Sequence analysis

WSsas: a web service for the annotation of functional residues through structural homologues

David Talavera*,†, Roman A. Laskowski and Janet M. Thornton
EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received on November 24, 2008; revised on February 9, 2009; accepted on February 23, 2009
Advance Access publication February 27, 2009
Associate Editor: Burkhard Rost

ABSTRACT

There are two common ways to automatically predict function: (i) by transference of features from one homologous protein (usually found by similarity searches) to another and (ii) using ab initio predictions. The majority of existing automatic annotation tools focus on predicting the function of whole domains or elements, such as kinase domains or transmembrane segments, e.g. InterPro (Mulder et al., 2007) or Gene3D (Yeats et al., 2008), or even entire proteins, e.g. ProFunc (Laskowski et al., 2005b). However, annotation of specific functional sites is important in order to improve our understanding of the essential residues involved in function and the way they act, e.g. SAS (Milburn et al., 1998). Furthermore, the increase of available variation data among individuals highlights the importance of functional sites and their neighbourhood. There are some tools that focus on this area: ScanProsite identifies functional residues which create an identifiable pattern in the sequence (de Castro et al., 2006); MSD-motif performs motif- or ligand-oriented web searches (Golovin and Henrick, 2008); and SAS uses sequence homologues containing structural information to predict functional and structural features and provides a graphical view of the results. However, the majority of these tools do not have programmatic access to perform new predictions [ScanProsite has access via REST protocol (Fielding, 2000) and MSD data can be retrieved via API]. Aiming to fill this gap, we have developed WSsas, a new annotation web service based on similarity searches and pairwise alignments.

1 INTRODUCTION

During the last decade the great number of sequencing projects has produced a huge amount of biological data. However, the string of nucleic or amino acids is frequently the only knowledge scientists have about these sequences. The lack of functional information is an important constraint for the interpretation of results. Consequently, enriching the raw data with functional annotation has become an important task for the scientific community. Not only are these annotations important for the aid of experimental researchers, but also for the performance of computational biology studies. Several projects or databases use human curators to increase our knowledge in the life sciences (Holliday et al., 2007; UniProt Consortium, 2008), e.g. UniProt entries contain feature fields annotating properties such as transit signals, ion-binding regions or alternative splicing changes. However, this careful annotation process is extremely time consuming. Thus, many computational tools aiming to automatically annotate genes and proteins have been developed, being one of the most successful fields in bioinformatics and theoretical biology (Fliceck et al., 2008; Mulder et al., 2007). Although automatic annotations are not exempt from errors, nor can they be as accurate as manually curated ones, their availability can be a good starting point for further experiments [for a recent review on annotation, see Reeves et al. (2009)].

To whom correspondence should be addressed.
*Present address: Faculty of Life Sciences, University of Manchester, Manchester M13 9PT, UK.
†Present address: Faculty of Life Sciences, University of Manchester, Manchester M13 9PT, UK.
WSsas: annotation of functional residues

![Diagram](image)

**Fig. 1.** (A) Design of WSsas. An input protein sequence is searched against all protein sequences in the PDB using FASTA. Residue information is retrieved from the PDBsum and CSA databases for all the FASTA hits. (B) Summary of the inputs and output of WSsas. The output is a file in XML format.

(Luscombe et al., 1997) and PROCHECK (Laskowski et al., 1993). PDBsum is used to extract information on binding residues and protein-protein interaction interfaces. The Catalytic Site Atlas is a database containing a set of well-curated catalytic sites and a set of homologous catalytic sites found by sequential and structural similarity. Both databases are based on three-dimensional information. This means that the functional information used in predictions is set in experimentally solved structures. It is worth mentioning that while PDBsum and CSA data are used by WSsas, these databases are not themselves annotated using WSsas as this would lead to a circular logic and would affect the reliability of both databases and the annotation tool.

SAS and WSsas are very similar tools; however, the way to use the annotation protocol is different. SAS is useful for researchers interested in a single protein sequence, or a small number of sequences, whereas the WSsas web service allows multiple searches or inclusion in workflows, since it is based on ‘computer-to-computer’ communication and its results are easy to retrieve via a script (Kappler, 2008; Labarga et al., 2007). At the moment, there is no restriction on the number of searches, but we may need to change this according to a fair-use policy if the tool is frequently used.

The query input for WSsas is a protein sequence. The output contains functional information (predicted functional residues and additional information about bound compounds), transference information (structures and residues used to transfer information) and alignment information [identity percentage, number of aligned residues and expectation (E) value]. Parameters that can be defined by the user are (Fig. 1B): the maximal E-value for selecting FASTA hits (0.001 by default); the minimum number of aligned residues between query and hits (80 residues by default); and the minimum identity percentage of the aligned stretches (30% by default). By changing alignment thresholds, users can perform a more or less stringent annotation transfer depending on their interests; e.g., increasing the percentage of identity or the number of aligned residues will reduce the number of annotated residues, but the reliability will be higher; conversely, decreasing significantly the thresholds will increase the number of annotated residues but could be a risky journey to the twilight zone. In addition, another way of increasing stringency of prediction exists: when the stringency flag is activated, the algorithm looks at the atomic contacts involved in the compound-binding or protein-protein interaction. Then, it discards annotations if the query residue does not contain the same functional atom type. Nevertheless, the accuracy of predictions relies absolutely on the availability of homologues with solved structure and their quality. Finally, an additional verbose flag allows listing all the relevant atomic interactions in the retrieved hits.

From the webpage http://www.ebi.ac.uk/thornton-srv/databases/WSsas/, it is possible to download an XML definition schema (XSD), a web service definition (WSDL) file and a Perl client for using the web service. The basic input for the Perl client is a file which contains a query sequence and the response received by the user is an easy to parse XML file. WSsas uses the SOAP protocol for information transfer.

3 CONCLUDING REMARKS

WSsas complements other annotation tools in our group, such as CSA, ProFunc and SAS. It predicts functional residues and their interaction partners on the basis of similarity to solved 3D structures while tolerating local variation. Further features may be added in the future.

ACKNOWLEDGEMENTS

The authors wish to thank Martin Eklund, Hamish McWilliam, Florian Reisinger, Ola Sjuth and one anonymous reviewer for helpful comments and suggestions on the web service and Daniela Wieser for a critical reading of the manuscript. This work was completed as part of the BioSapiens Network of Excellence.

Funding: European Commission within its FP6 Programme, under the thematic area ‘Life sciences, genomics and biotechnology for health’ (contract number LHSG-CT-2003-503265).

Conflict of Interest: none declared.

REFERENCES


