NNSSP

Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiply sequence alignments

Method description:

Yi and Lander (*) developed a neural-network and nearest-neighbor method with a scoring system that combined a sequence similarity matrix with the local structural environment scoring scheme of Bowie et al.(**) for predicting protein secondary structure. We have improved their scoring system by taking into consideration N- and C-terminal positions of a-helices and b-strands and also b-turns as distinctive types of secondary structure. Another improvement, which also significantly decrease the time of computation, is performed by restricting a data base with a smaller subset of proteins which are similar with a query sequence. Using multiple sequence alignments rather than single sequences and a simple jury decision method we achieved an over all three-state accuracy of 72.2%, which is better than that observed for the most accurate multilayered neural network approach, tested on the same data set of 126 non-homologous protein chains.

Input sequence for this program should be in fasta format with 80 or less sequence letters per line.

(*) Yi T-M., Lander E.S. (1993) Protein secondary structure prediction using nearest-neighbor methods. J.Mol.Biol.,232:1117-1129.

(**) Bowie J.U., Luthy R., Eisenberg D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science, 253, 164-170.)

Accuracy:

Overall 3-states (a, b, c) prediction gives ~67.6% correctly predic- ted residues on 126 non-homologous proteins using the jack-knife test procedure. Using multiple sequence alignments instead of single sequences increases prediction accuracy up to 72.2%.

SEE ALSO **"SSP"** program.

Example of NNSSP output: This output contains probabilities (Pa and Pb) of a and b structures in 0-9 scale. Probability of c is approximately 10 - Pa - Pb.

```
ADENYLATE KINASE ISOENZYME-3, /GTP:AMP$
 L=  214 SS content: a-  0.43 b=  0.05 c=  0.52
                    10        20        30        40        50
 PredSS     aaaaaaa           aaaaaa          aaaaaaaa       aa
 AA seq     RLLRAIMGAPGSGKGTVSSRITKHFELKHLSSGDLLRDNMLRGTEIGVLA
 Prob a     998886510000011122445454222111113467755542213323 35
 Prob b     000012210000011344223212222332210011100101011344 43
                    60        70        80        90       100
 PredSS     aaaa        aaaaaaaaaaaaaaa             aaaaaaaaa
 AA seq     KTFIDQGKLIPDDVMTRLVLHELKNLTQYNWLLDGFPRTLPQAEALDRAY
 Prob a     5454320111034678988887754555333421000111358888887 5
 Prob b     2222100121000111100000000011123341010111000000001 1
                   110       120       130       140       150
 PredSS         bb     aaaaaaaa   bb      bbbb
 AA seq     QIDTVINLNVPFEVIKQRLTARWIHPGSGRVYNIEFNPPKTMGIDDLTGE
 Prob a     3211111111146676664332111000110000000000011111111 1
 Prob b     1213564332122211012224553100147876421001333321110 1
                   160       170       180       190       200
 PredSS                aaaaaaaaaaaaaaaaaaaaaa   bbb          a
 AA seq     PLVQREDDRPETVVKRLKAYEAQTEPVLEYYRKKGVLETFSGTETNKIWP
 Prob a     2343321114678899999776557788888662112111111112333 5
 Prob b     1232100000111000000000000000000010136554211111122 1
                   210
 PredSS     aaaaaaa
 AA seq     HVYAFLQTKLPQRS
```

```
Prob a      46687764210111
Prob b      22211110110001
```

Reference:

Salamov A.A., Solovyev V.V.

Prediction of protein secondary sturcture by combining nearest-neighbor algorithms and multiply sequence alignments. J.Mol.Biol.,1995, 247, 11-15.

Parameters:

Input	
Sequence	Input file with a sequence. Input sequence for this program should be in fasta format with 80 or less sequence letters per line.
Output	
Result	Name of the output file.