PCPMer - the methods and mathematics

Petr Danecek

March 20, 2009

Abstract

Important regions shared amongst multiple protein sequences can be recognised by analyzing physicochemical properties (PCPs) of amino acids occurring in the sequences. PCPMer uses five-dimensional quantitative descriptors of the 20 naturally occurring amino acids to compare statistical distributions of the physicochemical properties, as observed in the analyzed sample against random background. Starting with one sequence, the software retrieves a set of related sequences (BLAST), generates a multiple sequence alignment (ClustalW), and returns a list of motifs together with a profile, an information on how conserved or variable is each column of the multiple sequence alignment. Sequence motifs are defined as continuous stretches of residues with either similar PCPs or statistically unusual distribution of PCPs. The motifs and variability data can be mapped on a PDB structure. Furthermore, the motifs can be used as an input for the data-mining MotifSearch tool, which will identify sequences in a protein database (Astral) sharing selected motifs in arbitrary order.

Contents

1 Introduction

On a logical level, the PCPMer software consists of four modules: multiple sequence alignment generator, MotifMaker, MotifSearch, and 3D Variability plotter. In this document are described the methods and mathematics used by PCPMer.

2 Quantitative descriptors

Each of the 20 naturally occuring amino acids is represented as a point in a five dimensional space, where the five dimensions roughly correspond to hydrophobicity/hydrophylicity (E1); size (E2); alpha-helix propensity (E3). The property E4 is partially related to the partial specific volume, number of codons and relative abundance of the amino acids; and E5 correlates weakly with β-strand propensity [1]. PCPMer was redesigned to work with any number of arbitrary descriptors, but for simplicity of formulation we will refer to them here as to "the five descriptors" or "the five PCPs".

The definition of the five quantitative descriptors:
`descriptors-VB2001-5.txt`
Perl script to generate vector library used by PCPMer:
`create-vectorlib`

3 Multiple Sequence Alignment Generator

Given a sequence containing unknown number and location of functionally important regions, additional information can be gained by comparing it with related sequences. In the ideal case, the sequence of interest comes from a pool of known sequences and the user submits the sequences either as a ClustalW input format or as an already completed multiple sequence alignment. However, the user often does not have the set of related sequences in advance. Starting with one sequence only, BLAST is used to find relevant sequences in a protein database and CD-HIT to filter out non-redundant representatives. A taxonomy tree of results is generated so that the user may select the sequences according to desired species, families, etc. for subsequent analysis.

http://landau.utmb.edu:8080/pcpmer/pcpmer/Tools/alignments.jsp

4 Motif Maker

For each column of the multiple sequence alignment are calculated average PCPs, standard deviations, relative entropy and physicochemical similarity of amino acids appearing in the given column of the alignment. These data form a

profile, which is used to recognise two types of motifs - based either on relative entropy [1] or physicochemical similarity [2].

The motifs are defined as continuous stretches of residues with relative entropy or physicochemical similarity values higher relative to the rest of the alignment (variable cut-off) or higher than a user specified threshold (fixed cut-off). The motif definition may contain restriction on the minimum motif length and allow certain number of non-significant residues to appear in the motifs, according to the user supplied parameters.

4.1 Motifs based on relative entropy

Because amino acids are distributed non-evenly in naturally occurring proteins, also the distribution of PCPs is not uniform. To compare the distribution of PCPs as found in a given column of the multiple sequence alignment to the background distribution of a random sample, the relative entropy (also known as Kullback-Leibler divergence) is used. The relative entropy is calculated for each of the five vectors E1 to E5 separately

$$K_p = \sum_{i=1}^{n} Q_p(i) \log_2 \frac{Q_p(i)}{P_p(i)}. \tag{1}$$

In the equation, Q_p are the discrete probability distributions of the five property values observed in the alignment and P_p are the probability distributions of a random sample. The index p iterates over the five properties E1 to E5 and the index i over the discrete probability distribution bins. By default, the 20 amino acids are grouped into 5 bins for each vector, so $n = 5$. Thus $Q_p(i)$ is the fraction of the component p observed in the bin i and $P_p(i)$ is the corresponding background frequency. In other words, $Q_p(i)$ gives the frequency at which a group of amino acids that are most closely matched in properties (according to p) occurs in a given column, and $P_p(i)$ is the corresponding expected random frequency calculated over the whole database.

Having the five entropy values for each of the vector, there are several ways how to calculate the value of the total relative entropy:

- The **maximum value** approach selects out of the five relative entropies the largest one.
- The **unscaled sum** approach, the sum of the five entropy values is taken as the total entropy.
- In the **scaled sum** approach, the five entropy values are first scaled to yield a "uniform" space, and then their sum is taken as the total entropy. Note that this approach has no theoretical justification and was implemented mainly for experimental reasons.

While the first two approaches are simple and intuitive, the scaled sum approach is best understood by example. The entropies calculated for absolutely

conserved columns of the 20 amino acids will have different maximum values for the five descriptors. In particular, the maximum value of the relative entropy for E1 is $K_{E1} = 0.74$ (Cys), while $K_{E2} = 0.62$ (Gly), $K_{E3} = 0.57$ (Pro), $K_{E4} = 1.00$ (Cys), and $K_{E5} = 0.49$ (Pro). The individual entropy values are in this approach first scaled so that the maximum value is for each entropy equal to 1.0. Thus the scale factors 1/0.74, 1/0.62, 1/0.57, 1 and 1/0.49 would be used for this set of descriptors (and background frequencies).

The total relative entropy is scaled so that the maximum value is 1.0 and thus any value obtained is always from the interval $[0, 1]$.

Example. Relative entropies for absolutely conserved columns.
The maximum value approach:

```
I: 0.28 0.21 0.20 0.22 0.38 .. 0.38
V: 0.28 0.21 0.39 0.22 0.39 .. 0.39
L: 0.28 0.21 0.39 0.22 0.25 .. 0.39
F: 0.28 0.45 0.45 0.25 0.25 .. 0.45
T: 0.45 0.21 0.20 0.25 0.39 .. 0.45
N: 0.21 0.21 0.45 0.43 0.45 .. 0.45
K: 0.21 0.32 0.20 0.22 0.45 .. 0.45
Q: 0.45 0.32 0.20 0.43 0.38 .. 0.45
R: 0.45 0.32 0.45 0.22 0.39 .. 0.45
E: 0.21 0.32 0.45 0.43 0.49 .. 0.49
S: 0.21 0.51 0.20 0.25 0.38 .. 0.51
A: 0.53 0.21 0.45 0.25 0.25 .. 0.53
M: 0.28 0.45 0.39 0.54 0.45 .. 0.54
H: 0.53 0.32 0.20 0.54 0.45 .. 0.54
D: 0.21 0.45 0.20 0.54 0.25 .. 0.54
W: 0.28 0.32 0.57 0.43 0.49 .. 0.57
P: 0.21 0.51 0.57 0.22 0.49 .. 0.57
G: 0.21 0.62 0.20 0.25 0.25 .. 0.62
Y: 0.74 0.45 0.57 0.25 0.38 .. 0.74
C: 0.74 0.21 0.45 1.00 0.39 .. 1.00
```

Unscaled sum approach:

```
I: 0.28 0.21 0.20 0.22 0.38 .. 0.47
L: 0.28 0.21 0.39 0.22 0.25 .. 0.49
K: 0.21 0.32 0.20 0.22 0.45 .. 0.51
T: 0.45 0.21 0.20 0.25 0.39 .. 0.54
V: 0.28 0.21 0.39 0.22 0.39 .. 0.54
G: 0.21 0.62 0.20 0.25 0.25 .. 0.55
S: 0.21 0.51 0.20 0.25 0.38 .. 0.56
D: 0.21 0.45 0.20 0.54 0.25 .. 0.60
F: 0.28 0.45 0.45 0.25 0.25 .. 0.61
A: 0.53 0.21 0.45 0.25 0.25 .. 0.61
N: 0.21 0.21 0.45 0.43 0.45 .. 0.63
Q: 0.45 0.32 0.20 0.43 0.38 .. 0.64
R: 0.45 0.32 0.45 0.22 0.39 .. 0.66
E: 0.21 0.32 0.45 0.43 0.49 .. 0.68
P: 0.21 0.51 0.57 0.22 0.49 .. 0.72
H: 0.53 0.32 0.20 0.54 0.45 .. 0.74
W: 0.28 0.32 0.57 0.43 0.49 .. 0.75
M: 0.28 0.45 0.39 0.54 0.45 .. 0.76
Y: 0.74 0.45 0.57 0.25 0.38 .. 0.86
C: 0.74 0.21 0.45 1.00 0.39 .. 1.00
```

Scaled sum approach:

```
I: 0.39 0.33 0.35 0.22 0.79 .. 0.53
L: 0.39 0.33 0.69 0.22 0.52 .. 0.55
K: 0.29 0.52 0.35 0.22 0.93 .. 0.59
T: 0.61 0.33 0.35 0.25 0.80 .. 0.60
V: 0.39 0.33 0.69 0.22 0.80 .. 0.62
D: 0.29 0.72 0.35 0.54 0.52 .. 0.62
G: 0.29 1.00 0.35 0.25 0.52 .. 0.62
S: 0.29 0.82 0.35 0.25 0.79 .. 0.64
A: 0.72 0.33 0.78 0.25 0.52 .. 0.67
F: 0.39 0.72 0.78 0.25 0.52 .. 0.68
Q: 0.61 0.52 0.35 0.43 0.79 .. 0.69
N: 0.29 0.33 0.78 0.43 0.93 .. 0.71
R: 0.61 0.52 0.78 0.22 0.80 .. 0.75
E: 0.29 0.52 0.78 0.43 1.00 .. 0.77
H: 0.72 0.52 0.35 0.54 0.93 .. 0.78
M: 0.39 0.72 0.69 0.54 0.93 .. 0.83
W: 0.39 0.52 1.00 0.43 1.00 .. 0.85
P: 0.29 0.82 1.00 0.22 1.00 .. 0.85
Y: 1.00 0.72 1.00 0.25 0.79 .. 0.96
C: 1.00 0.33 0.78 1.00 0.80 .. 1.00
```

Note that higher values of relative entropy do not necessarily imply conserved columns, but rather, unusal combinations of amino acids. For example, a column which contains only combinations of the Glu and Pro residues in the 1:1 ratio will be identified as unlikely, and therefore such a column will have large relative entropy. In contrast, columns containing only Leu or Ile residues in the ratio 1:1 will have smaller entropy values, as those amino acids are abundant and much more likely to occur.

Example. Relative entropy (unscaled sum approach) of a column containing the same number of Glu and Pro or Leu and Ile, calculated by `relentropy`.

```
relentropy G:1 P:1
P:50 G:50        ..       0.21 0.40 0.22 0.08 0.21 .. 0.40

relentropy I:1 L:1
I:50 L:50        ..       0.28 0.21 0.13 0.22 0.16 .. 0.36
```

Any gaps which ocurr in the multiple sequence alignment are in the relative entropy calculations treated the same way as an amino acid. The quantitative descriptors of the gap are set to the average values of descriptors of a random sample. This way is the relative entropy in Equation 1 lowered to 0 as the value of Q_p approaches the value of P_p.

The user can choose from multiple background frequencies specific to human, eucaryota, bacteria, virus and other protein databases.

4.2 Motifs based on similarity

To recognise columns conserved in a more intuitive sense, the concept of physicochemical similarity is used. The calculation of similarity is based on *physicochemical distances* between all pairs of amino acids appearing in the given

column of the alignment. The average distance D is given by

$$D = \frac{2}{N(N-1)} \sum_{i \le j}^{N} \sqrt{\sum_{p=E1}^{E5} (V_i^p - V_j^p)^2}, \tag{2}$$

where V_i^p, $p = E1, \ldots, E5$, are the five quantitative descriptors of i-th amino acid and N is the number of sequences in the multiple sequence alignment. The physicochemical distance is converted to a *physicochemical similarity* S with a fixed range of values

$$S = \frac{N_{\text{no gaps}}}{N} \exp(-0.1D). \tag{3}$$

The similarity is equal to 1 for absolutely conserved (identical) columns and 0 for the more diverse. The definition of similarity also contains a term which lowers its value when gaps are present in the column of the alignment.

Example. Physicochemical similarity of a column containing the same number of Glu and Pro or Leu and Ile, calculated by `relentropy`.

```
relentropy G:1 P:1
P:50 G:50        ..       similarity=0.24, PCdist=14.08

relentropy I:1 L:1
I:50 L:50        ..       similarity=0.46, PCdist=7.70
```

http://landau.utmb.edu:8080/pcpmer/pcpmer/Tools/SubmitFormMotifMaker.jsp

5 Motif Search

The profiles of motifs identified by MotifMaker can be used to look for sequences containing similar motifs in a protein database. Lorentzian scoring scheme is used to measure how well does a motif match a sequence in the protein database and a Bayesian method is then used to score the sequence according to the number and the quality of the motifs within the sequence [3]. The scores are calculated as follows.

5.1 "One motif against the sequence" score

To determine how well does a motif match a sequence, the motif is aligned to every position k of the sequence and the score values $S(k)$ are calculated (see below). The resulting score S is the maximum of the values calculated for each position k, that is

$$S = \max |S(k)|. \tag{4}$$

The score values $S(k)$ are based on the Lorentzian scores $S^p_{k,i}$ calculated for each position i within the motif as

$$S^p_{k,i} = \left[1 + \left(\frac{V^p_{k+i} - \langle V^p_i \rangle}{W\sigma^p_i + \Phi}\right)^2\right]^{-1}. \tag{5}$$

Here, k denotes the position in the sequence, to which the motif has been aligned; $i = 0, \ldots, n-1$ is the position index within the motif of the length n; V^p_{k+i}, $p = E1, \ldots, E5$, are the five quantitative descriptors of the amino acid at the position $k+i$ in the sequence; $\langle V^p_i \rangle$ is the average PCP value at the position i of the motif and σ^p_i is the corresponding standard deviation. W is the weight for standard deviation (by default set to 1.5) and the small positive shift Φ (set to 0.001) was added to prevent overflow during calculation when σ^p_i is zero.

The score value for the motif of the length n aligned at the position k in the sequence is then calculated as

$$S(k) = \sum_{\substack{p=E1,\ldots,E5 \\ i=0,\ldots,n-1}} S^p_{k,i} \Big/ \sum_{\substack{p=E1,\ldots,E5 \\ i=0,\ldots,n-1}} 1. \tag{6}$$

In the calculation are included only significant pairs (p, i), for which the relative entropy of the motif exceeds a certain threshold (default is set to 0.21).

The scores S are values from the interval 0 to 1, the score of 1 indicates a perfect match. To estimate the significance of a match, the scores are compared to the average motif score $\langle S \rangle_{\mathrm{db}}$ evaluated for all sequences in the protein database and the corresponding standard deviation σ_{db}. If the score is smaller than $\langle S \rangle_{\mathrm{db}} + 2\sigma_{\mathrm{db}}$, the motif is present with approximately the same scores in most other sequences in the database and such a match is considered non-significant. If the score is bigger than $(1 + \langle S \rangle_{\mathrm{db}} + 2\sigma_{\mathrm{db}})/2$, most other sequences in the database do not contain anything like the motif and such a match is considered significant.

5.2 "Multiple motifs against the sequence" score

The Bayesian method is used to decide if a given sequence is a sufficient match to a set of m motifs with the scores $\{S_1, \ldots, S_m\}$. The total score for the sequence is computed as

$$S_{\mathrm{tot}} = \sum_{i=1}^{m} \frac{S_i \; |\langle S \rangle_{\mathrm{aln}} - \langle S \rangle_{\mathrm{db}}|}{(\sigma_{\mathrm{aln}} + \sigma_{\mathrm{db}})^2 + \epsilon}. \tag{7}$$

The small positive shift ϵ (by default set to 10^{-10}) is added to prevent division by zero in case that the relative entropy threshold is too restrictive and excludes all pairs (p, i) in Equation 6.

5.3 Arbitrary substring search

The program can be used to search for arbitrary strings without a profile. This is accomplished by creating a fake alignment containing only one sequence, the

string itself. The program will generate a profile with the average PCPs equal to those of the amino acids in the string, the standard deviations set to zero, the relative entropies and similarities set to 1, and the physicochemical distances set to 0. After that, exactly the same method as above are used.

http://landau.utmb.edu:8080/pcpmer/pcpmer/Tools/SubmitFormMotifSearch.jsp

6 3D Variability Plotter

PCPMer can map motifs and the variability data (relative entropy or physicochemical similarity) onto a 3D structure. The program creates the mapping by running `ClustalW` for a selected chain of the PDB file and a selected sequence of the multiple sequence alignment. The variability data are displayed using a user defined color scale based on three values $v_{\min}$, v_{mid}, $v_{\max}$ and corresponding colors $c_{\min}$, c_{mid}, $c_{\max}$. The color c for the value v is determined trivially as

$$c = \begin{cases} v\,(c_{\mathrm{mid}} - c_{\min})/(v_{\mathrm{mid}}) + c_{\min} & \text{for } v < v_{\mathrm{mid}} \\ (v - v_{\mathrm{mid}})(c_{\max} - c_{\mathrm{mid}})/(v_{\max} - v_{\mathrm{mid}}) + c_{\mathrm{mid}} & \text{for } v \geq v_{\mathrm{mid}} \end{cases} \tag{8}$$

http://landau.utmb.edu:8080/pcpmer/pcpmer/Tools/SubmitFormVariability.jsp

References

[1] Mathura S. Venkatarajan and Werner Braun. New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical–chemical properties. *Journal of Molecular Modeling*, 7(12):445–453, December 2001.

[2] Petr Danecek and Catherine H Schein. Flavitrack analysis of the structure and function of West Nile non-structural proteins. *International Journal of Bioinformatics Research and Applications*, 2009.

[3] Venkatarajan S Mathura, Catherine H Schein, and Werner Braun. Identifying property based sequence motifs in protein families and superfamilies: application to DNase-1 related endonucleases. *Bioinformatics (Oxford, England)*, 19(11):1381–90, July 2003. PMID: 12874050.