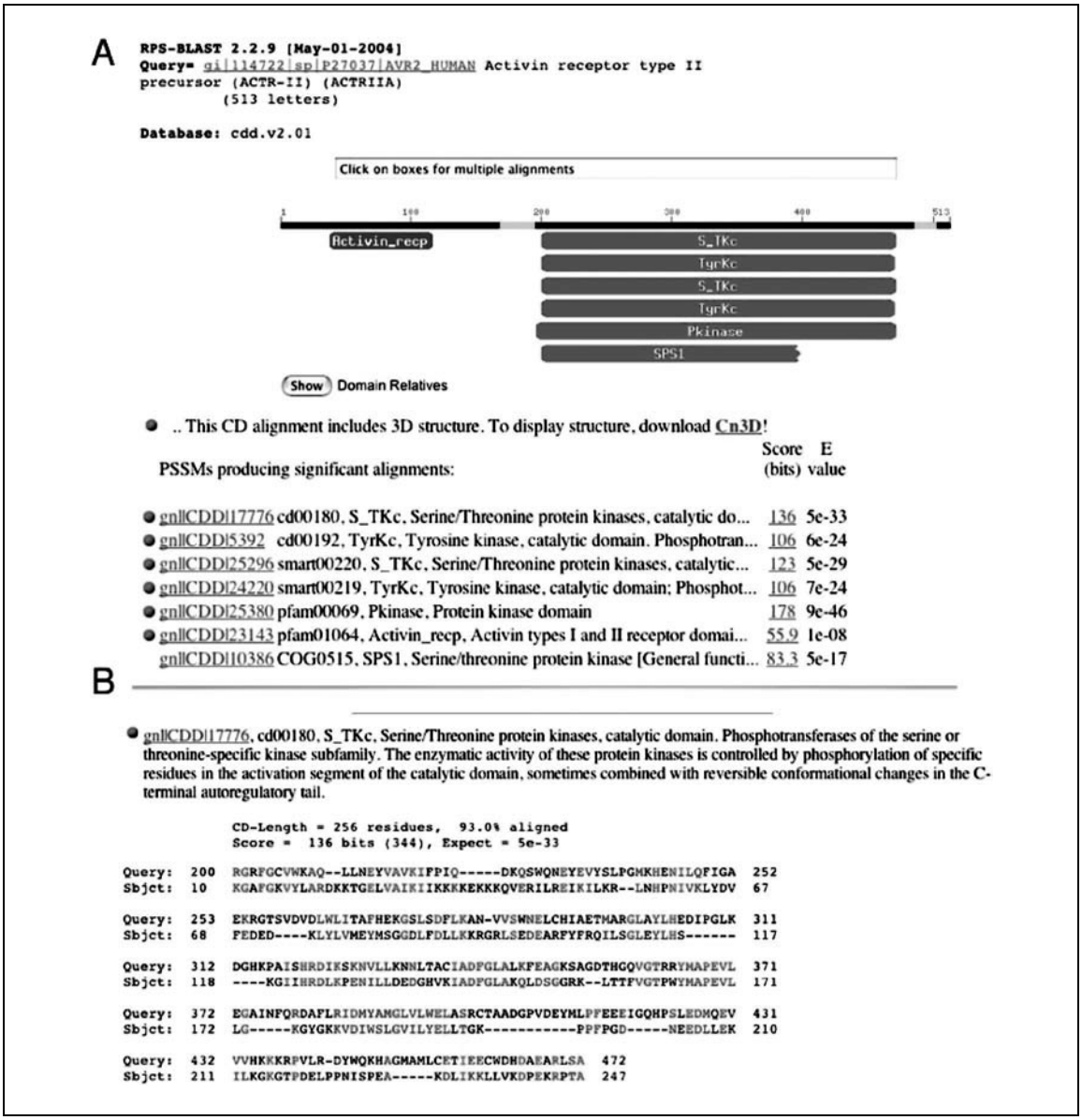


Figure 19.5.6 Results from CD-Search. (**A**) The output gives a graphic display and E-values of hits. (**B**) Pairwise alignment of the query with the top hit. This black and white facsimile of the figure is intended only as a placeholder; for full-color version of figure go to *http://www.interscience.wiley.com/c_p/colorfigures.htm*.

6. The CDD results page returns a graphic display of domains in the query, the E-values for each hit, and the alignments (Fig. 19.5.6). Analyze results (see Support Protocol 2).

GUIDELINES FOR UNDERSTANDING RESULTS OF ANALYSES FROM THE NCBI CD-SEARCH

SUPPORT PROTOCOL 2

See Alternate Protocol 1 for steps in carrying out this search.

1. The first few lines give details about the version of RPS-BLAST used, the definition line of the query sequence, and the database searched. Because the CDD has regular updates, we recommend noting these data for record-keeping purposes.
2. In the graphic display, the query sequence is represented as a black bar with a ruler above indicating its length. Low-complexity regions are marked in cyan.
3. The domain matches are shown below in different colors depending on the E-value of the hits. The best hits are marked in red, the second best in blue and so on. The CDD contains multiple PSSMs for the same structural domains; these may have differences in length and coverage of the query, and have correspondingly different E-values. Hits to conserved domains that are related by the Conserved Domain Architecture Retrieval Tool (CDART; Geer et al., 2002; also see Table 19.5.1) resource are indicated by the same color. In the example in Figure 19.5.6, the hits to different PSSMs representing kinase domains are all colored red.
4. A jagged edge seen in the hit display indicates incomplete coverage of the domain (a default minimum of 80% coverage is used by the CD search). For example the SPS1 domain in Figure 19.5.6 contains a jagged right edge, indicating that the alignment to the query does not contain this region of the domain.
5. Below the graphical display is a list of hits along with their E-values. A pink dot before the hit indicates the presence of a structure for the predicted hit. Clicking on the dot will display the structure. Prior installation of Cn3D software is required (*http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml*).
6. Scrolling down the page displays the pairwise alignment between the query and a representative sequence for the PSSM. Identical residues between the query and the hit sequence are colored red, similar residues are colored blue, and masked-out regions are in italics.

Viewing a multiple sequence alignment (MSA) of the query to a conserved domain

7. Clicking on the graphical display of a domain brings up a multiple sequence alignment of the query to selected PSSM training sequences (Fig. 19.5.7). By default, sequences with high similarity to the query protein are shown. This page gives information about the source of the domain alignment, PubMed references, description, and other relevant data. The default MSA viewing option is Hypertext; our guidelines for interpreting results assume this option is used.
8. Interpreting the MSA: aligned positions are in uppercase, unaligned positions are in lowercase, and gaps are indicated by dashes (-). Residues labeled "unaligned" are inserts between positions in the predicted consensus structure. Profiles and hidden Markov models (HMMs) are constructed to represent the conserved structural elements of a protein family or domain. Amino acids corresponding to additional structure (e.g., extended loop regions) will be displayed in lowercase. These lowercase letters are indicative of insertions relative to the consensus structure. Missing parts of the consensus structure are indicated with dash characters (--).

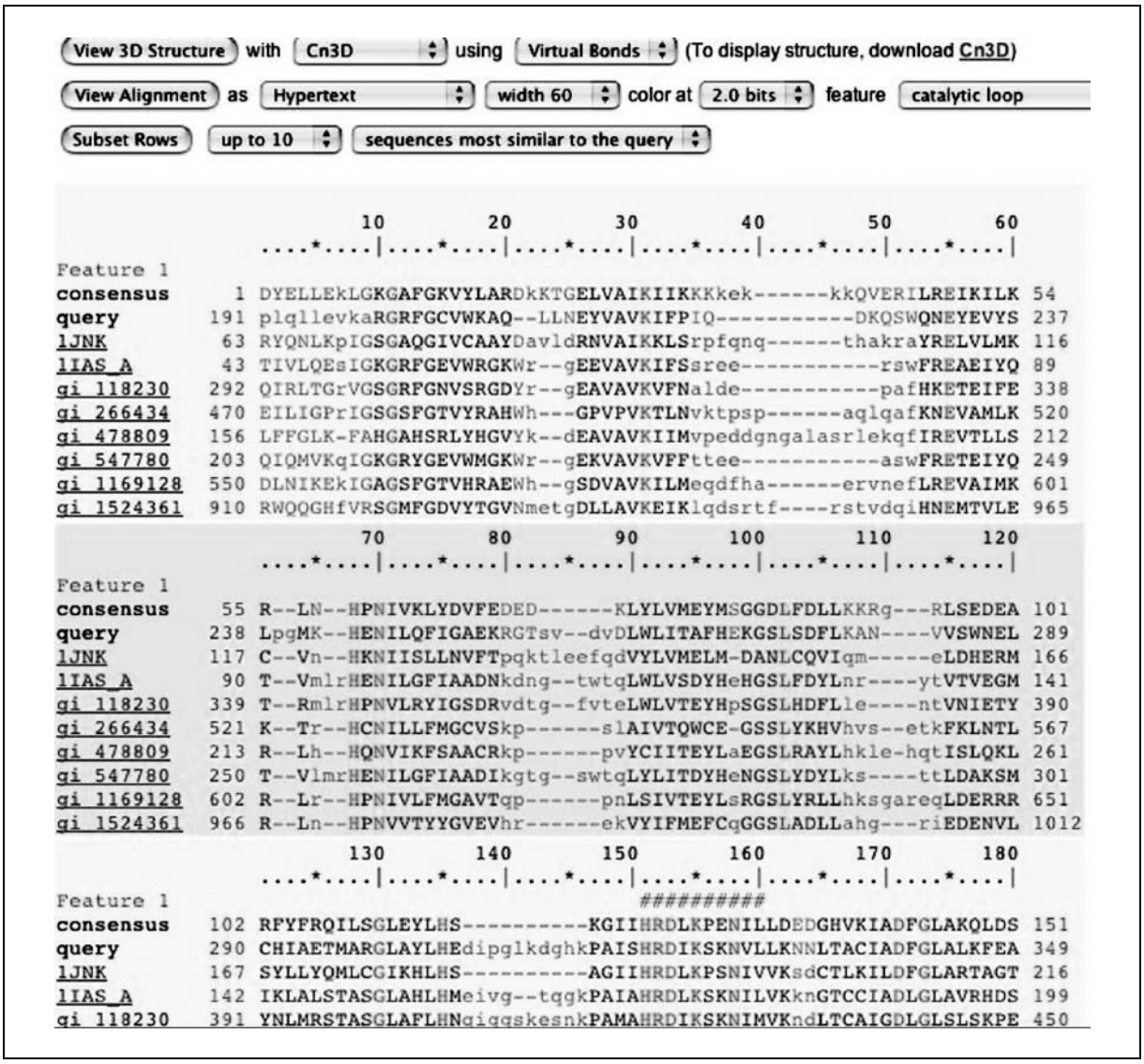


Figure 19.5.7 Multiple sequence alignment of AVR2_HUMAN kinase domain with a conserved kinase domain hit. The columns marked with # have been identified as central for kinase function.

9. For domains with a solved structure, one can view the structure using the View 3D Structure option. The coloring scheme of the alignment can be changed to indicate different levels of conservation. For instance using the option Identity from the "color bits" pull-down menu shows identical residues in the alignment in red, aligned residues in blue, and unaligned residues and columns containing many gaps in gray.

10. Features defining a family are highlighted with # above the corresponding columns in the MSA. The evidence for the feature is presented at the bottom of the MSA. This is useful to predict if the query sequence also contains residues important for the family. In our example, clicking on the `S_TKc` shows a catalytic loop as feature 1 and the 3-D structure as evidence.

Interpreting the significance of results

11. If E-values are weak, it may be helpful to examine the PSSM hits to see if all refer to the same structural or functional class, or if competing hypotheses for a region are presented by the CDD analysis. Checking for significant percent identity and coverage (as described in the SMART discussion earlier in this unit) is important to avoid spurious and misleading predictions.

12. The Advanced Search mode allows users to change the maximum allowed E-value cutoff. This is generally done if no matches are possible with more restrictive cutoffs. If a permissive threshold (e.g., E-value 1) is used, false positive matches should be expected.

13. Though the domains represented in CD are similar to those in PFAM and SMART, the search and scoring mechanisms are very different. While PFAM uses profile HMM scoring, CD-Search employs RPS-BLAST. Hence, the E-values that are reported from PFAM and CD-Search for the same PFAM domain may not be identical.

    To interpret the significance of E-values and the effect of profile/HMM lengths on E-values, see the guidelines described in Support Protocol 1.

PREDICTING STRUCTURAL DOMAINS AND SECONDARY STRUCTURE USING 3D-PSSM

ALTERNATE PROTOCOL 2

3D-PSSM is a Web server designed explicitly and exclusively for protein structure prediction, which incorporates a variety of advanced remote homolog detection and threading methods (Kelley et al., 2000). We present the 3D-PSSM as a supportive procedure in cases where other servers are unable to provide clues to molecular function or structure for regions of the protein. In these cases, the advanced methods incorporated in the 3D-PSSM workflow can often give insights not available via other servers for identifying homologous structures. In contrast to PFAM, the NCBI CDD, and many other servers that model both structural domains and domains having no known structure, any match to a 3D-PSSM domain implies a solved structure for that region. One of the great benefits of 3D-PSSM is the automatic construction of a homology model for any regions in the query matching one or more of the 3D-PSSM models.

The 3D-PSSM server is continually updated. As of February, 2005, 3D-PSSM contains a library of 9864 PSSMs representing structural domains.

Materials

See Basic Protocol 1.

1. Point the browser at *http://www.sbg.bio.ic.ac.uk/servers/3dpssm/*.
2. Click on Recognise a Fold (Fig. 19.5.8A) for the sequence submission form.
3. Enter a valid e-mail address.
4. Paste the query sequence in FASTA format. The query sequence must be <800 residues. The maximum number of residues allowed per submission by 3D-PSSM is restricted to 800. If the protein of interest has >800 residues, we suggest that it be submitted in two parts with overlap. For example, for a protein of 1000 residues, submit region 1 to 600 first, and 400 to 1000 next.
5. Click Submit. Results of the run will be returned by e-mail. The e-mail also contains a hyperlink to the Web version of the results. Results are stored only for 5 days in the server (we recommend downloading results from the Web site).
6. Click on the hyperlink in the e-mail to open a Web page with the results (Fig. 19.5.8B). Analyze results (see Support Protocol 3).

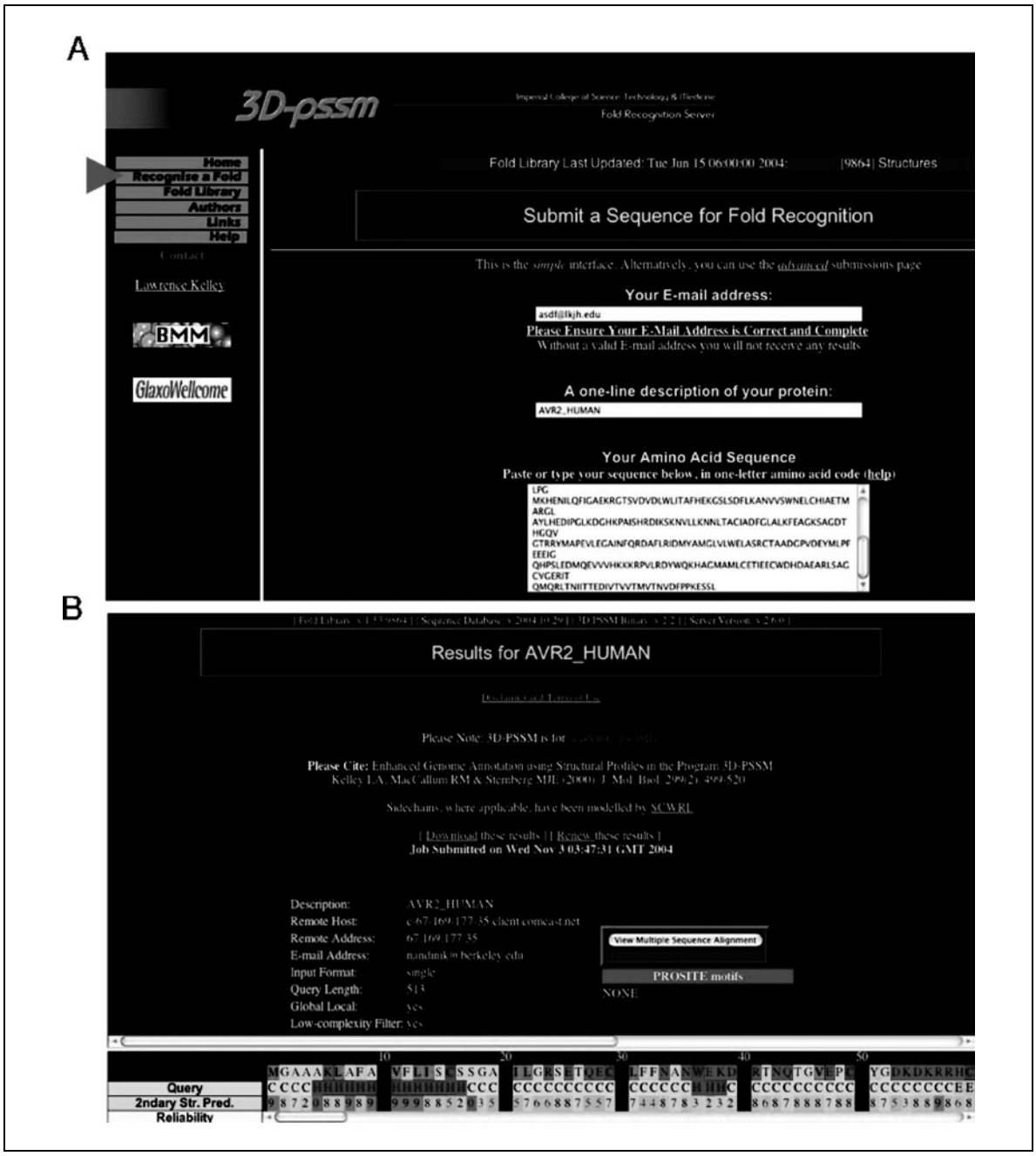


Figure 19.5.8 Submission form and results page from 3D-PSSM server. (**A**) The submission form; (**B**) the results page.

SUPPORT PROTOCOL 3

GUIDELINES FOR UNDERSTANDING RESULTS OF ANALYSES FROM THE 3D-PSSM SERVER

We recommend 3D-PSSM (see Alternate Protocol 2) as an alternative for domain identification for those cases where regions of a protein have no significant matches to domains in the CDD or SMART databases. In these cases, the advanced methods used by 3D-PSSM may be able to identify a weak but potentially significant match to a protein of solved structure and shed light on the functional role of the protein. Caution must be employed when examining the 3D-PSSM results where the E-value is only marginally significant (e.g., >0.001), as the potential for a false positive match increases as the E-value approaches 1.

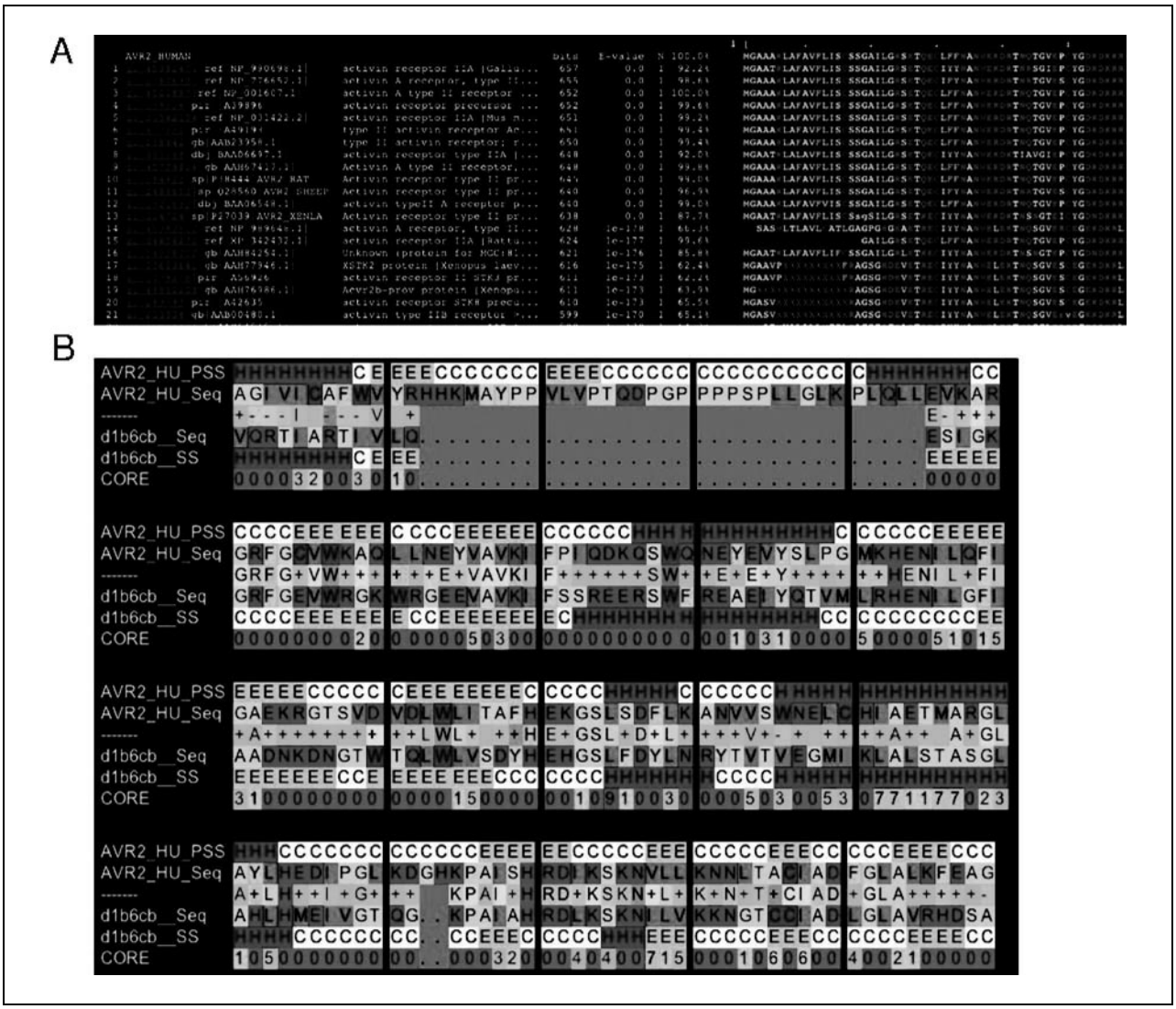


Figure 19.5.9 Sequence alignments produced by 3D-PSSM (**A**) Alignment of query to its homologs. (**B**) Pairwise alignment and comparison of secondary structure and solvent accessibility predictions for the query with that of the hit (structural domain). This black and white facsimile of the figure is intended only as a placeholder; for full-color version of figure go to *http://www.interscience.wiley.com/c_p/colorfigures.htm*.

The 3D-PSSM Web server produces a great deal of interesting and informative results, some of which are beyond the scope of this unit. In this protocol, we highlight a selection of the most interesting results. We encourage users to explore further.

1. The first page of the results is separated into two frames. The upper frame contains a summary of results. The lower frame shows the secondary structure of the query sequence predicted by PSIPRED (Jones, 1999; also see Table 19.5.1). We focus here on results in the upper panel.

2. Click on View Multiple Sequence Alignment to examine the alignment of the query and predicted sequence homologs (Fig. 19.5.9A). The 3D-PSSM Web server uses this alignment to construct a PSSM and score proteins of solved structure, as part of a bidirectional search for structural homologies. Each sequence in the alignment has a hyperlink to the corresponding GenBank file at NCBI. The scores and E-value of each hit, and percent ID to query are also available.

3. Return to the original results page and scroll down to view matches to structural domains. The top 20 structural hits are shown in a tabular format. The table contains the following information.

   a. Alignment.

      A link to the sequence alignment of the query to the hit (structural domain) and comparison of the secondary structure and solvent accessibility of residues between query and hit (Fig. 19.5.9B).

   b. Fold Library.

      The ID of the structural hit (`d[pdbcode][chain][region]`*), along with the pairwise ID between query and hit, are shown. Clicking on the ID brings up a page of the fold library containing information about the hit structural domain, including an MSA of the hit with its sequence and structural homologs.*

   c. Models.

      A link to the predicted 3-D model of the query protein based on the alignment to the hit which can be viewed by the protein visualization software RasMol (see Table 19.5.1) or Chime.

   d. E-value.

      The E-value of the hit is shown, colored according to level of confidence, red indicating very high confidence.

   e. Other information.

      The other information in the table includes SAWTED (Table 19.5.1) E-value for text matching between query and definitions of hits, SCOP classification of hit, and other data.

If the E-value to a structural hit is >.001, the following analyses are recommended to assess the credibility of the predicted homology

4. Check the sequence alignment. The query and structure should align with few insertions or gap characters, and have significant pairwise identity. As a rule of thumb, a pairwise identity of 25% or more over 80 amino acids can be used to infer homology (assuming very few gaps). As the percent identity drops below this level, the potential for a spurious match by chance alone increases. In these cases, it is critical to confirm that the alignment shows agreement at known critical residues.

5. Compare the agreement between predicted and known secondary structure. Compare the predicted secondary structure of the query with that of hit. The secondary structures are indicated as follows: `H` for Helix; `E` for Strand; `C` for Coil. Agreement between helical and sheet structure is generally more important than in the more variable coil regions.

6. Compare the agreement in the conserved core structure. The hydrophobic core region of the structural domain hit is marked in the MSA page as Core (Fig. 19.5.9B), numerically labeled to indicate how buried the residues are (`9` for very buried and making many contacts and `0` for exposed and making few contacts) and color coded to indicate important core residues (red for important core residues). Significant gaps or inserts in the conserved core elements of the structure should be considered evidence against the predicted homology.

   An important issue to keep in mind is that 3D-PSSM may only identify one structural domain in a multidomain protein. To enable identification of domains in the remaining region(s) of the query, separate submissions of unknown regions are necessary.

   There are a few other Web servers and databases enabling integrated searches, as shown in Table 19.5.2.

Table 19.5.2 Web Servers Providing Domain Identification Tools

Web server	Description	URL
SMART	Structural, functional, and localization prediction	*http://smart.embl-heidelberg.de/*
PFAM	Includes both structural and functional domains	*http://www.sanger.ac.uk/Software/Pfam/* *http://pfam.wustl.edu*
SUPERFAMILY	Based on the Astral PDB90 database of structural domains	*http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/hmm.html*
SAM-T02	The UCSC structure prediction HMM library	*http://www.cse.ucsc.edu/research/compbio/HMM-apps/T02-query.html*
TigrFAMs	From TIGR, a Web server of HMMs for structure (and function) prediction	*http://tigrblast.tigr.org/web-hmm/*
PhyloFacts	Includes subfamily HMMs for structural domains.	*http://phylogenomics.berkeley.edu/phylofacts/*
NCBI Conserved Domain	Includes protein domains with solved structures	*http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi*
3D-PSSM	Includes a secondary structure prediction to obtain predictions of 3-D fold	*http://www.sbg.bio.ic.ac.uk/~3dpssm/*

PREDICTING HELICAL TRANSMEMBRANE REGIONS AND SUBCELLULAR LOCALIZATION

BASIC PROTOCOL 2

The molecular function of a protein is context-dependent; the subcellular localization of a protein plays a large role in determining its partners and pathway interactions. The cellular sorting machinery makes use of sequence signals targeting a protein for secretion, targeting it to different organelles, or targeting it for insertion into the membrane. Bioinformatics methods take advantage of these motifs and patterns as well as other sequence attributes to predict where a protein is localized. We review these in this protocol.

Our primary recommended resource for this protocol makes use of a variety of software tools developed by the Center for Biological Sequence Analysis at the Technical University of Denmark (*http://www.cbs.dtu.dk/services/*). This resource enables the detection of helical transmembrane domains, signal peptides, and other organellar targeting signals. Due to the biological importance of transmembrane proteins, we discuss the detection of these domains first.

Prediction of helical transmembrane regions in protein of interest

Transmembrane predictions involve not only the location of probable membrane helices in a protein, but also the full topology of the protein with respect to the membrane (i.e., which regions of the protein span the membrane, which regions are cytoplasmic, and which are extracellular). Among the several Web servers available for transmembrane prediction, we describe the use of the TMHMM server. The TMHMM server uses sophisticated hidden Markov models methodologies to predict transmembrane regions in proteins, incorporating information on hydrophobicity, helix length, charge bias, and topological constraints into its model (Krogh et al., 2001). It has a user-friendly Web interface, is fast in its searches, and allows batch mode submissions of up to 4000 sequences.

Materials

Input data

Protein sequence file, in either raw or FASTA format (Fig. 19.5.1), or, if more than one sequence is used, a file containing all the input sequences in FASTA format can be used

1. Point the browser at the TMHMM server (*http://www.cbs.dtu.dk/services/TMHMM/*).
2. Paste in the sequence(s), or upload a file directly from the local computer (recommended when one wants to analyze large numbers of sequences). To upload a file, use the Browse button below the text "Submission of a local file in FASTA format." Each file can contain at most 4000 sequences.
3. Set display options. If submitting a single sequence, we recommend setting the display as "Extensive, with graphics." This will provide results in the most information-rich format (see below). The recommended option for multiple sequence search is "One line per protein." This returns the results in a tab-delimited format that can be saved as a text file to be opened in a program like Excel.
4. Click Submit. After a short time the results will be displayed.
5. Only 10 job (including batch mode) submissions per day are allowed per user.
6. Analyze results (see Support Protocol 4).

SUPPORT PROTOCOL 4

GUIDELINES FOR UNDERSTANDING RESULTS OF PREDICTIONS OF HELICAL TRANSMEMBRANE REGIONS AND SUBCELLULAR LOCALIZATION

For steps in predicting helical transmembrane regions and subcellular localization, see Basic Protocol 2.

1. Depending on the selection during submission, the results could be either in "long" or "short" format. The long format is shown in Figure 19.5.10.
2. View summary of results containing the following information:

   a. Length.

      Length of query protein.

   b. Number of predicted TMHs.

      Number of predicted TM helices in the query.

   c. Exp number of AAs in TMHs.

      Number of amino acids predicted to be in TM helix. If this number is greater than 18, then it is predicted to be a TM.

   d. Exp number, first 60 AAs.

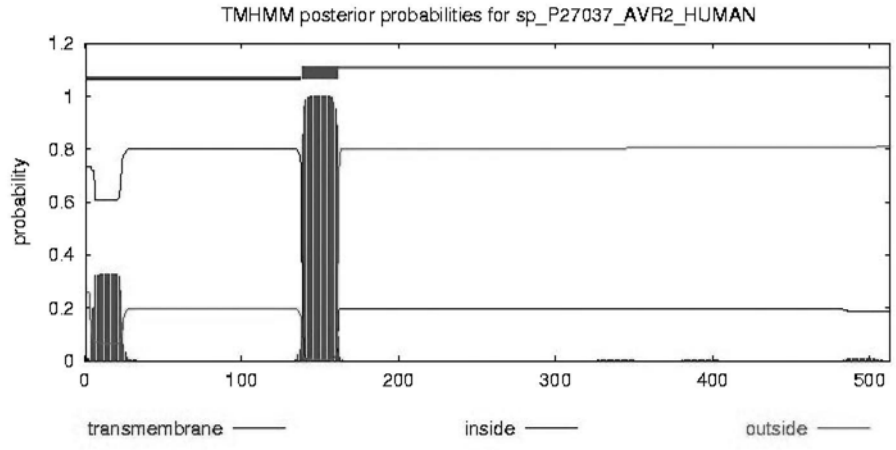
      Number of amino acids expected to be in TM helix in the first 60 residues. A warning `POSSIBLE N-term signal sequence` *is generated if this number exceeds 10.*

   e. Total prob of N-in.

      This represents the probability of N-terminal region being on the cytoplasmic side.

3. View plot. The plot summarizes the various signals used by the TM server. The *x* axis corresponds to the amino acid positions while the *y* axis corresponds to the likelihood of an individual position being predicted to be in a TM. Note that the *y* axis continues beyond a value of 1; the region above 1 is used to display the topology. Since the

```
# sp_P27037_AVR2_HUMAN Length: 513
# sp_P27037_AVR2_HUMAN Number of predicted TMHs:  1
# sp_P27037_AVR2_HUMAN Exp number of AAs in TMHs: 29.30047
# sp_P27037_AVR2_HUMAN Exp number, first 60 AAs:  6.25401
# sp_P27037_AVR2_HUMAN Total prob of N-in:        0.74053
sp_P27037_AVR2_HUMAN    TMHMM2.0        inside      1   138
sp_P27037_AVR2_HUMAN    TMHMM2.0        TMhelix   139   161
sp_P27037_AVR2_HUMAN    TMHMM2.0        outside   162   513
```



plot in postscript, script for making the plot in gnuplot, data for plot

Figure 19.5.10 TMHMM ouput for AVR2_HUMAN. Shown are results in the "extensive with graphics format." A summary of results is followed by a graphic display. The *x* axis on the graph represents amino acid positions in the query sequence and the *y* axis represents the probability of a residue to be in a TM. The peaks indicate positions with higher probability of being a TM domain. The predictions of TM and topology are indicated above a *y* axis value of 1. The region predicted to be inside (cytoplasmic) is represented with a blue line (marked as INSIDE in this figure) and the region on the outside (extracellular) is represented in pink (labeled OUTSIDE in this figure). This black and white facsimile of the figure is intended only as a placeholder; for full-color version of figure go to *http://www.interscience.wiley.com/c_p/colorfigures.htm*.

TMHMM server integrates a variety of signals from the protein, occasionally the predicted number of helices is lower than the number of peaks in the plot.

4. View raw data. Below the plot is a hyperlink to the raw data for the plot.

5. Cautions

   a. Accuracy of topology predictions

      The accuracy of topology prediction (i.e., which parts of the proteins are extracellular and which are cytoplasmic) is known to be one of the most challenging aspects of transmembrane predictions. Even when the location of transmembrane domains is predicted correctly, the topology can be reversed.

   b. Potential for false positives and false negatives.

      All TM prediction servers are known to have problems with false positive and false negative predictions. False positives arise often for hydrophobic stretches and signal peptides. False negatives can arise for any number of reasons, and are particularly common in noncanonical TM stretches containing some polar or charged residues.

   c. Incorporating prediction of globular structure.

      It can be helpful to integrate domain detection, particularly of globular domains, when interpreting TM predictions. If a predicted TM occurs in a region that has a significant match to a protein of solved structure, we recommend examining what is known about that structure. Some structures include transmembrane domains, but

these are few. The vast majority of solved structures represent globular (soluble) domains that do not include transmembrane domains. If a predicted TM domain has a significant match to a globular protein, and that region is not at the extreme N- or C-terminus of the domain, we would discount the likelihood of the TM domain.

ALTERNATE PROTOCOL 3

PREDICTING THE SUBCELLULAR LOCALIZATION OF A PROTEIN USING TargetP

Cells have evolved a complex machinery for enabling proteins to be targeted to specific sub-cellular locations. Targeted proteins have specific signals recognized by the machinery. For secreted proteins and those targeted to the organelles such as mitochondria or chloroplast, the targeting signals are amino-terminal peptides recognized by the translocation machinery (Schatz and Dobberstein, 1996).

The targeting signals for different locations have specific features. For instance, the amino-terminal signal peptide (~20 amino acids) that targets proteins for secretion starts with a positively charged segment followed by a central hydrophobic region and a polar segment containing the cleavage site recognized by the signal peptidase. Different approaches have been developed to predict protein localization (Emanuelsson and von Heijne, 2001). Most methods specialize in predicting one specific location (e.g., SignalP for predicting signal peptides). In this protocol we review TargetP, an integrated approach that distinguishes proteins targeted to different locations in the cell such as mitochondria, chloroplast, and the secretory pathway.

Materials

See Basic Protocol 2.

1. Point the browser at the TargetP server (*http://www.cbs.dtu.dk/services/TargetP/*).
2. Paste in the sequence(s), or upload a file directly from the local computer.
3. Identify the origin of the protein sequence (i.e., plant or animal). If it is of plant origin, TargetP will search for chloroplast localization signals.
4. Select the option to perform a cleavage site prediction.
5. Use default settings for specificity cutoffs. Click the "Submit sequence/file" button.
6. Analyze results (see Support Protocol 5).

SUPPORT PROTOCOL 5

GUIDELINES FOR UNDERSTANDING RESULTS PREDICTING THE SUBCELLULAR LOCALIZATION OF A PROTEIN USING TargetP

Results (see Alternate Protocol 3) include scores for different predicted cellular locations (Fig. 19.5.11).

SP : Signal peptide
mTP : Mitochondrial target peptide
cTP : Chloroplast (if plant sequence is used)
Other: For other locations

Along with the scores, a "reliability class" (RC) is also calculated. This gives information on the difference between the highest score and the second best.

```
RC 1: diff > 0.800
RC 2: 0.800 > diff > 0.600
RC 3: 0.600 > diff > 0.400
RC 4: 0.400 > diff > 0.200
RC 5: 0.200 > diff
```

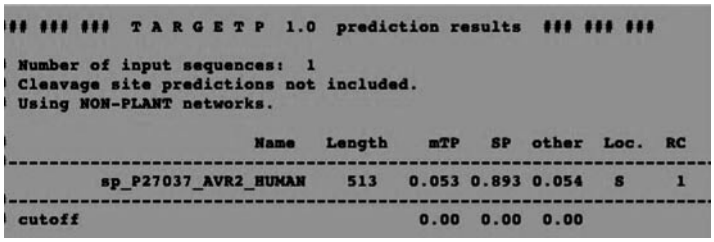


Figure 19.5.11 TargetP prediction of the subcellular localization of AVR2_HUMAN.

Table 19.5.3 Cellular Localization Prediction

Web server	Description	URL
TMHMM	The TMHMM transmembrane helix prediction server	*http://www.cbs.dtu.dk/services/TMHMM/*
DAS	The DAS transmembrane helix prediction server	*http://www.sbc.su.se/~miklos/DAS/*
SignalP	The SignalP signal peptide prediction server	*http://www.cbs.dtu.dk/services/SignalP/*
ChloroP	The ChloroP chloroplast targeting prediction server	*http://www.cbs.dtu.dk/services/ChloroP/*
PredictNLS	Nuclear localization signal prediction	*http://cubic.bioc.columbia.edu/predictNLS/*

In interpreting TargetP results, it is important to consider both the predicted location as well as the RC score. Lower RC values imply greater expected accuracy; an RC of 1 or 2 is a good indicator of reliability. In our example, the score for SP is 0.853 and the RC value is 1, both supporting the prediction of presence of a signal peptide in the protein (Fig. 19.5.11). TargetP is known to have greater specificity in prediction compared to other methods in this class, but this comes at a cost of sensitivity. A null prediction does not imply that the protein is cytoplasmic. By contrast, cleavage-site prediction is less precise. Finally it is critical to indicate whether the input sequence is of plant or animal origin so that the appropriate localization tests can be performed; it is otherwise difficult to distinguish between chloroplast and mitochondrial signals.

Other resources for prediction of transmembrane helices or subcellular localization can be found in Table 19.5.3.

PREDICTING KEY FUNCTIONAL RESIDUES AND MOTIFS USING THE PROSITE WEB SERVER

BASIC PROTOCOL 3

Not all residues in a protein are created equal. Residues in active sites, e.g., the catalytic triad of serine proteases, are critical for enzymatic activity. Other positions may be involved in substrate recognition and binding. Still other positions may be conserved for structural reasons. Over the years, biologists have identified a large number of key residues and conserved motifs that are useful in defining and recognizing protein families.

Our primary recommended resource for this protocol is PROSITE. The PROSITE Web server enables a biologist to determine whether any experimentally characterized motifs are present in a sequence of interest (Sigrist et al., 2002; Hulo et al., 2004). This resource enables the detection of different protein motifs, post-translational modification sites, and domains. Since domain detection is provided through other servers (presented earlier in this unit), this protocol focuses on the use of the PROSITE Web server for detecting the presence of conserved motifs and functional residues.

Materials

Input data

Sequence in raw or FASTA format (Fig. 15.9.1)

1. Point the browser to the PROSITE Web site (*http://us.expasy.org/prosite/*).
2. Input the sequence in FASTA format in the sequence box provided.
3. Choose the option to exclude patterns with high probability of occurrence. This will decrease the likelihood of spurious predictions.
4. Click the Quick Scan button to search for conserved motifs in the PROSITE database.
5. Analyze results (see Support Protocol 6).

SUPPORT PROTOCOL 6

GUIDELINES FOR UNDERSTANDING RESULTS OF SEARCHES DONE USING THE PROSITE WEB SERVER

The PROSITE Web server (see Basic Protocol 3) includes both pattern and profile search. Profiles are usually constructed to represent domains and protein families, while patterns are confined to the represention of short conserved motifs and critical residues. When a query protein is submitted, both patterns and profiles are searched and the results are returned as "hits by profile" (Fig. 19.5.12A) and "hits by patterns" (Fig. 19.5.12B). For each detected feature, the position and the sequence motif are provided along with a link to detailed information about each feature. Placing the mouse cursor over predicted domains in the graphical display highlights the corresponding region in the sequence.

Pattern detection has certain advantages relative to profile search. First, pattern detection is extremely fast. Second, and potentially of greater interest to biologists, the results are easily interpreted; the motifs are typically quite short, and critical residues defining the motif can be examined easily.

There are two main disadvantages to pattern detection, relative to profile search. First, a strict agreement with a pattern is required, resulting in lower sensitivity than the more permissive profile searches, which allow substitutions. Second, the results from pattern searches are not given scores; the pattern is either present or it is not. This makes is difficult to evaluate the significance of the match. In addition, some patterns have a high associated false-positive rate (this is more common in very short or highly variable patterns). For example, the post-translational modification site for N-glycosylation is Asn-Xaa-Ser/Thr (where Xaa represents any amino acid). However, the presence of this motif may not be sufficient to ensure glycosylation, as the process is also dependent on the structure of the protein (Sigrist et al., 2002).

The PROSITE Web server contains information to help reduce the likelihood of errors. First, each pattern includes information regarding the number of hits obtained while scanning the Swiss-Prot database with that pattern, the number of false positives and false negatives, and results from pattern search using database randomization. These data should be consulted to obtain an understanding of the likely significance of a match to this pattern. In addition, the PROSITE profile searches can provide additional support for a functional classification based on motif and pattern detection. For instance, the presence of a serine/threonine protein kinase signature pattern along with detection of kinase domain for AVR2_HUMAN strengthens the significance of the prediction.

An example of a six-amino-acid PROSITE pattern is shown below.

```
G-{EDRKHPFYW}-x(2)-[STAGCN]-{P} [G is the N-myristoylation site]
```

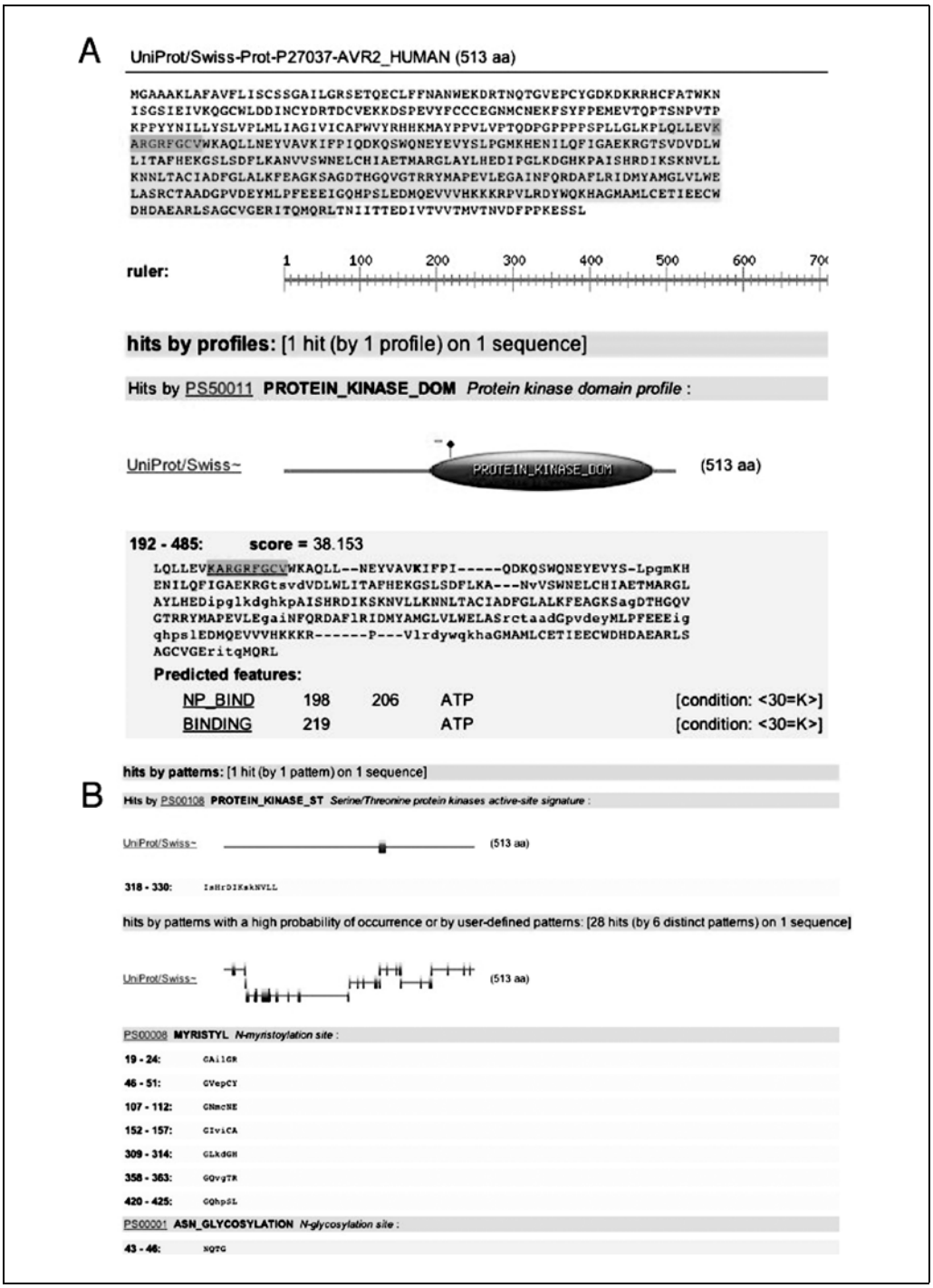


Figure 19.5.12 Prediction of patterns by PROSITE for AVR2_HUMAN. (**A**) results of profile search; (**B**) results of pattern search.

Residues in straight brackets (`[]`) designate residues that are favorable at a position. Residues in curly brackets ({}) designate those amino acids that are *not* allowed at a position (i.e., in this example, the amino acids `EDRKHPFYW` are not allowed at position 2). The sequence `x(2)` indicates any two amino acids.

Other resources for similar tasks can be found in Table 19.5.4.

Table 19.5.4 Predicting Key Functional Residues and Motifs

Web server	Description	URL
Evolutionary Trace	University of Cambridge implementation of Evolutionary Trace algorithm	*http://www-cryst.bioc.cam.ac.uk/~jiye/evoltrace/evoltrace.html*
MEME	The MEME motif discovery software	*http://meme.sdsc.edu/meme/Website/intro.html*
PROSITE	Detects previously characterized motifs characterizing protein families, as well as post-translational modifications	*http://www.expasy.ch/prosite*

SUPPORT PROTOCOL 7

HOMOLOG IDENTIFICATION

As a support protocol, we encourage biologists to gather homologs for their sequence of interest and to construct and analyze a multiple sequence alignment of these homologs. The BLAST family of methods (Altschul et al., 1990) is by far the most popular means for identifying homologs in database search; we refer readers to *UNIT 19.3* in this manual or the BLAST tutorial at the NCBI Website (*http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/tut1.html*).

Homolog identification is useful for a variety of reasons. First, homologs can give clues to the overall molecular function of a protein. When inferring molecular function by homology, it is critical to confirm that the query and hit have the same overall fold (i.e., are globally alignable), that they are orthologous, and that the database annotation of the hit is based on experiment and not homology (reviewed in Eisen, 1998, and in Sjölander, 2004).

Gathering homologs is also the first step in construction of a multiple sequence alignment and the (somewhat esoteric) art of alignment gazing. Alignment gazing can help a biologist discriminate between true homologs and spurious inclusions, and assists in the identification of key functional residues. Sequences inserting large numbers of amino acids, that do not align over long regions, or that do not agree at conserved positions (especially longer conserved motifs) should be removed from the alignment. The final set of high-confidence homologs can be used to supplement the structure-prediction efforts (i.e., by parallel analyses using all the servers and approaches described in this unit). In addition, the multiple sequence alignment can be used as input into motif-discovery systems such as MEME. This is particularly helpful when there is limited (or no) success in analysis of the query sequence. Consensus approaches such as these have been shown to dramatically improve prediction accuracy.

COMMENTARY

Background Information

Each of the methods presented in this unit has the potential to produce errors of different types. When examining results, it is important to keep in mind the types of errors that are possible with each method, and to attempt to separate those predictions that are highly credible from those that are potentially misleading.

Homology-based function prediction is particularly prone to systematic error; domain shuffling, gene duplication, and speciation produce families of proteins sharing some regions of similarity but with different functions and overall folds. Because of these issues, even a significant E-value between a query and a database hit is not sufficient for inference of molecular function. To avoid these errors, phylogenomic inference of a protein's molecular function in the context of its related family members is critical. A detailed description of this approach is beyond the scope of this unit; overviews of the issues and methodologies for phylogenomic analysis are presented in Eisen (1998) and Sjölander (2004).

For these reasons, this unit focuses less on obtaining a prediction of precise molecular

function and more on a structural annotation of a protein. The protocols presented here have been selected to enable a biologist to label regions in the protein as functional or structural domains, to identify key residues and motifs, and to predict a protein's cellular localization.

Most homolog-detection methods use either profile- or HMM-based strategies. Both profiles and HMMs are generalizations of multiple sequence alignment of homologs. They differ in how they estimate gap parameters; profiles have a fixed cost for insertions and deletions, while HMMs learn them from the training sequences used in the MSA, thus varying the cost for different positions. Both methods are dependent on precision of homolog gathering, alignment accuracy, and estimation of amino acid substitutions.

Domain-based approaches to protein annotation have a critical limitation. The profiles (and HMMs and PSSMs) used to detect the presence of these domains are typically optimized for remote homolog detection. This makes them very effective at providing clues to a protein's fold or function that might otherwise not be achievable. However, because some protein folds are able to provide a multiplicity of distinct functions, a significant score to a profile for a domain may mean only that the query has a similar *fold*, but may not actually imply a common molecular function. We strongly advise checking for agreement at key functional residues (where this information is available). The use of consensus approaches to predicting a protein's structure can help avoid errors, provided that the individual methods are orthogonal. For instance, since both SMART and the CDD include PFAM profile HMMs in their structure-prediction process, a consensus structure prediction between the NCBI CDD or SMART and PFAM is much less informative than a consensus of the NCBI CDD and the 3D-PSSM structure prediction Web server (which uses an entirely different approach).

Methods for predicting helical transmembrane domains often rely on weak nonspecific signals such as hydrophobicity, and hence have a fairly high frequency of false-positive as well as false-negative predictions (Chen et al., 2002). To boost accuracy in this task, we strongly encourage biologists to examine alternative hypotheses. Does the predicted transmembrane domain occur in a region predicted (with a significant E-value) to be a globular structure? Is it found at the amino-terminus (where it is more likely to be a signal peptide)? In either case, the probability of the region being a transmembrane domain should be discounted.

In summary, our recommendation for enhancing prediction accuracy involves integrating results and information from a variety of different resources, and biological data should be incorporated wherever possible.

Critical Parameters and Troubleshooting

On occasion, a Web server will report an error for certain input sequences. In some cases, hidden characters may make identification of the problem difficult. If this happens, try truncating the definition line to the name of the protein, inserting a carriage return after the definition line, and then resubmitting the sequence.

Literature Cited

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

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. 2002. The Pfam protein families database. *Nucleic Acids Res.* 30:276-280.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.* 31:365-370.

Chen, C.P., Kernytsky, A., and Rost, B. 2002. Transmembrane helix predictions revisited. *Protein Sci.* 11:2774-2791.

Eisen, J.A. 1998. Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. *Genome Res.* 8:163-167.

Emanuelsson, O. and von Heijne, G. 2001. Prediction of organellar targeting signals. *Biochim. Biophys. Acta* 1541:114-119.

Geer, L.Y., Domrachev, M., Lipman, D.J., and Bryant, S.H. 2002. CDART: Protein homology by domain architecture. *Genome Res.* 12:1619-1623.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., Durbin, R., Eyras, E., Gilbert, J., Hammond, M., Huminiecki, L., Kasprzyk, A., Lehvaslaiho, H., Lijnzaad, P., Melsopp, C., Mongin, E., Pettett, R., Pocock, M., Potter, S., Rust, A., Schmidt, E., Searle, S., Slater, G., Smith, J., Spooner, W., Stabenau, A., Stalker, J., Stupka, E., Ureta-Vidal, A., Vastrik, I., and Clamp, M. 2002. The Ensembl genome database project. *Nucleic Acids Res.* 30:38-41.

Hulo, N., Sigrist, C.J., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P., and Bairoch, A. 2004.

Recent improvements to the PROSITE database. *Nucleic Acids Res.* 32:D134-D137.

Jones, D.T. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* 292:195-202.

Kelley, L.A., MacCallum, R.M., and Sternberg, M.J. 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* 299:499-520.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* 305:567-580.

Letunic, I., Copley, R.R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., Ponting, C.P., and Bork, P. 2004. SMART 4.0: Towards genomic data integration. *Nucleic Acids Res.* 32:D142-D144.

Marchler-Bauer, A., Anderson, J.B., DeWeese-Scott, C., Fedorova, N.D., Geer, L.Y., He, S., Hurwitz, D.I., Jackson, J.D., Jacobs, A.R., Lanczycki, C.J., Liebert, C.A., Liu, C., Madej, T., Marchler, G.H., Mazumder, R., Nikolskaya, A.N., Panchenko, A.R., Rao, B.S., Shoemaker, B.A., Simonyan, V., Song, J.S., Thiessen, P.A., Vasudevan, S., Wang, Y., Yamashita, R.A., Yin, J.J., and Bryant, S.H. 2003. CDD: A curated Entrez database of conserved domain alignments. *Nucleic Acids Res.* 31:383-387.

Marchler-Bauer, A. and Bryant, S.H. 2004. CD-Search: Protein domain annotations on the fly. *Nucleic Acids Res.* 32:W327-W331.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247:536-540.

Schatz, G. and Dobberstein, B. 1996. Common principles of protein translocation across membranes. *Science* 271:1519-1526.

Schultz, J., Milpetz, F., Bork, P., and Ponting, C.P. 1998. SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc. Natl. Acad. Sci. U.S.A.* 95:5857-5864.

Sigrist, C.J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., and Bucher, P. 2002. PROSITE: A documented database using patterns and profiles as motif descriptors. *Brief. Bioinform.* 3:265-274.

Sjölander, K. 2004. Phylogenomic inference of protein molecular function: Advances and challenges. *Bioinformatics* 20:170-179.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. 2001. The COG database: New developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 29:22-28.

Contributed by Nandini Krishnamurthy
and Kimmen V. Sjölander
University of California
Berkeley, California