

CISPred: Consensus Integrated Protein Structure Prediction

Zheng Wang, Dr. Patricia Evans, and Dr. Virendra Bhavsar.
Faculty of Computer Science, University of New Brunswick, Fredericton, Canada.

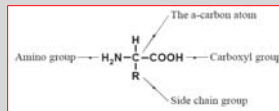


Abstract

Protein structure prediction is one of the most significant problems in bioinformatics. Currently, there are many tools that predict protein secondary structures, or find protein structural motifs and some specific structural segments. Sometimes the tool results are different or contradictory. CISPred is a consensus protein structure prediction system that integrates results of individual tools to provide consensus predictions. CISPred has an 83% average 3-state accuracy (Q3 score) on 109 CASP (Critical Assessments of Techniques for protein Structure Prediction experiment) sequences, and has an 89% average 3-state accuracy on 1785 randomly selected sequences.

Proteins

Proteins are large molecules. Every protein molecule is a long chain of 20 types of amino acids. The figure on the right shows the general formula of amino acids. The largest known protein is formed by up to 27,000 amino acids.



Each of the amino acid can be represented by a character. The string composed by these 20 characters is called an *amino acid sequence* as shown left. Amino acid sequences are the inputs of CISPred.

```
EDIIVVLYDYEAIHHEDLSPKQGDQMVVLEESGEWMAKSLATRKEYIP
SNYVARVDSLETBEWFKGSRKDAERQLLAPGNLMSFMRDSETKGYS
SLSVRYDPRGQDVTVK
```

Protein secondary structures

Proteins can have very complex three dimensional structures. However, all complex protein structures are composed from only 9 types of structural patterns. Each of these patterns can be represented by a character. The string composed by these 9 characters are named *protein secondary structure sequences* as shown right.

The following are some common structural patterns.

```
CCCCERSSCCSSSSCCCTCEEEERCCCTEEEEEETTCCEEEE
GGGEEETTSGGGSTTEETTCCHHHHHHHHSTTCCTTCEEEECSSSTSE
EEEEERCTTSCREEE
```



Alpha-helix



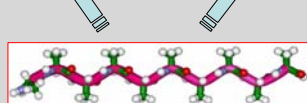
Beta-barrel



Parallel beta-sheet



Anti-parallel beta-sheet



Beta-strand

(Beta-sheets and beta-barrels are composed from Beta-strands)

Usually, all helices are represented by the letter "H", all sheets are represented by the letter "E", and all the other patterns are represented by the letter "C". The outputs of CISPred are secondary structure sequences containing "H", "E", and "C".

Motivations

Protein structures determine protein functions. However, to discover proteins structures by experimental methods such as X-ray crystallography or NMR analysis are expensive and slow. Furthermore, the number of known protein sequences is exponentially increasing. Therefore, the use of computer technologies becomes more and more indispensable to protein structure prediction.

A part of PSIPRED result

```
Conf: 97124684478999989898999999526899873356447887788
Pred: CCCCCCENNHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
AA: QAGITGRPEWIMLAGTALMGLTFLYFKMGVSDPDAKKFYAITTLVP
10 20 30 40 50
```

A part of SSPPRO result

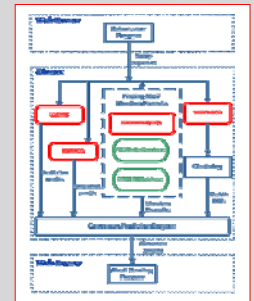
```
QAGITGRPEWIMLAGTALMGLTFLYFKMGVSDPDAKKFYAITTLVP
CCCCCCCCCENNHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
```

There are several existing tools that can predict protein secondary structures. However, there results are sometimes are contradictory for proteins.

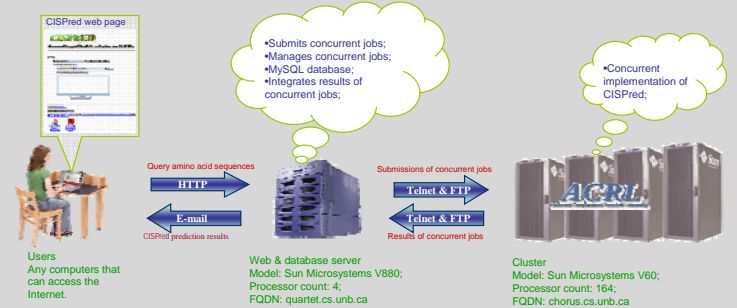
Methodology

CISPred integrates the following tools and databases to get consensus predictions:

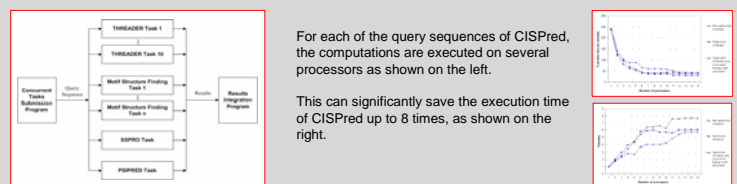
- **SSPRO** [1]: A protein secondary structure prediction tool using neural networks techniques and PSI-BLAST profiles.
- **PSIPRED** [2]: A protein secondary structure prediction tool using four separate neural networks.
- **THREADER** [3]: A tool that threads the sequence through structure libraries and selects the best alternatives.
- **PATMATMOTIFS** [4]: A tool that finds motifs from amino acid sequences.
- **PROSITE** [5]: A database containing information about protein motifs.
- **PDBFINDER** [6]: A database containing known protein amino acid sequences and protein secondary structure sequences.



System architecture of CISPred



Concurrent Implementation of CISPred



For each of the query sequences of CISPred, the computations are executed on several processors as shown on the left.

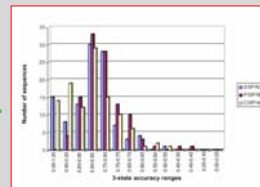
This can significantly save the execution time of CISPred up to 8 times, as shown on the right.

Experimental results

CISPred and the other two existing prediction tools PSIPRED and SSPPRO are tested on 109 CASP sequences and 1785 random sequences.

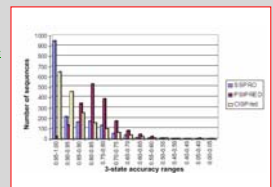
109 sequences

- Accuracy:
- SSPPRO: 82%
- PSIPRED: 78%
- CISPred: 83%



1785 sequences

- Accuracy:
- SSPPRO: 91%
- PSIPRED: 80%
- CISPred: 89%



References

- [1] G. Piatkowski, D. Przytycki, B. Rost, and P. Baldi. Improving the prediction of protein secondary structure in three- and eight classes using recurrent neural networks and profiles. *Proteins*, vol. 47, pp. 228-235, 2002.
- [2] D. T. Jones. Protein secondary structure prediction based on position specific scoring matrices. *Journal of Molecular Biology*, vol. 292, pp. 195-202, 1999.
- [3] D. T. Jones. THREADER: Protein sequence threading by double dynamic programming. In *Computational Methods in Molecular Biology* (S. Salzberg, D. Seaton, and S. Kasif, eds.), pp. 32, ch. 13. Amsterdam, Netherlands.
- [4] A. Batsch, P. Bucher, and K. Hofmann. The PROSITE database. In *status in 1997*. *Nucleic Acids Research*, vol. 25, pp. 217-221, 1997.
- [5] N. Hulo, C. J. A. Sigrist, V. Le Saux, et al. Recent improvements to the PROSITE database. *Nucleic Acids Research*, vol. 32, pp. D134-D137, 2004.
- [6] R. W. Hood, M. Schart, C. Sande, and G. Virendra. The PDBFINDER database: A summary of PDB, DSSP and HSSP information with added value. *CABIOS*, vol. 12, pp. 525-529, 1996.

