W116–W118 *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
doi:10.1093/nar/gkl282

CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues

Joe Dundas, Zheng Ouyang, Jeffery Tseng, Andrew Binkowski, Yaron Turpaz and Jie Liang*

Program in Bioinformatics, Department of Bioengineering, University of Illinois at Chicago, Chicago, IL 60612, USA

Received February 9, 2006; Revised March 4, 2006; Accepted April 5, 2006

ABSTRACT

Cavities on a proteins surface as well as specific amino acid positioning within it create the physicochemical properties needed for a protein to perform its function. CASTp (http://cast.engr.uic.edu) is an online tool that locates and measures pockets and voids on 3D protein structures. This new version of CASTp includes annotated functional information of specific residues on the protein structure. The annotations are derived from the Protein Data Bank (PDB), Swiss-Prot, as well as Online Mendelian Inheritance in Man (OMIM), the latter contains information on the variant single nucleotide polymorphisms (SNPs) that are known to cause disease. These annotated residues are mapped to surface pockets, interior voids or other regions of the PDB structures. We use a semi-global pair-wise sequence alignment method to obtain sequence mapping between entries in Swiss-Prot, OMIM and entries in PDB. The updated CASTp web server can be used to study surface features, functional regions and specific roles of key residues of proteins.

INTRODUCTION

Characterizing protein functions is an increasingly important challenging problem that has been approached from both the sequence and structure levels. The fact that only 4922 of the 35 000 Protein Data Bank (PDB) (1) structures contain any type of functional annotation illustrates the widening gap between our ability to resolve the proteins structure and our ability to locate functionally important residues and to obtain a comprehensive understanding of the structural basis of protein function. The 3D structure of a protein and its surface topography can provide important information for understanding protein function, if a broad knowledge base of the functionally important residues and where they are located on the protein structures is provided. This update of the CASTp web server incorporates functional information about a large set of annotated residues on PDB structures obtained from annotations in PDB, Swiss-Prot and Online Mendelian Inheritance in Man (OMIM).

This paper is organized as follows. We will first discuss our method for mapping annotated residues from Swiss-Prot and OMIM onto the PDB structure. We will then describe updates to the CASTp (2,3) web server for visualization of the annotated functional residues, with emphasis on mapping to surface pockets and interior voids. We will conclude with description of additional updates to the CASTp web server.

MATERIALS AND METHODS

Swiss-Prot mapping method

The numbered positions of annotated residues in the Swiss-Prot sequence often do not align to the same numbered positions of the sequence from the PDB structure. Therefore, a mapping of positions between the Swiss-Prot sequence and the PDB sequence must be obtained. We use a variation of the Needleman and Wunsch algorithm to identify if a sequence of a PDB structure can be found to match the sequence containing annotated residues from the Swiss-Prot database.

Specifically, every Swiss-Prot sequence containing one or more annotated residues and a link to a PDB structure was aligned to the corresponding sequence of the PDB structure. Standard annotations of Swiss-Prot used include

*To whom correspondence should be addressed. Tel: +1 312 355 1789; Fax: +1 312 413 2 18; Email: jliang@uic.edu
Present addresses:
Andrew Binkowski, Argonne National Laboratories, Argonne, IL 60439 USA
Yaron Turpaz, Affymetrix, Inc., Santa Clara, CA 95051, USA

post-translational modifications (MOD_RES), covalent binding of a lipid moiety (LIPID), glycosylation sites (CARBOHYD), post-translational formed amino acid bonds (CROSSLNK), metal binding sites (METAL), chemical group binding sites (BINDING), calcium binding regions (CA_BIND), DNA binding regions (DNA_BIND), nucleotide phosphate binding regions (NP_BIND), zinc finger regions (ZN_FING), enzyme activity amino acids (ACT_SITE) and any interesting single amino acid site (SITE). To ensure that the mapping is accurate, only alignments of two sequences with a sequence identity greater than ninety five percent were used. The annotated positions from Swiss-Prot are then transferred onto the PDB sequence, as long as the position is not aligned to a gap.

OMIM mapping method

Variant alleles that are known to be disease causing and are SNPs were selected from the OMIM (4). These OMIM entries that contain links to Swiss-Prot database were mapped onto the Swiss-Prot (5) sequence by measuring the relative distances in residue position between the OMIM alleles and then identifying the corresponding pairs of SNPs in the Swiss-Prot entry. If the Swiss-Prot entry identified the corresponding PDB entry, the sequence was extracted and aligned to the PDB structure using a semi-global pair-wise sequence alignment method. We follow Stitziel *et al*. (6,7) for the mapping between OMIM and PDB entries.

RESULTS

Mapping results

There are 113 928 annotated residues in 4, 922 structures labeled in PDB records. The transfer of 241 913 Swiss-Prot annotations added 226 177 unique annotations to 15 913 PDB structures. Of those structures, 13 094 did not previously have any annotation contained in the PDB records. Table 1 lists the type of Swiss-Prot annotations, number of PDB structures the annotation is found in, and the total number of annotated residues. Of the 15 661 BINDING residues, we were able to map 11 407 (81%) of them to a pocket or a void on the protein structure. We were also able to map 14 829 (74%) of the ACT_SITE sites of enzymes to an existing protein pocket. Additional computation can further raise these percentages (data not shown).

From the original set of 5467 nsSNPs in 1061 alleles, the mapping of OMIM disease mutations added 2128 annotated residues on 310 PDB structures. Of those 2128 variants, only 254 are mapped onto an annotation from either PDB or Swiss-Prot. This is reasonable, as it is possible that these mutations in some cases cause disease by disrupting the proteins structural stability rather than interrupting their functional interactions with other molecules. The database of all annotated residues from PDB, Swiss-Prot and OMIM can be downloaded from the CASTp web server.

Visualizing annotated residues in CASTp

In addition to file downloads, CASTp allows for interactive visualization of biologically important annotated residues by querying the CASTp server using a four letter PDB protein name, Swiss-Prot or GenBank identification. A new database of CASTp calculations of single chains of a multiple chain complex can also be queried by adding the chain identifier to the PDB protein name. Figure 1 shows the atoms of the charge relay system that resides in a functional pocket of serine protease/inhibitor (PDB 1a2c). The atoms of annotated residues that lie in the pocket are highlighted in red in contrast to the green pocket atoms. A table of all the annotated residues are also displayed on the right hand side of the browser window. This table reports the following information: the database from which the annotation was derived from, the annotation key word from the database, the position of the annotation on the sequence of the PDB structure, the three letter amino acid code of the annotated residue, the identifications of the pocket/pockets the annotated residue is located and a brief description of the annotation. If the user chooses to have the results emailed, a text file will be sent that contains all the information listed in the above table.

Table 1. Statistics of the Swiss-Prot annotated residues

Swiss-Prot key	#PDB	#Residues
ACT_SITE	6871	20 121
METAL	5014	37 824
BINDING	3199	13 987
CARBOHYD	2620	10 266
MOD_RES	2606	6556
SITE	1993	8003
NP_BIND	1748	58 777
DNA_BIND	464	33 978
CA_BIND	358	16 413
ZN_FING	295	19 273
CROSSLNK	230	467
LIPID	187	312

Column 1 reports the Swiss-Prot site key, column 2 lists the number of PDB structures the site was mapped to and column 3 lists the number of unique residues that were mapped to PDB structures.

Calculation requests

In addition to querying a database of single chain calculations, the 'Calculation Request' page allows the user to run a calculation on any combination of chains from a multiple chain complex. If the protein contains HET groups, the user is also given the option to include any combination of the HET groups in the calculation.

Improved visualization

For visualizing annotated residues, the JMOL plug-in (http://www.jmol.org) is now added as a visualization option. JMOL runs on Windows/Mac OS X/Linux and only requires a java enabled browser. The result is added functionality and a friendlier user interface.

The user is now also presented with a corresponding sequence map, where residues in highlighted pocket are highlighted in the same color as in the structural visualization. In addition, a user has finer control. The user is able to change the pocket colorings, the display of the PDB structure in wireframe, cartoon, strands or ribbons. The user can also send customized rasmol scripts to the Chime visualization.

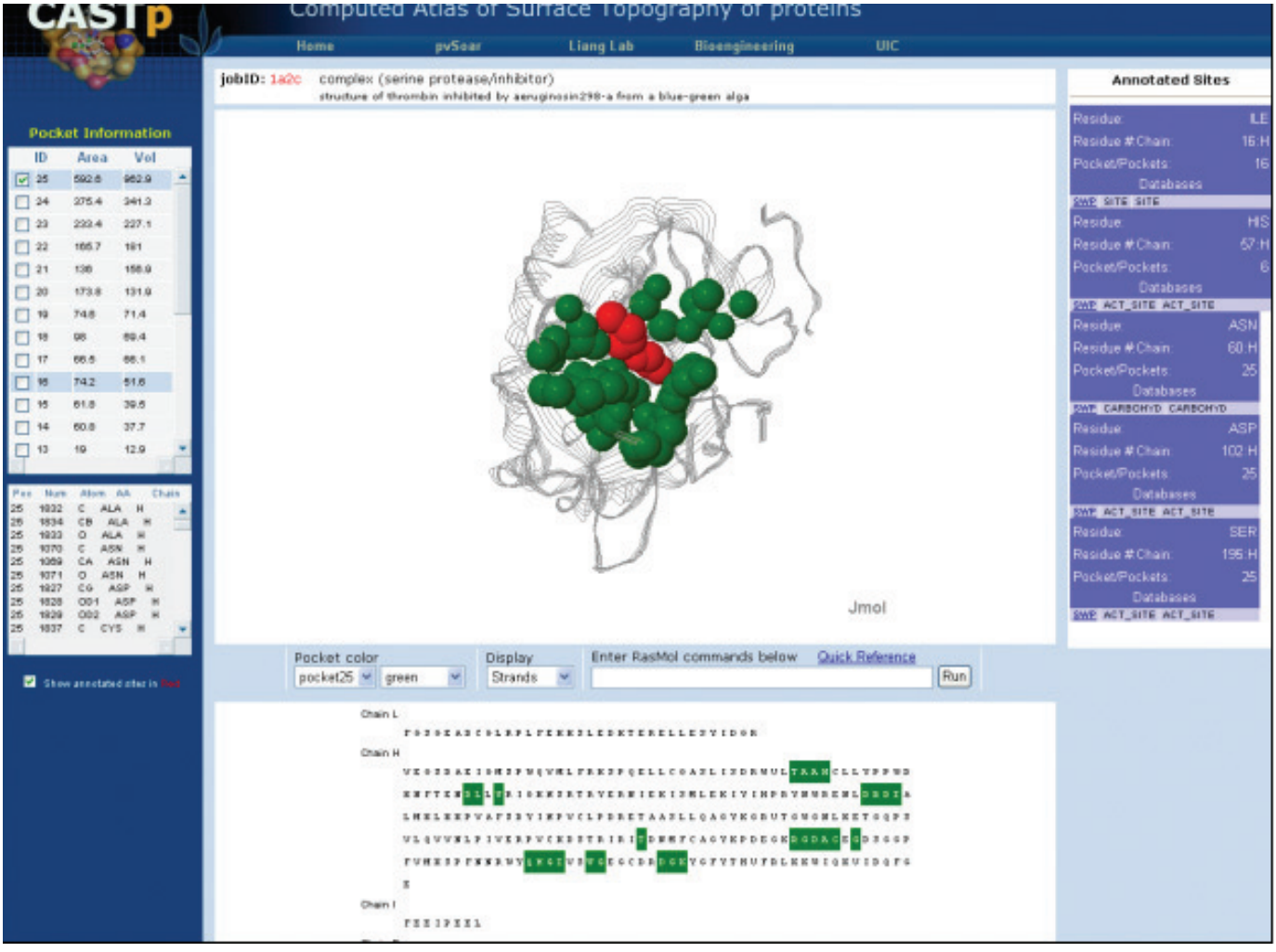

Figure 1. Chime visualization of serine protease/inhibitor (PDB 1a2c) showing atoms from residues in the functional pocket important for the charge relay system in red.

DISCUSSION

This paper describes major updates to the CASTp web server. Biologically important functional residues annotated from three sources are now mapped to PDB structures and visualization is provided. We believe these updates significantly increases the information content of CASTp and enhances our knowledge base needed for studying structural basis of protein functions.

AVAILABILITY

CASTp web server and the associated mapping database can be freely accessed on the World Wide Web at http://cast.engr.uic.edu.

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by grants from National Science Foundation (CAREER DBI0133856), National Institute of Health (GM68958),and Office of Naval Research (N00014-06-1-0100).

Conflict of interest statement. None declared.

REFERENCES

1. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
2. Binkowski,T.A., Naghibzadeh,S. and Liang,J. (2003) CASTp: computed atlas of surface topography of proteins. *Nucleic Acids Res.*, **31**, 3352–3355.
3. Liang,J., Edelsbrunner,H. and Woodward,C. (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci.*, **7**, 1884–1897.
4. McKusick,V.A. (1998) *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders, 12th edn*. Johns Hopkins University Press, Baltimore.
5. Gasteiger,E., Gattiker,A., Hoogland,C., Ivanyi,I., Appel,R.D. and Bairoch,A. (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.*, **31**, 3784–3788.
6. Stitziel,N., Tseng,Y.Y., Pervouchine,D., Goddeau,D., Kasif,S. and Liang,J. (2003) Structural location of disease-associated single-nucleotide polymorphisms. *JMB*, **327**, 1021–1030.
7. Stitziel,N., Binkowski,T.A., Tseng,Y.Y., Kasif,S. and Liang,J. (2004) topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. *Nucleic Acids Res.*, **32**, D520–D522.