Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Subject Section

SSR-viz - a toolbox to detect and visualize protein subfamily specific residues

Paul Zierep [1],*, Stefan GÃ¼nther [2]

[1] Department, Institution, City, Post Code, Country and
[2] Department, Institution, City, Post Code, Country.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Protein sub-families share a similar structure but differ in functionality such as substrate specificity. These differences can often be assigned to a limited set of residues, which are difficult to detect from sequence data alone. Here, we introduce a toolbox which allows to detect and visualize these residues based on a multiple sequence alignment. (todo Maybe Output ??) The applied algorithm can be adjusted to a given problem and allows users to perform a thorough analysis of their sequences.
Results: Substrate specificty of non-ribosomal peptide synthases were analysed. The detected residues are located in the binding side and are in accordance with previous performed studies.
Availability: The toolbox is written as a GUI in Python3 and is available via PyPi and the GitHub repository (todo). Standalone executalbes are also available for Windows and Ubuntu.
Contact: name@bio.com
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein sub-families differ from each other in very specific functions, such as substrate specificity, protein-protein interaction or reaction type. These functionalities can often be assigned to a limited set of residues (referred to as subfamily specific residues (SSRs) in this article). The identification of those SSRs is crucial to the understanding of protein sub-family diversity. SSRs with high statistical significance can form the basis for various experiments such as side directed mutagenesis and rational design.

Various approaches have been proposed to detect SSRs based on analysis of multiple sequence alignments (MSAs) of two or more protein subfamilies (todo cite). All these approaches share the common objective: To identify positions in an alignment which are conserved in one protein subfamily but differ between subfamilies. Some algorithms also include physico-chemical properties of the amino acids in order to distinguish true SSRs from residues which are due to evolutionary mutations.

In order to allow researchers a thorough study of their proteins we implemented an algorithm which follows this common objective but is also flexible and can be adjusted to the specific protein family observed.

Additionally, SSR-viz provides assistance for various tasks commonly encountered on the search for SSRs. The main features are: The easy mapping of the alignment with a CSV-file which holds the information of the subfamily class labels. The addition of a window function, which allows to observe single amino acid residues as well as general areas of interest in the alignment. The residues can be mapped to a protein structure, which facilitates the further investigation in a protein visualization tool such as PyMol (cite). The creation of a Jalview (cite) annotation file which allows to integrate the results in the Jalview Alignment Editor together with the original alignment.

2 Algorithm

$$a \tag{1}$$

$$a \tag{2}$$

$$a \tag{3}$$

Initially, each subfamily class in the alignment is converted into a position probability matrix (PPM), which represents the probability to find a specific amino acid at a given position (see eq. ...). This allows to compare the subfamily classes independent of the number of their sequences.

Each position in the alignment are assigned to a final score based on the combination of three independent scoring functions.

The conservedness of each position is defined by the normalized shannon entropy (see eq. ..). The difference between two classes is computed by summation of a exchange matrix. The matrix represents the probability of each amino acid in one class to be exchanged with another amino acid in the other classes. The more amino acids are exchanged the higher the score (see eq. ...). This matrix judges each amino acid exchange equally.

In order to differentiate the exchange of divers amino acids, an additionally scoring function is implemented similar to eq. 2, but this exchange matrix can be weighed with a substitution matrix such as Pam or Blossum. A purely physico-chemical substitution matrix is also available based on the research (?) of ....

3 Implementation

SSR-viz is based on a MSA as Input, additionally a CSV-File is needed which holds the subfamily specific label of each sequence. The CSV-File can be created using the **CSV_builder** tool. This tool allows also the label extraction from the sequence name using regular expressions. The CSV-File can have multiple labels, in case different types of functionality are investigated. The **CSV_builder** also excludes redundant sequences by default.

The detection of SSRs is handled by the **SSR_plot** tool. SSRs can be computed in three different kind of schemes: One-vs-one, which returns the SSRs to distinguish each class from another. One-vs-All, which returns SSRs to distinguish one class from all the others classes, and All-vs-All, which returns the SSRs to distinguish all the classes from each other.

The detection threshold for SSRs can be assigned in two ways. Either the top N SSRs can be returned independent of their significance or the SSRs can be returned based on their Z-score, which only returns positions that have a significant higher score then the other positions in the alignment.

SSR-viz supports three different kind of outputs: A mathplotlib-style plot in PDF Format, which is highly customizable and consists of a heatmap, which visualizes all the scores and a plot which shows the most significant SSRs. A window function can also be added to the plot, which allows to investigate the alignment for areas of general importance.

It is often desired to visualize the SSRs together with the MSA, therefore, an Jalview Annotation File can also be created, this File can be loaded into the Jalview Alignment Editor.

Finally a summary CSV-file can be created, this file shows the SSRs together with the conserved positions in each subfamily class.

It is often desired to investigate the SSRs in a structural context. Therefore, we also implemented a tool **Add_pdb**, which allows to map protein structure indices from a PDB File to the indices in the MSA. **Add_pdb** supports any PDB file, downloaded from PDB or custom exported from PyMol, as long as they can be aligned to the MSA.

4 Use case

Acknowledgements

Text Text Text Text Text Text Text Text. Bofelli *et al*., 2000 might want to know about text text text text

Funding

This work has been supported by the... Text Text Text Text.

References

Bofelli,F., Name2, Name3 (2003) Article title, *Journal Name*, **199**, 133-154.

Bag,M., Name2, Name3 (2001) Article title, *Journal Name*, **99**, 33-54.

Yoo,M.S. *et al*. (2003) Oxidative stress regulated genes in nigral dopaminergic neurnol cell: correlation with the known pathology in Parkinson's disease. *Brain Res. Mol. Brain Res.*, **110**(Suppl. 1), 76–84.

Lehmann,E.L. (1986) Chapter title. *Book Title*. Vol. 1, 2nd edn. Springer-Verlag, New York.

Crenshaw, B.,III, and Jones, W.B.,Jr (2003) The future of clinical cancer management: one tumor, one chip. *Bioinformatics*, doi:10.1093/bioinformatics/btn000.

Auhtor,A.B. *et al*. (2000) Chapter title. In Smith, A.C. (ed.), *Book Title*, 2nd edn. Publisher, Location, Vol. 1, pp. ???–???.

Bardet, G. (1920) Sur un syndrome d'obesite infantile avec polydactylie et retinite pigmentaire (contribution a l'etude des formes cliniques de l'obesite hypophysaire). PhD Thesis, name of institution, Paris, France.