Bioinformatics, 31(9), 2015, 1475–1477
doi: 10.1093/bioinformatics/btu836
Advance Access Publication Date: 18 December 2014
Applications Note

OXFORD

Sequence analysis

CDvist: a webserver for identification and visualization of conserved domains in protein sequences

Ogun Adebali[1,2,*], Davi R. Ortega[1,2,†] and Igor B. Zhulin[1,2,*]

[1]Computer Science and Mathematics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37861, USA and [2]Department of Microbiology, University of Tennessee, Knoxville, TN 37996, USA

*To whom correspondence should be addressed.
†Present address: Division of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA 91125, USA
Associate Editor: John Hancock

Received on September 22, 2014; revised on November 10, 2014; accepted on December 12, 2014

Abstract

Summary: Identification of domains in protein sequences allows their assigning to biological functions. Several webservers exist for identification of protein domains using similarity searches against various databases of protein domain models. However, none of them provides comprehensive domain coverage while allowing bulk querying and their visualization schemes can be improved. To address these issues, we developed CDvist (a comprehensive domain visualization tool), which combines the best available search algorithms and databases into a user-friendly framework. First, a given protein sequence is matched to domain models using high-specificity tools and only then unmatched segments are subjected to more sensitive algorithms resulting in a best possible comprehensive coverage. Bulk querying and rich visualization and download options provide improved functionality to domain architecture analysis.
Availability and implementation: Freely available on the web at http://cdvist.utk.edu
Contact: oadebali@vols.utk.edu or ijouline@utk.edu

1 Introduction

The identification of protein domains is a key feature of protein sequence analysis. Several databases, notably Pfam (Punta *et al.*, 2012), Simple Modular Architecture Research Tool (SMART) (Letunic *et al.*, 2009), Clusters of Orthologous Groups (COG) (Tatusov *et al.*, 2003), Conserved Domain Database (CDD) (Marchler-Bauer *et al.*, 2013) and others, develop and maintain domain models. Searching tools such as RPS-BLAST (Marchler-Bauer *et al.*, 2013), HMMER3 (Eddy, 2011) and HHpred/HHsearch (Hildebrand *et al.*, 2009; Soding, 2005) are used to match sequences to domain models present in a given database. The size of the protein sequence database grows dramatically, whereas its coverage by pre-computed domain models increases very slowly (Rekapalli *et al.*, 2012). Consequently, sensitive domain searches of sequences in bulk are necessary to improve computational coverage of the current and future protein sequence space. Despite the overwhelming success of the current state-of-the-art domain searching resources, three areas require further improvements: (i) combining tools with high specificity and tools with high sensitivity in a single framework, (ii) multiple query searches using highly sensitive (e.g. profile-to-profile) methods, (iii) visualization of most relevant information in a responsive and interactive way.

To address these issues, we have developed the Comprehensive Domain Visualization Tool (CDvist), a domain-searching webserver specialized in maximizing domain coverage of multidomain protein sequences with emphasis on visualization.

2 Implementation and features

Users submit protein sequences in FASTA format and each sequence is processed independently of each other on individual linux cluster nodes. Up to 500 queries per request are supported. The following

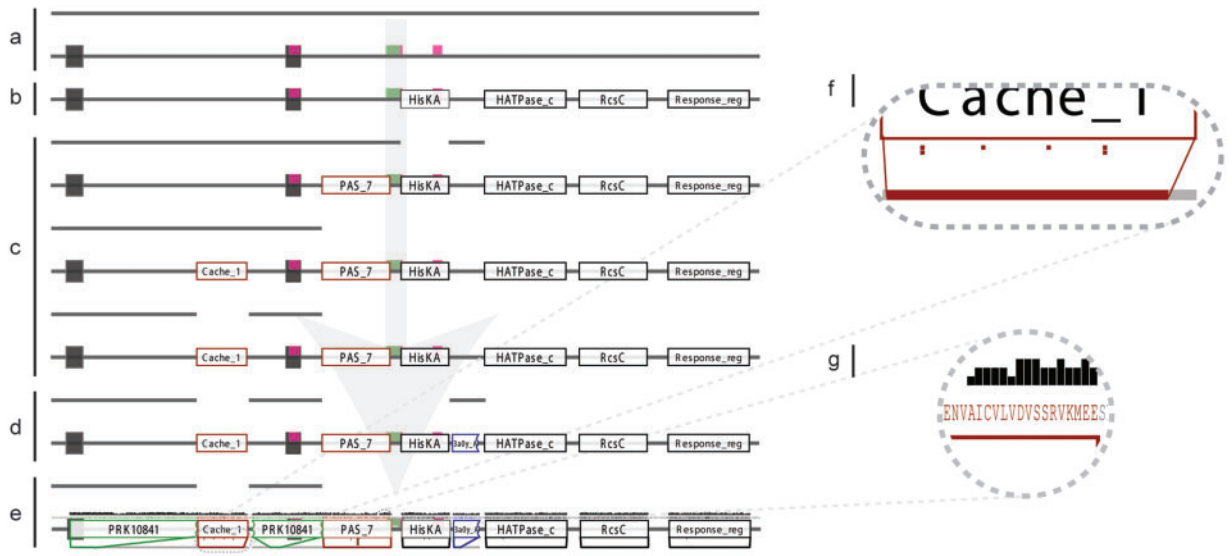

Fig. 1. Workflow and visualization example. (**a**) Primary sequence is used as input and transmembrane (gray), low complexity (magenta), and coiled-coil (green) regions are predicted. (**b**) HMMER3 scan against Pfam database is executed and first domain architecture is built. (**c–e**) HHblits followed by HHsearch is executed against (**c**) Pfam, (**d**) PDB and (**e**) CDD databases. (**f**) Domain coverage option: gray background represents the whole length of model whereas red bar displays the portion of the model that aligns with the query. Square points represent the domain positions that do not align with the query. (**g**) Alignment option. Sequence is displayed to scale, and each bar stands for alignment quality at that position. The absence of the bar at a given position indicates gap in the alignment on query side

domain search methods are implemented in CDvist: HMMER3 (Eddy, 2011), RPS-BLAST (Marchler-Bauer *et al.*, 2013; Schaffer *et al.*, 1999), HHSEARCH (Soding, 2005), and HHBLITS-HHSEARCH (Remmert *et al.*, 2012). Transmembrane regions are predicted by either TMHMM (Sonnhammer *et al.*, 1998) or Phobius (Kall *et al.*, 2007). Low complexity and coiled coil regions are predicted by SEG (Wootton, 1994) and Coils (Lupas, 1996) respectively. To improve domain coverage, rather than using the entire sequence, CDvist iteratively identifies regions without significant domain match (orphan segments) and submits each one of them to similarity search against a user-determined sequence of databases until the entire protein sequence is covered or all databases have been searched (Fig. 1). The key principle of this process is that tools that have high specificity—HMMER against Pfam and RPS-BLAST against CDD—are used first. Only then, the sequence segments that were not confidently matched to any model are used to build profiles and subjected to more sensitive profile-profile searches by HHsearch. Each algorithm can be turned on/off and the order of databases, and their significance thresholds, can be altered. This flexibility enables users to tailor the overall process for their specific purposes. Optional 'domain split' function splits the matched domain model if there is a considerable unaligned query region (5% by default) in the query-model alignment. This unaligned region is considered as an orphan segment and is used in the next run to search for potential domains.

A custom built JavaScript module powers the visualization on the client side with images in vector format (SVG) that are practical to edit, export as PDF and produce figures of publication quality. Results are displayed asynchronously for each query sequence submitted, which also allows the user to interact with the data before the completion of the entire request. Domain coverage bar provides information on what portion of the matched domain model is represented on the query sequence (Fig. 1f). Alignment quality is represented as vertical bar for each position of the alignment. Gaps in the alignment indicate that the corresponding part of the query is not aligned with the model (Fig. 1g). Scaled sequence information is mapped on the domain architecture, which is easily retrievable by zooming in on the browser. Drag feature allows user to align desired parts of batch data for further analysis. All this information is hosted in our webserver for over a week with a unique URL. Alternatively, the user can retrieve the HTML file to control the interactive feature visualizations locally on a web-browser. JSON formatted files containing the information used to draw the graphics in the website are available not only for each individual sequences but also for the entire input set as a single file. Finally, the log files for each run are available, which display the raw output of the whole process. Logs provide extra information on less significant hits which are not displayed visually. The databases are updated immediately upon their release.

3 Discussion

CDvist is designed to provide maximum domain coverage in protein sequences by bundling the best current domain search tools into a pipeline that exhaustively searches through a series of domain databases in an iterative fashion. This methodology yields the most comprehensive domain architecture for a given protein sequence. Rich visualization, download options and linear speed-up for bulk queries should be appealing to both biologists and bioinformaticians. This webserver would be especially useful for multi-domain proteins with rare or unique domain architectures and those prone to domain swap, where whole sequence similarity searches often yield uninformative and misleading results (Iyer *et al.*, 2001).

Funding

NIH GM072285 (to I.B.Z.). O.A. UT-ORNL Graduate Program of Genome Science and Technology (to O.A.).

Conflict of Interest: none declared.

References

Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.

Hildebrand,A. *et al.* (2009) Fast and accurate automatic structure prediction with HHpred. *Proteins*, **77** (Suppl. 9), 128–132.

Iyer,L.M. *et al.* (2001) Quod erat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences. *Genome Biol.*, **2**, RESEARCH0051.

Kall,L. *et al.* (2007) Advantages of combined transmembrane topology and signal peptide prediction–the Phobius web server. *Nucleic Acids Res.*, **35**, W429–W432.

Letunic,I. *et al.* (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res.*, **37**, D229–D232.

Lupas,A. (1996) Prediction and analysis of coiled-coil structures. *Methods Enzymol.*, **266**, 513–525.

Marchler-Bauer,A. *et al.* (2013) CDD: conserved domains and protein three-dimensional structure. *Nucleic Acids Res.*, **41**, D348–D352.

Punta,M. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.

Rekapalli,B. *et al.* (2012) Dynamics of domain coverage of the protein sequence universe. *BMC Genomics*, **13**, 634.

Remmert,M. *et al.* (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Schaffer,A.A. *et al.* (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.

Soding,J. (2005) Protein homology detection by HMM-HMM comparison. *Bioinformatics*, **21**, 951–960.

Sonnhammer,E.L. *et al.* (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. In: *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, **Vol. 6**, pp.175–182.

Tatusov,R.L., *et al.* (2003) The COG database: an updated version includes eukaryotes, *BMC Bioinformatics*, **4**, 41.

Wootton,J.C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.*, **18**, 269–285.