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 Günther [2]

[1] Pharmaceutical Bioinformatics, Institute of Pharmaceutical Science, Albert-Ludwigs-University, Hermann-Herder-Strasse 9, Freiburg 79104, Germany.

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Protein subfamilies share a common structure but differ in functionalities such as substrate specificity. These differences can often be assigned to a limited set of subfamily specific residues (SSRs). Often, only experimental evaluation, expert knowledge and detailed protein structure investigation, can lead to the correct assignment of these SSRs. Here, we introduce the toolbox SSR-viz, which allows to detect and visualize these SSRs based on a multiple sequence alignment (MSA) and a file containing the subfamily class labels. The applied algorithm can be adjusted to a given problem and allows users to perform a thorough analysis of their sequences.
Results: Substrate specificity of non-ribosomal peptide synthases were analyzed. 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 executables are also available for Windows and Ubuntu.
Contact: stefan.guenther@pharmazie.uni-freiburg.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein subfamilies 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 of proteins.

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 an 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

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. This allows to compare the subfamily classes independent of the number of their representing sequences. Each position in the alignment is 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 Equation (1)).

$$S(m) = 1 - \frac{-\sum(\alpha_m * \log_2(\alpha_m))}{max(S(\alpha))} \quad (1)$$

The difference between two classes is computed by summation of an 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. This matrix judges each amino acid exchange equally. In order to differentiate the exchange of divers amino acids, an additional scoring function is implemented. Again the exchange matrix is computed, 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 (cite). An detailed explanation of the algorithm including an hypothetical example is given in the Supplementary data.

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. In order to investigate the SSRs in a structural context, 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 it can be aligned to the MSA. A flow chart of the Implementation is shown in Figure 1.

4 Use case

The toolbox was applied to detect the SSRs for the adenylation domain of non-ribosomal peptide synthases (NRPS). Two subfamilies were chosen, with respective specificities for valine and glycine. For both subfamilies crystal structures are available (PDB: 3VNS, 4ZXI), so that the detected SSRs can be interpreted in a structural context. The 10 most important SSRs to distinguish the two subfamilies from each other were calculated (using the default parameters of ssrviz). Additionally the most important SSRs to distinguish all the major subfamilies (29 subfamiles, see supplementary data) were also calculated (see Table 1).

Fig. 1. General flow chart of the SSR-viz toolbox. a) The input consists of a MSA together with a CSV-file holding the corresponding subfamily class labels. Multiple protein structures can also be included in order to assign the structure indices. b) The output can be created as a mathplotlib PDF, a Jalview annotation file or a summary CSV-File. This file is used to assign the indices of the structure. c) Example assignment of SSRs for the NRPS adenylation domain with specificity for Valine and Glycine.

Table 1. 10 most important SSRs to distinguish subfamilies with glycine and valine specificity and all the subfamily classes

G vs V				All vs all	
Position	Score	G cons.	V cons.	Position	Score
486	0.74	I	A	486	0.59
749	0.70	W	F	749	0.57
478	0.68	T	S	739	0.56
601	0.66	W	F	591	0.55
748	0.65	I	T	490	0.54
591	0.64	Q	W	478	0.53
592	0.62	A	L	750	0.53
404	0.58	F	L	601	0.52
403	0.58	N	R	495	0.52
479	0.57	T	N	476	0.51

Note: G cons. stands for the conserved residue for the subfamily with specificity for Glycine, The Positions are based on the alignment supplied in the supplementary data.

All detected SSRs are in close proximity to the ligand (< 16 Angstrom). From the SSRs computed for the all-vs-all scheme, 6 are identical to those proposed by (cite Stachehauser) (749, 601, 591, 495, 490, 478). Interestingly, the most significant position found by our analysis (486) is not included in the Stachehauser code. Nevertheless, its key role can clearly be derived from its position in the binding pocket (see supplementary data and Figure 1). The two main positions are identical for both SSR schemes. These positions are also closest to the ligand in the crystal structure (see supplementary data). The subsequent positions differ between the schemes, highlighting the importance of a detailed subfamily specific investigation.

5 Conclusion

SSR-viz allows for a thorough analysis of protein subfamilies for SSRs based on different schemes and using various different substitution matrices. The diversity of the output options allows to further analyze the SSRs together with the alignment or in a structural context. We hope, that SSR-viz will improve the understanding of protein subfamily diversity by allowing the simple analysis of the responsible SSRs.

Funding

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

References