Protein analysis tools and services at IBIVU

Bernd W. Brandt[1*] and Jaap Heringa[1,2]

[1]Centre for Integrative Bioinformatics (IBIVU), VU University Amsterdam, De Boelelaan 1081A, Amsterdam, The Netherlands, http://www.ibi.vu.nl

[2]Netherlands Bioinformatics Centre, Geert Grooteplein 28, 6525 GA Nijmegen, The Netherlands, http://www.nbic.nl

Summary

During the last years several new tools applicable to protein analysis have made available on the IBIVU web site. Recently, a number of tools, ranging from multiple sequence alignment construction to domain prediction, have been updated and/or extended with services for programmatic access using SOAP. We provide an overview of these tools and their application.

1 Introduction

The Centre for Integrative Bioinformatics provides a number of algorithms, implemented as web-based tools, that can be used in different fields of protein analysis. These tools range from alignment construction, secondary structure prediction, repeat detection and prediction of specificity-determining residues to domain prediction. Recently, a number of these tools have been updated and changed to make programmatic access possible. Here, we summarize the new or updated tools and their field of application. An overview of all our tools is available at http://www.ibi.vu.nl/programs/.

1.1 PRALINE: multiple sequence alignment

PRALINE is a multiple alignment program that employs a profile-based progressive alignment protocol. The first version was published in 1999 [1] and has been improved and extended since. The sequence profiles can be constructed by using the other sequences in the alignment (preprofile processing) or by using hits generated by PSI-BLAST (PSI-BLAST preprofile processing or homology-extended alignment). PRALINE has now developed into a multiple sequence alignment tool kit with the possibility to take into account predicted transmembrane regions and secondary structure during the alignment process [2, 3]. For both transmembrane and secondary structure prediction a choice of predictors is available. Due to the profile-based progressive alignment protocol that can be combined with PSI-BLAST, PRALINE produces good alignments, but run times can be significant when aligning many sequences. As PRALINE is often used and not easily installed locally because of the many dependencies on third-party software, we developed a SOAP interface for users who wish to run PRALINE many times. The SOAP interface, described by a Web Service Description Language (WSDL) file, provides methods for asynchronous access.

1.2 Multi-Harmony: detection of sub-family specific residues

Many protein families contain sub-families with functional specialization, such as binding different ligands or being involved in different protein-protein interactions. A small number of

[*] To whom correspondence should be addressed. Email: bwbrandt@few.vu.nl

doi:10.2390/biecoll-jib-2011-168

amino acids generally determine functional specificity. Multi-Harmony is an interactive web server for detecting sub-type-specific sites starting from a multiple alignment and a definition of the sub-families (groups) [4]. This server combines two methods to predict these specificity-determining residues: multi-group Sequence Harmony and a newly implemented multi-RELIEF [4]. Predictions are provided in a table, which can be filtered and sorted interactively, and as annotations in the Jalview multiple sequence alignment editor [5]. Additionally, the predictions, which pass the user-supplied filter criteria, can be transferred as an annotation track to the alignment in Jalview and to Jmol [6] to view the positions in three-dimensional context.

Multi-Harmony now has a SOAP interface available which returns the predictions per position and per predictor. It also is possible to drive the program in a REST-like way. This way, many automated runs can be performed, while the same result files as produced by the web server are available. This latter option has been used by users to generate predictions for large sets of alignments.

1.3 Domain prediction

Most proteins contain more than one structural domain. For accurate sequence analysis, similarity searching, making constructs for protein in vitro production, and understanding protein cellular function, it is essential to have an indication of the location of these domains. We have developed two domain predictors which work on sequence alone. In addition, we implemented an existing profile comparer as web server which adds considerable functionality as compared to the stand-alone program.

1.3.1 DOMAINATION

DOMAINATION is a domain delineation method based on sequence similarity searching with PSI-BLAST [7]. It starts with a PSI-BLAST search on a single query sequence. Based on the distribution of hits, the query sequence is split in one or more domains, which can be continuous or discontinuous. The identified domains are used in new searches until no new domains are found. This way the detection of distant homologues is also improved. DOMAINATION has been implemented as a web-based tool (not published) to automate the searches. The progress of DOMAINATION through the iterations can be followed in real-time on the web server. In addition, the domain and repeat “families” from Pfam present in the query protein are indicated. When the run is finished, the web-server output contains a complete domain overview of the sequence. Clicking on a domain highlights this domain in the FASTA sequence. This tool also has a SOAP web service, which is asynchronous as runs can last a long time.

1.3.2 SCOOBY-DOmain

Scooby-DOmain (*s*equen*c*e hydr*o*ph*o*bicit*y* predicts *domain*s) is a fast and simple method to identify globular regions in a protein sequence [8]. Similar to DOMAINATION, domain prediction is based on protein sequence alone. Scooby-DOmain predictions are based on the observed length and hydrophobicity of domains from proteins with known tertiary structure present in the CATH domain database [9]. The prediction method employs an A*-search to identify sequence regions that form a globular structure and those that are unstructured [8]. The method can be run on a single query sequence and does not rely on homology searches. However, when given an alignment, the sensitivity and accuracy of the predictions improved in a benchmark study [8]. This tool now has a SOAP synchronous and asynchronous service. The latter service currently supports the submission of up to 200 protein identifiers at once.

1.3.3 webPRC

Profile–profile methods are well suited to detect remote evolutionary relationships between protein families. By running a query alignment or a Hidden Markov Model (HMM) against a domain database these methods can also be used to delineate known domains in a protein. WebPRC [10] provides a web-based interface to the Profile Comparer [11] for alignment-based searching of public domain databases. The Profile Comparer (PRC) is a program for aligning and scoring profile HMMs [11]. Since PRC compares profile HMMs instead of sequences, it can be used to find distant homologues. The webPRC server converts an input alignment to an HMM and runs this against a selection of well-known domain databases (including Pfam [12], CATH [9] and NCBI's Conserved Domain Database [13]). PRC only reports profile HMM alignments and does not produce multiple sequence alignments. The webPRC interface facilitates the identification and evaluation of “hit” alignments by providing the results both as multiple sequence alignments and aligned HMMs and by including the domain annotation from the domain databases as well as links to the entries in these databases. The query-hit alignments can also be viewed in the Jalview alignment editor and as logos based on the aligned HMMs or on the aligned multiple sequence alignments. Due to its interactive nature, this tool does not have a SOAP service. However, large scale analyses can be carried out with the stand-alone version of PRC [11] (http://supfam.org/PRC/).

2 Methods

Most tools have been published and more information can be found in their publications. Although the DOMAINATION method is published [7], the DOMAINATION web server has not been published. All servers contain documentation pages that describe the method and how to use the server.

The SOAP (http://www.w3.org/TR/soap/) servers have been implemented with Perl and SOAP::Lite and Web Service Description Language (WSDL, http://www.w3.org/TR/wsdl) files are available. The general system architecture is depicted in Figure 1. The REST-like interface of multi-Harmony uses HTTP status codes to communicate the result URL and run status. First a “Location:” redirect header is sent to provide the URL of the results page. Polling this URL returns a “202 Accepted”, while the program is running and a “200 OK” when it has finished. In case of errors a “500 Error” is returned. The output is the same as the output of the normal web interface; hence only the plain-text output can be retrieved. An example of a Perl client script for using multi-Harmony this way is available upon request.

3 Results and discussion

3.1 Using the IBIVU services

An overview of the tools described above is available at http://www.ibi.vu.nl/programs/. The user can choose to execute the tool via a web browser or via a SOAP client for PRALINE, multi-Harmony, DOMAINATION or Scooby-DOmain. For checking the SOAP services and for viewing the input and output, the WSDL files can easily be loaded with a general SOAP tool, for example soapUI (http://www.soapui.org/). The European Bioinformatics Institute provides tutorials for programming web services in different languages (*e.g.* Perl or Python) and protocols (SOAP or REST) at http://www.ebi.ac.uk/Tools/webservices/tutorials/06_programming. For background information on web services in bioinformatics, we refer to reader to [14, 15].

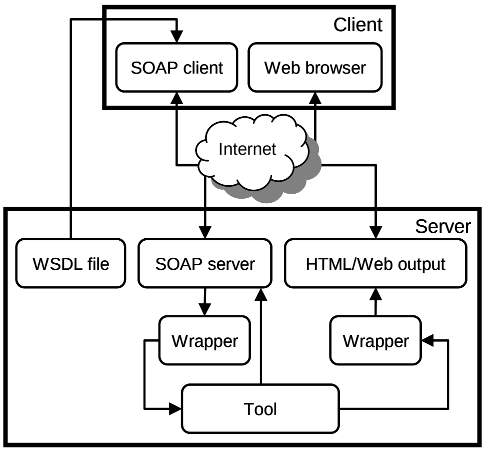


Figure 1: The tool specific WSDL file is used by the user to produce the SOAP client, which sends a request to the remote SOAP server. This request is forwarded to the respective tool via a wrapper. The SOAP server processes the tool's output and sends back the results or a ticket, which can be polled until the tool run has been completed. For several tools, the result also contains a URL. In these cases, the output is also available as an interactive HTML web page, which can be viewed in the user's web browser.

PRALINE and DOMAINATION can only be run in asynchronous mode since their run times are always longer than a few minutes and can vary greatly depending on the input. Scooby-DOmain and multi-Harmony can be run in synchronous mode as responses are generally fast. For bulk queries of up to 200 protein identifiers, the Scooby-DOmain service provides an asynchronous service.

Methods, inputs and outputs are defined in the WSDL files. In short, inputs are as follows: DOMAINATION and Scooby-DOmain require a protein identifier (UniProt or RefSeq) or FASTA sequence. Scooby-DOmain can also be run on an alignment. PRALINE requires a set of FASTA sequences and multi-Harmony a multiple alignment and list of group sizes. More information is available on the web server and in the WSDL files.

Although the current services are primarily developed for programmatic access, they can also be used in workflows. Figure 2 shows an example workflow, which combines several services to obtain predictions of three different tools. Here, Scooby-DOmain and webPRC are run on a on a multiple sequence alignment.

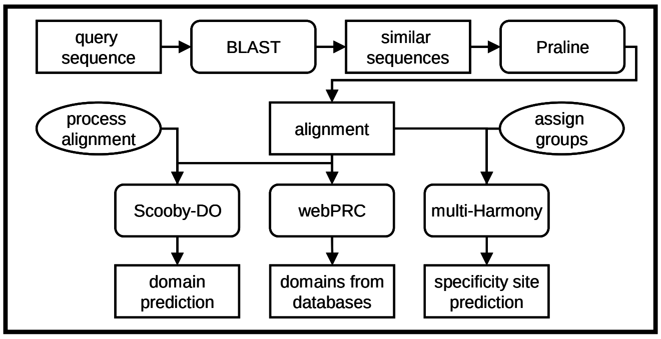


Figure 2: Example workflow. A query sequence is BLASTed and hits are extracted. These sequences are aligned with the Praline service. The alignment can then be used for three other tools: Scooby-DOmain, webPRC and multi-Harmony. For Scooby-DOmain, the alignment may need to be pre-processed ("process alignment", ***cf*****. [8]). Before multi-Harmony can be used, the sequence groups need to be assigned first ("assign groups").**

3.2 Example: DOMAINATION

We use the DOMAINATION SOAP service to generate a domain prediction for the WD repeat-containing protein 5 of mouse (RefSeq: NP_543124). With this asynchronous service, we first call `getTicketById` and provide the protein identifier. DOMAINATION will now return a ticket. This ticket is needed in the call to `getStatus`, which will return the status. After waiting until the values of `<completed>` and `<statusCode>` are 1, `getResults` can be called to retrieve the domain prediction results (alternatively, `<statusString>` can be checked for successful completion). Additionally, we can view the full (dynamic) HTML results with graphics in a web browser and download additional information such as the FASTA records of all domains by using the result URL present in the SOAP result. The domain prediction graph, which can be viewed this way, is shown in Figure 3.

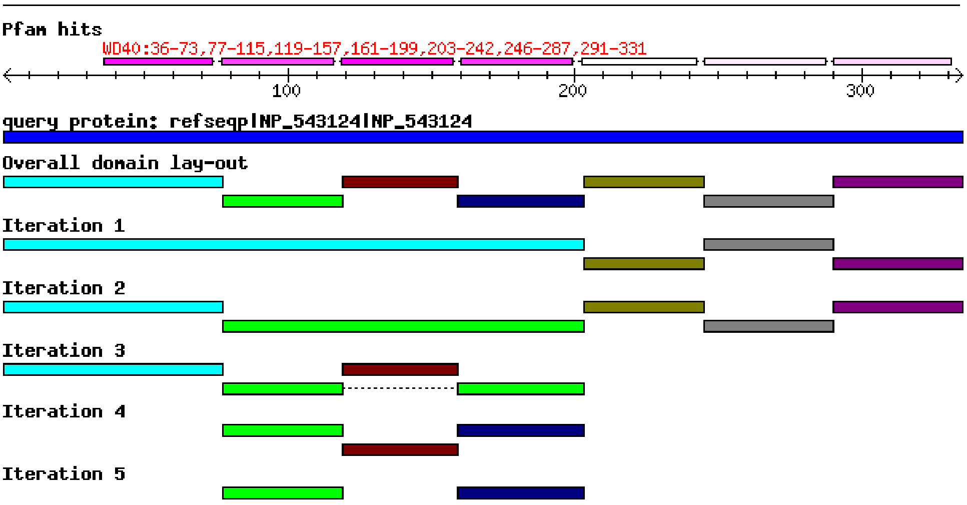
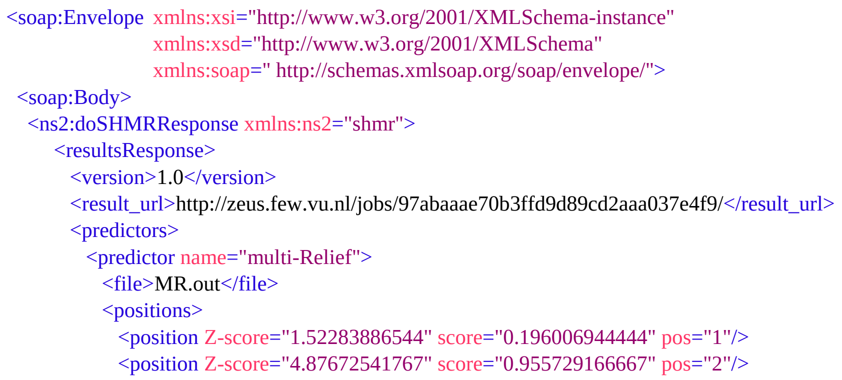

Pfam hits
WD40:36-73,77-115,119-157,161-199,203-242,246-287,291-331
100
200
300
query protein: refseqp|NP_543124|NP_543124
Overall domain lay-out
Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 5

Figure 3: The domain lay-out of WDR5 as present on the web output. Above the query protein (blue) the Pfam hits (repeats and domain classes) are shown. As these seven hits are from the same Pfam family, they are presented in one track. The track "Overall domain lay-out" summarizes all domains found in all DOMAINATION iterations. Domains are coloured based on the start position of the domain and numbered per iteration. This is followed by the results per iteration. Each domain formed during an iteration is used in the next iteration. Domains that are not split again are not used in the next iteration (*e.g.* the last three domains in iteraton 2).

The Pfam database [12] contains an alignment of the separate repeats (accession: PF00400). The search with WDR5 against the Pfam-A database is carried out on-the-fly using the Pfam web service. Pfam also locates seven domains. The domain prediction by DOMAINATION agrees well with the Pfam-A matches (Pfam: 36-73, 77-115, 119-157, 161-199, 203-242, 246-287, 291-331; DOMAINATION: 1-76, 77-115, 119-158, 159-202, 203-244, 245-289, 290-334). Since this prediction method depends on homology searches, it is not surprising that WDR5 is delineated into 7 domains (which can be seen as mini-domains). The predicted locations are present in the FASTA files, in the annotated domain graph on the web server and in the SOAP output.

3.3 Example: multi-Harmony

We here briefly illustrate the multi-Harmony server. This server can be used to predict sub-type-specific sites, also known as specificity determining positions (SDP). These positions play a role in the functional specialisation of proteins in different sub-families. The input consists of an alignment (at least 4 sequences) and the sizes of the different groups in the alignment (at least 2 sequences per group). We use the example input from the multi-Harmony web server (http://www.ibi.vu.nl/programs/shmrwww/) and call the SOAP `doSHMR` method as defined in the WSDL. After a while, the SOAP server sends the response with the predictions to the SOAP client (Figure 4). The response also includes a URL, which can be used to analyze the results interactively online (*e.g.* for interesting cases). In addition, this URL can be prepended to the file names supplied by the `<file>` tag to download the prediction files in plain-text format.



```
<soap:Envelope xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
               xmlns:xsd="http://www.w3.org/2001/XMLSchema"
               xmlns:soap=" http://schemas.xmlsoap.org/soap/envelope/">
 <soap:Body>
  <ns2:doSHMRResponse xmlns:ns2="shmr">
     <resultsResponse>
       <version>1.0</version>
       <result_url>http://zeus.few.vu.nl/jobs/97abaaae70b3ffd9d89cd2aaa037e4f9/</result_url>
       <predictors>
         <predictor name="multi-Relief">
           <file>MR.out</file>
           <positions>
             <position Z-score="1.52283886544" score="0.196006944444" pos="1"/>
             <position Z-score="4.87672541767" score="0.955729166667" pos="2"/>
```

Figure 4: The beginning of the multi-Harmony SOAP response output. The method “doSHMR” has returned the results (doSHMRResponse). The <result_url> provides a URL, which can be used for interactive analysis of the results in a web browser. This is followed by the predictions grouped per predictor (multi-Relief and multi Sequence-Harmony), the name of plain-text prediction file and the predicted scores per position.

3.4 Other protein analysis tools at IBIVU

The IBIVU programs page (http://www.ibi.vu.nl/programs/) also provides several other tools related to protein analysis, such as secondary structure prediction and repeat detection. These tools do not have special programmatic access (yet). Another tool that has programmatic access (URL-API) is SEQATOMS [16]. It can be used to identify missing regions in proteins in the PDB in their sequence context by masking all residues missing from the structure.

3.5 Conclusion

Several of IBIVU’s most used tools now have SOAP services available. This facilitates large(r) scale analysis with these tools. For PRALINE, multi-Harmony and DOMAINATION, where users may want to analyse (some) results interactively with a web browser, the SOAP result provides a URL that points to the tool’s output page on our web server.

Acknowledgements

We would like to thank Florian Reisinger (EMBL-EBI) for helpful discussions on web services, feedback on WSDL and testing the DOMAINATION service in Java JAX-WS. Part of this work was supported by ENFIN, a Network of Excellence funded by the European Commission within its FP6 Programme, under the thematic area ‘Life sciences, genomics and biotechnology for health’, contract number LSHG-CT-2005-518254.

References

[1] J. Heringa. Two strategies for sequence comparison: profile-preprocessed and secondary structure-induced multiple alignment. *Computers & Chemistry*, 23(3-4):341-364, 1999.

[2] V. A. Simossis and J. Heringa. PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. *Nucleic Acids Research*, 33(suppl 2):W289-294, 2005.

[3] W. Pirovano, K. A. Feenstra and J. Heringa. PRALINE^{TM}: a strategy for improved multiple alignment of transmembrane proteins, *Bioinformatics*, 24(4):492-497, 2008.

[4] B. W. Brandt, K. A. Feenstra and J. Heringa. Multi-Harmony: detecting functional specificity from sequence alignment. *Nucleic Acids Research*, 38(suppl 2):W35-W40, 2010.

[5] A. M. Waterhouse, J. B. Procter, D. M. A. Martin, M. Clamp, G. J. Barton. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189-1191, 2009.

[6] A. Herráez. Biomolecules in the computer: Jmol to the rescue. *Biochemistry and Molecular Biology Education*, 34(4):255-261, 2006.

[7] R. A. George and J. Heringa. Protein domain identification and improved sequence similarity searching using PSI-BLAST. *Proteins*, 48(4):672-681, 2002.

[8] C. N. I. Pang, K. Lin, M. A. Wouters, J. Heringa and R. A. George. Identifying foldable regions in protein sequence from the hydrophobic signal. *Nucleic Acids Research*, 36(2):578-588, 2007.

[9] A. L. Cuff, I. Sillitoe, T. Lewis, O. C. Redfern, R. Garratt, J. Thornton and C. A. Orengo. The CATH classification revisited—architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Research*, 37(suppl 1):D310-314, 2009.

[10] B. W. Brandt and J. Heringa. webPRC: the Profile Comparer for alignment-based searching of public domain databases. *Nucleic Acids Research*, 37(suppl 2):W48-W52, 2009.

[11] M. Madera. Profile Comparer: a program for scoring and aligning profile hidden Markov models. *Bioinformatics*, 24(22):2630-2631, 2008.

[12] R. D. Finn, J. Tate, J. Mistry, P. C. Coggill, S. J. Sammut, H. R. Hotz, G. Ceric, K. Forslund, S. R. Eddy, E. L. L. Sonnhammer and A. Bateman. The Pfam protein families database. *Nucleic Acids Research*, 36(suppl 1):D281-D288, 2008.

[13] A. Marchler-Bauer, S. Lu, J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, F. Lu, G. H. Marchler, M. Mullokandov, M. V. Omelchenko, C. L. Robertson, J. S. Song, N. Thanki, R. A. Yamashita, D. Zhang, N. Zhang, C. Zheng and S. H. Bryant. CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Research*, 39(suppl 1):D225-229, 2011.

[14] P. B. T. Neerincx and J. A. M. Leunissen. Evolution of web services in bioinformatics, *Briefings in Bioinformatics*, 6(2):178-188, 2005.

[15] H. Stockinger, T. Attwood, S. N. Chohan, R. Côté, P. Cudré-Mauroux, L. Falquet, P. Fernandes, R. D. Finn, T. Hupponen, E. Korpelainen, A. Labarga, A. Laugraud, T.

Lima, E. Pafilis, M. Pagni, S. Pettifer, I. Phan and N. Rahman. Experience using web services for biological sequence analysis. *Briefings in Bioinformatics*, 9(6):493-505, 2008.

[16] B. W. Brandt, J. Heringa and J. A. M. Leunissen. SEQATOMS: a web tool for identifying missing regions in PDB in sequence context, *Nucleic Acids Research*, 36(suppl 2):W255-W259, 2008.