CTL-Epitope

This program is designed for prediction of CTL epitopes of length=9 in protein sequences.

Datasets

For training data we used set of epitopes of length 9 from MHCBN database (Bhasin *et al*, (2003) *Bioinformatics*, 19,666). CTL epitopes which possess binding and activity and sequence length 9 were selected from the database without non-standard amino acid codes and no sequence duplication.

To construct negative dataset we found all sequences from SWISS-PROT database that contain at least one of the epitopes (1717 sequences). From these sequences all the overlapping fragments of length 9 were obtained. From this set of overlapping peptides those were removed, which overlapped with epitope sequences. The remained sequences were filtered so that any of the pair of sequences have no more tan one amino acid in common out of 9 positions. The epitope sequences (932) are the positive set, all the other sequence fragments comprise the negative set (131710). To test the performance the overall data set was splitted randomly on the training and testing sets. The training set comprises 112380 sequences (704 positive). The testing set comprise of 20262 sequences (228 out of them were positive).

Algorithm

To classify sequences the following scores were implemented. (1) Weight matrix scores for each peptide position for PSSM (position specific scoring matrix) formed by positive set sequences, they presented ; (2) positive and negative sequence sets are scanned for the sequence similarity by BLOSUM62 matrix with query sequence and top 5 sequences from both sets separately is determined (5 top from positive set, 5 top from negative set). The similarity scores for positive set ranked by their value and formed additional 5 classification parameters. The similarity scores for negative set ranked by their value and formed another 5 classification parameters. Overall 19 parameters are implemented (9 PSSM positional weights, 5 top positive set similarity scores and 5 top negative set similarity scores). The separation is performed by Linear Discriminant Analysis.

Error estimates

Error estimates on the test set were calculated:

The prediction quality (fraction of correctly predicted sequences) q=0.839058.
npos=228 (epitope sequences)
npos_true=178
npos_false=50
nneg=20034 (non-epitope sequences)
nneg_true=16823
nneg_false=3211
Quality: all=0.839
Positive set =0.781
Negative set=0.840

Input data:

Protein sequence in 20-letter alphabet in FASTA format.

Input Parameters:

- List Output: if this check box is set checked, output data contain list of predicted peptides with their locations in the sequence and scores.
- Threshold: This parameter specifies at which score value will separate positive examples (predicted epitopes, score >= threshold) and negative examples (non-epitopes, score < threshold). By default, threshold=0 (recommended).

Output data:

For each position of the sequence (except eight C-terminal positions) the program output whether the polypeptide of length 9 starting at this position is predicted as cytotoxic T lymphocyte epitope(*) or not ( ). If List Output checkbox is checked, list of predicted epitopes is printed out.

Output example

```
# CTL-epitope-Finder ver. 1.1:
# Program for prediction of  putative cytotoxic T-lymphocyte (CTL) epitopes
# Softberry Inc., 2005
# N-terminal positions of positive peptides (length=9) marked by '*'
# THRESHOLD=0.000
# SEQUENSE LENGTH=191
# NUMBER OF POSITIVE PREDICTIONS=20
# Epitope prediction:
>HCV_core
    .    10   .    20   .    30   .    40   .    50   .    60
MSTNPKPQKKNNRNTNRRPQDVKFPGGGQIVGGVYLLPRRGPRLGVRATRKTSERSQPRG
*     *                   *       *    * *            *
    .    70   .    80   .    90   .    100  .    110  .    120
RRQPIPKARQPEGRAWAQPGYPWPLYGNEGLGWAGWLLSPRGSRPSWGPTDPRRRSRNLG
 *     *              *    *  *            *
    .    130  .    140  .    150  .    160  .    170  .    180
KVIDTLTCGFADLMGYIPLVGAPLGGAARALAHGVRVLEDGVNYATGNLPGCSFSIFLLA
          *                                 *  * *       ***
    .    190  .    200  .    210  .    220  .    230  .    240
LLSCLTIPASA


# Output positive peptide list
# Start-End [score]: SEQUENCE
  1-  9 [+13.193]: MSTNPKPQK
  7- 15 [ +0.630]: PQKKNNRNT
 28- 36 [+24.625]: GQIVGGVYL
 36- 44 [+27.123]: LLPRRGPRL
 41- 49 [+25.420]: GPRLGVRAT
 43- 51 [+24.164]: RLGVRATRK
 57- 65 [ +2.835]: QPRGRRQPI
 62- 70 [ +4.587]: RQPIPKARQ
 68- 76 [ +1.264]: ARQPEGRAW
 83- 91 [ +2.128]: WPLYGNEGL
 88- 96 [+20.329]: NEGLGWAGW
 91- 99 [ +3.308]: LGWAGWLLS
104-112 [ +6.383]: RPSWGPTDP
132-140 [+14.183]: DLMGYIPLV
164-172 [ +1.569]: YATGNLPGC
167-175 [ +1.402]: GNLPGCSFS
169-177 [+25.489]: LPGCSFSIF
177-185 [ +5.293]: FLLALLSCL
178-186 [ +5.299]: LLALLSCLT
179-187 [ +1.837]: LALLSCLTI
```