PSite

Search for of prosite patterns with statistical estimation

Method description:

The method is based on statistical estimation of expected number of a prosite pattern in a given sequence. It uses the PROSITE database (author: Amos Bairoch,1995) of functional motifs. If we found a pattern which has expected number significantly less than 1, it can be supposed that the analyzed sequence possesses the pattern function. Presented version 1 is the simplest version that search for patterns without any deviation from a given Prosite consensus. In the following version we will include this possibility. In the output of PSite we can see a prosite pattern, its position in the sequence, accession number, ID, Description in the PROSITE database as well as Document number where is pattern characteristics outlined. It must be noted that patterns which started at the beginning or end of protein sequence will be recognized along the whole sequence in this version. It may be useful for analysis of ORF or 6 frame translation sequences.

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

Acknowledgments: We acknowledge Ilgam Shahmuradov and Igor Rogozin which took part in development some applications of this method for nucleotide consensuses searching and Asya Salihova for protein sites searching on IBM PC.

Example of PSite output:

```
 PSite V1 - search for Prosite patterns
         10         20         30         40         50         60
 RLLRAIMGAPGSGKGTVSSRITKHFELKHLSSGDLLRDNMLRGTEIGVLAKTFIDQGKLI
         70         80         90        100        110        120
 PDDVMTRLVLHELKN*TQYNWLLDGFPRTLPQAEALDRAYQIDTVINLNVPFEVIKQRLT
        130        140        150        160        170        180
 ARWIHPGSGRVYNIEFNPPKTMGIDDLTGEPLVQREDDRPETVVKRLKAYEAQTEPVLEY
        190        200        210        220        230        240
 YRKKGVLETFSYTETNKIWPHVYAFLQTKLPDANKDDALDQREWSAAAAWLAAAAALDLN
        250        260        270        280        290        300
 AGCPAAALAAAAAGSAACAAAAAFAAAAAACCAACAAAAAAACAAAADAACGAYAYACAP

ID   GLYCOSAMINOGLYCAN; RULE.
AC   PS00002;
DE   Glycosaminoglycan attachment site.
DO   PDOC00002;
PA   S-G-x-G.
 Sites found:  1 Expected number:   0.0272 95% confidential interval:   0
  #  Start  End  Expected  Site sequence
  1    12    15   0.0272  SGKG
ID   EF_HAND; PATTERN.
AC   PS00018;
DE   EF-hand calcium-binding domain.
DO   PDOC00018;
PA   D-x-[DNS]-{ILVFYW}-[DENSTG]-[DNQGHRK]-{GP}-[LIVMC]-[DENQSTAGC]-x(2)-
PA   [DE]-[LIVMFYW].
 Sites found:  1 Expected number:   0.0004 95% confidential interval:   0
  #  Start  End  Expected  Site sequence
  1   212   224   0.0004  DANKDDALDQREW
ID   ADENYLATE_KINASE; PATTERN.
AC   PS00113;
DE   Adenylate kinase signature.
DO   PDOC00104;
PA   [LIVMFYW](3)-D-G-[FY]-P-R-x(3)-[NQ].
 Sites found:  1 Expected number:   0.0000 95% confidential interval:   0
  #  Start  End  Expected  Site sequence
  1    81    92   0.0000  WLLDGFPRTLPQ
```

Reference:

Solovyev V.V., Kolchanov N.A. 1994,
Search for functional sites using consensus
In Computer analysis of Genetic macromolecules. (eds. Kolchanov N.A., Lim H.A.), World Scientific, p.16-21.