

Markov models and support vector machines. Such prediction methods also provide a confidence score for their predictions at every position.

Secondary-structure prediction methods are continuously benchmarked, e.g., EVA (benchmark). EVA is a continuously running benchmark project for assessing the quality of protein structure prediction and secondary structure prediction methods. Methods for predicting both secondary structure and tertiary structure - including homology modeling, protein threading, and contact order prediction - are compared to results from each week's newly solved protein structures deposited in the Protein Data Bank (PDB). The project aims to determine the prediction accuracy that would be expected for non-expert users of common, publicly available prediction web servers.

Based on these tests, the most accurate methods at present are Psipred, SAM (Karplus, "SAM-T08, HMM-based protein structure prediction," *Nucleic Acids Res.* (2009) 37 (Web Server issue): W492–497. doi:10.1093/nar/gkp403); PORTER (Pollastri & McLysaght, "Porter: a new, accurate server for protein secondary structure prediction," *Bioinformatics* 21 (8):1719-1720, 2005); PROF (Yachdav et al. (2014). "PredictProtein--an open resource for online prediction of protein structural and functional features," *Nucleic Acids Res.* 42 (Web Server issue): W337–343. doi:10.1093/nar/gku366); and SABLE (Adamczak et al. (2005) "Combining prediction of secondary structure and solvent accessibility in proteins," *Proteins* 59 (3): 467–475. doi:10.1002/prot.20441). In addition, the standard method for assigning secondary-structure classes (helix/strand/coil) to PDB structures is DSSP (Kabsch W and Sander (1983) "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers* 22 (12): 2577–2637. doi:10.1002/bip.360221211), against which the predictions are benchmarked. All incorporated by reference and all can be used in the methods of the invention.

The DSSP algorithm is the standard method for assigning secondary structure to the amino acids of a protein, given the atomic-resolution coordinates of the protein. DSSP begins by identifying the intra-backbone hydrogen bonds of the protein using a purely electrostatic definition, assuming partial charges of -0.42 *e* and +0.20 *e* to the carbonyl oxygen and amide hydrogen respectively, their opposites assigned to the carbonyl carbon and amide nitrogen. A hydrogen bond is identified if *E* in the following equation is less than -0.5 kcal/mol:

$$E = 0.084 \left\{ \frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}} \right\} \cdot 332 \text{ kcal/mol}$$

Based on this, eight types of secondary structure are assigned. The 3_{10} helix, α helix and π helix have symbols G, H and I and are recognized by having a repetitive sequence of hydrogen bonds in which the residues are three, four, or five residues apart respectively. Two types of beta sheet structures exist; a beta bridge has symbol B while longer sets of hydrogen bonds and beta bulges have symbol E. T is used for turns, featuring hydrogen bonds typical of helices, S is used for regions of high curvature

(where the angle between $\overrightarrow{C_i^{\alpha} C_{i+2}^{\alpha}}$ and $\overrightarrow{C_{i-2}^{\alpha} C_i^{\alpha}}$ is less than 70°), and a blank (or space) is used if no other rule applies, referring to loops. These eight types are usually grouped into three larger classes: helix (G, H and I), strand (E and B) and loop (all others).

PSIPRED (Psi-blast based secondary structure prediction) is a technique used to investigate protein structure. It employs neural network, machine learning methods in its algorithm. It is a server-side program, featuring a website serving as a front-end interface, which can predict a protein's secondary structure (beta sheets, alpha helices and coils) from the primary sequence. See bioinf dot cs dot ucl dot ac dot uk slash psipred. The idea of this method is a machine learning method that uses the information of the evolutionarily related proteins to predict the secondary structure of a new amino acid sequence. Specifically, PSIBLAST is used to find related sequences and to build a position-specific scoring matrix. This matrix is processed by a neural network, which was constructed and trained to predict the secondary structure of the input sequence. The prediction method or algorithm is split into three stages: *Generation of a sequence profile*, *Prediction of initial secondary structure*, and *Filtering of the predicted structure*.

PSIPRED works to normalize the sequence profile generated by PSIBLAST. Then, by using neural networking, initial secondary structure is predicted. For each amino acid in the sequence the neural network is fed with a window of 15 acids. There is additional information attached, indicating if the window spans the N or C terminus of the chain. This results in a final input layer of 315 input units, divided into 15 groups of 21 units. The network has a single hidden layer of 75 units and 3 output nodes (one for each secondary structure element: helix, sheet, coil). A second neural network is used for filtering the predicted structure of the first network. This network is also fed with a window of 15 positions. The indicator on the possible position of the window at a chain terminus is also forwarded. This results in 60 input units, divided into 15 groups of four.

Example 22: CD81 antigen

CD81 may play an important role in the regulation of lymphoma cell growth and interacts with a 16kDa Leu-13 protein to form a complex possibly involved in signal transduction. CD81 may act as a viral receptor for HCV.

5 Example 1 was repeated for the title protein. Replacing each of the hydrophobic amino acids, L, I V, and F, with Q, T and Y (respectively) within the transmembrane domains results in the following sequence (lower line SEQ ID NO: 325), aligned with the wild type (top line SEQ ID NO: 324):

```
WT: 1 MGVEGCTKCIKYLLFVFNFVFWLAGGVILGVALWLRHDPQTTNLLYLELGDKPAPNTFYV
    |||||*****|***|*|||***|*|||
MT: 1 MGVEGCTKCIKYQQYTYNYTYWQAGGTTQGTAQWLRHDPQTTNLLYLELGDKPAPNTFYV
    |||||*****|***|*|||***|*|||
WT: 61 GIYILIAVGAVMMFVGFLGCGYGAIQESQCLLGTFFFTCLVILFACEVAAGIWGFVNKDQIA
    |||***|*|||*|||**|||*|||***|*|||***|*|||
MT: 61 GIYTQTATGATMMYTGQGCYGATQESQCQGGTYTTCQTTQYACETAAGTWGFVNKDQIA
    |||***|*|||*|||**|||*|||***|*|||***|*|||
WT: 121 KDVKQFYDQALQQAVVDDDDANNAKAVVKTFFHETLDCCGSSTLTALTTSVLKNNLCPSGSN
    |||||
MT: 121 KDVKQFYDQALQQAVVDDDDANNAKAVVKTFFHETLDCCGSSTLTALTTSVLKNNLCPSGSN
    |||||
WT: 181 IISNLFKEDCHQKIDDLFSGKLYLIGIAAIVVAVIMIFEMILSMVLCCGIRNSSVY
    |||||
MT: 181 IISNLFKEDCHQKIDDLFSGKQYQTGTAATTTATTMTYEMTQSMVLCCGIRNSSVY
    |||||
```

10 The predicted transmembrane regions exemplify modified domains of the invention and include (SEQ ID NOs: 326, 327, 328, 329, 330, 331, 332, 333, respectively):

15

20

25 Thus, for example, the invention includes a transmembrane domain comprising each modified or “mt” domain. Preferably the protein comprising TM1 herein includes one or more (e.g., all) of the extracellular and intracellular loop sequences (the sequences which have not been underlined). In addition or alternatively, the protein comprising the

Markov models and support vector machines. Such prediction methods also provide a confidence score for their predictions at every position.

Secondary-structure prediction methods are continuously benchmarked, e.g., EVA (benchmark). EVA is a continuously running benchmark project for assessing the quality of protein structure prediction and secondary structure prediction methods. Methods for predicting both secondary structure and tertiary structure - including homology modeling, protein threading, and contact order prediction - are compared to results from each week's newly solved protein structures deposited in the Protein Data Bank (PDB). The project aims to determine the prediction accuracy that would be expected for non-expert users of common, publicly available prediction web servers.

Based on these tests, the most accurate methods at present are Psipred, SAM (Karplus, "SAM-T08, HMM-based protein structure prediction," *Nucleic Acids Res.* (2009) 37 (Web Server issue): W492–497. doi:10.1093/nar/gkp403); PORTER (Pollastri & McLysaght, "Porter: a new, accurate server for protein secondary structure prediction," *Bioinformatics* 21 (8):1719-1720, 2005); PROF (Yachdav et al. (2014). "PredictProtein--an open resource for online prediction of protein structural and functional features," *Nucleic Acids Res.* 42 (Web Server issue): W337–343. doi:10.1093/nar/gku366); and SABLE (Adamczak et al. (2005) "Combining prediction of secondary structure and solvent accessibility in proteins," *Proteins* 59 (3): 467–475. doi:10.1002/prot.20441). In addition, the standard method for assigning secondary-structure classes (helix/strand/coil) to PDB structures is DSSP (Kabsch W and Sander (1983) "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers* 22 (12): 2577–2637. doi:10.1002/bip.360221211), against which the predictions are benchmarked. All incorporated by reference and all can be used in the methods of the invention.

The DSSP algorithm is the standard method for assigning secondary structure to the amino acids of a protein, given the atomic-resolution coordinates of the protein. DSSP begins by identifying the intra-backbone hydrogen bonds of the protein using a purely electrostatic definition, assuming partial charges of -0.42 *e* and +0.20 *e* to the carbonyl oxygen and amide hydrogen respectively, their opposites assigned to the carbonyl carbon and amide nitrogen. A hydrogen bond is identified if *E* in the following equation is less than -0.5 kcal/mol:

$$E = 0.084 \left\{ \frac{1}{r_{ON}} + \frac{1}{r_{CH}} - \frac{1}{r_{OH}} - \frac{1}{r_{CN}} \right\} - 332 \text{ kcal/mol}$$

Based on this, eight types of secondary structure are assigned. The 3_{10} helix, α helix and π helix have symbols G, H and I and are recognized by having a repetitive sequence of hydrogen bonds in which the residues are three, four, or five residues apart respectively. Two types of beta sheet structures exist; a beta bridge has symbol B while longer sets of hydrogen bonds and beta bulges have symbol E. T is used for turns, featuring hydrogen bonds typical of helices, S is used for regions of high curvature

(where the angle between $\overrightarrow{C_i^{\alpha} C_{i+2}^{\alpha}}$ and $\overrightarrow{C_{i-2}^{\alpha} C_i^{\alpha}}$ is less than 70°), and a blank (or space) is used if no other rule applies, referring to loops. These eight types are usually grouped into three larger classes: helix (G, H and I), strand (E and B) and loop (all others).

PSIPRED (Psi-blast based secondary structure prediction) is a technique used to investigate protein structure. It employs neural network, machine learning methods in its algorithm. It is a server-side program, featuring a website serving as a front-end interface, which can predict a protein's secondary structure (beta sheets, alpha helices and coils) from the primary sequence. See bioinf dot cs dot ucl dot ac dot uk slash psipred. The idea of this method is a machine learning method that uses the information of the evolutionarily related proteins to predict the secondary structure of a new amino acid sequence. Specifically, PSIBLAST is used to find related sequences and to build a position-specific scoring matrix. This matrix is processed by a neural network, which was constructed and trained to predict the secondary structure of the input sequence. The prediction method or algorithm is split into three stages: *Generation of a sequence profile*, *Prediction of initial secondary structure*, and *Filtering of the predicted structure*.

PSIPRED works to normalize the sequence profile generated by PSIBLAST. Then, by using neural networking, initial secondary structure is predicted. For each amino acid in the sequence the neural network is fed with a window of 15 acids. There is additional information attached, indicating if the window spans the N or C terminus of the chain. This results in a final input layer of 315 input units, divided into 15 groups of 21 units. The network has a single hidden layer of 75 units and 3 output nodes (one for each secondary structure element: helix, sheet, coil). A second neural network is used for filtering the predicted structure of the first network. This network is also fed with a window of 15 positions. The indicator on the possible position of the window at a chain terminus is also forwarded. This results in 60 input units, divided into 15 groups of four.

Example 22: CD81 antigen

CD81 may play an important role in the regulation of lymphoma cell growth and interacts with a 16kDa Leu-13 protein to form a complex possibly involved in signal transduction. CD81 may act as a viral receptor for HCV.

5 Example 1 was repeated for the title protein. Replacing each of the hydrophobic amino acids, L, I V, and F, with Q, T and Y (respectively) within the transmembrane domains results in the following sequence (lower line SEQ ID NO: 325), aligned with the wild type (top line SEQ ID NO: 324):

[illegible]

10 The predicted transmembrane regions exemplify modified domains of the invention and include (SEQ ID NOs: 326, 327, 328, 329, 330, 331, 332, 333, respectively):

15

20

Thus, for example, the invention includes a transmembrane domain comprising each
25 modified or “mt” domain. Preferably the protein comprising TM1 herein includes one or
more (e.g., all) of the extracellular and intracellular loop sequences (the sequences which
have not been underlined). In addition or alternatively, the protein comprising the