Published in final edited form as:

J Struct Funct Genomics. 2010 December ; 11(4): 241–245. doi:10.1007/s10969-010-9097-4.

Domain view: a web tool for protein domain visualization and analysis

Xiaokang Pan, Craig A. Bingman, Gary E. Wesenberg, Zhaohui Sun, and George N. Phillips Jr

Center for Eukaryotic Structural Genomics, Department of Biochemistry, University of Wisconsin at Madison, 433 Babcock Dr, Madison, WI 53706, USA

George N. Phillips: Phillips@biochem.wisc.edu

Abstract

The identification of sequence-based protein domains and their boundaries is often a prelude to structure determination. An accurate prediction of disordered regions, secondary structures and low complexity segments of target protein sequences can improve the efficiency of selection in structural genomics and also aid in design of constructs for directed structural biology studies. At the Center for Eukaryotic Structural Genomics (CESG) we have developed DomainView, a web tool to visualize and analyze predicted protein domains, disordered regions, secondary structures and low complexity segments of target protein sequences for selection of experimental protein structure attempts. DomainView consists of a relational database and a web graphical-user interface. The database was developed based on MySQL, which stores data from target protein sequences and their domains, disordered regions, secondary structures and low complexity segments. The program of the web user interface is a Perl CGI script. When a user searches for a target protein sequence, the script displays the combinational information about the domains and other features of that target sequence graphically on a web page by querying the database. The graphical representation for each feature is linked to a web page showing more detailed annotation information or to a new window directly running the corresponding prediction program to show further information about that feature.

Keywords

Structure genomics; Target selection; Domain; Database; Graphical-user interface; Software

Introduction

One definition of a protein domain is a continuous region of a protein sequence and structure that can evolve, function, and exist to some degree independently of the rest of the protein chain. Domain analysis is a very important component of structural genomics, especially in eukaryotic systems where there are larger proteins with multiple domains. One focus of structural genomics initiatives is to solve structures for novel domains and thereby increase

the coverage of fold and structure space [1]. An accurate identification of protein domain boundaries can significantly increase the successful chance of target selection in protein structure determination.

Transmembrane domains usually comprise multiple transmembrane alpha helices. Each transmembrane helix is sometimes called a “domain” because an alpha-helix in a membrane can be folded independently from the rest of the protein, similar to domains of water-soluble proteins. However, because of their hydrophobic properties, the conventional experimental approach, such as X-ray crystallography or NMR (Nuclear Magnetic Resonance) cannot be easily applied to determine their 3D structures. A signal domain (signal peptide) is a short (3–60 amino acids long) peptide chain that directs the post-translational transport of a protein and also plays an important role in protein targeting and protein translocation. Therefore, it is very important to identify putative transmembrane and signal domains of a protein sequence in the design of constructs for expression.

Protein disordered regions at the N- and C- termini (or even within domains) often lead to difficulties in protein expression, purification and crystallization. It is therefore essential to be able to predict which regions of a target protein are potentially disordered or unstructured and then remove them during the design process.

Low complexity regions in a protein sequence are sub-sequences of biased composition or repeats, which tend to have extended or linker type structures and thus inhibit protein structural determination. The removal of such regions can increase the chance of a successful structural determination.

Prediction of protein secondary structure from a primary protein structure (alpha-helix, beta-sheet, coil, etc.) is an important step in the design process. As a general rule, it seems counterproductive to begin or end a sequence to be expressed in the midst of a strongly predicted secondary structural element and these tend to stabilize tertiary structure.

During the protein target selection process in structural genomics, it behooves researchers to perform analysis of domains, secondary structures, disordered regions, and low complexity segments of target protein sequences, and the presence of the entire target fragment which lacks homology to proteins in PDB and TargetDB. Although many programs and databases have been developed and publicly available for researchers to predict domains, secondary structures, disordered regions and low complexity segments, it is not convenient for them to apply such programs and resources together for the analysis of protein sequences in target selection in the same time.

Using database and web programming techniques, we developed a web tool, DomainView. This tool integrates existing programs [2–8] for the prediction of domains, disordered regions, low complexity regions and secondary structures and graphically displays the combinational features of a target protein sequence on a web page so that researchers at CESG can analyze the domains, secondary structure, disordered regions and low complexity segments of that target protein sequence conveniently.

Methods

General system design

DomainView was designed as a web tool with a graphical user interface driven by a relational database, which stores output data generated by the programs that predict protein features. It also run the programs directly for further information (Fig. 1).

Data generation

Most recent HMM database and HMMPfam program [2] were downloaded from website <http://www.pfam.sanger.ac.uk>. HMMPfam program was run by using multiple target protein sequences against the HMM database in our local machine with default parameters. The hits with score larger than 10 and E-value less than 0.001 were selected as Pfam domains. Conserved domain database [3] was downloaded from website <http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>. RPSBlast program was also run by using multiple target protein sequences against the conserved domain database with default parameters. The hits with E-value less than 0.001 were used as CDD domains. SignalP was applied to predict signal peptides using “-t euk -m hmm -f summary -trunc 70” as parameters [4]. PSIPRED was run to find secondary structures using “-b 0 -j 3 -h 0.001” as parameters [5]. PONDR, HMMTOP, and SEG [6–8] were applied to find disordered regions, transmembrane domains, and low complexity segments, respectively. We run all those three programs with default parameters.

Database creation

A MySQL database was created to store data for target protein sequences as well as their domains, secondary structures, disordered regions, low complexity regions and fragments. A set of data pipelines, written in Perl, was developed to transfer data from the output of the prediction programs. Each data pipeline parses the output of an application program such as HMMPfam, CDD, HMMTOP, SignalP, PONDR, SEG, PSIPRED, etc. and loads them into the database via the DBI (database interface) and MySQL DBD (database driver).

Web interface development

A Perl CGI script DomainView was written to retrieve general information of a target protein sequence and its features from the database and display the result graphically. The script communicates the database via DBI and MySQL DBD. It then calls GD (Graphics Design, <http://www.search.cpan.org/dist/GD/>) to draw graphs and generate graphical images. The CGI script was set up in the CGI-BIN directory of an Apache server.

Results

A web tool, DomainView, has been developed for researchers at CESG to visualize and analyze domains, secondary structures, disordered regions and low complexity segments of target protein sequences in structural genomics. The software is available upon request. DomainView consists of a relational database and a web graphical-user interface.

Database implementation

A MySQL-based relational database was designed and created to store and retrieve data of target sequences as well as their domains, secondary structures, disordered regions and low complexity regions. Figure 2 shows a diagram of the database schema. There are five tables which were named ORF_info, Domain, Secondary_structure, Disorder and Low_complexity, respectively. Table ORF_info contains six fields: “go_id” (target identifier), “organism”, “sequence”, “accession” (genbank accession), “gene_name” (gene name), and “createdate” (create date) for each target. Field “sequence” is the required information in addition to the primary key field. Table Domain includes eight fields: “name” (domain name), “go_id”, “type”, “seq_start” (sequence start position), “seq_end” (sequence end position), “score”, “evaluate” (E-value), and “createdate” (create date) for each domain within a target protein sequence. The combination of fields “name”, “go_id”, “type”, “seq_start” and “seq_end” was used as the primary key because they are all important for a unique domain within a protein sequence. Field “type” can be “Pfam”, “CDD”, “Transmembrane”, or “Signal” domains. There are four fields: “id” (identifier number), “go_id”, “prediction” (prediction string of E, H and C representing for strand, helices and coils, respectively) and “confidence” (confidence value associated with the prediction for each residue) in table Secondary_structure. Three fields, “id”, “go_id” and “rates” were designed in table Disorder. Field “rates” is a set of numbers from 0 to 10. Each number represents the disorder score of a residue in the target protein sequence. When the number is 0, the corresponding residue is ordered while the number is 1, the residue is totally disordered. There are also three fields: “id”, “go_id” and “sequence” in table Low_complexity. Field “sequence” contains a protein sequence with low complexity regions marked by “x” in the position of each residue. In table ORF_info, field “go_id” is the primary key, while in other tables it is the foreign key, which links all of the tables together.

A set of data pipelines, consisting of a set of Perl scripts, was developed to transfer data from the outputs of domain prediction programs into the database. Each script was developed independently to parse the output of a prediction program for information about domains, disordered regions, secondary structures, fragments, low complexity, signal peptides or transmembrane domains of the targets and then load them into the MySQL database.

Web graphical-user interface

A Perl CGI script, DomainView, was coded to retrieve general information about protein sequences as well as their domains, secondary structures, disordered regions and low complexity segments from the relational database and then display them graphically on a web page. The graphical representation for each feature such as protein sequence, Pfam domains, CDD domains, transmembrane domains, signal domains, disordered regions, secondary structures, or low complexity segments links to another web page showing further information about that feature. The second web page may either be the official annotation website of the prediction program or a new web page displaying results generated directly from the prediction program. For example, when target GO.13577 was searched, a graphical view for the target protein sequence as well as domains and other features was displayed as shown in Fig. 3. The first track shows the protein sequence information of the target. The

second track depicts Pfam domains with regions (the blue bar) on the protein sequence. Each region is described briefly with domain annotation, score and E-value for Pfam. The brief domain annotation description links to the Pfam official website for detailed domain annotation information. The third track describes that the CDD domains in similar way except that domains are represented by purple rectangles and the brief domain annotation description links to the CDD official annotation website. The fourth track simply displays transmembrane domains represented by brown rectangles. In fifth track, signal peptide prediction is shown in a brief description, which links to another window running the SignalP program and showing the detailed result in both graphics and description. The sixth track displays a chart to show the disorder probability of each residue of the protein sequence from 0 (ordered) to 1 (disordered) predicted by PONDR. The seventh track describes the secondary structure of the residues of the target protein sequence. The helices are represented by the red cylinders, the strands are depicted by blue bars with arrows and the coils are presented by black lines. The confidence level (from 0 to 10) associated with the secondary structure prediction for each residue is shown below the secondary structure graph. The eighth track shows the prediction of two low complexity regions suggested by program SEG. The last track displays a green bar showing the presence of the entire target fragment which lacks homology to proteins in PDB and TargetDB.

Discussion

It is essential for structural genomics researchers to do analysis of domains, signal peptides, disordered regions, secondary structures, and low complexity segments of target protein sequences in target selection. Many programs and resources have been developed to predict domains, signal peptides, disordered regions, secondary structures, and low complexity segments, respectively. Visualization tools showing the individual result of each program are also available [2–8].

DomainView is a web tool for domain visualization and analysis of protein sequences. It integrates results from multiple existing programs and resources for the predictions of domains, secondary structures, disordered regions and low complexity segments. These results are conveniently displayed in graphical form. The most widely used programs and resources such as Pfam, CDD, HMMTOP, SignalP, PSIPRED, PONDR and SEG were selected for the display with each prediction result shown in a track. In some cases, the representing graph for a domain or another feature links a new window showing more detailed information about that feature. This design gives users a bird's-eye view, first for multiple features of a target protein sequence, and then lets them also see more detailed information about each feature. This provides a very convenient way for researchers to quickly find and study the combined information about domains, secondary structures, disordered regions and low complexity segments of protein sequences in target selection.

At CESG, we have had many partially-good targets such as those that show N¹⁵ heteronuclear single quantum coherence NMR spectra that are not good enough for a structure determination but close enough to deserve more work to improve them (HSQC $\pm$). In many cases, the targets yielding HSQC $\pm$ spectra have some highly disordered regions. Potentially solvable targets can be often be obtained by extracting structural domains and

removing disordered regions and low complexity segments. DomainView provides researchers with a powerful tool in selecting protein target sequences.

In the future, DomainView will be enhanced to let users paste a protein sequence and run the prediction programs in combination directly on the web. The combined results for domains, disordered regions, secondary structures and low complexity segments will be displayed on the fly. This will provide researchers at CESG with a direct and simple way to visualize and analyze domains and other features of a new target protein sequence through a single integrated interface.

Acknowledgments

We thank CESG IT core staff and other staff who supported this development work by providing computer services and making suggestions. This project is supported by NIH/NIGMS Grant U54 GM074901.

References

1. Brenner SE. Target selection for structural genomics. *Nat Struct Biol.* 2000; 7(Suppl):967–969. [PubMed: 11104002]
2. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL, et al. The Pfam protein family database. *Nucleic Acid Res.* 2004; 32:138–141.
3. Marchler-Bauer A, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, Hurwitz DI, Jackson JD, Jacobs AR, Lanczycki CJ, Liebert CA, Liu C, Madej T, Marchler GH, Mazumder R, Nikolskaya AN, Panchenko AR, Rao BS, Shoemaker BA, Simonyan V, Song JS, Thiessen PA, Vasudevan S, Wang Y, Yamashita RA, Yin JJ, Bryant SH. CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.* 2003; 31:383–387. [PubMed: 12520028]
4. Emanuelsson O, Brunak S, von Heijne G, Nielsen H. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc.* 2007; 2:953–971. [PubMed: 17446895]
5. Jones DT. Protein secondary structure prediction based on position-specific scoring mathrics. *J Mol Biol.* 1999; 292:195–202. [PubMed: 10493868]
6. Garner E, Cannon P, Romero P, Obradovic Z, Dunker AK. Predicting disordered regions from amino acid sequence: common themes despite differing structural characterization. *Genome Inform.* 1998; 9:201–214.
7. Tusnady GE, Simon I. The HMMTOP transmembrane topology prediction server. *Bioinformatics.* 2001; 17(9):849–850. [PubMed: 11590105]
8. Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. *Comput Chem.* 1993; 17:149–163.

Fig. 1.

Diagram of system design. *Blue* compartments are the parts we developed while the *black* ones are the existing programs or databases

Fig. 2.

Diagram of database schema. Each rectangle contains a table with a table name (the *top one*) and a set of field names. The field with *underlined* word is a primary key. The field with *italic* word is a foreign key

Fig. 3.

A screen shot of the graphical-user interface, displaying combined information of a protein sequence as well as its domains, disordered regions, secondary structures, low complexity segments and fragments in various graphical representations