CHAPTER

8

Additional Bioinformatic Analyses Involving Protein Sequences*

OUTLINE

8.1 PROTEIN STRUCTURE

Proteins have four levels of structure: primary, secondary, tertiary, and quaternary.

Primary structure is simply the amino-acid sequence of the polypeptide, and is determined by the sequence of codons in the gene encoding the polypeptide. Therefore, the open reading frame (ORF)-prediction programs predict the primary structure of the encoded proteins.

Secondary structure is the hydrogen (H)-bonded three-dimensional local conformation. The two most common secondary structures are the **α-helix** and **β-pleated sheet**. In addition, four other commonly occurring secondary structures are the **3_{10}-helix**, **π-helix** (**pi helix**), **β-turn**, and **Ω-loop** (**omega loop**). There are still other regions in proteins whose secondary structure can not be classified under any established categories; these have been traditionally referred to as **random coils**, but can be more appropriately referred to as **unstructured regions**.

An α-helix (radius = 2.3 Å) is a right-handed helix that has 3.6 amino acids per helical turn (100° turn/residue),

*The opinions expressed in this chapter are the author's own and they do not necessarily reflect the opinions of the FDA, the DHHS, or the Federal Government.

and the structure is stabilized by H-bonds formed between the C═O of residue n and the N—H of residue n + 4; both these groups are part of the helical backbone and not the side chains (R groups) that protrude out of the backbone. The pitch of the helix (vertical distance in one complete helical turn) is 5.4 Å; hence, the rise per residue along the helix axis is 1.5 Å. In an α-helix, the H-bonds are intrachain and parallel to the axis of the helix. The α-helix is a **3.6_{13}-helix**, where 3.6 is the number of residues per turn and 13 is the number of atoms in the H-bonded loop. *The α-helix is the most abundant secondary structure found in globular proteins, and it accounts for 32—38% of all residues. The average length of an α-helix is 10 residues*.

A less common helical secondary structure found in proteins is the **3_{10}-helix** (radius = 1.9 Å), which has 3 amino acids per turn (120° turn/residue) and 10 atoms in the H-bonded loop. In a 3_{10}-helix, H-bonds involve residues n and n + 3 (instead of n + 4 as in the α-helix), and the backbone conformational angles are slightly different from those of the α-helix. The pitch of the helix is 6.0 Å; hence, the rise per residue along the helix axis is 2.0 Å. The length of the 3_{10}-helix may vary from 3 to 10 residues. The ideal 3_{10}-helix is rare and when it occurs, it tends to be at the C- and N-termini; the 3_{10}-helix has been described in channels and membrane proteins.[1]

Like the α-helix and 3_{10}-helix, the π-helix (radius = 2.8 Å) is also a right-handed helix. There are 4.4 residues per turn (81.8° turn/residue) and 16 atoms in the H-bonded loop; hence, the π-helix is a **4.4_{16}-helix**. The structure is stabilized by H-bonds formed between the C═O of residue n and the N—H of residue n + 5 (compared to n + 4 in the α-helix, and n + 3 in the 3_{10}-helix). The pitch of the helix is 4.8 Å; hence, the rise per residue along the helix axis is 1.1 Å. A π-helix can be derived from an α-helix by the insertion of a single amino acid. Such insertion tends to destabilize the α-helix. As a result, the formation of π-helix is tolerated only if it provides some selective advantage to the protein. One such possibility involves affecting the functional site of proteins. Consistent with this hypothesis, the π-helix is typically found near the functional site of proteins. About 15% of known protein structures contain a π-helix. Naturally occurring π-helices are typically 7—10 residues in length, but are mostly composed of 7 residues; they are found at the end of a regular α-helix or within an α-helix—that is, a π-helix is flanked by α-helices.[2]

Two or more (two to seven) α-helices can wrap around each other creating **coiled coils**, which are **superhelical** (**supersecondary**) **structures**. In most coiled coils, the α-helices are wrapped around each other into a left-handed helical supercoil. The α-helical coiled coil is a common structural motif in proteins that facilitate subunit oligomerization. Coiled coils can be composed of parallel or antiparallel helices. An example of a functional protein with coiled coils is the Fos-Jun heterodimer, known to regulate gene expression. Another example is tropomyosin. Each strand of a coiled coil has a repeat of seven residues (heptads; *a-b-c-d-e-f-g*). In these heptads, the first and the fourth residues (*a* and *d*) are hydrophobic; they face the helical interface and facilitate hydrophobic interactions. Good candidate amino acids at these positions are isoleucine, leucine, and valine. The other residues are hydrophilic and exposed to the solvent. Of these, the fifth and the seventh residues (*e* and *g*) confer specificity between the two helices through electrostatic interactions. Good candidate amino acids at these positions are the charged amino acids, such as aspartic acid, glutamic acid, lysine, and arginine. Discontinuities in the heptad pattern are quite frequent. Algorithms that predict coiled coils scan the sequence for the regular patterns and heptad signatures using a window size of 14, 21, or 28 amino acids.

In contrast to the helices, a β-pleated sheet (β-sheet) involves two or more polypeptide chains and the H-bonds are formed between residues that are part of different polypeptide chains. Therefore, in a β-pleated sheet, the H-bonds are interchain and are perpendicular to the polypeptide backbones. Each polypeptide chain involved in the formation of a β-pleated sheet is a **β-strand**; a β-pleated sheet can be two stranded or multi-stranded. As the name suggests, the β-pleated sheet has a zigzag appearance. *After the α-helix, the β-sheet is the major secondary-structural element in globular proteins, accounting for 20—28% of all residues*.

In a **β-turn** (also called β-bend) the direction of the polypeptide chain is sharply reversed. The name β-turn owes its origin to the fact that they often connect antiparallel β-sheets. *A β-turn is composed of four amino acids*[a]. The **Ω loop**, as a secondary-structural motif in globular proteins, was first described in 1986.[3] These are a six-amino-acid or longer backbone motif. The polypeptide reverses its direction over the course of this six- (or more) amino-acid-long, omega-shaped loop region[b].

The **tertiary structure** of a protein is the overall folded structure in three-dimensional (3D) space. The tertiary structure is formed by the interactions between

[a]Depending on the number of amino acids involved, other tight turns are named as the **δ-turn** (involves two amino acids), **γ-turn** (involves three amino acids), **α-turn** (involves five amino acids), and **π-turn** (involves six amino acids).[4]

[b]The existence of a variety of morphologies of loops (4 to 20 residues in length) as secondary-structural motifs has been reported in proteins, such as **strap loops** (linear), **omega loops** (nonlinear and planar), **zeta loops** (nonlinear and non-planar, i.e. globular).[5]

the side-chain R-groups, such as ionic interactions, hydrophobic interactions, H-bonds, and disulfide bonds. The amino-acid sequence (the primary structure) primarily dictates how a protein should fold into a 3D tertiary structure. However, proper folding is now known to be achieved with the help of **chaperone** molecules. In folded conformation (tertiary structure), most proteins contain specific **domains** that are discrete structural and functional units of the protein (discussed later).

Quaternary structure of proteins refers to the overall structure of multimeric proteins—that is, proteins composed of two or more subunits, each subunit being a monomer. Quaternary structures are stabilized by non-covalent interactions as well as disulfide linkages. Proteins with molecular weight >100 kD mostly contain more than one polypeptide chain, and hence have a quaternary structure.

The secondary, tertiary, and quaternary structures of proteins are maintained by non-covalent forces, such as H-bonds, electrostatic interactions, and van der Waals forces.

8.2 PEPTIDE BOND, PEPTIDE PLANE, BOND ROTATION, DIHEDRAL ANGLES, AND RAMACHANDRAN PLOT

Amino acids are linked together by peptide bonds. Peptide bonds are **amide linkages** between the $-NH_2$ and $-COOH$ groups of neighboring amino acids. The peptide bond (C—N) has a partial double-bond character. Thus, it is rigid and planar and not free to rotate. The plane on which it lies is called the **peptide plane** or **amide plane**. Peptide bonds are *trans* bonds—that is, the carbonyl oxygen and amide hydrogen are in *trans* position. However, the N—C_α and C_α—C bonds are not rigid and they can freely rotate, being only limited by the size and character of the R-groups. The angle of rotation (also called **torsion angle** or **dihedral angle**) around the N—C_α bond is called **phi** (φ) and that around the C_α—C bond is called **psi** (ψ) (Figure 8.1A). These two angles largely determine the 3D shape of the polypeptide backbone of the protein.

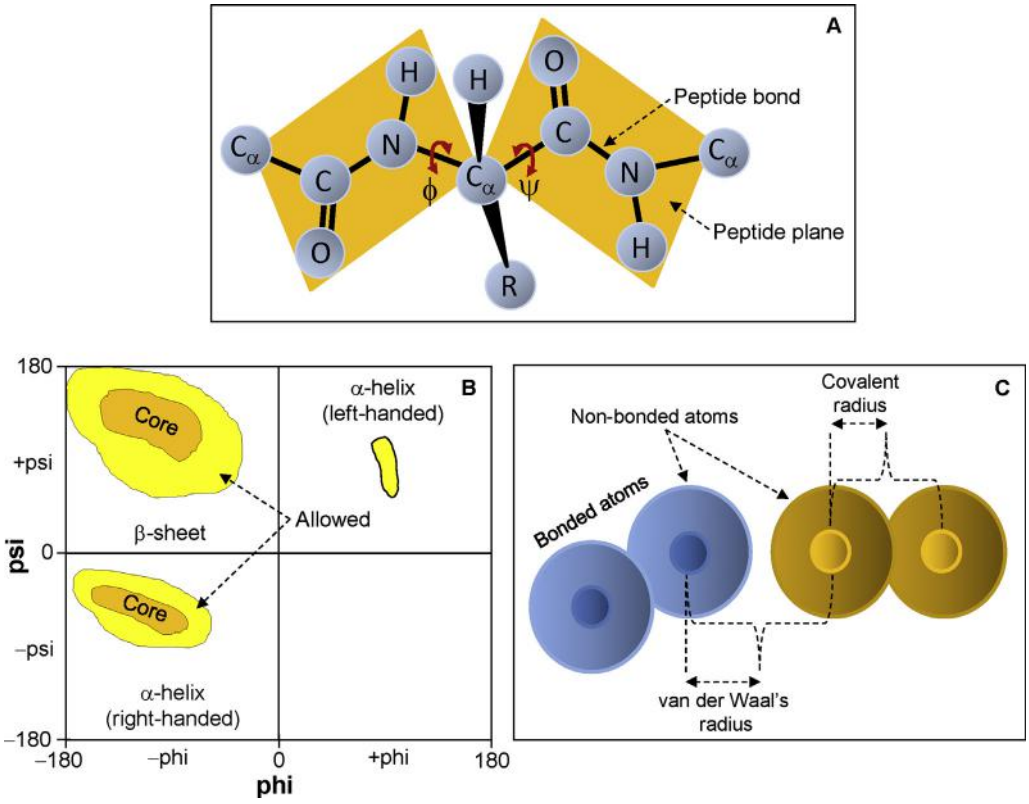


FIGURE 8.1 **Peptide bond, peptide plane, and the Ramachandran plot.** (A) Peptide bond, peptide plane, phi and psi angles, and bond rotation involving two amino acids. The N—C_α and C_α—C bonds are not rigid and can freely rotate, being only limited by the size and character of the R-groups. (B) Diagram of a typical Ramachandran plot (φ/ψ plot). The regions marked "Core" correspond to conformations that do not have any steric hindrance. The yellow areas labeled "Allowed" correspond to conformations that could be possible if the atoms could come a little closer together. The white areas represent conformations that are sterically unfavorable (see text). (C) In computing a Ramachandran plot, atoms are treated as hard spheres whose dimensions correspond to their van der Waals radii. The van der Waals radius and covalent radius are depicted for comparison.

Although φ and ψ are less restricted in terms of rotation, the bulkiness of R-groups of the amino acids tends to impose some restrictions on the rotation through steric hindrance. This makes certain combinations of φ and ψ preferred. The φ/ψ plot of the amino acid residues in a peptide is called the **Ramachandran plot**. It involves plotting the φ values on the *x*-axis and the ψ values on the *y*-axis to predict the possible conformation of the peptide. The angle spectrum in each axis is from −180° to +180°. In computing a Ramachandran plot, atoms are treated as hard spheres whose dimensions correspond to their van der Waals radii. Any angle that results in the collision of the spheres is regarded as sterically unfavorable; hence, such conformations are also sterically not allowed. Figure 8.1B shows a simplified diagram of a Ramachandran plot. The regions marked "Core" correspond to conformations that do not have any steric hindrance. The yellow areas labeled "Allowed" correspond to conformations that could be possible if slightly shorter van der Waals radii are used in the calculation. In other words, if the atoms could come a little closer together, then these conformations would be possible. The white areas represent conformations that are sterically unfavorable. The van der Waals radius and covalent radius are depicted in Figure 8.1C. The residues with a less bulky side chain or no side chain, such as glycine (no side chain), can have many possible combinations of φ and ψ (e.g. in a polyglycine backbone) resulting in a larger allowable area on the plot in all four quadrants, whereas residues with bulky side chains, such as proline or phenylalanine, have fewer possible combinations of φ and ψ, hence a smaller allowable area on the plot.

The φ and ψ angles for each residue in a helical structure are very similar, and that is what confers regularity to the helical structure. *Positive angles correspond to clockwise rotation and negative angles correspond to anticlockwise rotation*. The ideal values of φ/ψ were determined to be as follows: right-handed α-helix −57°/−47°; left-handed α-helix +57°/+47°; right-handed 3_{10} helix −74°/−4°; right-handed π-helix −57°/−70°; parallel β-sheet (uncommon) −119°/+113°; antiparallel β-sheet (common) −139°/+135°. The actual values differ somewhat from these idealized values. Recent experimental data have demonstrated that both φ and ψ can undergo large rotations, which are usually coupled. See Hovmöller, et al.[6] for more details on experimental determination of main-chain conformations in 1042 protein subunits.

Online tools are available from several sources for the analysis of Ramachandran plots of proteins. One such tool is available at the **Uppsala Ramachandran Server** (http://eds.bmc.uu.se/ramachan.html). This service is based on the Moleman2 program.[7]

8.3 PREDICTION OF PHYSICOCHEMICAL PROPERTIES OF A PROTEIN

The physicochemical properties of a protein can be deduced from its sequence. The **ExPASy** (Expert Protein Analysis System; http://www.expasy.org/) bioinformatics resource portal of the Swiss Institute of Bioinformatics (SIB) provides many protein-analysis tools. One such tool is **ProtParam**,[8] which analyzes the physicochemical properties of proteins based on the sequence. ProtParam can be accessed directly at http://web.expasy.org/protparam/, or it can be accessed by first accessing ExPASy, then clicking the "Resources A..Z" link on the left, and finding ProtParam from the resource list. Mouse Slco1a6 protein was analyzed in ProtParam; the results are presented and explained in Figure 8.2. *ProtParam analyzes the sequence as is and does not take into account any post-translational modifications*. The output parameters are explained in the "Documentation" link on the ProtParam home page (http://web.expasy.org/protparam/protparam-doc.html).

8.4 PREDICTION OF PROTEASE DIGESTIBILITY

The protease digestibility prediction tool in ExPASy is called **PeptideCutter**,[8] which can be accessed directly at http://web.expasy.org/peptide_cutter/. Alternatively, it can be accessed by first accessing ExPASy, then clicking the "Resources A..Z" link on the left, and finding PeptideCutter from the resource list. There is a list of many proteases on the PeptideCutter home page. Specific enzymes can be selected from this list to map their cleavage sites in the protein. For example, analyzing mouse Slco1a6 protein in PeptideCutter to find only the pepsin cleavage sites (at pH >2) revealed that there are a total of 179 such sites (not shown). PeptideCutter can return the output as table, as a map of cleavage sites on the sequence itself, or both. *The analysis output marks the amino acid residue; the actual cleavage occurs at the right-hand side (C-terminal side) of this marked residue*. PeptideCutter also predicts potential cleavage sites of some chemicals in a given protein sequence.

8.5 HYDROPHOBICITY, HYDROPHILICITY, AND ANTIGENICITY PREDICTION, AND THE HYDROPATHY PLOT

The R-group of an amino acid determines whether it is hydrophobic or hydrophilic. Hydropathy is a

```
Molecular weight: 74145.2,
Theoretical pI: 8.36

Amino acid composition:
Ala (A)   33       4.9%, Arg (R)   21     3.1%, Asn (N)   29     4.3%
Asp (D)   19       2.8%, Cys (C)   30     4.5%, Gln (Q)   13     1.9%
Glu (E)   36       5.4%, Gly (G)   57     8.5%, His (H)    8     1.2%
Ile (I)   58       8.7%, Leu (L)   73    10.9%, Lys (K)   42     6.3%
Met (M)   24       3.6%, Phe (F)   39     5.8%, Pro (P)   30     4.5%
Ser (S)   49       7.3%, Thr (T)   43     6.4%, Trp (W)    7     1.0%
Tyr (Y)   23       3.4%, Val (V)   36     5.4%,

Extinction coefficients:
Extinction coefficients are in units of  M-1 cm-1, at 280 nm measured in
water.
   Ext. coefficient    74645
        Abs 0.1% (=1 g/l)   1.007, assuming all pairs of Cys residues
        form cystines
   Ext. coefficient    72770
        Abs 0.1% (=1 g/l)   0.981, assuming all Cys residues are
        reduced

Estimated half-life:
The N-terminal of the sequence considered is M (Met)
The estimated half-life is:
       30 hours (mammalian reticulocytes, in vitro)
       >20 hours (yeast, in vivo)
       >10 hours (Escherichia coli, in vivo)

Instability index:
The instability index (II) is computed to be 37.21
This classifies the protein as stable

Aliphatic index: 96.76

Grand average of hydropathicity (GRAVY): 0.267
```

FIGURE 8.2 **Partial ProtParam analysis output for Slco1a6.** The actual analysis contains more information. ProtParam analyzes the sequence as is and does not take into account any post-translational modifications. The **extinction coefficient** (E) indicates how much light a protein absorbs at a certain wavelength (e.g. 280 nm). It is useful to have an idea about the E value of a protein when purifying it. An approximate $E(Prot)_{280}$ = Tyr*E(Tyr) + Trp*E(Trp) + cystine*E(cystine); where E(Tyr) = 1490, E(Trp) = 5500, E(cystine) = 125 (cysteine does not absorb appreciably at wavelengths > 260 nm but cystine does). The approximate Abs_{280} = E(Prot)/MW (MW = molecular weight). For proteins rich in cysteines that do not form cystine (e.g. metallothionein), this calculation may have 10% or more error. ProtParam predicts an estimated **half-life** based on the "N-end rule," which relates the in vivo half-life of a protein to the identity of its N-terminal residues.[9] Note that ProtParam does not consider post-translational modifications, so the N-terminal-end-based rule does not account for any N-terminal modifications, which might significantly alter the predicted half-life. The **instability index** provides an estimate of the stability of the protein in a test tube. Statistical analysis of 12 unstable and 32 stable proteins has revealed that the occurrence of certain dipeptides is significantly different in the unstable proteins compared with the stable ones.[10] Based on the statistically determined weight value of instability, an instability index can be calculated. An instability index value < 40 predicts the protein to be stable; a value > 40 predicts that the protein may be unstable. The **aliphatic index** (X) of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine). X = X(Ala) + a*X(Val) + b*[X(Ile) + X(Leu)]; where X(Ala), X(Val), X(Ile), and X(Leu) are mole percent (100*mole fraction). The coefficients a and b are the volume of the valine side chain (a = 2.9) and of the Leu/Ile side chains (b = 3.9) relative to the side chain of alanine.[11] The **GRAVY** value for a peptide or protein is calculated as the sum of hydropathy values (Kyte and Doolittle) of all the amino acids, divided by the number of residues in the sequence. The hydropathy is discussed later in the chapter. A positive GRAVY value indicates that the protein is hydrophobic and a negative value indicates that it is hydrophilic.

measure of the hydrophobicity or hydrophilicity of an amino acid. Proteins are composed of both hydrophobic and hydrophilic amino acids, but the localization of these amino acids in the protein is related to the subcellular localization of the proteins (see Chapter 1 for a discussion on this subject). For example, proteins that are localized in an aqueous environment have hydrophobic amino acids (and their hydrophobic R groups) located towards the center of the molecule, away from water. In contrast, an integral membrane

protein always has a stretch of about 20 hydrophobic amino acids on the surface to enable it to pass through the membrane lipid bilayer. All hydrophilic amino acids are pushed to the outside of the membrane.

The hydropathy of amino acids is assigned specific values to create a **hydropathy scale**. There are different hydropathic scales; each scale assigns slightly different hydrophobicity or hydrophilicity values to the amino acids. Using a specific hydropathic scale the overall hydropathic character of a polypeptide can be determined, which is revealed by its **hydropathy plot**. Therefore, the hydropathy plot shows the hydrophobicity and hydrophilicity along the length of a polypeptide. Hydropathy is an important determinant of protein folding. One of the most widely used hydropathy plots is that of Kyte and Doolittle (1982).[12] The standard Kyte and Doolittle plot is a hydrophobicity plot. The plot is based on the consideration of the hydrophobic and hydrophilic properties of the 20 amino acids, shown in Table 8.1. Computation of the hydropathy plot requires setting a window size; the default is usually set at 7. The computation starts with the first window of amino acids (#1–7), the average hydrophobicity score of the first window is calculated and plotted as the midpoint of the window. Then the window moves by one amino acid, the second window spans amino acids #2–8, and the average hydrophobicity score of the second window is calculated and plotted as the midpoint of the window. This reiterative process continues until the last window at the end of the protein[c]. The averages are then plotted on a graph. The *y*-axis represents the hydrophobicity scores and the *x*-axis represents the window number/position of the amino acids. ExPASy provides **ProtScale**[8] (http://web.expasy.org/protscale/) that can be accessed to run the hydropathy plots. In addition to ExPASy, there are many more links providing online tools for the analysis of hydropathy plots of proteins. These links can be obtained by simply Googling the term.

In a hydrophobicity plot, hydrophilic amino acids receive negative values, whereas in a hydrophilicity plot, hydrophobic amino acids receive negative values.

Figure 8.3A shows the hydrophobicity plot of mouse Slco1a6 protein with a window size of 7. It is a transmembrane protein. Changing the window size to 21 clearly makes the transmembrane regions prominent (Figure 8.3B). A window size of 19 can also be used to visualize the transmembrane domains. Peaks above the line corresponding to 0 represent the hydrophobic regions and peaks below this line represent hydrophilic regions of the protein. The default window size in a Kyte and Doolittle plot is usually set at 7 or 9. An inverse Kyte and Doolittle plot will reverse these regions—that is, hydrophilic amino acids will be above the 0 axis and hydrophobic amino acids will be below the 0 axis.

Another widely used hydropathy plot, based on the Hopp and Woods hydropathy scale, is the Hopp and Woods hydrophilicity/antigenicity plot.[13] In this plot, hydrophilic amino acids get positive scores and hydrophobic amino acids get negative scores (Table 8.1). The Hopp and Woods hydropathy scale was developed for predicting potential antigenic sites in a polypeptide, which are likely to be rich in charged and polar residues. The default window size is usually set at 6 or 7; the regions of high hydrophilicity are likely to be antigenic sites. Figure 8.3C shows the Hopp and Woods plot of mouse Slco1a6 with a window size of 7.

TABLE 8.1 Hydrophobicity and Hydrophilicity Scores of Different Amino Acids

Amino Acid	Kyte–Doolittle	Hopp–Woods
Alanine	1.8	−0.5
Arginine	−4.5	3.0
Asparagine	−3.5	0.2
Aspartic acid	−3.5	3.0
Cysteine	2.5	−1.0
Glutamine	−3.5	0.2
Glutamic acid	−3.5	3.0
Glycine	−0.4	0.0
Histidine	−3.2	−0.5
Isoleucine	4.5	−1.8
Leucine	3.8	−1.8
Lysine	−3.9	3.0
Methionine	1.9	−1.3
Phenylalanine	2.8	−2.5
Proline	−1.6	0.0
Serine	−0.8	0.3
Threonine	−0.7	−0.4
Tryptophan	−0.9	−3.4
Tyrosine	−1.3	−2.3
Valine	4.2	−1.5

[c]Effective length of a polypeptide for hydropathy analysis = total # of windows of the desired size = total # of amino acids in the protein − window size + 1. For example, Slco1a6 has 670 amino acids. Hence, the effective length of Slco1a6 for hydropathy analysis = total # of windows of the desired size = 670 − 7 + 1 = 664. In other words, after the 664th amino acid, there are no more windows of 7 amino acids.

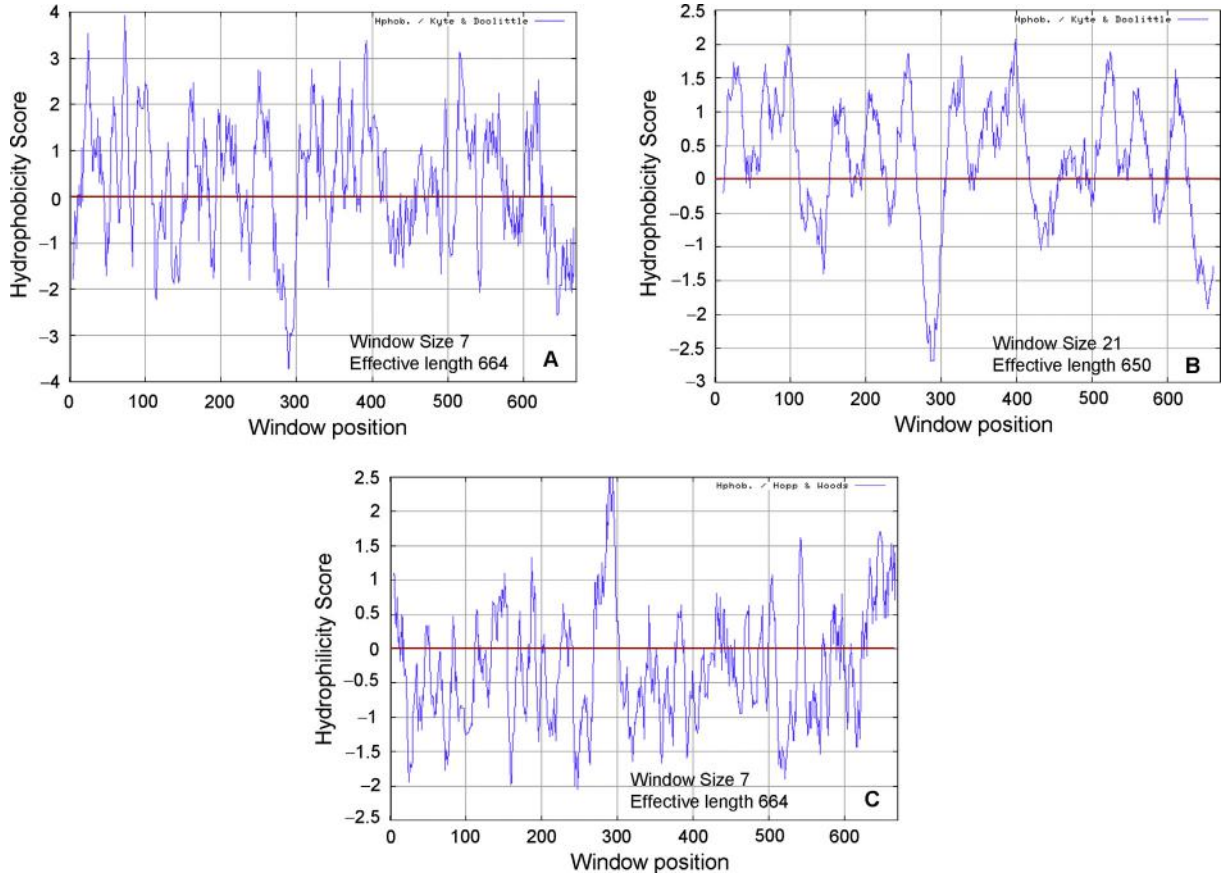


FIGURE 8.3 **Hydropathy plots.** Kyte and Doolittle plots and Hopp and Woods plot run in ProtScale at ExpaSy. (A) Kyte and Doolittle hydrophobicity plot of mouse Slco1a6 protein with a window size of 7. As a result, the effective length is 664—that is, after the 664th amino acid, another 7-amino-acid window is not available (the protein length is 670 amino acids). Peaks above the line corresponding to 0 represent the hydrophobic regions and peaks below this line represent hydrophilic regions of the protein. (B) Slco1a6 is a transmembrane protein. Thus, increasing the window size to 21 clearly makes the transmembrane regions prominent. This change makes the effective length 650. (C) Hopp and Woods hydrophilicity/antigenicity plot with a window size of 7. Peaks above the line corresponding to 0 represent the hydrophilic regions and peaks below this line represent hydrophobic regions of the protein.

When designing peptide antibodies, a Hopp and Woods hydropathy plot can be used to determine the regions of the polypeptide that are expected to have good antigenicity and thus trigger an antibody response in an animal treated with adjuvant-coupled peptide containing those sequence(s). Recently, Jääskeläinen et al. (2010)[14] investigated the prediction accuracy of 56 hydropathy scales by correlating predicted values with the accessible surface area in known 3D structures of proteins. They found that some epitopes are located among the most exposed regions, thereby reinforcing the utility of the hydropathy scales in predicting the antigenic regions of a protein.

Another metric of the overall hydrophobicity/hydrophilicity of a polypeptide is the **GRAVY** (grand average of hydropathy) score. The GRAVY value of a polypeptide is calculated by adding the hydropathy values of all the constituent amino acids and dividing the sum by the length of the sequence. *A positive GRAVY value indicates that the protein is hydrophobic and a negative value indicates that it is hydrophilic.*[12] Therefore, membrane proteins have higher GRAVY scores than globular proteins. ProtParam calculates the GRAVY score (Figure 8.2). The GRAVY score of mouse Slco1a6 is 0.267, indicating that it is a hydrophobic protein.

8.6 PREDICTION OF POST-TRANSLATIONAL MODIFICATION AND SORTING

Proteins can be post-translationally modified in many different ways, such as *N*-glycosylation, *O*-glycosylation and many other post-translational modifications. Proteins are also sorted (targeted) to various subcellular compartments either during translation (co-translational) or following translation (post-translational).

TABLE 8.2 Some Online Analysis Tools for Prediction of Post-Translational Protein Modifications, Protein Sorting, Localization Signals.

Online Tool	URL
CBS Prediction Servers (Center for Biological Sequence Analysis, Technical University of Denmark DTU)	http://www.cbs.dtu.dk/services/*
PSORT (Protein Sorting)	http://psort.hgc.jp/†
Gene Infinity	http://www.geneinfinity.org/sp/sp_proteinptmodifs.html‡

*Check CBS access policy to prediction servers at http://www.cbs.dtu.dk/cgi-bin/nph-access.
†PSORT program was coded by Kenta Nakai, Ph.D., Human Genome Center, Institute for Medical Science, University of Tokyo, Japan. Various scientists and their collaborators involved in developing different versions of the PSORT program are acknowledged on the PSORT home page.
‡Check the Terms of Service on the Gene Infinity home page.

For example, a large number of secretory proteins, membrane-bound proteins, and proteins in the endoplasmic reticulum are sorted co-translationally, whereas proteins targeted to the nucleus, mitochondria, and chloroplast are sorted post-translationally. Protein sorting requires specific signal sequences. In eukaryotic proteins, signal sequences are present at the N-terminal end of the protein. A comprehensive list of online analysis tools for the prediction of various post-translational protein modifications as well as protein sorting and localization signals can be found at the resources listed in Table 8.2.

8.7 SECONDARY-STRUCTURE PREDICTION

Efforts to predict protein secondary structures began long before the first protein structures were solved. Two of the earliest methods, the Chou–Fasman method and the GOR method, developed in the 1970s, have been widely used and are still being used.

8.7.1 The Chou–Fasman and GOR Methods

The Chou–Fasman and GOR (Garnier–Osguthorpe–Robson) methods were developed in the 1970s, and are among the oldest secondary-structure prediction methods. They are still widely used. The latest version of the GOR method is GOR V.[15] Both the Chou–Fasman and GOR methods are based on the analysis of the propensity of different amino acids to be in α-helix, β-strand, or β-turn. In these methods, the relative frequencies of amino acids in helix, strand, and turn are calculated based on known protein structures solved by X-ray crystallography. These relative frequency values are used to calculate the probability that an amino acid will appear in a helix, strand, or turn in a protein.

The application of the Chou–Fasman method is simple in principle. The sequence is scanned to identify regions of high helix or strand probability. For α-helix, a window size of six amino acids is used. If four contiguous residues out of six have P(α-helix) > 100, that segment is called as a helix. Once the helix is predicted, it is extended on both sides until at least four contiguous residues with P(α-helix) < 100 are found. That region is called as the end of the helix. For β-strand, a window size of five amino acids is used. The sequence is scanned to identify regions where at least three contiguous residues out of five have a value of P(β-strand) > 100. That region is called as a β-strand, and is extended on both sides until a set of three contiguous residues that have an average P(β-strand) < 100 is reached. That region is called as the end of the β-strand. If the residues in a region show the propensity of being in both α-helix and β-strand, the prediction is made based on the following principle: if $\Sigma[P(\alpha\text{-helix})] > \Sigma[P(\beta\text{-strand})]$, the region is called as a α-helix, otherwise a β-strand. Turns are also evaluated in four-residue windows, and are identified if $P(\beta\text{-turn}) > 0.000075$, where $P(\beta\text{-turn}) = f(i)*f(i+1)*f(i+2)*f(i+3)$. Table 8.3 shows the relative propensity values of amino acids as used by the Chou–Fasman method. Online Chou–Fasman and GOR prediction tools can be accessed from many sources (Table 8.4; see also CFSSP link in Table 8.5).

Like the Chou–Fasman method, the original GOR method also uses the propensity of amino acids to be in a helix, strand, turn, or coil. However, the GOR method uses a 17-residue window size and calculates the propensity of the residues in that window to be in each of the four states. The state with the highest score is predicted to be the state of the central residue (9th residue) of that window. Because the state of an amino acid is often influenced by the states of the neighboring amino acids, the GOR method takes into account the interactions of the neighboring residues.

With the availability of more sequences and more solved protein structures, some of the older methods have been revised and improved, such as GOR II, III, and IV.

8.7.2 Advances in Secondary-Structure Prediction

As the atomic detail of the structure of integral membrane proteins was determined in the mid-1980s, the homology-modeling method was developed as a

TABLE 8.3 Amino-Acid Relative Propensity Values Used by the Chou–Fasman Method

Amino Acid	P (α-helix)	P (β-strand)	P (β-turn)	f(i)	f(i + 1)	f(i + 2)	f(i + 3)
Alanine	142	83	66	0.06	0.076	0.035	0.058
Arginine	98	93	95	0.070	0.106	0.099	0.085
Asparagine	67	89	156	0.161	0.083	0.191	0.091
Aspartic acid	101	54	146	0.147	0.110	0.179	0.081
Cysteine	70	119	119	0.149	0.050	0.117	0.128
Glutamic acid	151	037	74	0.056	0.060	0.077	0.064
Glutamine	111	110	98	0.074	0.098	0.037	0.098
Glycine	57	75	156	0.102	0.085	0.190	0.152
Histidine	100	87	95	0.140	0.047	0.093	0.054
Isoleucine	108	160	47	0.043	0.034	0.013	0.056
Leucine	121	130	59	0.061	0.025	0.036	0.070
Lysine	114	74	101	0.055	0.115	0.072	0.095
Methionine	145	105	60	0.068	0.082	0.014	0.055
Phenylalanine	113	138	60	0.059	0.041	0.065	0.065
Proline	57	55	152	0.102	0.301	0.034	0.068
Serine	77	75	143	0.120	0.139	0.125	0.106
Threonine	83	119	96	0.086	0.108	0.065	0.079
Tryptophan	108	137	96	0.077	0.013	0.064	0.167
Tyrosine	69	147	114	0.082	0.065	0.114	0.125
Valine	106	170	50	0.062	0.048	0.028	0.053

TABLE 8.4 Some Online Chou–Fasman and GOR Prediction Tools

Chou–Fasman and GOR Prediction Tool	URL
University of Virginia	http://fasta.bioch.virginia.edu/fasta_www2/fasta_www.cgi?rm = misc1* (select Chou–Fasman or GOR method)
ProtScale at ExPASy	http://web.expasy.org/protscale/[8] (select Chou–Fasman or GOR method)
Center for Informational Biology, Japan	http://cib.cf.ocha.ac.jp/bitool/MIX/† (select Chou–Fasman or GOR method)

*©1988, 2006, by William R. Pearson and the University of Virginia.
†The home page cites the papers based on which the method implemented in this server was developed. The Chou–Fasman and GOR papers are cited elsewhere in the text.

way of predicting secondary structures. In **homology modeling**, the secondary structure of the target protein is predicted based on the known structure of homologous proteins (template). Hence, homology modeling is based on sequence similarity/identity; obviously, the higher the sequence similarity/identity between the target and the template, the greater is the chance of accuracy of prediction. Nevertheless, homology modeling may not accurately predict the side chains and folds, making the overall predictions less accurate.

With advances in computation techniques, increase in the number of database entries, and increased knowledge of various protein folds, the concept of protein **sequence–structure threading** developed in the 1990s. In **protein threading** (**fold recognition**), target sequence is mapped to known template structures from the database. The sequence–structure compatibility is assessed by a scoring function. The method is based on the premises that, (1) there is a far lower number of unique folds among proteins than there are known proteins, and (2) information on the physicochemical properties of amino acids and knowledge of their occurrence in different structural environments provide important clues to their potential occurrence among different types of folds. Energy functions are an important consideration because energetics is very important in folding. *During computation of threading, the threading with minimum energy is assumed to represent the most likely fold structure.*

TABLE 8.5 Some Online Tools for the Analysis of Possible Secondary Structure of a Protein

Online Tool	Comments and URL
APSSP	http://imtech.res.in/raghava/apssp/*
CFSSP (Chou–Fasman[16] Secondary-Structure Prediction)	http://www.biogem.org/tool/chou-fasman/†
GOR IV	GOR IV[17]; GOR I, the original GOR[18] (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html)‡[19]
HMMSTR	HMM-based[20,21] (http://www.bioinfo.rpi.edu/bystrc/hmmstr/server.php)
JPred 3	Combines the analysis from multiple prediction algorithms, such as DSC, JNET, PHD, and PREDATOR[22] (http://www.compbio.dundee.ac.uk/www-jpred/)
NPS@ (Network Protein Sequence Analysis)	This site contains links to a number of prediction tools including GOR and PHD. However, GOR and PHD are mentioned here separately as well. Pay attention to those that were developed in the late 1990s. Compare the output from these tools‡[19] (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_server.html)
PHD	Neural-network-based[23–25] (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_phd.html)
PredictProtein	Meta-server that combines the analysis from multiple prediction algorithms such as Jpred, PHD, PROF, and PSIPRED. It is a good secondary-structure prediction program** (https://www.predictprotein.org/)
PROTEUS 2	Combination of HMM- and neural-network-based prediction[26] (http://wishart.biology.ualberta.ca/proteus2/)
PSIPRED	Combination of homology modeling and neural-network-based prediction. It is a good secondary-structure prediction program[27] (http://bioinf.cs.ucl.ac.uk/psipred/)
Quick2D	Provides an overview of secondary-structure features like α-helices, extended β-sheets, coiled coils, transmembrane helices, and disordered regions. Predictions by PSIPRED, JNET, Prof(Rost), Prof (Ouali), Coils, MEMSAT2, HMMTOP, DISOPRED2 and VSL2†† (http://toolkit.tuebingen.mpg.de/quick2_d)
SCRATCH Protein Predictor	The SCRATCH software suite includes predictors for a number of parameters, such as secondary structure, relative solvent accessibility, disordered regions, domains, individual residue contacts, tertiary structure, and more[28] (http://scratch.proteomics.ics.uci.edu/index.html)
SSPro 4.0	Bidirectional recurrent neural network (BRNN)-based[29,30] (http://download.igb.uci.edu/sspro4.html)
SYMPRED	SYMPRED can be run using any combination of the following programs: PHD, PROF, SSPro2.01, YASPIN, JNet, and PSIPRED. The consensus of the outputs is derived through dynamic programming to achieve a higher level of prediction accuracy[31] (http://www.ibi.vu.nl/programs/sympredwww/)
SOPMA	An improved self-optimized prediction method (SOPM)[32] (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopma.html)
YASPIN	Neural-network-based[33] (http://www.ibi.vu.nl/programs/yaspinwww/)

*An advanced version of the PSSP server.[34]

†© 2012, BioGem.Org.

‡Service supported by Ministère de la recherche (ACI IMPBio, ACC-SV13), CNRS (IMABIO, COMI, GENOME) and Région Rhône-Alpes (Programme EMERGENCE). The "Abstract" link can be clicked to obtain all the original references.

**The website provides a link to the entire PredictProtein team.

††© 2008, Dept. of Protein Evolution at the Max Planck Institute for Developmental Biology, Tübingen.

TABLE 8.6 Some Online Prediction Tools for Coiled Coils and Zippers

Online Tool	Comments and URL
ExPASy COILS	COILS compares the input sequence to a database of known parallel two-stranded coiled coils and derives a similarity score. By comparing this score to the scores in globular and coiled-coil proteins, COILS calculates the probability that the sequence will adopt a coiled-coil conformation[35] (http://embnet.vital-it.ch/software/COILS_form.html)
Paircoil2 at MIT	New version of the Paircoil program, which uses pairwise residue probabilities to detect coiled-coil motifs. Paircoil2 achieves 98% sensitivity and 97% specificity on known coiled coils[36] (http://groups.csail.mit.edu/cb/paircoil2/paircoil2.html)
2ZIP	Combines a standard coiled-coil-prediction algorithm with an approximate search for the characteristic leucine repeat. No further information from homologs is required for prediction[37] (http://2zip.molgen.mpg.de/)

Advances in protein-threading algorithms have allowed more accurate fold prediction. Secondary-structure prediction has further benefited from the introduction of methods like neural networks, hidden Markov models (HMMs), and the ability to train new models on an extensive set of sequence and structural data.

There are a number of online tools available for the analysis of possible secondary structure of a protein. ExPASy provides links to many of these tools. The links in Table 8.5 are cited because the analysis can be done in real time using most of these tools and the output is quickly obtained. There are many more online secondary-structure predictions tools that are not cited here.

These tools predict various secondary structures that different parts of the polypeptide can assume, such as the α-helix, 3_{10}-helix, π-helix, extended strand, β-turn, random coil, or ambiguous state. Analyzing a polypeptide sequence using different prediction tools may not produce the same results. For example, analyzing mouse Slco1a6 using four of these tools produces the following results: the prediction of α-helix varies between ~23 and 38%, that of extended strand varies between ~11 and 27%, and that of random coil varies between 42 and 51%. It is therefore advisable to analyze the sequence using multiple programs. Some of the standard notations in the output are as follows: α-helix (H/h), 3_{10}-helix (G/g), π-helix (I/i), extended strand (E/e), β-turn (T/t), random coil (C/c).

Online tools for the prediction of coiled coils and zippers are shown in Table 8.6. The direct link for ExPASy COILS is given in the table. It can also be accessed by first accessing ExPASy (http://www.expasy.org/), then accessing COILS by clicking "Resources A..Z".

8.7.3 Predicting the Accuracy of Secondary-Structure Prediction

A widely used metric to determine the overall accuracy of secondary-structure prediction is the **Q3 score**. A Q3 score is a measure of the quality of prediction of all three states (helix, strand, and coil), and it represents the percentage of residues that are correctly predicted (the states of the residues). The Q3 score can range from 0 to 1; 1 being the perfect prediction (100%). Currently, almost all secondary-structure-prediction algorithms achieve a Q3 score of 0.75 or higher. It should be remembered that Q3 is not an absolute measure of the prediction accuracy; there are other measures as well.

8.8 PREDICTION OF DOMAINS AND MOTIFS

A domain is part of the tertiary structure of protein. Each domain is a discrete globular unit that folds independently of the rest of the protein. Domains have specific functional roles. Domains can be composed of as few as 20–25 amino acids, but frequently much more than 25. The average number of domains in a protein is usually two to three, but can be more. By shuffling a finite number of domains, nature has created proteins with diverse functions during evolution. Thus, proteins with similar functions are expected to contain conserved regions that are associated with the function; the rest of the protein sequence may be different. Examples of some familiar domains are the **SH3 (Src-homology 3) domain**, which is around 50 amino acids and involved in protein–protein interactions; the **chromo (chromatin organization modifier) domain**, which is 30–70 amino acids and involved in the assembly of protein complexes on chromatin; and the **death domain**, which is around 80–100 amino acids and involved in apoptotic signal transduction.

As opposed to domains, a specific functional element of the protein that usually does not fold independently of the rest of the protein is called a **motif**, such as a sequence motif or a structural motif (e.g. a stretch of secondary structure). Domains contain within themselves specific motifs that are critical to domain function. Some examples of structural motifs in proteins are various loop and turns, such as omega loops, beta turns, helix–loop–helix, and helix–turn–helix. Sometimes the terms domain and motif are used interchangeably in the context of proteins, such as "coiled-coil" domain/motif, "leucine-zipper" domain/motif.

The domain analysis of Slco1a6 using **InterProScan** (http://www.ebi.ac.uk/Tools/pfa/iprscan/)[38] at the European Molecular Biology Laboratory's European Bioinformatics Institute (EMBL-EBI) is shown in Figure 8.4 and Figure 8.5. At the default setting, all

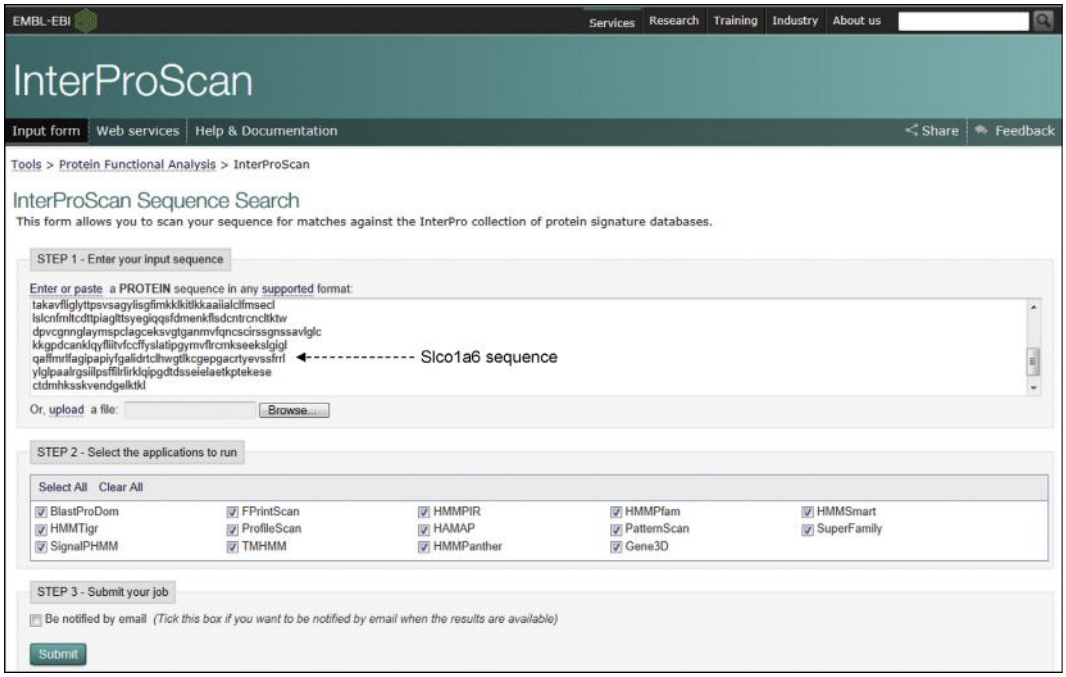


FIGURE 8.4 **InterProScan home page at EMBL-EBI from where the search and analysis can be launched.** The page shows that at the default setting all applications are checked; each one scans the input sequence against a specific database.

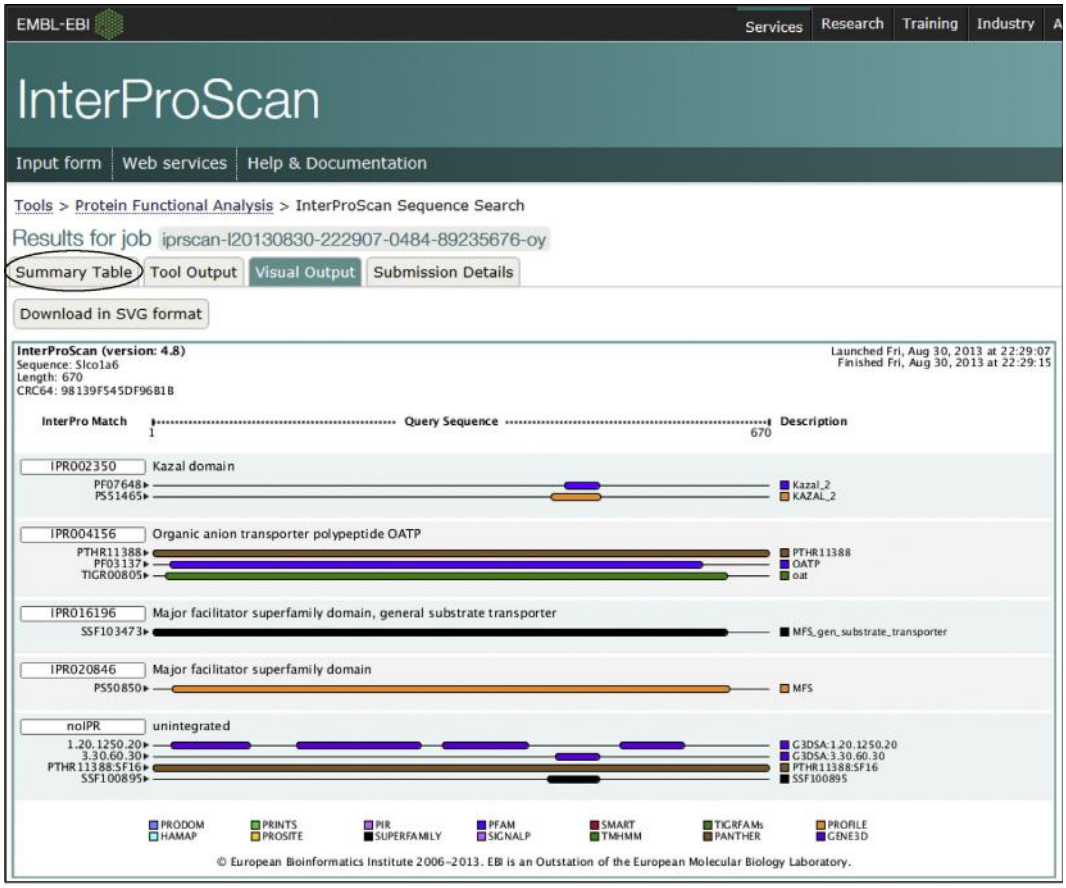


FIGURE 8.5 **The graphical display of InterProScan analysis.** Two major domains identified are Kazal and MFS. More information on these domains can be obtained from various links under the "Summary Table" tab. The predictions from different databases may not be identical (see text). Nevertheless, these tools are very important in identifying specific signatures in protein sequence.

applications are checked; each one scans the input sequence against a specific database (see "Help & Documentation" for details; Figure 8.4). The graphical display of the analysis is shown in Figure 8.5. Two major domains identified are **Kazal** and **MFS** (see Box 8.1). Clicking "Summary Table" shows various links for more information on the domains and their distribution. The predictions from different databases may not be identical; for example, PROFILE predicts the Kazal domain spanning from residue 433 to 488, whereas Pfam predicts the Kazal domain spanning from residue 447 to 486. PROFILE predicts the MFS domain spanning from residue 21 to 627, whereas SuperFamily predicts the MFS domain spanning from residue 1 to 625. Despite small differences in prediction, these tools are very important in identifying specific sequence signatures in protein sequence.

The domain analysis of Slco1a6 using the NCBI CDD is shown in Figure 8.6, Figure 8.7, and Figure 8.8. **CDD** (**Conserved Domain Database**) of NCBI provides annotation of protein sequences with the location of conserved-domain footprints and functional sites inferred from these footprints. CDD is built on NCBI-curated domains and data imported from Pfam, SMART, COG, PRK, and TIGRFAM.[39] CDD can be accessed directly at http://www.ncbi.nlm.nih.gov/cdd, or from the NCBI home page. Figure 8.6 shows the CDD home page. Clicking "**CD-Search**" (circled) takes the user to the search launch page, shown in Figure 8.7. Submitting the Slco1a6 sequence in FASTA format under default settings returns the analysis shown in Figure 8.8. The result can be displayed in a "concise format" that displays the best hits, or "full format" that displays all hits. Figure 8.8 shows the concise format. Like InterProScan, CDD analysis also shows that Slco1a6 contains **Kazal** (**Kazal_SLC21**) and **MFS** domains. However, the predicted MFS domain is shorter (21–270) than that predicted by InterProScan (PROFILE).

It should be remembered that the domain/motif prediction is predicated on sequence alignment. Just like with any other

BOX 8.1

KAZAL AND MFS DOMAINS

The activity of proteases in cells is under tight control to prevent any unintended tissue damage. Cells produce various types of proteases along with peptide protease inhibitors to regulate the protease activity. Serine protease* activities are regulated by serine protease inhibitors, which are distributed in a wide range of organisms from all kingdoms of life. Pancreatic acinar cells produce two types of serine protease inhibitors; one is the **Kunitz** inhibitors (e.g. PTI, or pancreatic trypsin inhibitor) that remain in the pancreatic cells, and the other is **Kazal** inhibitors (e.g. PSTI, or pancreatic secretory trypsin inhibitor) that are secreted with the zymogens in the pancreatic juice. Some other examples of Kazal-type inhibitors are avian ovomucoid, acrosin inhibitor, and elastase inhibitor. Kazal-type inhibitors are the most studied protease inhibitors, and they contain one or more Kazal-type domains. The typical **Kazal domain** is a small α/β fold, consisting of one α-helix surrounded by an adjacent three-stranded β-sheet and loops of peptide segments†.[40]

The major facilitator superfamily (**MFS**) is the largest known superfamily of **secondary transporters** found in living organisms. Secondary transporters do not use ATP directly for transport, but use an already-existing electrochemical gradient‡. More than 70 families are known; members of each family transport a different set of related compounds, such as simple monosaccharides, oligosaccharides, amino acids, peptides, vitamins, enzyme cofactors, drugs, nucleobases, nucleosides, nucleotides, and organic and inorganic anions and cations. MFS proteins are single-polypeptide secondary transporters, and the **MFS domain** consists of either 12 or 14 transmembrane helices connected by hydrophilic loops**.[42,43] Secondary active transport can move materials against the concentration gradient, and can also transport just one substrate (uniporter), or two substrates in the same direction (symporter), or in the opposite direction (antiporter).

*Serine proteases contain a reactive serine in their active site and this serine is crucial for their function. Trypsin, chymotrypsin, and elastase are three important eukaryotic serine proteases; subtilisin is an important bacterial serine protease. Trypsin is involved in the activation of pancreatic zymogens. Serine proteases also constitute over one-third of all proteases[41]

†http://www.ebi.ac.uk/interpro/entry/IPR002350; http://prosite.expasy.org/PDOC00254

‡An electrochemical gradient is a gradient of electrochemical potential, which is generated by the differential distribution of electrical potential and chemical concentration across the membrane. Differential distribution of ions across the membrane, for example sodium ions, generates an electrochemical gradient. It consists of two components: the electrical potential difference caused by the uneven distribution of the charge, and the concentration difference caused by the uneven distribution of sodium itself. The electrochemical gradient generates potential energy because the ions involved are ready to move across the membrane. However, the ions cannot pass through the membrane lipid bilayer without the help of an active transport mechanism. The MFS transporters convert this potential energy into kinetic energy when they transport the respective substrates

**http://www.ebi.ac.uk/interpro/entry/IPR016196;jsessionid; http://pfam.sanger.ac.uk/clan/CL0015

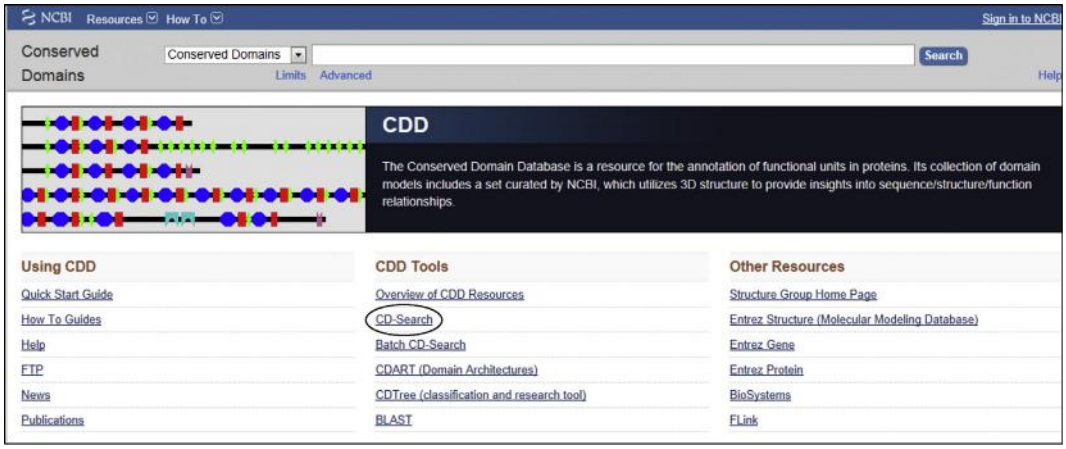


FIGURE 8.6 **The Conserved Domain Database (CDD) home page.** Clicking CD-search (circled) takes the user to the search and analysis launch page (Figure 8.7).

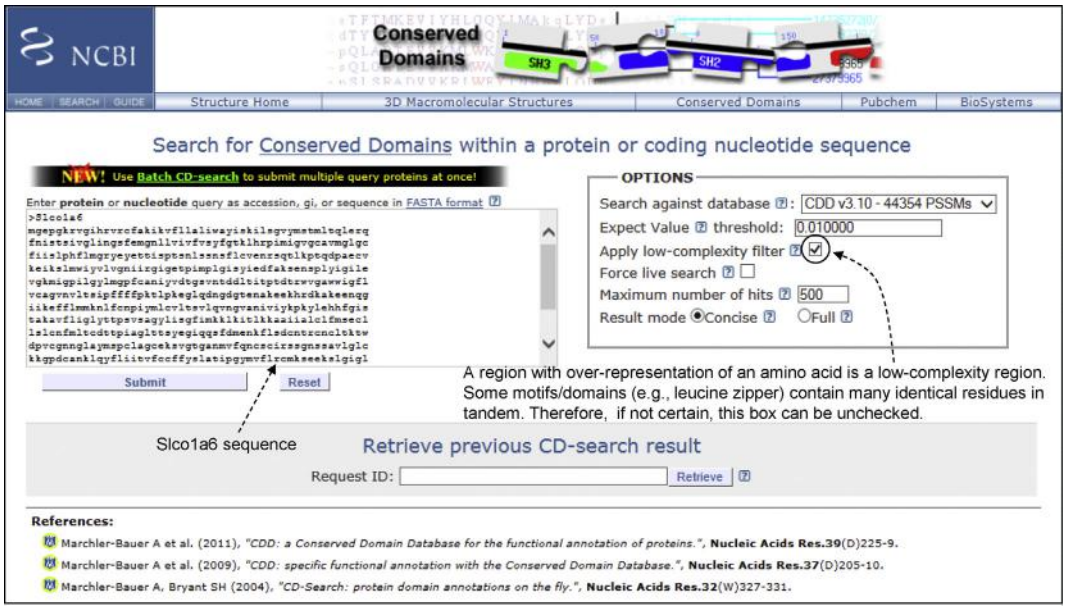


FIGURE 8.7 **The CDD search and analysis launch page.** Submitting the Slco1a6 sequence in FASTA format under default settings returns the analysis shown in Figure 8.8. In the default settings, the "low-complexity" filter is on. This can be turned off.

predictions, there is an element of uncertainty—that is, a domain may be falsely predicted or a true domain may be missed, particularly conformational domains.

Another good online tool for domain analysis is **PROSITE** (http://prosite.expasy.org/prosite.html).[44,45] PROSITE scan (**ScanProsite**) of Slco1a6 produces the following results: **Kazal** domain spanning residues 433–488 and **MFS** domain spanning residues 21–627 (not shown).

8.8.1 Transmembrane-Helix Prediction

Because domain analysis shows the existence of an MFS domain in Slco1a6, a specific search for the transmembrane (TM) helices can be done. There are a number of good online TM-helix-prediction tools, as shown in Table 8.7.

RHYTHM produces a nice graphical output of TM helices, showing the amino-acid sequence in each helix. Figure 8.9 shows the gist of TM-helix prediction by all four prediction tools. TMHMM (version 2.0) predicted 11 TM helices, whereas RHYTHM, OCTOPUS, and Phobius each predicted 12 TM helices (Figure 8.9). The graphical outputs of RHYTHM and OCTOPUS are shown in Figure 8.10. In the span of residue 110 to residue 240 (approximately), TMHMM predicted one TM helix, whereas RHYTHM, OCTOPUS, and Phobius predicted two. As a result, the assignment of inside and

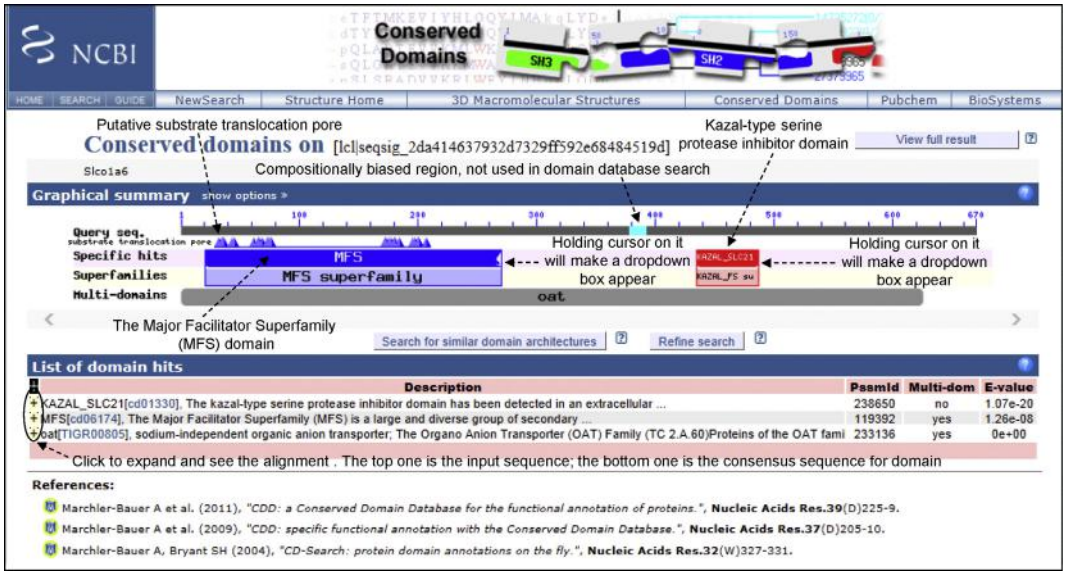


FIGURE 8.8 **Result of CDD domain analysis.** The result is displayed in the "concise format." Analysis shows that Slco1a6 contains Kazal (Kazal_SLC21) and MFS domains. The predicted MFS domain is shorter (21–270) than that predicted by InterProScan (see text). Holding the cursor over MFS or Kazal_SLC21 produces a drop-down box that contains detailed description of the specific hit.

TABLE 8.7 Some Online Tools for Transmembrane-Helix Prediction

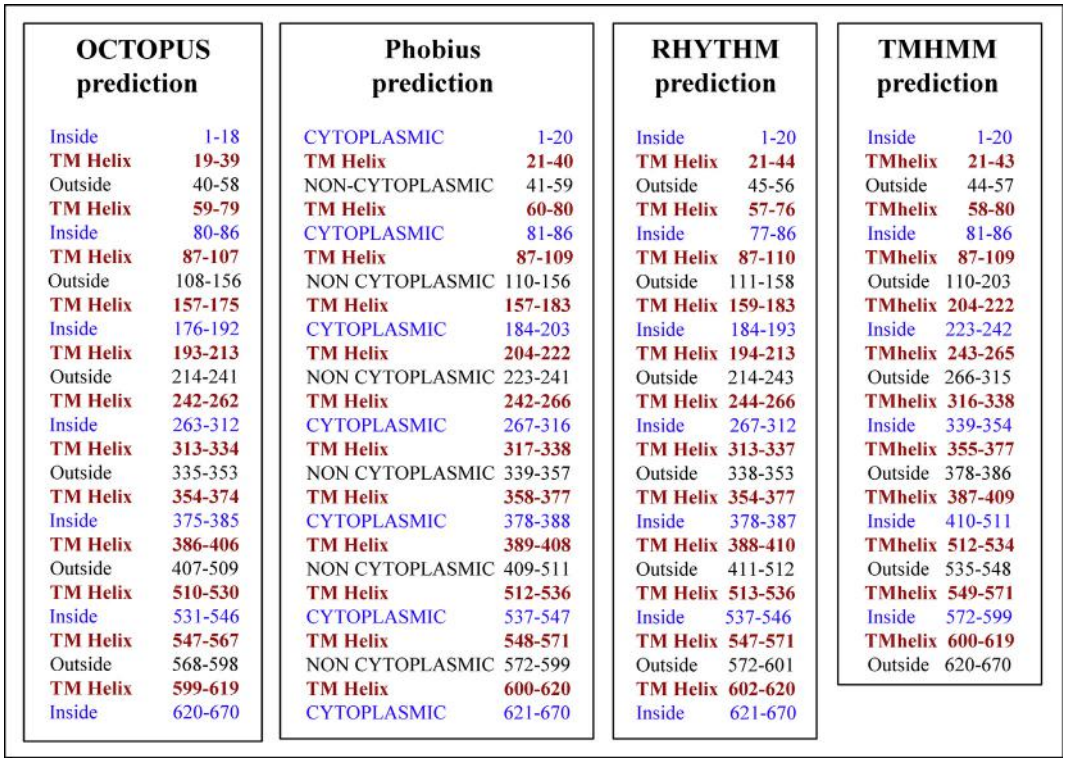
Online Tool	Comments and URL
TMHMM	Hidden-Markov-model-based[46] (http://www.cbs.dtu.dk/services/TMHMM/)
RHYTHM	Utilizes the structural information from ever-growing data sets and evolutionary information from conserved-sequence patterns in a representative data set of membrane proteins[47] (http://proteinformatics.charite.de/rhythm/)
Phobius	Hidden-Markov-model-based[48] (http://phobius.sbc.su.se/)
OCTOPUS	Artificial-neural-network-based[49] (http://octopus.cbr.su.se/)

outside segments is reversed between the TMHMM prediction and those of the other three programs from residue 214/223 onwards. However, TMHMM is a widely used, good TM-helix-prediction program, and TMHMM prediction is focused on TM helices only and not necessarily on the cytoplasmic and the extracellular segments. Overall, the TM helices were predicted correctly by all four programs. Nevertheless, this example further underscores the fact that it is a good idea to run an analysis simultaneously using multiple programs.

8.9 VIEWING THE 3D STRUCTURE OF PROTEINS (AND OTHER BIOLOGICAL MACROMOLECULES)

The 3D structures of many proteins and other biological macromolecules have been determined using various techniques of modern structural biology. These structures are deposited in the **PDB** (**Protein Data Bank**) database and are given a PDB ID. The PDB ID is a four-character unique identifier, consisting of numbers and letters, assigned to a protein or other biological macromolecule submitted to the PDB. The PDB is an archive of the structure of proteins and other biological macromolecules; the structures have been determined using techniques like X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy, and cryo-electron microscopy. After structural information is submitted to the PDB, the submission is annotated and publicly released by the **wwPDB** (http://www.wwpdb.org/). As of July 30, 2013, there were 92,689 structures in the PDB. PDB IDs are usually written in uppercase. Some examples of PDB IDs are 2HHD (human hemoglobin, deoxy form), 9INS (pig insulin), and 2VRY (mouse neuroglobin). The PDB can be searched by simply typing the description, or partial sequence, or the PDB ID (if known).

FirstGlance in Jmol (http://bioinformatics.org/firstglance/fgij/index.htm) is a user interface to the free molecular visualization program named **Jmol** (http://jmol.sourceforge.net/). **Jmol** is a free and



OCTOPUS prediction		Phobius prediction		RHYTHM prediction		TMHMM prediction	
Inside	1-18	CYTOPLASMIC	1-20	Inside	1-20	Inside	1-20
TM Helix	19-39	TM Helix	21-40	TM Helix	21-44	TMhelix	21-43
Outside	40-58	NON-CYTOPLASMIC	41-59	Outside	45-56	Outside	44-57
TM Helix	59-79	TM Helix	60-80	TM Helix	57-76	TMhelix	58-80
Inside	80-86	CYTOPLASMIC	81-86	Inside	77-86	Inside	81-86
TM Helix	87-107	TM Helix	87-109	TM Helix	87-110	TMhelix	87-109
Outside	108-156	NON CYTOPLASMIC	110-156	Outside	111-158	Outside	110-203
TM Helix	157-175	TM Helix	157-183	TM Helix	159-183	TMhelix	204-222
Inside	176-192	CYTOPLASMIC	184-203	Inside	184-193	Inside	223-242
TM Helix	193-213	TM Helix	204-222	TM Helix	194-213	TMhelix	243-265
Outside	214-241	NON CYTOPLASMIC	223-241	Outside	214-243	Outside	266-315
TM Helix	242-262	TM Helix	242-266	TM Helix	244-266	TMhelix	316-338
Inside	263-312	CYTOPLASMIC	267-316	Inside	267-312	Inside	339-354
TM Helix	313-334	TM Helix	317-338	TM Helix	313-337	TMhelix	355-377
Outside	335-353	NON CYTOPLASMIC	339-357	Outside	338-353	Outside	378-386
TM Helix	354-374	TM Helix	358-377	TM Helix	354-377	TMhelix	387-409
Inside	375-385	CYTOPLASMIC	378-388	Inside	378-387	Inside	410-511
TM Helix	386-406	TM Helix	389-408	TM Helix	388-410	TMhelix	512-534
Outside	407-509	NON CYTOPLASMIC	409-511	Outside	411-512	Outside	535-548
TM Helix	510-530	TM Helix	512-536	TM Helix	513-536	TMhelix	549-571
Inside	531-546	CYTOPLASMIC	537-547	Inside	537-546	Inside	572-599
TM Helix	547-567	TM Helix	548-571	TM Helix	547-571	TMhelix	600-619
Outside	568-598	NON CYTOPLASMIC	572-599	Outside	572-601	Outside	620-670
TM Helix	599-619	TM Helix	600-620	TM Helix	602-620		
Inside	620-670	CYTOPLASMIC	621-670	Inside	621-670		

FIGURE 8.9 **Transmembrane-helix prediction at a glance by RHYTHM, OCTOPUS, Phobius, and TMHMM.** TMHMM (version 2.0) predicted 11 TM helices, whereas RHYTHM, OCTOPUS and Phobius predicted 12 (see text for details). This example underscores the fact that it is a good idea to run an analysis simultaneously using multiple programs.

open-source software program written in Java for viewing chemical structure in 3D. It runs on various operating systems, such as Windows, MacOS, and Unix, and is also downloadable. The Jmol website has a user-friendly tutorial. **FirstGlance in Jmol** provides an easy way to look at the 3D structures of proteins, DNA, RNA, and their complexes, including with animation. In order to use **FirstGlance in Jmol**, one has to know the PDB ID of the macromolecule or have the data as PDB file format. On the **FirstGlance in Jmol** website, help is displayed automatically with links to further information about structural biology terms and concepts. The website also provides links to a "Gallery of Interactive Molecules" and a "Snapsot Gallery." Therefore, between the **Jmol** tutorial and **FirstGlance in Jmol** helpful links, the beginner will find it quite easy to understand the output.

8.10 ALLERGENIC PROTEIN DATABASES AND PROTEIN-ALLERGENICITY PREDICTION

Substances that cause allergic reactions are called **allergens**. Almost all allergens are proteins and they induce allergic response in susceptible individuals. Because allergic reactions result from complex interactions between the allergenic proteins and the immune system (see footnote on epitopes), and because allergic reactions are seen only in susceptible individuals, the allergenic potential of proteins is difficult to predict.

8.10.1 WHO/IUIS Allergen Nomenclature and Database of Allergenic Proteins

The World Health Organization/International Union of Immunological Societies (WHO/IUIS) Allergen Nomenclature Subcommittee is responsible for developing a systematic Linnaean nomenclature of allergens and maintaining a database of confirmed allergenic proteins.[50,51] A Linnaean nomenclature of an organism consists of a genus and a species term. The allergen name is normally made up of the first three letters of the genus name, first one letter from the species name, and a number that represents the order of its identification. In some instances, this rule has to be modified, such as Asp fl 13 (from *Aspergillus flavus*) and Asp f 13 (from *Aspergillus fumigatus*). Note that for *Aspergillus flavus* Asp fl 13, two letters from the species name, instead of one letter, have been used.

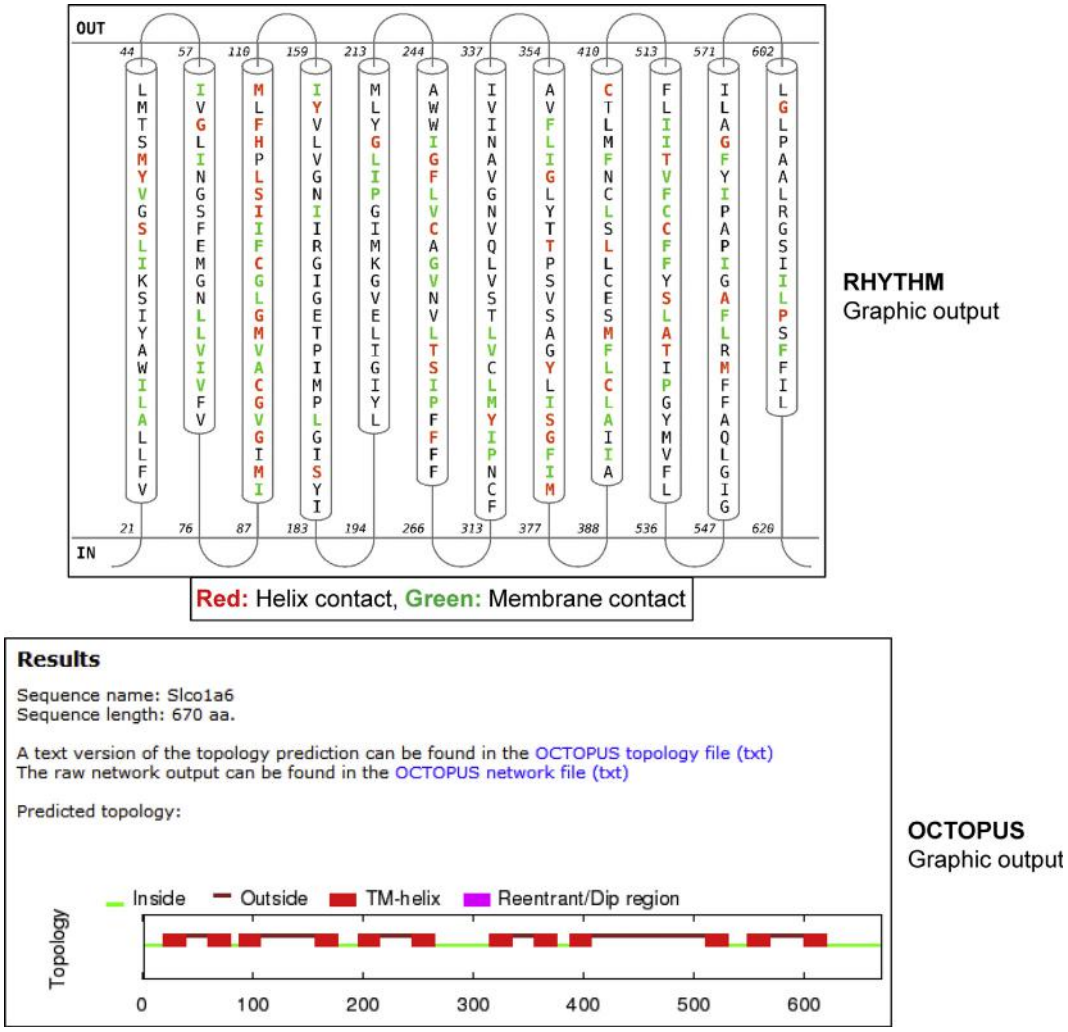


FIGURE 8.10 **The graphical outputs of RHYTHM and OCTOPUS.** The RHYTHM graphical output shows the relative length of the predicted helices and the amino-acid sequence of each predicted helix, as well as the residues that are in contact with the membrane and the residues involved in helix contact.

The WHO/IUIS allergen database contains information of approved and officially recognized allergens—that is, for a protein to be designated an allergen by WHO/IUIS, the allergenicity of the protein should be clinically documented. The database can be quickly searched for an allergen or an allergen source on the home page (http://www.allergen.org/index.php). Alternatively, an advanced search can be performed on the search page by clicking the "**Search**" tab or using the direct link http://www.allergen.org/search.php. By clicking the "**Tree View**" tab or using the direct link http://www.allergen.org/treeview.php, a list of allergens in fungi, plants, and different animal phyla can be directly obtained. An allergen record shows much important information about the allergen, such as the source, the evidence of allergenicity, allergenicity reference in PubMed, information on whether the allergen is a food allergen or not, any isoallergens and variants, and finally the sequence in both GenBank and UniProt.

8.10.2 Other Databases of Allergenic Proteins

In addition to the WHO/IUIS database, there are a number of other databases of allergenic proteins. Three of these databases are described in Chapter 5 (the **Structural Database of Allergenic Proteins (SDAP)**, **Allergenonline**, and **Allermatch**). Both the SDAP and Allergenonline databases are periodically updated; they both list more than 1500 allergenic proteins from food and non-food sources. Many allergens listed in these databases do not have IUIS designations yet. For a more comprehensive list of currently available allergen databases and allergen semantics, see Gendel[52] and other publications by the same author referenced in the paper.

8.10.3 Linear Epitopes, Conformational Epitopes, and Allergenicity

Although a protein acts as an allergen, the immune system actually recognizes smaller sections of the protein to trigger an allergic response. These small segments of the allergenic protein are called allergenic determinants, or **epitopes**[d]. The cognate antibody (IgE) binds to these allergenic epitopes to trigger the allergic response. Epitopes can be linear or conformational. In a **linear epitope**, the amino-acid sequence is continuous, whereas in a **conformational epitope**, the 3D conformation of the protein brings two separate sequence segments together to create the epitope. Conformational epitopes are usually destroyed when the protein is denatured, but linear epitopes are not affected by denaturation. Because many food allergens are stable in heat processing and digestion, it has been proposed that linear epitopes are more important than conformational epitopes for food allergens. However, the allergenicity of some foods, such as cow's milk and egg, is partly due to the IgE-binding conformational epitopes of their constituent proteins, such as α- and β-casein in cow's milk and ovomucoid in egg. Individuals whose immune system reacts to these conformational epitopes tend to grow out of the allergy as they get older, but reaction to the linear epitopes results in persistent allergy.[53–55] Conformational epitopes are also important for environmental allergens that are primarily inhaled.[56]

8.10.4 Allergenicity-Prediction Paradigm

Bioinformatics tools have been developed to identify the allergenic potential of an unknown protein by comparing its sequence to the sequences of known allergenic proteins in the database. A paradigm for assessing the allergenic potential of a protein in food was developed by the Food and Agricultural Organization/World Health Organization (FAO/WHO) as part of a multi-step safety-assessment process for foods produced through agricultural biotechnology.[57] The FAO/WHO paradigm uses two criteria: (1) an exact match of 6 contiguous amino acids, and (2) an overall sequence identity of more than 35% in a sliding window of 80 amino acids. Any protein that satisfies one or both of these criteria should trigger additional investigation to confirm whether the protein may truly have allergenic potential.

At the time the FAO/WHO paradigm was developed, it was already known that the smallest IgE-binding epitopes in an allergen could be only six-amino-acids long, as had been reported for Ara h 1 and Ara h 2.[58,59] The findings in these publications were based on epitope mapping with synthetic peptides that reacted with serum IgE from individuals with documented peanut hypersensitivity. Also, a publication by Burkhard Rost[60] had described the basis for a 35% identity cutoff and 80-amino-acid window threshold in pairwise sequence alignment. The author reported that protein pairs with similar structure (and function) are likely to have $>35\%$ sequence identity. The author analyzed more than a million sequence alignments between protein pairs of known structure. The goal was to distinguish between true and false positives for low levels of similarity. The author noted that sequence alignments could unambiguously distinguish between protein pairs of similar and non-similar structure when the pairwise sequence identity was $>40\%$ for long alignments. The signal, however, became blurred when the sequence identity was between 20 and 35%; this $20-35\%$ range was termed the **twilight zone** of sequence identity. The pairwise sequence identity by itself is not meaningful without the context of a length-dependent threshold. In other words, a significant sequence identity can only be defined in the context of an optimum window of sequence length, which was determined to be around 80 amino acids. Such a requirement for a length threshold (around 80 amino acids) to determine a significant sequence identity had been described earlier by Sander and Schneider[61] and was also discussed by Rost.

8.10.5 Allergenicity-Prediction Servers

The bioinformatic tools to analyze the sequence of a protein according to FAO/WHO rules are available from multiple sources, such as SDAP, and Allermatch.

[d]An epitope, also called an antigenic determinant, is a region of the antigen (protein) that binds a secreted antibody, such as immunoglobulin G (IgG), or a membrane receptor on a lymphocyte, such as the T-cell receptor (TCR). Normally, such binding results in a humoral (antibody-mediated) immune response or a cellular (T-cell-mediated) immune response. Allergy is a special type of immune response that occurs in some individuals whose immune system overreacts to certain environmental substances that do not bother most other people. During an allergic response, IgE binds to the IgE receptor on mast cells (in tissues) and basophils (in circulation). When two or more IgEs bound to receptors on the mast cells or basophils are cross-linked by the allergen through the allergenic epitope, these cells are activated. Both mast cells and basophils contain special cytoplasmic granules that store many mediators of inflammation. The extracellular release of these mediators following activation of these cells is known as **degranulation**. A well-known mediator of inflammation released by mast cells is histamine. The released mediators of inflammation trigger allergy symptoms.

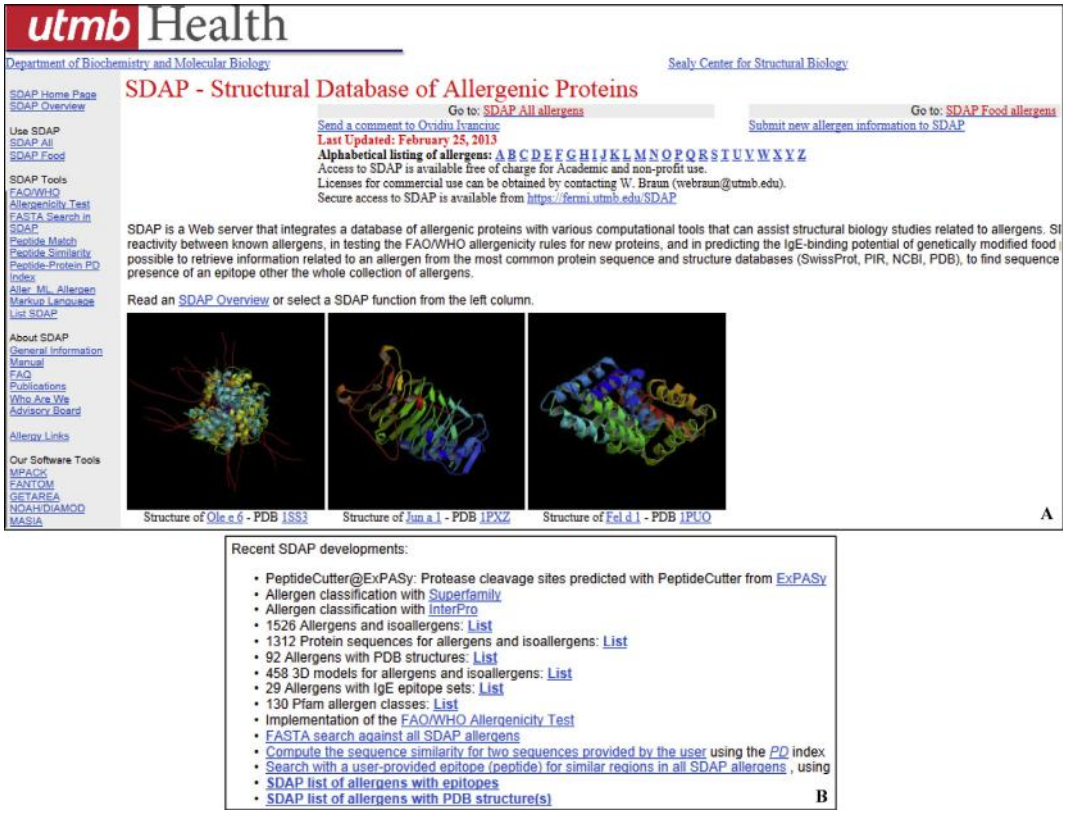


FIGURE 8.11 **The SDAP database home page.** (A) Partial (upper) screenshot of the SDAP database home page. Note the panel with links on the left-hand side, including links to SDAP tools. (B) Further down the home page is the "Recent SDAP developments" section (as of August 2013).

Allergenonline allows searching for an eight- (instead of six-) contiguous-amino-acid exact match. This change is based on the argument that searching for an exact match of six contiguous amino acids has the potential of generating many false positives.

In this section, we will focus on the information available from the SDAP database and analysis tools available on the **SDAP**[62,63] (https://fermi.utmb.edu/SDAP/) and **AlgPred**[64] (http://www.imtech.res.in/raghava/algpred/) servers. Figure 8.11A shows a partial (upper) screenshot of the SDAP database, whereas Figure 8.11B shows recent SDAP developments, as of August 2013.

On the panel on the left there are various links. One such link is "FAO/WHO Allergenicity Test." Clicking this link takes the user to the screen shown in Figure 8.12. The search for allergenicity of a protein can be launched from this page. Hitting the "Search" button returns a list of allergenic protein sequences that share one or more segments of six-contiguous-amino-acid identity with the input sequence. For demonstration, the sequence of mouse Slco1a6 has been pasted in the box (Figure 8.12) and analyzed using FAO/WHO rules. In this example, a total of six different segments of Slco1a6 (each segment is six-contiguous-amino-acids long) were found to match with segments of six different allergens from the database (Figure 8.13A and B). Figure 8.13A is a partial screenshot as displayed in the output. Figure 8.13B lists the other five hits between Slco1a6 and five different allergenic proteins. For these five hits, the screenshots of alignment are not shown, to save space. No sequence identity 35% or greater was found in a sliding window of 80 amino acids. *In practice, it is more common to have one or more six-contiguous-amino-acid sequence matches than to have >35% sequence identity in a sliding window of 80 amino acids.*

In the situation when there are six-contiguous-amino-acid segment matches between the input protein sequence and various allergenic proteins in the database, additional sequence comparison can be performed. For example, the distribution of these six-contiguous-amino-acid sequence segments can be verified using BLASTP against a curated protein database, such as UniProtKB/Swiss-Prot. The goal is to find out if these six-amino-acid sequence segments widely occur in various proteins that are not known to be allergenic. Additionally, the input sequence can be further analyzed using other prediction tools, such as **AlgPred**. Figure 8.14A shows that AlgPred offers several different approaches for predicting the allergenic potential of a protein (the input sequence). Five different

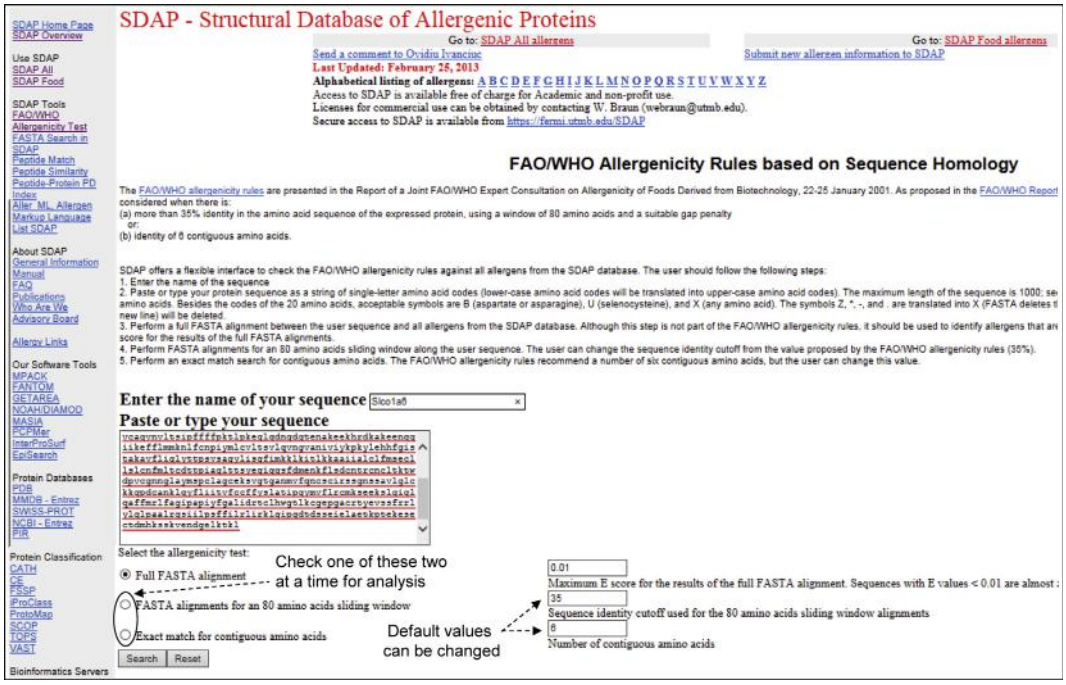


FIGURE 8.12 **FAO/WHO Rule-Based Allergenicity Prediction at the SDAP database.** The search for allergenicity of a protein according to FAO/WHO rules can be launched from this page. The default settings are 6 for contiguous amino acids, and 35 for % cutoff in a sliding window of 80 amino acids. These values can be changed by the user if needed. Selecting any one of these two options and hitting the "Search" button returns the results of the analysis. The sequence of mouse Slco1a6 has been pasted in the box for analysis according to FAO/WHO rules.

```
Alignment 1
Sequence 1: Slco1a6
Sequence 2: Allergen Lyc e 4.0101, Sequence: CAA75803
Sequence 1 -----------TTISPT-----------------------------------
Sequence 2 MGVNTYTYESTTTISPTRLFKALVLDFDNLVPKLLSQHVKNNETIEGDGG

Sequence 1 --------------------------------------------------
Sequence 2 VGSIKQMNFVEGGPIKYLKHKIHVIDDKNLETKYSLIEGDILGEKLESIT

Sequence 1 --------------------------------------------------
Sequence 2 YDIKFEANDNGGCVYKTTTEYHTKGDHVVSEEEHNVGRRENHEYFQGCRS

Sequence 1 ----------------------------
Sequence 2 IPSRESFCLRLNIDEKESGLHVRNYACT
```

A

Other 5 hits

Allergen name	Sequence ID	6 amino acid exact match
Hel a 2,	O81982:	GIIKEF
Bos d 6,	P02769:	AGCEKS
Ves v 6.0101,	G8IIT0:	NSSAVL
Ani s 7.0101,	ABL77410:	RCMKSE
Hev b 10.0103,	CAC13961:	AALRGS

B

FIGURE 8.13 **Results of the FAO/WHO Rule-Based Allergenicity Prediction of Slco1a6.** A total of six different segments of Slco1a6, each six-contiguous-amino-acids long, were found to match with six different allergens from the database. (A) A partial screenshot of the six-contiguous-amino-acid hit, as displayed in the output. (B) The other five hits between Slco1a6 and five different allergenic proteins. No sequence identity 35% or greater was found in a sliding window of 80 amino acids.

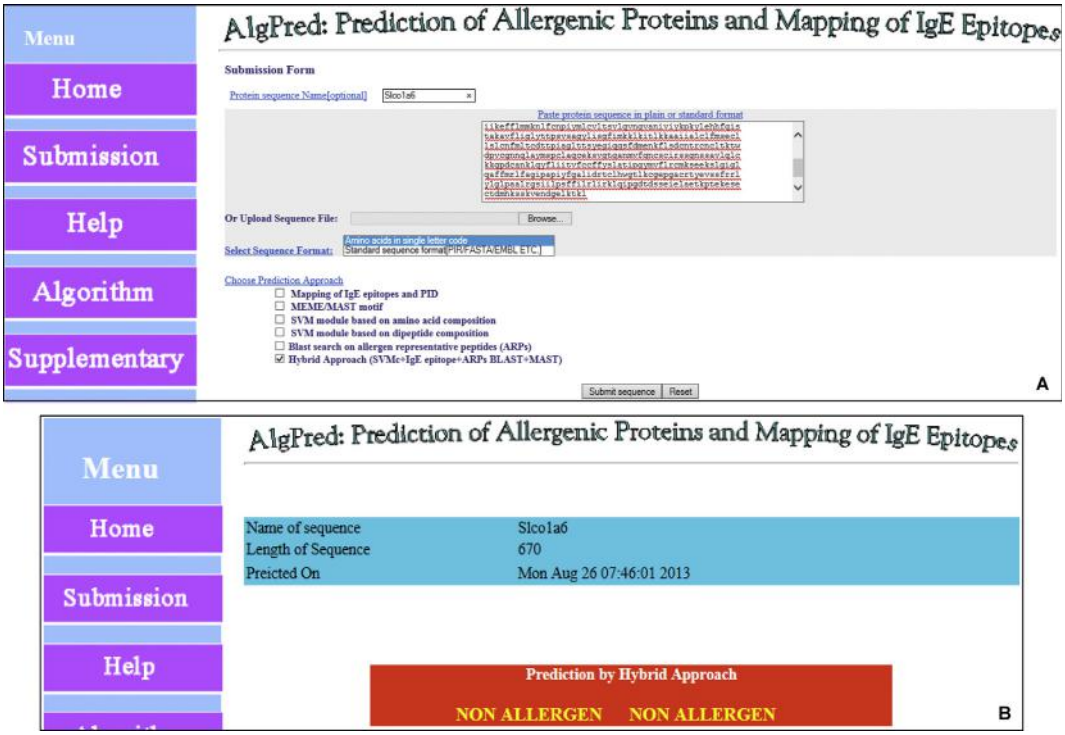


FIGURE 8.14 **Analysis of the input sequence using AlgPred.** (A) AlgPred offers several different approaches for predicting the allergenic potential of a protein (the input sequence). The hybrid approach that combines all five other approaches was chosen for the prediction (box checked). (B) The hybrid approach predicts Slco1a6 as a non-allergen. The same approach can be used to predict the potential allergenicity of a non-food protein.

approaches can be chosen for the prediction (listed on the home page), or the combination of all five in the "Hybrid Approach". Figure 8.14B shows that the hybrid approach predicts Slco1a6 as a non-allergen. The same approach can be used to predict the potential allergenicity of a non-food protein. It should be remembered that the sequence-based approach of allergenicity prediction is one of many tools utilized to assess whether a protein has the potential to be allergenic.

In addition to predicting the allergenic potential of a protein, there are a number of online T-cell and B-cell epitope-prediction tools that can be used to predict T-cell and B-cell epitopes, both continuous and discontinuous, in an input protein sequence. Such prediction methods take into account many aspects of protein structure, such as amino-acid properties (e.g. hydrophilicity and antigenicity, solvent accessibility, secondary structure, flexibility), amino-acid sequence, 3D structure wherever available, and information about the known epitopes from databases. The machine-learning prediction methods include the hidden Markov model (HMM), artificial neural network (ANN), and support vector machine (SVM). The SVM was found to be a better predictor compared to the other machine-learning prediction methods.[65] Some easily accessible online T-cell and B-cell epitope-prediction tools are available from the following sources:

http://www.imtech.res.in/raghava/
http://www.cbs.dtu.dk/services/
http://tools.immuneepitope.org/main/.

8.11 INTRINSICALLY DISORDERED PROTEIN ANALYSIS

Intrinsically disordered proteins (**IDPs**), also known as **intrinsically unstructured proteins** (**IUPs**), are characterized by the lack of a stable tertiary structure under physiological conditions. The lack of structural order in a protein goes against the traditional wisdom that protein function depends on a stable tertiary structure (the structure–function paradigm). It has long been realized that proteins possess configurational adaptability (e.g. induced fit). However, the presence of disordered segments in a functional protein became apparent when the crystal structures of various proteins became available. Techniques, such as NMR, X-ray crystallography, and circular dichroism helped uncover the disordered/unstructured state of certain proteins (e.g. missing

electron density of certain segments; hence, missing segments in X-ray crystallography). For these proteins, the intrinsically disordered state is necessary for function; some of these proteins fold only in complex with the substrate. It has been estimated that at least 50% of eukaryotic proteins possess at least one long (>40-amino-acid) loop, while this fraction is lot lower in prokaryotes and Archaea. *Protein disorder is found within loops*. Coiled coils may also assume disorder as they only assume globular structure when the coiled-coil partners interact with one another. *IDPs play an important role in signaling, recognition, and regulation*; recognition and regulation may involve processes like substrate recognition, catalysis, transport, DNA and RNA binding, and gene regulation. The presence of flexible structure and flexible structural segments helps accommodate a greater spectrum of binding targets, and also allows the IDP–target interaction to be short-lived, which is crucial for proper regulation. Because IDPs play an important role in signaling and regulation, they are much more abundant in eukaryotes than prokaryotes.[66–68]

8.11.1 IDP Databases

There are a number of databases of IDPs available; three are indicated in Table 8.8, along with their respective URLs.

Figure 8.15 shows a screenshot of the DisProt database home page. It is a curated database. The current

TABLE 8.8 IDP Databases

	URL
DisProt	http://www.disprot.org/[69]
IDEAL	http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL/[70]
MobiDB	http://mobidb.bio.unipd.it/[71]

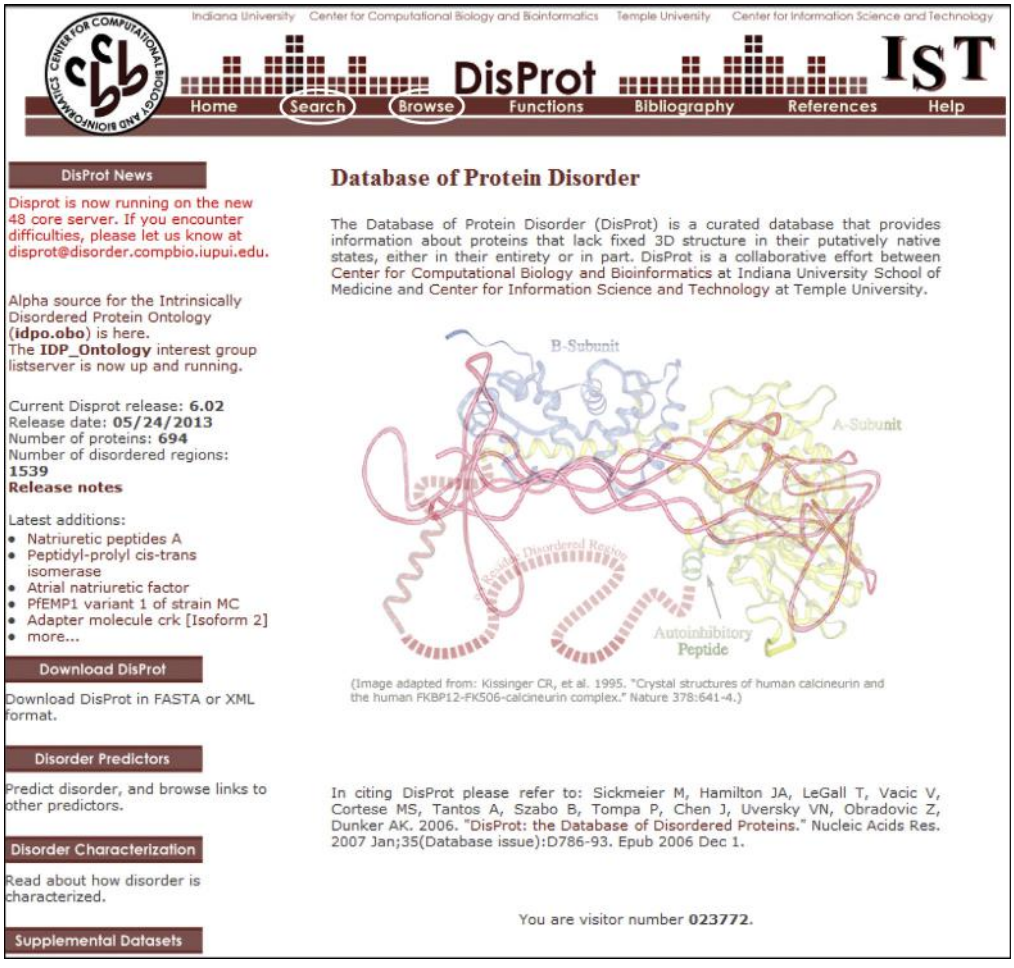


FIGURE 8.15 **Screenshot of the DisProt database home page.** On the left it displays the release number and the number of entries in the database. The entire database can be browsed by clicking the "Browse" link from the home page (circled). Alternatively, clicking the "Search" link (circled) takes the user to the search page, where a specific search can be launched (see text for details).

version (release 6.02) of the database has 694 proteins and a total of 1539 disordered regions. Clicking the "Search" link (circled) takes the user to the search page. An unknown sequence can be searched for the presence of a potential disordered segment by local-similarity search with other known disordered proteins from the database. Alternatively, a search can be launched by typing a keyword. In the absence of any specific search term, simply typing the keywords "signaling" or "regulation" will return a series of relevant entries from the database. An entry can be clicked to obtain more information, such as general information about the protein, sequence, percentage of the sequence that is disordered, map of the ordered and disordered segments, details of the disordered segments, and the references. The entire database can also be browsed by clicking the "Browse" link from the home page (circled). The other databases can also be searched/browsed in a similar fashion.

8.11.2 IDP Prediction

A number of online tools are also available to analyze a protein sequence for the existence of potentially disordered segments. Some of these tools are mentioned in Table 8.9, along with their respective URLs.

Figure 8.16 shows the DisProt disorder-prediction launch page. The sequence is pasted in the box, the desired analysis algorithm is checked, and the sequence is submitted for analysis. The Slco1a6 sequence was analyzed separately using VSL2B, VLXT, and PONDR-FIT. Because three different screenshots could not be

TABLE 8.9 Online Tools for IDP Prediction

Online Tool	Comments and URL
PONDR-FIT	Artificial-neural-network-based meta-predictor developed by combining several individual disorder predictors, such as PONDR-VLXT, PONDR-VSL2, PONDR-VL3, FoldIndex, IUPred, and TopIDP[72] (http://www.disprot.org/metapredictor.php)
DisEMBL	Artificial-neural-network-based. Trained for predicting several definitions of disorder, such as *loops/coils* as defined by DSSP*[73]; *hot loops*, i.e. the loops with a high B-factor from X-ray crystal structure†; *missing coordinates* (disordered regions) in X-ray structure as defined by REMARK465 entries in PDB, which indicate missing residues listed[74] (http://dis.embl.de/)
DISOPRED2	The link for PSIPRED analysis workbench is http://bioinf.cs.ucl.ac.uk/psipred/?disopred = 1. Check the box for DISOPRED2 in order to predict disordered protein
RONN	Bio-basis function neural network (BBFNN)-based. In BBFNN, the prediction is based on the likelihood of disorder determined by the alignment of the target sequence to a large group of sequences of known folding state (including known state of disorder)[75] (http://www.strubi.ox.ac.uk/RONN)

**DSSP (Dictionary of Secondary Structure of Proteins) is a program and database developed to standardize secondary-structure assignment for proteins of known 3D structure (hence entries in PDB database). DSSP describes eight states of protein secondary structure with single-letter codes: G (3/10 helix), H (α-helix), I (pi-helix), B (β-bridge), E (extended strand in β-sheet), S (bend), T (H-bonded turn), and C (coil).*
†In X-ray crystallography, the B-factor (temperature factor) is a measure of the extent of oscillation or vibration of an atom around the position specified in the model. So, a higher B-factor means more spread-out (lower) electron density, which indicates greater flexibility and disorder of the region.

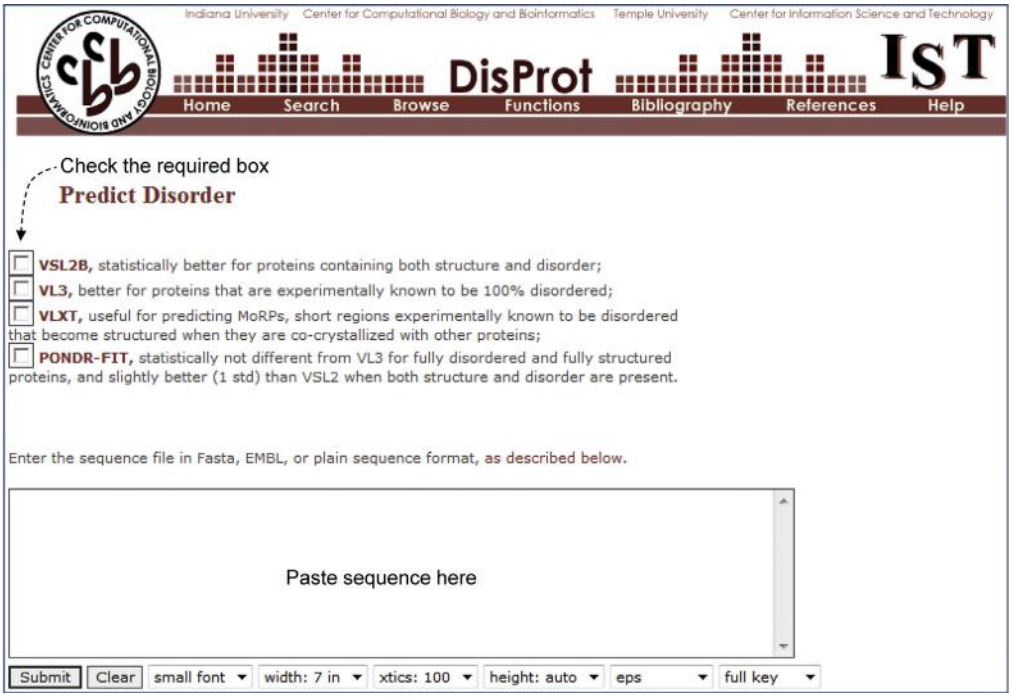


FIGURE 8.16 **The DisProt disorder-prediction launch page.** Providing options for analysis using PONDR-VSL2B, PONDR-VL3, PONDR-VLXT, and PONDR-FIT.

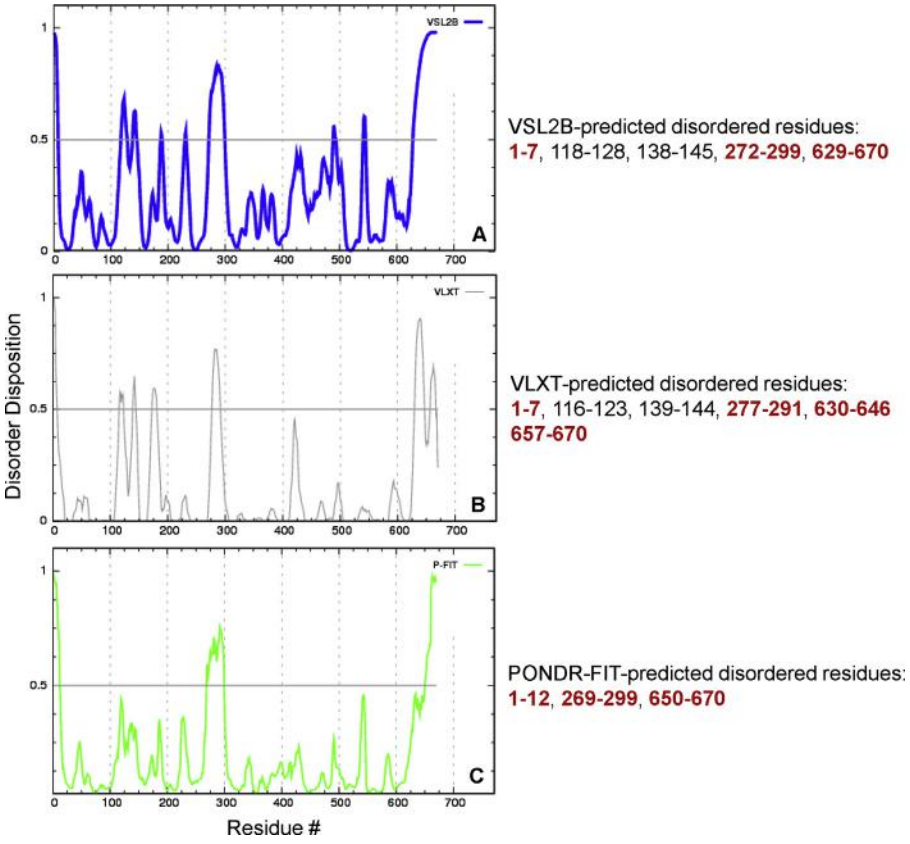


FIGURE 8.17 The Slco1a6 sequence, analyzed separately using VSL2B, VLXT, and PONDR-FIT. The graphical outputs of the analysis are shown. All three algorithms predict three regions of Slco1a6 to be potentially disordered: a very small region at the N-terminal end (around 1–10), a region in the middle (around 270–300), and at the C-terminal end (around 630–670). These predicted common residues are shown in red.

accommodated in one figure, only the graphical outputs of the analysis are shown, in Figure 8.17. All three algorithms predict three regions of Slco1a6 to be disordered. These predicted common residues are shown in red (Figure 8.17).

A separate analysis using RONN predicted three regions of disorder: 120–147, 272–299, and 630–670 (output not shown). Another analysis, using DisEMBL, predicted two regions of disorder: 279–296 and 640–670. Thus, different analysis programs consistently predicted two segments of Slco1a6 as potentially disordered regions: around 275–300 and around 635–670. Both these regions of Slco1a6 are part of the inside (cytoplasmic) segments, as predicted by RHYTHM, OCTOPUS, and Phobius (Figures 8.9 and 8.10).

References

1. Vieira-Pires RS, Morais-Cabral JH. *J Gen Physiol* 2010;**136**:585–92.
2. Cooley RB, et al. *J Mol Biol* 2010;**404**:232–46.
3. Leszczynski JF, Rose GD. *Science* 1986;**234**:849–55.
4. Chou KC. *Anal Biochem* 2000;**286**:1–16.
5. Ring CS, et al. *J Mol Biol* 1992;**224**:685–99.
6. Hovmöller S, et al. *Acta Crystallogr D Biol Crystallogr* 2002;**58** (Pt 5):768–76.
7. Kleywegt GJ, Jones TA. *Structure* 1996;**4**:1395–400.
8. Gasteiger E, et al. In: Walker JM, editor. *The proteomics protocols handbook*. Totowa, NJ: Humana Press; 2005. p. 571–607.
9. Varshavsky A. *Genes Cells* 1997;**2**:13–28.
10. Guruprasad K, et al. *Protein Eng* 1990;**4**:155–61.
11. Ikai AJ. *J Biochem* 1980;**88**:1895–8.
12. Kyte J, Doolittle RF. *J Mol Biol* 1982;**157**:105–32.
13. Hopp TP, Woods KR. *Proc Natl Acad Sci USA* 1981;**78**:3824–8.
14. Jääskeläinen S, et al. *Int J Data Min Bioinform* 2010;**4**:735–54.
15. Kloczkowski A, et al. *Proteins* 2002;**49**:154–66.
16. Chou PY, Fasman GD. *Biochemistry* 1974;**13**:222–45.
17. Garnier J, et al. *Methods Enzymol* 1996;**266**:540–53.
18. Garnier J, et al. *J Mol Biol* 1978;**120**:97–120.
19. Combet C, et al. *Trends Biochem Sci* 2000;**25**:147–50.
20. Bystroff C, Shao Y. *Bioinformatics* 2002;**18**(Suppl. 1):S54–61.
21. Bystroff C, et al. *J Mol Biol* 2000;**301**:173–90.
22. Cole C, et al. *Nucl Acids Res* 2008;**35**(Suppl. 2):W197–201.
23. Rost B, Sander C. *J Mol Biol* 1993;**232**:584–99.
24. Rost B, Sander C. *Proc Natl Acad Sci USA* 1993;**90**:7558–62.
25. Ouali M, King RD. *Protein Sci* 2000;**9**:1162–76.
26. Montgomerie S, et al. *BMC Bioinformatics* 2006;**7**:301.
27. Jones DT. *J Mol Biol* 1999;**292**:195–202.
28. Cheng J, et al. *Nucl Acids Res* 2005;**33**:72–6 (Web Server issue).
29. Baldi P, et al. *Bioinformatics* 1999;**15**:937–46.
30. Pollastri G, et al. *Bioinformatics* 2001;**17**(Suppl. 1):S234–42.
31. Simossis VA, Heringa J. *Comput Biol Chem* 2004;**28**:351–66.
32. Geourjon C, Deleage G. *Comput Appl Biosci* 1995;**11**:681–4.
33. Lin K, et al. *Bioinformatics* 2005;**21**:152–9.
34. Raghava GPS. *CASP* 2000;**4**:75–6.
35. Lupas A, et al. *Science* 1991;**252**:1162–4.
36. McDonnell AV, et al. *Bioinformatics* 2006;**22**:356–8.
37. Bornberg-Bauer E, et al. *Nucl Acids Res* 1998;**26**:2740–6.

38. Quevillon E, et al. *Nucl Acids Res* 2005;**33**:W116–20 (Web Server issue).
39. Marchler-Bauer A, et al. *Nucl Acids Res* 2013;**41**:D348–52 (Database issue).
40. Rimphanitchayakit V, Tassanakajon A. *Dev Comp Immunol* 2010;**34**:377–86.
41. Cera ED. *IUBMB Life* 2009;**61**:510–5.
42. Pao SS, et al. *Microb Mol Biol Rev* 1998;**62**:1–34.
43. Reddy VS, et al. *FEBS J* 2012;**279**:2022–35.
44. Sigrist CJ, et al. *Nucl Acids Res* 2013;**41**:D344–7 (Database issue).
45. Sigrist CJA, et al. *Brief Bioinform* 2002;**3**:265–74.
46. Krogh A, et al. *J Mol Biol* 2001;**305**:567–80.
47. Rose A, et al. *Nucl Acids Res* 2009;**37**:W575–80 (Web Server issue).
48. Käll L, et al. *J Mol Biol* 2004;**338**:1027–36.
49. Viklund H, Elofsson A. *Bioinformatics* 2008;**24**:1662–8.
50. Marsh DG, et al. *Bull World Health Org* 1986;**6**:767–70.
51. King TP, et al. *J Allergy Clin Immunol* 1995;**96**:5–14.
52. Gendel SM. *Regulat Toxicol Pharmacol* 2009;**54**:S7–10.
53. Vila L, et al. *Clin Exp Allergy* 2001;**31**:1599–606.
54. Wang J, Sampson HA. *J Clin Invest* 2011;**121**:827–35.
55. Roth-Walter F, et al. *Mol Nutr Food Res* 2013;**57**:536–44.
56. Taylor SL. *Annu Rev Pharmacol Toxicol* 2002;**42**:99–112.
57. FAO/WHO. Evaluation of allergenicity of genetically modified foods. Available online at: <http://www.who.int/foodsafety/publications/biotech/en/ec_jan2001.pdf>; 2001.
58. Burks AW, et al. *Eur J Biochem* 1997;**245**:334–9.
59. Stanley JS, et al. *Arch Biochem Biophys* 1997;**342**:244–53.
60. Rost B. *Protein Eng* 1999;**12**:85–94.
61. Sander C, Schneider R. *Proteins* 1991;**9**:56–68.
62. Ivanciuc O, et al. *Bioinformatics* 2002;**18**:1358–64.
63. Ivanciuc O, et al. *Nucl Acids Res* 2003;**31**:359–62.
64. Saha S, Raghava GPS. *Nucl Acids Res* 2006;**34**:W202–9.
65. Bhasin M, Raghava GPS. *Vaccine* 2004;**22**:3195–204.
66. Tompa P. *Trends Biochem Sci* 2002;**27**:527–33.
67. Tompa P. *Trends Biochem Sci* 2012;**37**:509–16.
68. Uversky VN, Dunker AK. *Biochim Biophys Acta* 2010;**1804**:1231–64.
69. Sickmeier M, et al. *Nucl Acids Res* 2007;**35**:D786–93 (Database issue).
70. Fukuchi S, et al. *Nucl Acids Res* 2012;**40**:D507–11 (Database issue).
71. Di Domenico T, et al. *Bioinformatics* 2012;**28**:2080–1.
72. Xue B, et al. *Biochim Biophys Acta* 2010;**1804**:996–1010.
73. Kabsch W, Sander C. *Biopolymers* 1983;**22**:2577–637.
74. Linding R, et al. *Structure* 2003;**11**:1453–9.
75. Yang ZR, et al. *Bioinformatics* 2005;**21**:3369–76.