*W642–W644* *Nucleic Acids Research, 2006, Vol. 34, Web Server issue*
*doi:10.1093/nar/gkl323*

# SNP@Domain: a web resource of single nucleotide polymorphisms (SNPs) within protein domain structures and sequences

**Areum Han, Hyo Jin Kang, Yoobok Cho[1], Sunghoon Lee, Young Joo Kim and Sungsam Gong***

National Genome Information Center (NGIC), Korea Research Institute of Bioscience and Biotechnology
52 Eoeun-dong, Yuseong-gu, Daejeon 305-333, Korea and [1]Object Interaction Technologies, Inc., Daejeon, Korea

Received February 14, 2006; Revised March 1, 2006; Accepted April 13, 2006

## ABSTRACT

**The single nucleotide polymorphisms (SNPs) in conserved protein regions have been thought to be strong candidates that alter protein functions. Thus, we have developed SNP@Domain, a web resource, to identify SNPs within human protein domains. We annotated SNPs from dbSNP with protein structure-based as well as sequence-based domains: (i) structure-based using SCOP and (ii) sequence-based using Pfam to avoid conflicts from two domain assignment methodologies. Users can investigate SNPs within protein domains with 2D and 3D maps. We expect this visual annotation of SNPs within protein domains will help scientists select and interpret SNPs associated with diseases. A web interface for the SNP@Domain is freely available at http://snpnavigator.net/ and from http://bioportal.net/.**

## INTRODUCTION

To facilitate the identification of disease-associated single nucleotide polymorphisms (SNPs) from a large number of SNPs, it is important to select functionally relevant SNPs (1). There are many SNP annotation servers and databases, such as FESD (http://combio.kribb.re.kr/FESD/), PicSNP (http://plaza.umin.ac.jp/~hchang/picsnp/), SNPper (http://snpper.chip.org/) and SNPs3D (http://www.snps3d.org). These are useful for selecting SNPs without a priori biological knowledge (2–13). They help biologists focus on specific genomic/proteomic regions or gene sets providing functional annotations and visualization.

The SNPs in conserved protein regions have been thought to be strong candidates that can alter protein functions (8,11). However, up to now, there have been no web servers that provide extensive protein domain annotation of SNPs. Currently, Ensembl (14) provides domain annotation of SNPs assigned by Pfam (15), PROSCAN (16) and PFscan (17). However, these protein domains are all sequence-based functional domains that are based on protein sequence profiles. Structure-based approaches define domains according to the compactness and conservation of protein structural regions (18) while sequence-based domain databases constructed based on sequence similarity of proteins implied evolutionary relationships (19,20). If a structure-based domain family and sequence-based domain family are defined at the same location over the same set of protein chains, they should map exactly to each other in a protein. However, it has been known that they have conflicts (19,20).

SCOP (21) is a representative structure-based classification database for Protein Data Bank (PDB) (22). They list all the proteins with known structures and organize them hierarchically. Pfam (15) is a representative sequence-based domain database that contains hidden Markov model-based profiles of many common protein domains constructed using multiple sequence alignments. Previously, Elofsson's group (19) reported that 70% of SCOP families exist in Pfam, while 57% of Pfam families exist in SCOP. Recent research conducted by Zhang's group (20) shows that 80% of SCOP domains overlap with at least one Pfam family. These SCOP domain families correspond to 99.7% of the Pfam families. Although the overlaps increased (SCOP, from 70 to 80%; and Pfam, from 57 to 99.7%), partial mapping between SCOP and Pfam domain could still occur. Zhang's group reported that only 62% of the cases of one-to-one mapping of a SCOP domain to a Pfam domain agreed by 90% or more of their coverage (20).

Since a non-synonymous SNP can correspond to an amino acid change, it is necessary to have a good protein domain annotation and visualization server. Here, we introduce the

*To whom correspondence should be addressed. Tel: +82 42 879 8549; Fax: +82 42 879 8519; Email: ssgong@kribb.re.kr

SNP@Domain server providing information for SNPs found within protein domains. SNP@Domain contains all the human SNPs from dbSNP (23) that match SCOP and Pfam domain sequences that are assigned to Ensembl database proteins. A 2D map of Pfam and SCOP domains with SNPs is provided. Additionally, a 3D map of SNPs within domains is provided if protein structures are available.

## METHODS AND USAGE

### Identifying SNPs within protein domains

We annotated protein domains to human proteins in the Ensembl database (ftp://ftp.ensembl.org/pub/human-25.34e/data/mysql/homo_sapiens_snp_25_34e) and mapped whole SNPs from dbSNP (http://www.ncbi.nlm.nih.gov/SNP/) (23). Since the Ensembl database provides Pfam domain annotation information, we performed a structure-based domain assignment using the PDB-ISL method (24,25) using SCOP version 1.65 and Ensembl human proteins. Domains were classified by keeping BLAST-matched regions having an *E*-value 1e − 4 or lower. In total, 17 639 SNPs within SCOP and 28 238 SNPs within Pfam domains were identified. Furthermore, 4226 (12.39%) human proteins had at least one SNP within SCOP domain regions and 6781 (19.88%) human proteins had at least one SNP within Pfam domain regions. Two useful annotations of SNPs were parsed with Perl scripts, and subsequently imported into a MySQL relation database including (i) the effects of SNPs predicted by the Sorting Intolerant from Tolerant Server (SIFT; http://blocks.fhcrc.org/sift/SIFT.html) (11) and (ii) the relationships between SNPs and diseases from the Online Mendelian Inheritance in Man (http://www.ncbi.nlm.nih.gov/Omim/) (26) database.

### Two- and three-dimensional maps of SNPs within protein domains

SNP@Domain is a web-based tool that was constructed using Java Server Pages and Perl Common Gateway Interface scripts. SNP@Domain provides three query interfaces as shown in Figure 1: (i) SNP identifier (rs number), (ii) gene identifier (Ensemble protein ID) and (iii) domain identifier (SCOP concise classification strings ID or Pfam ID). SNP@Domain also supports keyword searches with gene and/or domain names. When the user accesses it with a queried SNP or a gene name, the 2D image map of SNPs within protein domains is displayed as shown in Figure 2. This 2D image map utilizes the Generic Genome Browser (Gbrowse; http://www.gmod.org), originally developed by Stein's group (27). The 2D map has four kinds of horizontal tracks corresponding to SCOP domains, Pfam domains, synonymous and non-synonymous SNPs within a protein. For convenience, synonymous SNPs and non-synonymous SNPs are displayed separately. The queried SNPs are highlighted in the map so they can be easily distinguished. Each SNP in the 2D map links to detailed information of the SNP such as chromosomal position, class, validation, alleles, effects predicted by SIFT server and relationships with disease(s), if available. If the structure of the protein is available in the PDB, SNP@Domain

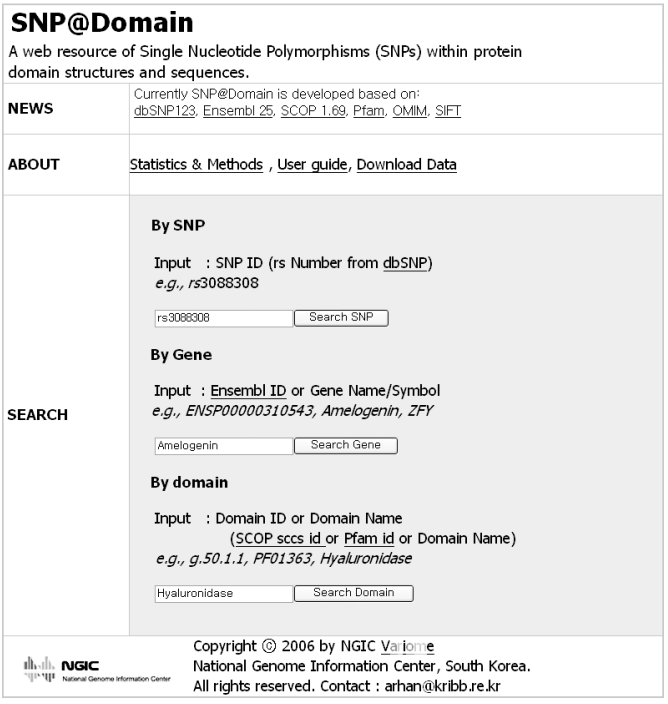


**Figure 1.** Search interface of SNP@Domain. The user is able to search SNP domain annotations with three inputs including (i) SNP identifier (rs number), (ii) Gene identifier (Ensembl protein ID) or name/symbol, and (iii) Domain identifier (SCOP concise classification strings ID or Pfam ID) or name.

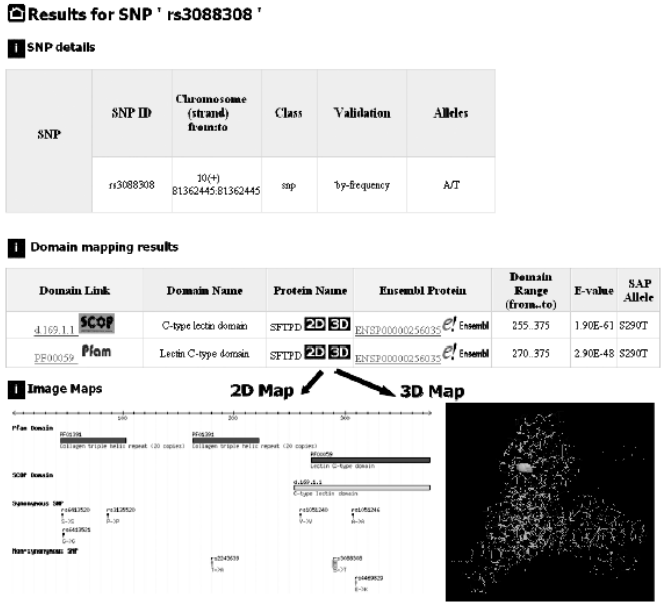


**Figure 2.** An example of detail information and image maps of an SNP within protein domains. Following the user's query to the SNP (rs number = 'rs3088308'), the SNP's detail information including chromosomal location, class, validation and alleles were displayed. And a summary of domain mapping results and a corresponding 2D image map were shown up. Four tracks of the 2D image map were displayed including (i) Pfam domain, (ii) SCOP domain, (iii) synonymous SNPs and (iv) non-synonymous SNPs within the protein. The 3D image map of the SNP is also available.

provides a 3D view of the protein highlighting the amino acids affected by SNPs. To avoid sequence conflicts between an Ensembl protein sequence and a PDB sequence, SNP@Domain carries out a BLAST with a query of Ensemble protein sequence against a protein sequence from PDB and parsed hits. We use MDL Chime plugin (http://www.mdli.com/downloads/) for visualizing 3D structures of proteins which was developed based on RasMol (http://www.umass.edu/microbio/rasmol/) (28).

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## ACKNOWLEDGEMENTS

We thank Maryana and Jong Bhak for editing the manuscript. This project was supported by the Korean Ministry of Science and Technology (MOST) under grant number M10508040002-05N0804-00210 and M10407010001-05N0701-00100. Y.B.C. is supported by Biogreen21 program (20050401-034-791-006-03-00 and 20050301-034-481-006-02-00). Funding to pay the Open Access publication charges for this article was provided by M10407010001-05N0701-00100 grant of MOST.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Wjst,M. (2004) Target SNP selection in complex disease association studies. *BMC Bioinformatics*, **5**, 92.
2. Kang,H.J., Choi,K.O., Kim,B.D., Kim,S. and Kim,Y.J. (2005) FESD: a Functional Element SNPs Database in human. *Nucleic Acids Res*., **33**, D518–D522.
3. Chang,H. and Fujita,T. (2001) PicSNP: a browsable catalog of nonsynonymous single nucleotide polymorphisms in the human genome. *Biochem. Biophys. Res. Commun*., **287**, 288–291.
4. Riva,A. and Kohane,I.S. (2002) SNPper: retrieval and analysis of human SNPs. *Bioinformatics*, **18**, 1681–1685.
5. Yue,P., Melamud,E. and Moult,J. (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 166.
6. Reumers,J., Schymkowitz,J., Ferkinghoff-Borg,J., Stricher,F., Serrano,L. and Rousseau,F. (2005) SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. *Nucleic Acids Res*., **33**, D527–D532.
7. Dantzer,J., Moad,C., Heiland,R. and Mooney,S. (2005) MutDB services: interactive structural analysis of mutation data. *Nucleic Acids Res*., **33**, W311–W314.
8. Zhang,F. and Zhao,Z. (2005) SNPNB: analyzing neighboring-nucleotide biases on single nucleotide polymorphisms (SNPs). *Bioinformatics*, **21**, 2517–2519.
9. Stitziel,N.O., 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.
10. Bao,L., Zhou,M. and Cui,Y. (2005) nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res*., **33**, W480–W482.
11. Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*., **31**, 3812–3814.
12. Ramensky,V., Bork,P. and Sunyaev,S. (2002) Human non-synonymous SNPs: server and survey. *Nucleic Acids Res*., **30**, 3894–3900.
13. Doron-Faigenboim,A., Stern,A., Mayrose,I., Bacharach,E. and Pupko,T. (2005) Selecton: a server for detecting evolutionary forces at a single amino-acid site. *Bioinformatics*, **21**, 2101–2103.
14. Hubbard,T., Andrews,D., Caccamo,M., Cameron,G., Chen,Y., Clamp,M., Clarke,L., Coates,G., Cox,T., Cunningham,F. *et al*. (2005) Ensembl 2005. *Nucleic Acids Res*., **33**, D447–D453.
15. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res*., **32**, D138–D141.
16. Bairoch,A., Bucher,P. and Hofmann,K. (1996) The PROSITE database, its status in 1995. *Nucleic Acids Res*., **24**, 189–196.
17. Bucher,P., Karplus,K., Moeri,N. and Hofmann,K. (1996) A flexible motif search technique based on generalized profiles. *Comput. Chem*., **20**, 3–23.
18. Veretnik,S., Bourne,P.E., Alexandrov,N.N. and Shindyalov,I.N. (2004) Toward consistent assignment of structural domains in proteins. *J. Mol. Biol*., **339**, 647–678.
19. Elofsson,A. and Sonnhammer,E.L. (1999) A comparison of sequence and structure Protein domain families as a basis for structural genomics. *Bioinformatics*, **15**, 480–500.
20. Zhang,Y., Chandonia,J.M., Ding,C. and Holbrook,S.R. (2005) comparative mapping of sequence-based and structure-based protein domains. *BMC Bioinformatics*, **6**, 77.
21. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*., **247**, 536–540.
22. 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 Acid Res*., **28**, 235–242.
23. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res*., **29**, 308–311.
24. Teichmann,S.A., Chothia,C., Church,G.M. and Park,J. (2000) Fast assignment of protein structures to sequences using the intermediate sequence library PDB-ISL. *Bioinformatics*, **16**, 117–124.
25. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
26. Hamosh,A., Scott,A.F., Amberger,J., Bocchini,C., Valle,D. and McKusick,V.A. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res*., **30**, 52–55.
27. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al*. (2002) The Generic Genome Browser: a building block for a Model Organism System Database. *Genome Res*., **12**, 1599–1610.
28. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci*., **20**, 374–376.