NAME

saps - Statistical Analysis of Protein Sequences

SYNOPSIS

saps [-dtv] [-T] [-s species] [-H] [-a XY...] [-o outfname] [-p] [-b/B libname] [-l lstfname] seqfname(s)

DESCRIPTION

SAPS evaluates by statistical criteria a wide variety of protein sequence properties. Properties considered include compositional biases; clusters and runs of charge and other amino acid types; different kinds and extents of repetitive structures; locally periodic motifs; and anomalous spacings between identical residue types. The statistics are computed for any single (or appropriately concatenated) protein sequence input. Statistically significant sequence features highlighted by SAPS in the input sequence may suggest promising regions for experimental investigation. The program also finds application in the description of conserved features of families of proteins as well as in the inverse problem of deriving protein groupings based upon sequence features.

Short sequences are subject to larger statistical fluctuations than longer sequences. The statistical evaluations of SAPS are reliable only for sequences of at least about 200 residues. Shorter sequences may in some cases be appropriately concatenated and analyzed as a representative combined sequence (e.g., histones, or Ras family proteins).

The SAPS program was developed in the group of Prof. Samuel Karlin at Stanford University. The program is available via anonymous ftp from gnomic.stanford.edu. Correspondence relating to SAPS should be addressed to Volker Brendel at the Department of Mathematics, Stanford University, Stanford CA 94305, U.S.A.; phone: (415) 723-9256; fax: (415) 725-2040; email: volker@gnomic.stanford.edu. Users of the program should cite the following reference:

Brendel, V., Bucher, P., Nourbakhsh, I., Blaisdell, B.E., Karlin, S. (1992)
Methods and algorithms for statistical analysis of protein sequences.
Proc. Natl. Acad. Sci. USA 89: 2002-2006.

OPTIONS

-d Generate documented output.

-t Generate terse output.

-v Generate verbose output.

-T Append computer-readable summary output to file `saps.table'.

-s species
Use species.q for quantile comparisons.

-H Count H as positive charge.

-a XY...
Analyze spacings of amino acids X, Y, ...

-o outfname
Redirect output to file outfname [default: stdout].

-p Read protein sequence data from stdin.

-b libname
Read protein sequence data from library file libname.

-l lstfname
Read protein sequence data from files specified in LST_lstfname.

seqfname(s)
Read protein sequence data from file(s) seqfname(s).

USAGE

Input File Format

Input to SAPS consists of individual protein sequences of lengths not exceeding 10,000 residues. Input is supplied by the arguments seqfname(s), -p, -l lstfname, and -b/B libname.

A. seqfname(s)

Individual sequences are supplied via the files seqfname(s) in minimal EMBL format: the first line of the

file is a descriptor line which will be printed, following lines (if any) are annotation, the first line of the sequence is immediately preceded by a line beginning with the delimiter `SQ', and subsequent symbols are A-Z (one-letter-code symbols) as part of the sequence or irrelevant characters (like numbers and blanks); non-standard symbols for ambiguous or missing residues are ignored. Lines should not exceed 512 characters. SWISS-PROT files may be used without change in the distributed format (for such files, also the DE line is printed by default). Optionally, the input may also include the corresponding coding sequence; if so, the coding sequence should precede the SQ line and should commence with a `CQ' delimiter on a single line.

Example (SWISS-PROT entry for Drosophila cut protein):

```
     ID   HMCU_DROME     STANDARD;      PRT;  2175 AA.
     (any number of comment lines that not beginning with `CQ' or `SQ')
      CQ
           1  ATGCAGCCAA CATTGCCACA AGCCGCTGGG ACAGCCGATA TGGATCTGAC
             (sequence continued)
        6481  GCGGTAACCA CTGCAGCAGC AACTGCGGCA GCCGGTTGGA ACTACTAA
      SQ
           1  MQPTLPQAAG TADMDLTAVQ SINDWFFKKE QIYLLAQFWQ QRATLAEKEV
             (sequence continued)
        2161  AVTTAAATAA AGWNY
```

B. -p

This option allows to read input formatted as described under A to be read from stdin. One possible use for this option is in conjunction with a file reformatting program such as ReadSeq (D.G. Gilbert; available via anonymous ftp from ftp.bio.indiana.edu, directory molbio/readseq). Thus, for a protein data file in any format recognized by ReadSeq, one may run saps with the command `readseq -p -f4 seqfname | saps -p', for example.

C. -l lstfname

There are two other possible inputs to SAPS that can be used alternatively or in conjunction with sequence file input as described above. If the -l lstfname command line flag is specified, input is taken from files in minimal EMBL

format, the names of which are specified in the file LST_lstfname. A list file must be named with a prefix LST_ and arbitrary suffix lstfname. It must have two lines of comments indicated by a # symbol in the first position followed by lines giving the names of input files in minimal EMBL format, one per line. Memory limitations on the system may limit the number of input files that can be specified in this way.

Example:

```
#'HELIX.*LOOP.*HELIX' proteins:
#
ARLC_MAIZE
ARRS_MAIZE
ASH1_RAT
```

D. -b libname

Library files (invoked by the command line flag -b libname) contain one or more sequence files assembled in LIB format: one-line descriptors beginning with > in the first position followed by the sequence in free format (non-one-letter-code symbols again being ignored; up to 10,000 characters per line). Memory limitations on the system may limit the number of input files that can be specified in this way.
Example:

```
>SW;ARLC_MAIZE: ANTHOCYANIN REGULATORY LC PROTEIN (GENE NAME: LC).
MALSASRVQQAEELLQRPAERQLMRSQLAAAARSINWSYALFWSISDTQP(sequence continued)
>SW;ARRS_MAIZE: ANTHOCYANIN REGULATORY R-S PROTEIN (GENE NAME: R-S).
MAVSASRVQQAEELLQRPAERQLMRSQLAAAARSINWSYALFWSISDTQP(sequence continued)
>SW;ASH1_RAT: ACHAETE-SCUTE HOMOLOGUE 1 (GENE NAME: MASH-1).
MESSGKMESGAGQQPQPPQPFLPPAACFFATAAAAAAAAAAAAAQSAQQQ(sequence continued)
```

Running SAPS on each of the above three sequences could thus be done in any of the following ways (assuming that the list file under C is named LST_hlh and that the library file under D is named LIBhlh):

```
a) saps ARLC_MAIZE ARRS_MAIZE ASH1_RAT > OUTPUT
b) saps -b LIBhlh > OUTPUT
c) saps -l hlh > OUTPUT
```

E. -B libname

This input flag assumes a library file of coding (rather than protein) sequences. The format is as above, except that the sequences are assumed to be nucleotide sequences in the ACGTN alphabet (all other characters ignored).

Output format

Output is directed to standard output. To run SAPS on the sequence file HMCU_DROME, for example (see above), one might type the command `saps HMCU_DROME | more' or `saps HMCU_DROME > OUTPUT'. The output format can be modified by the flags -d, -t, or -v, and -T:

-d The output will come with documentation that annotates each part of the program; this flag should be set when SAPS is used for the first time as it provides helpful explanations with respect to the statistics being used and the layout of the output.

-t This flag specifies terse output that is limited to the analysis of the charge distribution and of high scoring segments.

-v This flag specifies verbose output with more detail than normally required.

-T This flag is used in conjunction with the analysis of sets of proteins (specified typically with the -b libname or -l lstfname options); if specified, the file `saps.table' is appended with computer-readable lines describing the input files and their significant features.

Compositional analysis

The residue composition of the input protein may be evaluated relative to standard sets of proteins grouped by species, size class, subcellular location, function, or other criteria. Specifically, the composition of the input protein is compared with the quantile table of residue usage for the the user-specified standard set. Extremal usages which fall in the tails of the reference distribution are indicated for individual amino acids, charged and hydrophobic residues. The reference set is selected with the command line flag `-s species'. The following options for `species'

are currently supported: BACSU (Bacillus subtilis); CHICK (chicken); DROME (Drosophila melanogaster); ECOLI (Escherichia coli); HUMAN (human); MOUSE (mouse); RAT (rat); XENLA (frog); YEAST (Saccharomyces cerevisiae); swp23s (random sample of proteins from SWISS-PROT, Release 23.0). By default, a sequence file ending in _SPECIES is evaluated with the quantile table SPECIES (if among the ones listed above); otherwise swp23s is used. For each reference set, only proteins of lengths at least 200 residues were included; redundant entries were culled (for lists of SWISS_PROT file names composing each set and the quantile tables see directory SAPS/Inc).

Classification of histidine

By default, SAPS treats only lysine (K) and arginine (R) as positively charged residues. If the command line flag `-H' is set, then histidine (H) is also treated as positively charged in all parts of the program involving the charge alphabet.

Analysis of specified amino acid distribution

Clusters of particular amino acid types may be evaluated by means of the same tests that are used to detect clustering of charged residues (binomial model and scoring statistics). These tests are invoked by setting the `-a' flag; for example, to test (separately) for clusters of alanine (A) and serine (S), set `-a AS'. The binomial test is also programmed for certain combinations of amino acids: AG (flag `-a a'), PEST (flag `-a p'), QP (flag `-a q'), ST (flag `-a s').

FILES

SAPS.SSPA/Inc/(files)
SAPS.SSPA/README

SAPS.SSPA/saps.1 (this file)
SAPS.SSPA/testpro
SAPS.SSPA/testdnapro
SAPS.SSPA/testout
SAPS.SSPA/testout_dflag
SAPS.SSPA/testout_dna

NOTES

A hardcopy of this manual page is obtained by `man -t saps'.

AUTHOR

Volker Brendel <volker@gnomic.stanford.edu>