BMC Bioinformatics

Research

An integrated approach to the interpretation of Single Amino Acid Polymorphisms within the framework of CATH and Gene3D

Jose MG Izarzugaza*[1,2], Anja Baresic[1], Lisa EM McMillan[1], Corin Yeats[1], Andrew B Clegg[1], Christine A Orengo[1], Andrew CR Martin[1] and Alfonso Valencia[2]

Address: [1]Institute of Structural and Molecular Biology, Darwin Building, University College London, Gower Street, London WC1E 6BT, UK and [2]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), C/Melchor Fernandez Almagro 3, E28029 Madrid, Spain

Email: Jose MG Izarzugaza* - jmgonzalez@cnio.es; Anja Baresic - anya@biochem.ucl.ac.uk; Lisa EM McMillan - mcmillan@biochem.ucl.ac.uk; Corin Yeats - yeats@biochem.ucl.ac.uk; Andrew B Clegg - clegg@biochem.ucl.ac.uk; Christine A Orengo - orengo@biochem.ucl.ac.uk; Andrew CR Martin - a.martin@biochem.ucl.ac.uk; Alfonso Valencia - avalencia@cnio.es

* Corresponding author

Abstract

Background: The phenotypic effects of sequence variations in protein-coding regions come about primarily via their effects on the resulting structures, for example by disrupting active sites or affecting structural stability. In order better to understand the mechanisms behind known mutant phenotypes, and predict the effects of novel variations, biologists need tools to gauge the impacts of DNA mutations in terms of their structural manifestation. Although many mutations occur within domains whose structure has been solved, many more occur within genes whose protein products have not been structurally characterized.

Results: Here we present 3DSim (3D Structural Implication of Mutations), a database and web application facilitating the localization and visualization of single amino acid polymorphisms (SAAPs) mapped to protein structures even where the structure of the protein of interest is unknown. The server displays information on 6514 point mutations, 4865 of them known to be associated with disease. These polymorphisms are drawn from SAAPdb, which aggregates data from various sources including dbSNP and several pathogenic mutation databases. While the SAAPdb interface displays mutations on known structures, 3DSim projects mutations onto known sequence domains in Gene3D. This resource contains sequences annotated with domains predicted to belong to structural families in the CATH database. Mappings between domain sequences in Gene3D and known structures in CATH are obtained using a MUSCLE alignment. 1210 three-dimensional structures corresponding to CATH structural domains are currently included in 3DSim; these domains are distributed across 396 CATH superfamilies, and provide a comprehensive overview of the distribution of mutations in structural space.

Conclusion: The server is publicly available at http://3DSim.bioinfo.cnio.es/. In addition, the database containing the mapping between SAAPdb, Gene3D and CATH is available on request and most of the functionality is available through programmatic web service access.

Background

The most common biologically-relevant mutations are single base changes often referred to as **single nucleotide polymorphisms** (SNPs). These account for about 90% of sequence polymorphisms in humans [1] at an overall frequency of about one per 1000 bases [2]. Traditionally, SNPs are classified as coding or non-coding according to their genomic location – coding SNPs are further sub-classified according to the protein product expressed. Non-Synonymous SNPs (nsSNPs) are those that alter the amino acid sequence of the protein product, either through amino acid substitution (a 'single amino acid polymorphisms', SAAP), or by the generation of truncation mutations. By contrast, synonymous SNPs (also referred to as silent or sSNPs) are those that do not alter the amino acid sequence of the protein product.

Not all synonymous SNPs are neutral since they may still affect the expression of gene products or protein translation by introducing alterations into regulatory regions, interfering with splice sites or impinging on other regulatory mechanisms [3,4]. Similarly, not all nsSNPs are associated with pathological diseases, since some changes are, by nature, milder than others, and diseases commonly involve complex sets of alterations.

Strictly the term 'SNP' is defined as a mutation which occurs in at least 1% of a 'normal' population. Thus SNPs are expected to have a neutral non-deleterious or low-penetrance phenotypic effect whereas the term **pathogenic deviation** (PD) refers to those mutations that generally occur at much lower frequencies in the population and have a severe effect on phenotype.

The most commonly used database for storing information on SNPs is dbSNP [5], which currently contains several million validated SNPs from humans and other species. Other sources of genomic information about SNPs include Ensembl [6] and the HapMap Project [7].

Several efforts have been devoted to the prediction of the pathogenicity of amino acid mutations, resulting from single nucleotide changes. These methods make use of a set of characteristics which may be based both on sequence and structure, to determine whether a mutation can affect protein function and therefore be, potentially, associated with disease. This is an area of active research as shown by the considerable number of publications on the subject during the last few years [8-18].

Several efforts, SAAPdb [19] among others, have been devoted to compiling this information and to providing a sequence and structural analysis, where possible, aiming to determine the origin of the pathogenicity shown. In this type of repository, the term SNP is used to refer to

essentially phenotypically silent mutations, while PD is used for mutations known to have a severe effect on phenotype, i.e. any single base change reported to correlate with disease. Online Mendelian Inheritance in Man (OMIM) [20] is a collection of information about inherited disease and contains data on PDs. However a great deal more information is held and maintained by individual research groups in locus-specific mutation databases or LSMDBs [21]. Like PDs, nsSNPs are point mutations, but by definition they occur in at least 1% of a 'normal' population. They are expected to have a neutral non-deleterious or low-pentrance phenotypic effect whereas PDs are known to be detrimental. By mapping these SAAPs (a term we use for both PDs and mutations resulting from nsSNPs) onto protein structures, we can begin to understand how protein structure might be affected by mutant residues, and so begin to explain the functional effect (if any) of the mutation. SAAPdb provides potential explanations for both PDs, derived from various sources, and SNPs, derived from dbSNP [5].

The CATH [22] structural domain database is a manually curated classification of domain structures found in the Protein Data Bank (PDB) [23], grouped according to evolutionary relationships and structural features. Hidden Markov Models (HMMs) are derived from alignments of these structural exemplars and used by Gene3D [24] to identify homologues within the protein sequences of UniProt [25], RefSeq [20] and Ensembl [6].

Here we present 3DSim (3D Structural Implication of Mutations), a system mapping single amino-acid polymorphisms onto structures of CATH domains. For sequences with no known structure, the Gene3D resource of domain structure annotations is used to map the sequence onto the closest homologous domain of known structure in CATH. Thus 3DSim is of particular interest when no structural information is available for a protein in which mutations are known to occur as it uses sequence homology to map to the closest representative structure. This provides a comprehensive overview of the distribution of mutations in structural space, as well as a visualization tool for pinpointing the locations of mutations on individual structures rendered in Jmol http://www.jmol.org/, as well as links to detailed information on each sequence, structure and mutation. The 3DSim application, which was designed with the aim of being very intuitive, easy to use and user-friendly, is publicly available at http://3DSim.bioinfo.cnio.es/. Several worked examples are available, along with a 6-minute video tutorial. In addition, for those advanced users needing intensive programmatic access to the information stored, the underlying database containing the mappings between SAAPdb, Gene3D and CATH is available on

request, and most of the functionality is available as web services implemented in SOAP.

Results and discussion

The mapping between SAAPdb and Gene3D

SAAPdb contains polymorphism data for 11956 sequences without a structure. Almost all of these could be mapped to Gene3D: 11904 identical sequences were found in the Gene3D database. Of the remaining 52, 17 had sequences with the same length and associated uniprot accession, leaving only 35 for which a reliable match could not be obtained directly.

The mapping between Gene3D and CATH

Where no structural data are available, the best representative CATH domain is selected on the basis of homology. For each of the 2179 superfamilies in CATH, a database of all CATH domains was built. For each of the 11904 Gene3D domain sequences mapped to CATH structural superfamilies for which there is information about mutations in SAAPdb (see previous section), a BLAST [26] search was run against the corresponding superfamily database. The closest relative found (i.e. the one with the lowest e-value and highest sequence identity) was used to cluster the sequences. Sequences with a sequence identity less than 20% were placed in separate clusters. This process yielded 2091 different groups. The groups (including the sequence of the representative structure) were then aligned using MUSCLE [27] and the resulting alignments used to transfer the mutations from Gene3D sequences to CATH domain representative structures. At the end of the pipeline we were able to display information on 6514 point mutations, 4865 of them known to be associated with disease, mapping to 396 CATH superfamilies. The complete pipeline is described in Figure 1 and details are provided in the Methods section.

Description of web application

The initial input for the system is a CATH superfamily identifier for which the user wants to retrieve information on mapped mutations. Alternatively, the database can be searched using Uniprot accession codes or CATH domain identifiers. The user can either manually introduce the desired superfamily identifier in the provided form, or browse the superfamilies in CATH in order to access the information. After this initial step, information about the
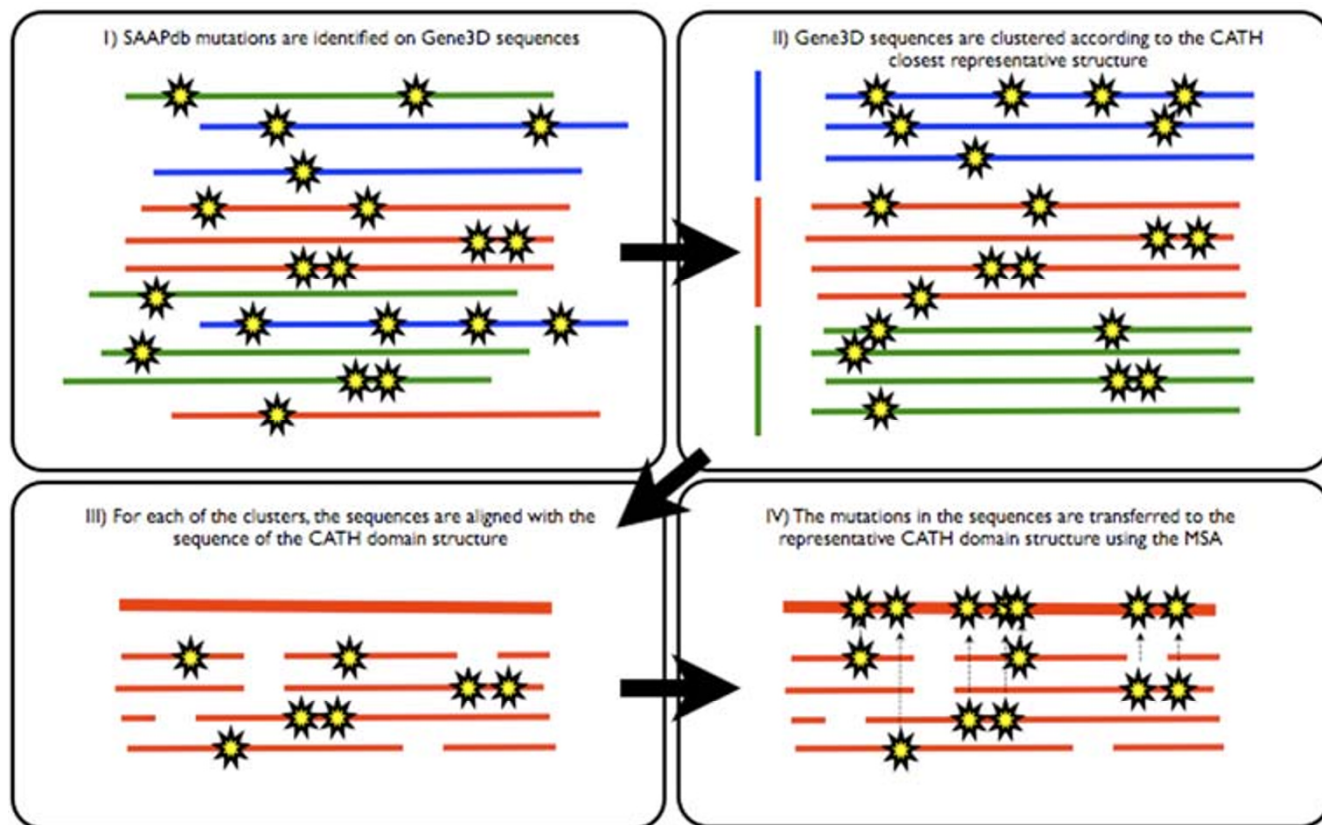

Figure 1
The mapping pipeline. Illustration of the mapping process between SAAPdb and CATH via Gene3D

selected CATH superfamily is displayed along with the CATH domains for which there is information about mutations in SAAPdb.

In addition, for users that are interested in a general overview of the distribution of mutations within structural superfamilies in CATH, one can obtain a list of the superfamilies with known mutations and analyze domains in that superfamily.

Once the user has selected a CATH domain, 3DSim displays both an interactive Jmol plug-in that allows the visualization of the mutations projected onto the three-dimensional structure of the representative CATH domain and a table displaying all the information available for that given domain in terms of available mutations, sequence and structure positions of the mutations, pathogenicity information, and similarity (BLAST sequence identity) between the sequences in Gene3D and the representative CATH domain sequence.

This similarity index provides the user with a hint about the reliability of the homology based transference of mutations between sequences in Gene3D and the structures in CATH. As a rule of thumb, the higher the similarity the more reliable the transference of mutations is. Tweaking this index is of particular interest when there are few mutations in the close relatives for a given structural family and looser constrains need to be taken into account to allow more mutations in the analysis. By default, the server rejects those mutations transferred from sequences obtaining a BLAST sequence identity of less than 20%, but – due to the interactive approach of the server – the user can decide to establish more stringent constrains depending on the study case.

In addition, the site is linked to several external annotation providers (including CATH, Gene3D, SAAPdb, Modbase, PDBsum and UniProt) where more information about the mutations, the proteins and the structures can be gathered. In particular, SAAPdb provides information about the structural implications of mutations. This information can be related, in some cases, to the pathogenic character of the mutations and provides an insight into the mechanism of molecular function for several proteins.

Figure 2 shows a worked example of the different views available through the server's graphical user interface.

Description of web services
In order to allow remote programmatic access to the information contained in the database, we have developed a total of nine SOAP web services, powered by the Perl SOAP::Lite toolkit http://www.soaplite.com/. These allow users to retrieve in simple XML format:

• all known mutations for a given CATH domain, grouped by UniProt ID.

• the total number of mutations in a CATH domain.

• all the CATH domains which are associated with a given UniProt ID.

• the amino-acid sequence of a given CATH domain.

• all CATH domains in a CATH superfamily, queried by the four-part CATH code.

• the superfamily to which a given CATH domain belongs.

• the description and representative structure associated with a given CATH superfamily.

• all the mutations in SAAPdb for a given UniProt accession.

• the total number of mutations in SAAPdb for a given UniProt accession.
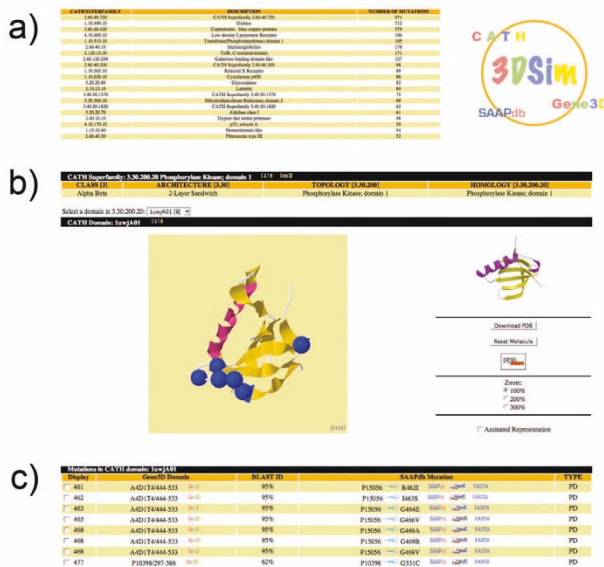

Figure 2
The web interface. A collage showing a worked example of the different sections available through the server's graphical user interface. a) Browsable list of superfamilies b) Example of structure displaying known pathogenic mutations c) Explanatory table of the mutations

These services were designed in such a way as to facilitate construction of computational analysis pipelines. For example, a user starting with a UniProt protein of interest could retrieve a list of all the domains found in that protein, then the CATH superfamilies to which each domain belongs, then all the other domains found in each superfamily, and filly all known mutations in those related domains, by chaining together four web service calls.

More information is provided at http://3DSim.bio info.cnio.es/webservices.html. The page contains example Perl code for querying the web-services, and examples of output from each one.

Database update
The database storing the information presented by both the webserver and the webservice intrinsically depends on the other databases providing the source information (i.e. CATH, Gene3D and SAAPdb), each one being updated at its own pace. This fact, in addition to the computationally expensive calculations needed to compute the mapping between Gene3D and the representative structures in CATH, makes it impossible to schedule an automatic updating calendar. Therefore, the database will be updated based on a release system, where new versions will be made public as regularly as possible.

Typical usage example
As an illustrative example, here we present the case of the ATP binding subunit of the kinases (CATH superfamily 1.10.510.10) which is accessible through the server's web page http://3DSim.bioinfo.cnio.es. This superfamily corresponds to the Phosphotransferase domain I homology group in CATH, and is subdivided into a number of different domains. However, for this particular example, we will focus only on the domain with the highest number of mutations (24), 1rw8A02. Of these 24 mutations, only three come from the sequence which maps directly to the domain. The remaining 21 come from homologous sequences with 40–65% sequence identity identified via

Gene3D (Table 1). Figure 3 shows the structure with the pathogenic deviations coloured in blue. This image can be obtained directly from the server, and is one of the main features available for the analysis of the distribution of mutations within structures. Additional links to other structure-based databases such as PDBsum [28] are provided in order to enhance the information provided, for this particular case, the position of the catalytic site, involved in binding of ATP, is described to be near residues from 333 to 338. Visual inspection of the position of the pathogenic deviations reveals that they tend to cluster around this catalytic core of the structure. Indeed, the higher the similarity in terms of BLAST identity between the CATH domain and the Gene3D sequence, the closer these positions are to the binding core and hence, more reliable the observations are.

This PDB chain (1rw8A) maps to the UniProtKB/Swiss-Prot accession P36897 and the information provided by the UniProt record (accessible through the web server's cross references) shows that it corresponds to the TGF-beta receptor type-1 precursor in humans (TGFR1_HUMAN) for which there is a level of association with disease, in particular to Furlong syndrome also known as Loeys-Dietz syndrome type 1A (LDS1); [29]. LDS1 is an aortic aneurysm syndrome with widespread systemic involvement. The disorder is characterized by arterial tortuosity and aneurysms, craniosynostosis, hypertelorism, and bifid uvula (cleft palate). Other findings include exotropy, micrognathia and retrognathia, structural brain abnormalities, intellectual deficit, congenital heart disease, translucent skin, joint hyperlaxity and aneurysm with dissection throughout the arterial tree. The mutations listed as pathogenic deviations (R487P, M318R and D400G) in the server for this protein, which has a 100% identity between the Gene3D sequence and the representative structure of the CATH domain are reported in the literature [29] as involved in LDS1 development.

Table 1: Mutations mapped to 1rw8A02. SwissProt accession P36897 maps directly to PDB code 1rw8 chain A and to CATH domain 1rw8A02 which represents residues 285–500. Other SwissProt entries containing known pathogenic deviations (PDs) are mapped to this domain via Gene3D and the mutations are mapped to the 1rw8 structure.

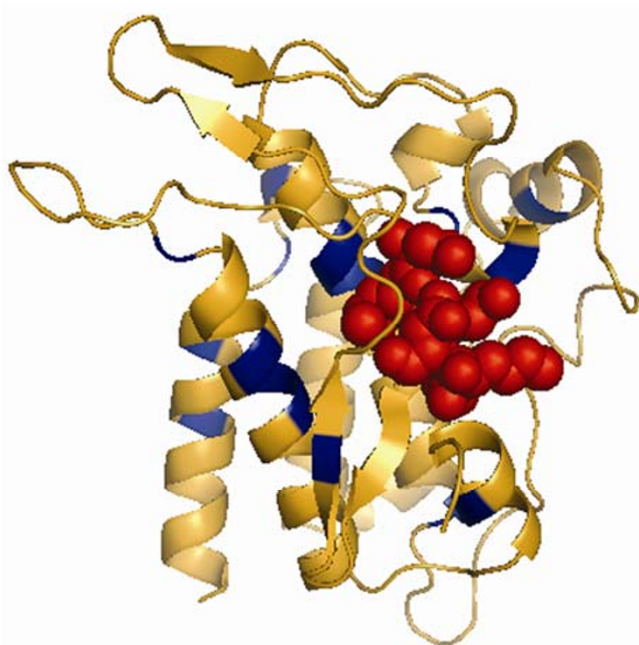| SwissProt Accession	Domain Range	Sequence Identity	Mutations
P36897	285–500	100%	M318R, D400G, R487P
O00238	284–499	65%	R486W
P36894	314–529	62%	A338D, C376Y, M470T
P37023	282–497	60%	C344Y, R374W, M376R I398N, W399S, R411P R411Q, R411W, R484W
P37173	330–546	40%	Y336N, A355P, G357W S449F, E526Q, R528C R528H, R537C

Figure 3
Three-dimensional structure of CATH domain
1rw8A02. 1rw8A02 is member of the Phosphotransferase
superfamily in CATH (1.10.510.10). 24 positions are
reported to be pathogenic mutations (PDs, blue).

Conclusion
We have presented 3DSim (3D Structural Implication of
Mutations), a system that enables the localization and vis-
ualization of single amino acid polymorphisms projected
onto protein structures based on homologous relation-
ships captured in the CATH and Gene3D databases. This
provides a comprehensive overview of the distribution of
mutations in structural space.

Although there are other servers mapping mutations to
structure already (reviewed by Uzun *et al.* [30]) the server
presented here has several unique features not available in
existing servers. Firstly, the similar treatment of SNPs and
the rarer more harmful PDs allows users to inspect and
compare both kinds of mutation through the same inter-
face, including explanatory metadata where available. Sec-
ondly, the localization of these SAAPs within the CATH
hierarchy allows users to query and explore the distribu-
tion of mutations at various levels of structural classifica-
tion. Thirdly, the mapping of sequences onto
homologous CATH domains via Gene3D helps users pre-
dict the effects of polymorphisms in proteins whose struc-
ture has not been solved. Finally, the availability of the
data via web services and database dumps enables power
users to include this information efficiently in their own
analyses. These facilities allow the independent integra-
tion of our data in any other pipeline or workflow.

The server has been running internally since we started
working on the analysis of point mutations in protein
families [31,32] and is accessible at http://3DSim.bio
info.cnio.es/. Examples and documentation are also avail-
able, together with a tutorial video and samples of outputs
of the main functions. This website is available to all users
with no login requirement. It is likely that we will include
additional features related with the structural interpreta-
tion of mutations and their relationship with disease, after
receiving feedback from external users.

In summary the 3DSim server provides up-to-date, com-
plete information automatically to map mutations in the
domain sequences of proteins annotated in Gene3D onto
protein structures classified in the CATH database.

Methods
Obtaining the mutations from SAAPdb
SAAPdb [19] is a database of single amino acid polymor-
phisms (SAAPs) from several resources, such as dbSNP
[5], ADABase [33], G6PD [34] HAMSTeRs [35], IARC p53
Database [36], LDLR [37], OMIM http://
www.ncbi.nlm.nih.gov/omim/, OTC [38], SOD1db [39]
and ZAP70Base [33], mapped to protein structure, where
available in the PDB [23]. As of October 2008, SAAPdb
contains 9060 unique pathogenic deviations (PDs: SAAPs
associated to a disease) and 2532 unique single nucle-
otide polymorphisms (SNPs: SAAPs with no known path-
ogenic effect) successfully mapped to the UniProtKB [25]
sequences in Gene3D [24].

Both pathogenic deviations and single nucleotide poly-
morphisms were only taken into account if the alteration
introduced was non-silent, that is, if the mutation is both
in a coding residue and the resulting amino acid is differ-
ent from the native one. Where mutations are recorded
both as neutral and disease-associated, the mutations
were considered pathogenic.

Gene3D domain assignments
The process by which homologues of CATH domains are
identified in sequences, and presented in the Gene3D
database, has been described previously [40]. For this par-
ticular dataset, the CATH v3.2.0 Hmmer HMM library was
scanned against the UniProt (Swiss-Prot and TrEMBL)
sequence database in collaboration with the SIMAP data-
base [41]. FASTA files of each superfamily were generated
by extracting the subsequences of the domains belonging
to each superfamily.

Gathering the sequences from Gene3D
For each CATH superfamily a library of one or more
HMMs is generated using the SAM Target2K procedure
[22]. These HMMs are then searched against UniProt in
collaboration with the SIMAP resource at the Munich

Information Centre for Protein Sequences [41]. The hits are resolved into a single set of non-overlapping domains for each sequence, using the in-house DomainFinder 2.0 protocol. The resulting domain subsequences were then extracted and dumped into the relevant superfamily FASTA file.

Gathering the sequences of the CATH domains
For each of the 2097 superfamiles in CATH [22], all CATH domains were recovered along with the corresponding amino acid sequences directly from CATH's Oracle database. A total of 86463 CATH domains were found. Afterwards, all CATH domains in the same CATH superfamily were grouped together in order to build a BLAST database of the sequences of three-dimensional structures specific to each of the CATH superfamilies.

Generation of the groups of Gene3D sequences represented by the same CATH domain
In order to assign the closest CATH domain to each of the Gene3D sequences belonging to the same CATH superfamily, we queried each of the sequences in Gene3D against a database of CATH domains in that superfamily using BLAST. The best BLAST hit for each of the Gene3D sequences – provided the identity between the hit and the query was greater than 20% – was considered the closest CATH domain and hence the CATH domain was assigned as the structural representative of this sequence. After performing this classification for the whole set of sequences, all Gene3D sequences represented by the same CATH domain were grouped together and all the sequences within a group considered similar. A total of 2091 unique groups were generated.

Alignment of the CATH domain groups using MUSCLE
During the previous step of the pipeline, the sequences contained in each of the groups of Gene3D sequences represented by the same CATH domain were considered similar. However, in order to collapse all the mutations from the Gene3D sequences onto the representative CATH domain sequence, the equivalence between pairs of residues needed to be established. To perform this task, multiple sequence alignments were constructed using the alignment package MUSCLE.

Mapping SAAPdb mutations to CATH domain representative structures
The alignments generated by MUSCLE during the previous step of the pipeline were used to transfer the mutations, both pathogenic (PDs) and neutral (SNPs), from the sequences in Gene3D to the corresponding CATH structural representatives.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
Conceived the idea: AV, JMGI, CO, ACRM. Gathered the data and generated the mapping: JMGI, AB, LM, CY. All authors designed the server and its functionalities. Implemented the server: JMGI. Implemented the database: JMGI. Implemented the webservices: JMGI, AC. Wrote the paper: all authors. All authors read and approved the manuscript. CNIO covered the publication expenses.

References
1. Collins FS, Brooks LD, Chakravarti A: **A DNA Polymorphism Discovery Resource for Research on Human Genetic Variation.** *Genome Research* 1998, **8(12):**1229-1231.
2. Taillon-Miller P, Gu Z, Li Q, Hillier L, Kwok PY: **Overlapping genomic sequences: a treasure trove of single-nucleotide polymorphisms.** *Genome Research* 1998, **8(7):**748-754.
3. Schattner P, Diekhans M: **Regions of extreme synonymous codon selection in mammalian genes.** *Nucleic Acids Research* 2006, **34(6):**1700-1710.
4. Sauna ZE, Kimchi-Sarfaty C, Ambudkar SV, Gottesman MM: **Silent polymorphisms speak: how they affect pharmacogenomics and the treatment of cancer.** *Cancer Research* 2007, **67(20):**9609-9612.
5. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Research* 2001, **29:**308-311.
6. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Herrero J, Holland R, Howe K, Howe K, Johnson N, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Melsopp C, Megy K, Meidl P, Ouverdin B, Parker A, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Severin J, Slater G, Smedley D, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wood M, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Flicek P, Kasprzyk A, Proctor G, Searle S, Smith J, Ureta-Vidal A, Birney E: **Ensembl 2007.** *Nucleic Acids Res* 2007, **35(Database issue):**D610-D617.
7. Consortium H: **The International HapMap Project.** *Nature* 2003, **426:**789-796.
8. Bromberg Y, Yachdav G, Rost B: **SNAP predicts effect of mutations on protein function.** *Bioinformatics* 2008, **24(20):**2397-2398.
9. Mort M, Ivanov D, Cooper DN, Chuzhanova NA: **A meta-analysis of nonsense mutations causing human genetic disease.** *Human Mutation* 2008, **29(8):**1037-1047.
10. Torkamani A, Schork NJ: **Accurate prediction of deleterious protein kinase polymorphisms.** *Bioinformatics* 2007, **23(21):**2918-2925.
11. Yue P, Moult J: **Identification and analysis of deleterious human SNPs.** *Journal of Molecular Biology* 2006, **356(5):**1263-1274.
12. Gabdoulline RR, Ulbrich S, Richter S, Wade RC: **ProSAT2-Protein Structure Annotation Server.** *Nucleic Acids Res* 2006, **34(Web Server):**W79-83.

13. Ferrer-Costa C, Gelpí JL, Zamakola L, Parraga I, de la Cruz X, Orozco M: **PMUT: a web-based tool for the annotation of pathological mutations on proteins.** *Bioinformatics* 2005, **21(14):**3176-3178.
14. Ferrer-Costa C, Orozco M, de la Cruz X: **Sequence-based prediction of pathological mutations.** *Proteins* 2004, **57(4):**811-819.
15. Wang Z, Moult J: **Three-dimensional structural location and molecular functional effects of missense SNPs in the T cell receptor Vbeta domain.** *Proteins* 2003, **53(3):**748-757.
16. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function.** *Nucleic Acids Research* 2003, **31(13):**3812-3814.
17. Ferrer-Costa C, Orozco M, de La: **Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties.** *Journal of Molecular Biology* 2002, **315(4):**771-786.
18. Wang Z, Moult J: **SNPs, protein structure, and disease.** *Human Mutation* 2001, **17(4):**263-270.
19. Hurst JM, McMillan LEM, Porter CT, Allen J, Fakorede A, Martin ACR: **SAAPdb web resource: a large scale structural analysis of mutant proteins.** *Human Mutation* 2009 in press.
20. Sayers EWW, Barrett T, Benson DAA, Bryant SHH, Canese K, Chetvernin V, Church DMM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LYY, Helmberg W, Kapustin Y, Landsman D, Lipman DJJ, Madden TLL, Maglott DRR, Miller V, Mizrachi I, Ostell J, Pruitt KDD, Schuler GDD, Sequeira E, Sherry STT, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TAA, Wagner L, Yaschenko E, Ye J: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2009, **37(Database issue):**D5-D15.
21. Claustres M, Horaitis O, Vanevski M, Cotton RG: **Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases.** *Genome Research* 2002, **12(5):**680-688.
22. Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, Sillitoe I, Yeats C, Thornton JM, Orengo CA: **The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution.** *Nucleic Acids Research* 2007:D291-D297.
23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Research* 2000, **28:**235-242.
24. Yeats C, Lees J, Reid A, Kellam P, Martin N, Liu X, Orengo C: **Gene3D: comprehensive structural and functional annotation of genomes.** *Nucleic Acids Research* 2008:D414-D418.
25. Consortium U: **The Universal Protein Resource (UniProt).** *Nucleic Acids Research* 2007:D193-D197.
26. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Research* 1997, **25(17):**3389-3402.
27. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5:**113.
28. Laskowski RA, Chistyakov VV, Thornton JM: **PDBsum more: new summaries and analyses of the known 3D structures of proteins and nucleic acids.** *Nucleic Acids Research* 2005:D266-D268.
29. Loeys BL, Chen J, Neptune ER, Judge DP, Podowski M, Holm T, Meyers J, Leitch CC, Katsanis N, Sharifi N, Xu LL, Myers LA, Spevak PJ, Cameron DE, De Backer JD, Hellemans J, Chen Y, Davis EC, Webb CL, Kress W, Coucke P, Rifkin DB, De Paepe AMD, Dietz HC: **A syndrome of altered cardiovascular, craniofacial, neurocognitive and skeletal development caused by mutations in TGFBR1 or TGFBR2.** *Nature Genetics* 2005, **37(3):**275-281.
30. Uzun A, Leslin CM, Abyzov A, Ilyin V: **Structure SNP (StSNP): a web server for mapping and modeling nsSNPs on protein structures with linkage to metabolic pathways.** *Nucleic Acids Research* 2007, **35:**W384-W392.
31. Izarzugaza JMG, Redfern OC, Orengo CA, Valencia A: **Cancer associated mutations are preferentially distributed in protein kinase functional sites.** 2009 in press.
32. Izarzugaza JMG, Redfern OC, Orengo CA, Valencia A: **Distribution of pathogenic mutations within the representative structures in the CATH hierarchy.** 2009 in press.
33. Piirilä H, Väliaho J, Vihinen M: **Immunodeficiency mutation databases (IDbases).** *Human Mutation* 2006, **27(12):**1200-1208.
34. Kwok CJ, Martin ACR, Au SWN, Lam VMS: **G6PDdb, an Integrated Database of Glucose-6-phosphate Dehydrogenase (G6PD) Mutations.** *Hum Mutat* 2002, **19:**217-224.
35. Kemball-Cook G, Tuddenham E, Wacey A: **The factor VIII Structure and Mutation Resource Site: HAMSTeRS version 4.** *Nucl Acids Res* 1998, **26:**216-219.
36. Petitjean A, Mathe E, Kato S, Ishioka C, Tavtigian SV, Hainaut P, Olivier M: **Impact of Mutant p53 Functional Properties on TP53 Mutation Patterns and Tumor Phenotype: Lessons from Recent Developments in the IARC TP53 Database.** *Hum Mutat* 2007, **28:**622-629.
37. Leigh SEA, Foster AH, Whittall RA, Hubbart CS, Humphries SE: **Update and Analysis of the University College London low Density Lipoprotein Receptor Familial Hypercholesterolemia Database.** *Ann Hum Genet* 2008, **72:**485-498.
38. Tuchman M, Jaleel N, Morizono H, Sheehy L, Lynch MG: **Mutations and Polymorphisms in the Human Ornithine Transcarbamylase gene.** *Hum Mutat* 2002, **19:**93-107.
39. Wroe R, Wai-Ling Butler A, Andersen PM, Powell JF, Al-Chalabi A: **ALSOD: the Amyotrophic Lateral Sclerosis Online Database.** *Amyotroph Lateral Scler* 2008, **9:**249-250.
40. Yeats C, Lees J, Reid A, Kellam P, Martin N, Liu X, Orengo C: **Gene3D: Comprehensive Structural and Functional Annotation of Genomes.** *Nucleic Acids Res* 2008, **36:**D414-D418.
41. Rattei T, Tischler P, Arnold R, Hamberger F, Krebs J, Krumsiek J, Wachinger B, Stümpen V, Mewes W: **SIMAP-structuring the Network of Protein Similarities.** *Nucleic Acids Res* 2008, **36:**D289-D292.