CFSSP: Chou and Fasman Secondary Structure Prediction server

T. Ashok Kumar
Department of Bioinformatics, Noorul Islam College of Arts and Science, Kumaracoil - 629180,
E-Mail: *ashok@biogem.org*

ABSTRACT

CFSSP (Chou & Fasman Secondary Structure Prediction Server) is an online protein secondary structure prediction server. This server predicts regions of secondary structure from the protein sequence such as alpha helix, beta sheet, and turns from the amino acid sequence. The output of predicted secondary structure is also displayed in linear sequential graphical view based on the probability of occurrence of alpha helix, beta sheet, and turns. The method implemented in CFSSP is Chou-Fasman algorithm, which is based on analyses of the relative frequencies of each amino acid in alpha helices, beta sheets, and turns based on known protein structures solved with X-ray crystallography. CFSSP is freely accessible via ExPASy server or directly from BioGem tools at *http://www.biogem.org/tool/chou-fasman.* CFSSP server is written in Perl, which runs through CGI.

Key words: CFSSP, ExPASy, BioGem Tools, Secondary Structure, Chou and Fasman.

INTRODUCTION

Successful prediction of protein structure from the amino acid sequence is one of the challenging tasks in bioinformatics and structural biology; it is highly important in medicine (for example, in drug design) and biotechnology (for example, in the design of novel enzymes). Although experimental structure determination has improved, information about the three dimensional structure is still available for only a small fraction of known proteins. Structure prediction of soluble proteins using experimental methods is still a challenging task due to the vast number of degrees of freedom in the molecule. An intermediate but useful step is to predict the protein secondary structure, that is, each residue of a protein sequence is assigned a conformational state, either helix (H), strand (E) or coil (C). The information provided by this assignment is valuable both in *ab initio* tertiary structure prediction and as additional restraints for fold recognition algorithms (Cuff and Barton, 2000). In addition, it can also be used in protein function prediction (Paquet *et al.*, 2000).

The Chou-Fasman method was among the first secondary structure prediction algorithms developed and relies predominantly on probability parameters determined from relative frequencies of each amino acid's appearance in each type of secondary structure (Chou and Fasman, 1974). The original Chou-Fasman parameters, determined from the small sample of structures solved in the mid-1970s, produce poor results compared to modern methods, though the parameterization has been updated since it was first published. The Chou-Fasman method is roughly 56-60% accurate in predicting secondary structures (Mount, 2004).

The evolutionary conservation of secondary structures can be exploited by simultaneously assessing many homologous sequences in a multiple sequence alignment, by calculating the net secondary structure propensity of an aligned column of amino acids. In concert with larger databases of known protein structures and modern machine learning methods such as neural networks and support vector machines, these methods can achieve up 80% overall accuracy in globular proteins (Dor and Zhou, 2006). The theoretical upper limit of accuracy is around 90% (Dor and Zhou, 2007), partly due to idiosyncrasies in DSSP assignment near the ends of secondary structures, where local conformations vary under native conditions but may be forced to assume a single conformation in crystals due to packing constraints. Limitations are also imposed by secondary structure prediction's inability to account for tertiary structure; for example, a sequence predicted as a likely helix may still be able to adopt a beta-strand conformation if it is located within a beta-sheet region of the protein and its side chains pack well with their neighbors. Dramatic conformational changes related to the protein's function or environment can also alter local secondary structure.

METHODS

The algorithm implemented in the CFSSP server is Chou-Fasman algorithm. The Chou-Fasman method (1985) is a combination of such statistics-based methods and rule-based methods (Chou and Fasman, 1989). Here are the steps of the Chou-Fasman algorithm:

Table 1: Conformational Parameters for α-Helical, β-Sheet, and β-Turn Residues in 29 Proteins.[a]

Residue[b]	P_α	α-Type	Residue[c]	P_β	β-Type	Residue	P_t
Glu$^{(-)}$	1.51	H_α	Val	1.70	H_β	Asn	1.56
Met	1.45		Ile	1.60		Gly	1.56
Ala	1.42		Tyr	1.47		Pro	1.52
Leu	1.21		Phe	1.38	h_β	Asp$^{(-)}$	1.46
Lys$^{(+)}$	1.16	h_α	Trp	1.37		Ser	1.43
Phe	1.13		Leu	1.30		Cys	1.19
Gln	1.11		Cys	1.19		Tyr	1.14
Trp	1.08		Thr	1.19		Lys$^{(+)}$	1.01
Ile	1.08		Gln	1.10		Gln	0.98
Val	1.06		Met	1.05		Thr	0.96
Asp$^{(-)}$	1.01	I_α	Arg$^{(+)}$	0.93	i_β	Trp	0.96
His$^{(+)}$	1.00		Asn	0.89		Arg$^{(+)}$	0.95
Arg$^{(+)}$	0.98	i_α	His$^{(+)}$	0.87		His$^{(+)}$	0.95
Thr	0.83		Ala	0.83		Glu$^{(-)}$	0.74
Ser	0.77		Ser	0.75	b_β	Ala	0.66
Cys	0.70		Gly	0.75		Met	0.60
Tyr	0.69	b_α	Lys$^{(+)}$	0.74		Phe	0.60
Asn	0.67		Pro	0.55	B_β	Leu	0.59
Pro	0.57	B_α	Asp$^{(-)}$	0.54		Val	0.50
Gly	0.57		Glu$^{(-)}$	0.37		Ile	0.47

[a] Chou and Fasman (1974)

[b] α-helix assignments: H_α (strong α former), h_α (α former), I_α (weak α former), i_α (α indifferent), b_α (α breaker), B_α (strong α breaker)

[c] β-sheet assignments: H_β (strong β former), h_β (β former), I_β (weak β former), i_β (β indifferent), b_β (β breaker), B_β (strong β breaker).

i. Search for Helical Regions

Any segment of six residues or longer in a native protein with $\langle P_\alpha \rangle \geq 1.03$ as well as $\langle P_\alpha \rangle > \langle P_\beta \rangle$, and satisfying conditions *i.a.* through *i.d.*, is predicted as helical.

a. Helix Nucleation. Scan the peptide and identify regions four helical residues (h_α, or H_α) out of six residues along the polypeptide chain. Weak helical residues (I_α,) count as 0.5 h_α, (i.e., three h_α and two I_α residues out of six could also nucleate a helix). Helix formation is unfavorable if the segment contains ⅓ or more helix breakers (b_α or B_α), or less than ½ helix formers.

b. Helix Termination. Extend the helical segment in both directions until terminated by tetrapeptides with $\langle P_\alpha \rangle < 1.00$. The following helix breakers can stop helix propagation: b_4, b_3i, b_3h, b_2i_2, b_2ih, b_2h_2, bi_3, bi_2h, bih_2, and i_4. Once the helix is defined, some of the residues (especially h or i) in the tetrapeptides may be incorporated at the helical ends. The notations i, b, h in the tetrapeptide breakers also include I, B, and H, respectively. Adjacent *β* regions can also terminate *α* regions.

c. Pro cannot occur in the inner helix or at the C-terminal helical end.

d. Helix Boundaries. Pro, $\text{Asp}^{(-)}$, $\text{Glu}^{(-)}$ prefer the N-terminal helical end. $\text{His}^{(+)}$, $\text{Lys}^{(+)}$, $\text{Arg}^{(+)}$ prefer the C-terminal helical end. I_α, assignments are given to Pro and Asp (near the N-terminal helix) as well as Arg (near the C-terminal helix) if necessary to satisfy condition *i.a.*

ii. Search for *β*-Sheet Regions

Any segment of five residues or longer in a native protein with $\langle P_\beta \rangle \geq 1.05$ as well as $\langle P_\beta \rangle > \langle P_\alpha \rangle$, and satisfying conditions *ii.a.* through *ii.d.*, is predicted as *β* sheet.

a. β-Sheet Nucleation. Scan the peptide and identify regions of three *β* residues (h_β or H_β) out of five residues along the polypeptide chain. *β*-sheet formation is unfavorable if the segment contains ⅓ or more *β*-sheet breakers (b_β or B_β), or less than ½ *β*-sheet formers.

b. β-Sheet Termination. Extend the sheet in both directions until terminated by tetrapeptides with $\langle P_\beta \rangle < 1.00$. Once the sheet is defined, some of the residues (especially h or i) in the tetrapeptides may be incorporated at the helical ends. The notations i, b, h in the tetrapeptide breakers also include I, B, and H, respectively. Adjacent *α* regions can also terminate *β* regions.

c. Glut occurs rarely in the *β* region. Pro occurs rarely in the inner *β* region.

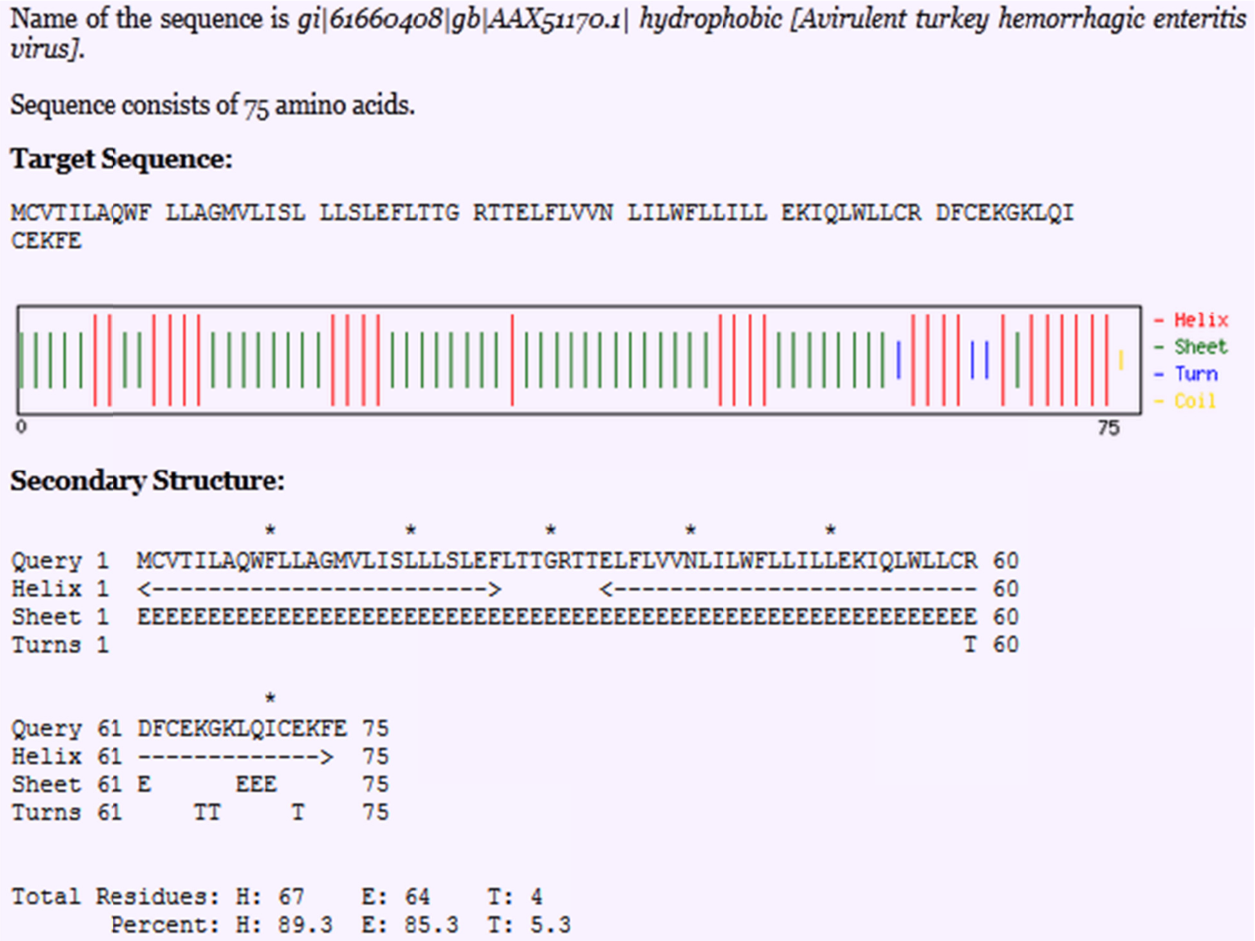
d. β-Sheet Boundaries. Charged residues occur rarely at the N-terminal *β*-sheet end, and infrequently at the inner *β* region and C-terminal *β* end. Trp occurs mostly at the N-terminal *β*-sheet end and rarely at the C-terminal *β*-end.

iii. Search for *β*-turn Regions

Proline and glycine are both common in turns. A turn is predicted only if the turn probability is greater than the helix or sheet probabilities and a probability value based on the positions of particular amino acids in the turn exceeds a predetermined threshold. After both *α*-helix and *β*-sheet regions have been predicted, the Chou-Fasman algorithm compares the relative probabilities of regions to resolve predictions that overlap. The conformational parameters for coil are not employed; coil is predicted by default. However, in most cases it will be found adequate to use only the former, breaker, indifferent assignments, and the termination tetrapeptides to locate the secondary structural regions of proteins.

IMPLEMENTATION

The CFSSP web server is presented to the user as a single page form. User can input the protein sequence in standard *fasta* file format. The characters in the given sequence are filtered from unknown characters and white spaces. By default, the first line in the sequence is read as protein name and remaining as protein sequence. The predicted secondary structure regions of the amino sequence are represented in graphical and characters as follows: α-helix (<–>), β-sheet (E), β-turns (T).



Name of the sequence is *gi|61660408|gb|AAX51170.1| hydrophobic [Avirulent turkey hemorrhagic enteritis virus]*.

Sequence consists of 75 amino acids.

Target Sequence:

```
MCVTILAQWF LLAGMVLISL LLSLEFLTTG RTTELFLVVN LILWFLLILL EKIQLWLLCR DFCEKGKLQI
CEKFE
```

Secondary Structure:

```
                  *         *         *         *         *
Query 1  MCVTILAQWFLLAGMVLISLLLSLEFLTTGRTTELFLVVNLILWFLLILLEKIQLWLLCR 60
Helix 1  <------------------------>       <-------------------------- 60
Sheet 1  EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE 60
Turns 1                                                             T 60

                  *
Query 61 DFCEKGKLQICEKFE 75
Helix 61 ------------->  75
Sheet 61 E      EEE      75
Turns 61     TT     T    75


Total Residues: H: 67    E: 64    T: 4
       Percent: H: 89.3  E: 85.3  T: 5.3
```

Fig 1: The predicted secondary structure of protein of Avirulent turkey hemorrhagic enteritis virus.

REFERENCES

1. Mount,D.M. (2004) *Bioinformatics: Sequence and Genome Analysis*, 2nd edn. Cold Spring Harbor Laboratory Press, New York.
2. Chou,P.Y. and Fasman,G.D. (1974) Prediction of protein conformation. *Biochemistry*, **13** (2), 222–245.

3. Dor,O. and Zhou,Y. (2006) Achieving 80% tenfold cross-validated accuracy for secondary structure prediction by large-scale training. *Proteins*, **66** (4), 838–845.
4. Cuff,J.A. and Barton,G.J. (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins*, **40**, 502–511.
5. Paquet,J.Y. *et al.* (2000) Topology prediction of Brucella abortus Omp2b and Omp2a porins after critical assessment of transmembrane beta strands prediction by several secondary structure prediction methods. *J. Biomol. Struct. Dyn.*, **17**, 747–757.
6. Peter Prevelige,Jr. and Fasman,G.D. (1989) Chapter 9: Chou-Fasman Prediction of the Secondary Structure of Proteins: The Chou-Fasman-Prevelige Algorithm. In Fasman,G.D., *Prediction of Protein Structure and the Principles of Protein Conformation*, Plenum, New York, pp.391-416