BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 19 2010, pages 2458–2459
doi:10.1093/bioinformatics/btq447

Sequence analysis

Advance Access publication August 4, 2010

DMDM: domain mapping of disease mutations

Thomas A. Peterson, Asa Adadey, Ivette Santana-Cruz, Yanan Sun, Andrew Winder and Maricel G. Kann*

[1]Department of Biological Sciences, University of Maryland, Baltimore County, Baltimore, MD 21250, USA

Associate Editor: Burkhard Rost

ABSTRACT

Summary: Domain mapping of disease mutations (DMDM) is a database in which each disease mutation can be displayed by its gene, protein or domain location. DMDM provides a unique domain-level view where all human coding mutations are mapped on the protein domain. To build DMDM, all human proteins were aligned to a database of conserved protein domains using a Hidden Markov Model-based sequence alignment tool (HMMer). The resulting protein-domain alignments were used to provide a domain location for all available human disease mutations and polymorphisms. The number of disease mutations and polymorphisms in each domain position are displayed alongside other relevant functional information (e.g. the binding and catalytic activity of the site and the conservation of that domain location). DMDM's protein domain view highlights molecular relationships among mutations from different diseases that might not be clearly observed with traditional gene-centric visualization tools.

Availability: Freely available at http://bioinf.umbc.edu/dmdm

Contact: mkann@umbc.edu

Received on June 3, 2010; revised on July 16, 2010; accepted on July 31, 2010

1 INTRODUCTION

The domain mapping of disease mutations (DMDM) database provides an aggregated view of all human coding disease-related mutations and SNPs for each protein domain. Domains are the structural, functional and evolutionary units of proteins (Holm and Sander, 1996; Murzin *et al*., 1995; Orengo *et al*., 1997). Most proteins contain multiple domains in a variety of domain combinations; and different domain combinations are associated with different protein functions (Bornberg-Bauer *et al*., 2005; Doolittle, 1995; Nikitin and Lisacek, 2003). Thus, the aggregated view of all human mutations at the domain level is an extremely useful tool for visualizing the molecular events that lead to diseased and healthy states in organisms. DMDM will also aid scientists to generate new hypotheses concerning the role of protein domains in key complex biological systems.

2 METHODS

HMMer's semi-global implementation (Eddy, 1996) was used to search for complete domains in human proteins from the RefSeq (Pruitt *et al*., 2007) and SWISS-PROT (Boeckmann *et al*., 2003) databases. Hidden Markov models for protein domains from SMART (Letunic *et al*., 2006), COG (Tatusov *et al*., 2003), CDD (Marchler-Bauer *et al*., 2007) and Pfam (Finn *et al*., 2008) were built using multiple sequence alignments from CDD with the `hmmerbuild` tool. The human disease mutations and SNPs mapped onto these domains were extracted from the OMIM (McKusick, 2007), SWISS-PROT and the dbSNP (Sherry *et al*., 2001) databases.

3 DMDM NAVIGATION LAYERS

The data in DMDM can be visualized at three levels: gene, protein and protein domain. A search within DMDM can be performed at any of these three levels, or by disease name, using multiple search options. For instance, users may search by description, which is useful when only a keyword about the molecular entity is known, or by any gene or protein identifier. The results of the search in any of the layers consist of a summary of the information at the top of the page that includes a description, identifiers and external links to the gene, protein or domain. The summary is followed by either a graphical display of the information and/or tables of domain and mutational information, with key identifiers and relevant links. Proteins, both in the gene and protein layers, are depicted as a scaled bar that indicates the amino acid positions; the corresponding domains are shown below the bar. By selecting a region of the protein, information and links about the subset of mutations found around that region are displayed on a separate page along with a graphical display of each mutation.

The domain layer, an example of which is illustrated in Figure 1, displays three levels of information: sequence logos, mutational data and conserved functional features/sites for each domain position. Multiple sequence alignment information, obtained from CDD for each conserved domain model, is displayed using sequence logos [WebLogo software (Crooks *et al*., 2004)]. Mutational data for each human protein with one or more domains is represented by histograms under each position on the sequence logo. The third level, which was extracted from the CDD manual annotation and which is displayed below the histogram bars, provides the functional information for each domain position. The height of the histogram's bars represents the number of mutations found at individual domain positions for all human proteins that match that domain. Polymorphisms are represented in blue and disease mutations in red.

Redundant mutations that share location, amino acid types and gene, but that are in different proteins are counted only once. When selecting a position on the histogram bar, a list of all mutations in that position, including redundancies, are displayed on a separate page. The upper left boxes in domain pages can be used to locate a particular protein position in the display. The check boxes provide control over the display of functional features shown for each domain position. The histogram bars provide a

*To whom correspondence should be addressed.

 © The Author 2010. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

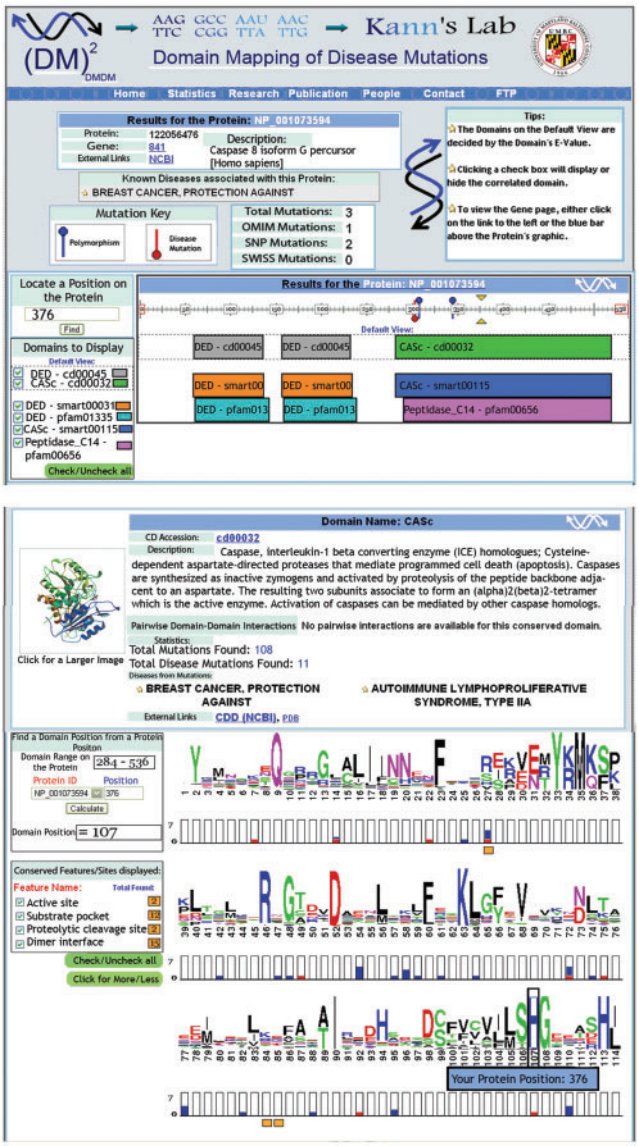

Fig. 1. Screen shots of two DMDM queries using: (**a**) the NP_001073594 protein (GI:122056476) and (**b**) the CASc (cd00032) domain. The above display only shows a subset of the actual results available to users on the DMDM site.

unique display of all the available information regarding human mutations, polymorphisms and disease mutations that were mapped to a particular domain position.

4 CONCLUSIONS AND FUTURE DEVELOPMENTS

The DMDM database is an online resource created for displaying human mutations with other relevant functional information at the protein domain level. This domain-centric database provides an ideal framework for studying biological processes relevant to human health, as well as for the integration of other molecular events, such as protein translational modifications and alternative splicing events affecting protein domains. Updates to DMDM will be performed every six months. In the future, DMDM will be expanded to map and integrate additional experimental data from the next generation of sequence, gene expression and proteomic experiments.

ACKNOWLEDGEMENTS

The authors wish to thank Russ B. Altman, Donna Maglott, Aron Marchler-Bauer and Mileidy W. Gonzalez for their helpful comments on the manuscript and feedback on the webpage design.

Funding: National Institutes of Health (NIH) [1K22CA143148 to M.G.K. (PI); R01LM009722 to M.G.K. (collaborator)].

Conflict of Interest: none declared.

REFERENCES

Boeckmann,B. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Bornberg-Bauer,E. *et al.* (2005) The evolution of domain arrangements in proteins and interaction networks. *Cell. Mol. Life Sci.*, **62**, 435–445.

Crooks,G.E. *et al.* (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.

Doolittle,R.F. (1995) The multiplicity of domains in proteins. *Annu. Rev. Biochem.*, **64**, 287–314.

Eddy,S.R. (1996) Hidden Markov models. *Curr. Opin. Struct. Biol.*, **6**, 361–365.

Finn,R.D. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.

Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science,* **273**, 595–603.

Letunic,I. *et al.* (2006) SMART 5: domains in the context of genomes and networks. *Nucleic Acids Res.*, **34**, D257–D260.

Marchler-Bauer,A. *et al.* (2007) CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res.*, **35**, D237–D240.

McKusick,V.A. (2007) Mendelian Inheritance in Man and its online version, OMIM. *Am. J. Hum. Genet.*, **80**, 588–604.

Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Nikitin,F. and Lisacek,F. (2003) Investigating protein domain combinations in complete proteomes. *Comput. Biol. Chem.*, **27**, 481–495.

Orengo,C.A. *et al.* (1997) CATH—a hierarchic classification of protein domain structures. *Structure,* **5**, 1093–1108.

Pruitt,K.D. *et al.* (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.

Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

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