BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 19 2013, pages 2496–2497
doi:10.1093/bioinformatics/btt411

Sequence analysis

Advance Access publication July 18, 2013

SAPA tool: finding protein regions by combination of amino acid composition, scaled profiles, patterns and rules

Josef Maier[1,*], Alexei A. Adzhubei[2] and Wolfgang Egge-Jacobsen[3,4,*]

[1]IStLS, 78727 Oberndorf, Germany, [2]Engelhardt Institute of Molecular Biology, Moscow 119991, Russia, [3]Department of Molecular Biosciences, Glyconor Mass Spectrometry, University of Oslo, 0316 Oslo and [4]Norbrain Mass Spectrometry Facility, Unit for Genome Dynamics, Department of Microbiology, Oslo University Hospital, 0372 Oslo, Norway

Associate Editor: Burkhard Rost

ABSTRACT

Summary: Functional modules within protein sequences are often extracted by consensus sequence patterns representing a linear motif; however, other functional regions may only be described by combined features such as amino acid composition, profiles of amino acid properties and randomly distributed short sequence motifs. If only a small number of functional examples are well characterized, the researcher needs a tool to extract similar sequences for further investigation.

Availability and Implementation: We provide the web application 'SAPA tool', which allows the user to search with combined properties, ranks the extracted target regions by an integrated score, estimates false discovery rates by using decoy sequences and provides them as a sequence file or spreadsheet. Source code, user manual and the web application implemented in Perl, HTML, CSS and JavaScript and running on Apache are freely available at http://sapa-tool.uio.no/sapa/

Contact: josef.maier@istls.de or w.m.egge-jacobsen@imbv.uio.no

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on December 5, 2012; revised on June 30, 2013; accepted on July 12, 2013

1 INTRODUCTION

Finding functional regions in protein sequences is an important annotation task. Although many features of proteins are encoded by interacting amino acid residues positioned apart from each other in sequence, some exist as linear motifs represented by a continuous stretch of residues. Estimations for eukaryotes assume that such linear regions cover up to 30% of all protein sequences (Dunker *et al*., 2008). They have diverse functions in protein folding, protein complex assembly, protein–protein interactions, ligand binding, signalling and as sites of post-translational modifications (Dunker *et al*., 2008; Fuxreiter *et al*., 2007).

Many such sites may be extracted with distinct sequence patterns, simple or complex rules for amino acid neighbourhoods or position-specific matrices, which describe a specific amino acid sequence and its allowed variations, as collected by the INTERPRO (Hunter *et al*., 2009) or the PROSITE database (Hulo *et al*., 2008; Sigrist *et al*., 2010). Functional regions may also be predicted by a specific amino acid composition and similarity to a scaled amino acid profile or index, which defines for each residue a weight for being part of a sequence with a specific property. Such scaled profiles are collected by the AAINDEX database (Kawashima and Kanehisa, 2000; Kawashima *et al*., 2008) and may be applied by scanning tools like the application pepwindow of the EMBOSS program suite (Rice *et al*., 2000).

In silico identification and analysis of hydroxyproline-rich glycoproteins from *Arabidopis thaliana* were recently performed by the BIO-OHIO software, a tool, which combines the searches for regions of biased amino acid composition (e.g. 50% PAST) and that for repeated small patterns like AP, PA, SP and TP (Showalter *et al*., 2010).

A functional region may also be characterized by low complexity because of stretches of the same amino acid or the repeated occurrence of short sequence motifs as compiled by the RepSeq database (Depledge *et al*., 2007). Regions having a specific amino acid composition may be mapped using the oddcomp application of EMBOSS or generally searched for by the software SEG (Wootoon and Federhen, 1993).

Although there are several servers and tools separately performing profiling sequences using amino acid scales, analysis of amino acid composition and scanning motifs, there is no application that combines all three search strategies in a flexible way. We programmed, therefore, the web application 'SAPA tool', where the user can upload protein sequences, search with combined strategies and download the found and scored target sequences. It was named after a typical overrepresented motif (SAPA) in bacterial glycopeptides of *Neisseria gornorrhoeae*, for which it was first programmed.

For demonstration, we used the SAPA tool for retrieving the possibly *O*-glycosylated sequence regions from proteins of *Mycobacterium tuberculosis* (see Supplementary information). Starting from 21 known examples, the tool was used to extract putative target regions having glycopeptide-like composition from the proteome of *M.tuberculosis* H37Rv (Camus *et al*., 2002; Cole *et al*., 1998).

2 FEATURES

Protein sequences are imported from an uploaded multiple sequence file (FASTA format), from the NCBI protein database or from pasted sequences, which may alternatively be used as

*To whom correspondence should be addressed.

control sequences. The region of the proteins to be scanned may be restricted. The set minimum size of the target sequence has to comply with all settings. Overlapping or adjoining targets will be fused.

Three decoy methods (riffled, shuffled, reversed) allow the user to make a set of decoy sequences, which are scanned together with the imported proteins for estimation of a cumulative local false discovery rate (FDR) (Elias *et al*., 2005; Wang *et al*., 2009). The FDR for a specific target score is two times the observed number of decoy targets above or equal to that score, divided by the number of all observed targets above or equal to that score. Although the reversed method conserves local amino acid composition completely, the random method would scramble it; however, it removes the natural sequence redundancy and underestimates the FDR. We designed a riffling method, which mixes the sequence analogous to the riffled shuffle of a stack of playing cards, mostly conserves sequence redundancy and destroys local amino acid composition (for details see the programs decoy help file, manual and the supplement).

The user selects minimum occurrence percentages for up to six single amino acids or three groups of related amino acids, which must be found in selected target sequences.

Up to three scaled AAINDEX amino acid profiles can be used for scoring and/or selecting target sequences, which can be set to remain below or above a specified mean score value.

Targets are also optionally selected depending on their motif content. The motifs are defined by an extended version of the PROSITE pattern syntax (Hulo *et al*., 2008; Sigrist *et al*., 2010), and are combined with 'AND', 'NOT' or 'OR' operators.

Each target sequence is scored by the information content of each amino acid matched by a composition setting, the scores of the appropriately re-scaled and weighted AAINDEX scales and the information content of the defined motifs. Details of the scoring scheme are described in the manual and the supplement. The scores of each protein are the sum score of their target scores.

The result screen summarizes the applied search settings, the scoring scheme and target/protein extraction results. A result table shows all targets sorted by their scores and contains protein sequence icons with highlighted target regions, where the scores are encoded by different colour intensities. When clicking on an icon, a pop-up window shows the sequence with the highlighted regions. All setting and result tables may be downloaded as a multiple Excel spreadsheet file and protein sequences as a FASTA-formatted sequence file.

The SAPA tool enables easy subsetting of any protein list using compositional, profile and motif data, scores the targets appropriately and estimates their FDR. The example provided in the Supplementary information shows how the SAPA tool can extract sequences with properties derived from *O*-glycosylated peptides. Often, only a few experimentally confirmed example sequences are available. Similar sequences, when extracted, can be experimentally investigated for more information, which allows a new and better informed round of sequence retrieval by the SAPA tool. For example, mass-spectrometric analysis of fragmented peptide ions provides quantitative information on their amino acid composition via their immonium ions; this information can be used to extract the respective subset of sequences, resulting in higher confidence levels for peptide identification. More usage examples and references for them are described in Section 4 of the Supplementary Data.

3 IMPLEMENTATION

The web application was written in Perl, HTML, CSS and JavaScript and runs on a web server enabled for Perl CGI (e.g. Apache) (http://sapa-tool.uio.no/sapa/). Images of sequence icons are dynamically produced by the Perl GD package, spreadsheet files by Spreadsheet::WriteExcel. The online manual offers usage examples and screen views. The application can be downloaded and installed locally, as described in the manual.

ACKNOWLEDGEMENTS

We thank Gisela Schmidt for design of the results page and for providing JavaScript, PHP, HTML and CSS code.

Funding: This work was supported by Norwegian Research Council grants 186032 and by funds from Glyconor, the Department of Molecular Biosciences and the Center for Molecular Biology and Neurosciences of the University of Oslo; the RFBR grant 13-04-91458 НИЗ_а.

Conflict of Interest: none declared.

REFERENCES

Camus,J.C. *et al*. (2002) Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology*, **148**, 2967–2973.

Cole,S.T. *et al*. (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, **393**, 537–544.

Depledge,D.P. *et al*. (2007) RepSeq—a database of amino acid repeats present in lower eukaryotic pathogens. *BMC Bioinformatics*, **8**, 122.

Dunker,A.K. *et al*. (2008) The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics*, **9** (**Suppl. 2**), S1.

Elias,J.E. *et al*. (2005) Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat. Methods*, **2**, 667–675.

Fuxreiter,M. *et al*. (2007) Local structural disorder imparts plasticity on linear motifs. *Bioinformatics*, **23**, 950–956.

Hulo,N. *et al*. (2008) The 20 years of PROSITE. *Nucleic Acids Res*., **36**, D245–D249.

Hunter,S. *et al*. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res*., **37**, D211–D215.

Kawashima,S. and Kanehisa,M. (2000) AAindex: amino acid index database. *Nucleic Acids Res*., **28**, 374.

Kawashima,S. *et al*. (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res*., **36**, D202–D205.

Rice,P. *et al*. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*., **16**, 276–277.

Showalter,A.M. *et al*. (2010) A bioinformatics approach to the identification, classification, and analysis of hydroxyproline-rich glycoproteins. *Plant Physiol*., **153**, 485–513.

Sigrist,C.J. *et al*. (2010) PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res*., **38**, D161–D166.

Wang,G. *et al*. (2009) Decoy methods for assessing false positives and false discovery rates in shotgun proteomics. *Anal. Chem*., **81**, 146–159.

Wootoon,J.C. and Federhen,S. (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Computers Chem*., **17**, 149–163.