Human Mutation
OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

SM2PH-db: An Interactive System for the Integrated Analysis of Phenotypic Consequences of Missense Mutations in Proteins Involved in Human Genetic Diseases

Anne Friedrich,[1†] Nicolas Garnier,[2†] Nicolas Gagnière,[1] Hoan Nguyen,[1] Laurent-Philippe Albou,[1] Valérie Biancalana,[3] Emmanuel Bettler,[2] Gilbert Deléage,[2] Odile Lecompte,[1] Jean Muller,[1] Dino Moras,[1] Jean-Louis Mandel,[3,4] Thierry Toursel,[5] Luc Moulinier,[1] and Olivier Poch[1*]

[1]Département de Biologie et Génomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire (UMR7104), Centre National de la Recherche Scientifique/Institut National de la Santé et de la Recherche Médicale/Université de Strasbourg, Illkirch, France; [2]Institut de Biologie et Chimie des Protéines (UMR 5086); Centre National de la Recherche Scientifique/Université de Lyon, Lyon, France; [3]Laboratoire de Diagnostic Génétique, CHRU, Faculté de Médecine et laboratoire de Génétique Médicale EA3949, Université de Strasbourg, Strasbourg, France; [4]Département de Neurobiologie et Génétique, Institut de Génétique et de Biologie Moléculaire et Cellulaire (UMR7104), Centre National de la Recherche Scientifique/Institut National de la Santé et de la Recherche Médicale/Université de Strasbourg, Illkirch, France; [5]Association Franç aise contre les Myopathies, Evry, France

Communicated by Mauno Vihinen
Received 6 May 2009; accepted revised manuscript 2 November 2009.
Published online 17 November 2009 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/humu.21155

ABSTRACT: Understanding how genetic alterations affect gene products at the molecular level represents a first step in the elucidation of the complex relationships between genotypic and phenotypic variations, and is thus a major challenge in the postgenomic era. Here, we present SM2PH-db (http://decrypthon.igbmc.fr/sm2ph), a new database designed to investigate structural and functional impacts of missense mutations and their phenotypic effects in the context of human genetic diseases. A wealth of up-to-date interconnected information is provided for each of the 2,249 disease-related entry proteins (August 2009), including data retrieved from biological databases and data generated from a Sequence–Structure–Evolution Inference in Systems-based approach, such as multiple alignments, three-dimensional structural models, and multidimensional (physicochemical, functional, structural, and evolutionary) characterizations of mutations. SM2PH-db provides a robust infrastructure associated with interactive analysis tools supporting in-depth study and interpretation of the molecular consequences of mutations, with the more long-term goal of elucidating the chain of events leading from a molecular defect to its pathology. The entire content of SM2PH-db is regularly and automatically updated thanks to a computational grid data federation facilities provided in the context of the Decrypthon program.
Hum Mutat 31:127–135, 2010. © 2009 Wiley-Liss, Inc.

KEY WORDS: SM2PH-db; human genetic disease; mutation impact; genotype–phenotype relationship; structural homology model

Introduction

The completion of the human genome has provided a large volume of data that represents the basis for the characterization of all human genes [International Human Genome Sequencing Consortium, 2004]. It has also paved the way to the systematic study of its variability [Ring et al., 2006; The International HapMap Consortium, 2003], which is related to the evolution of the human species, required for its constant development and adaptation, but also to the emergence of human diseases. Human variability is naturally expressed through, for example, genetic shuffling during meiosis, but also by all sorts of accidental DNA changes, ranging from the substitution of a single nucleic acid to major chromosomal rearrangements.

A major source of interindividual human variation results from the substitution of one residue by another one, called single nucleotide polymorphism (SNP). SNPs are highly abundant and distributed throughout the genome [Stranger et al., 2007], and in addition, SNPs are the mutations that are the most related to human diseases [Antonarakis et al., 2000]. SNPs can be linked to the emergence of or to the predisposition to disease and influence its severity, progression, as well as its drug sensitivity. Deleterious SNPs occur in both coding and noncoding regions. In noncoding regions, the variation mostly affects gene expression by disrupting functional sites at the transcriptional level (e.g., transcription factor binding sites) [Kim et al., 2008], or result in splicing defects

Additional Supporting Information may be found in the online version of this article.

†The first two authors contributed equally to this article.

Present address for Nicolas Garnier: Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India.

Present address for Jean Muller: Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany.

*Correspondence to: Olivier Poch, Département de Biologie et Génomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire (UMR7104), 67 404 Illkirch, France. E-mail: poch@igbmc.fr

[Krawczak et al., 2007]. In coding regions, and in particular protein-coding genes, deleterious SNPs are mainly non synonymous SNPs (nsSNPs), also called missense mutations, which result in the modification of the amino acid sequence of the encoded protein. nsSNPs have been linked to a wide variety of diseases, for example by affecting protein function, by reducing protein solubility or by destabilizing protein structure [Chasman and Adams, 2001]. All these perturbations can be considered as the primary molecular phenotype associated with the missense mutation, having a cascade of consequences and finally leading to the emergence of a genetic disease.

The elucidation of the complex relationships between genotypic and phenotypic variations is a major challenge in the postgenomic era. Indeed, this step is crucial to a better understanding of gene functions and networks, aimed at revealing the mechanism of diseases and the development of specific therapeutic solutions.

With the current amount of information available in various biological databases, including sequences, structures, functions, pathways, interactions, variations, etc. [Galperin and Cochrane, 2009], and the subsequent development of in silico analysis tools, it is now possible to better understand and/or predict the correlation between a missense mutation and its associated molecular phenotypes. However, to gain further insight into the mechanisms of diseases, these molecular phenotypes have to be linked to several levels of phenotypic consequences, from the cell to the organism, ideally on the basis of ontologies. Unfortunately, the access to genotypic data with precise phenotypic descriptions represents a bottleneck in the elucidation of these complex relationships. Briefly, two main classes of databases incorporate mutations and the description of their consequences [Horaitis and Cotton, 1999]: (1) central databases, such as OMIM [McKusick, 2007], HGMD [Stenson et al., 2008], and UniProtKB/Swissprot [Yip et al., 2008], that include mutations related to a wide range of genes with limited genotypic/phenotypic descriptions; (2) locus-specific databases (LSDBs), which are specific to a gene (or gene family) and typically contain much more precise genotypic and phenotypic descriptions. Around 700 LSDBs are listed on the Human Genome Variation Society Web site [Horaitis et al., 2007] as being currently available on the Web. However, the LSDBs have very different data formats, although some efforts are being undertaken toward their homogenization, in particular the development of generic tools to facilitate LSDB implementation, such as the Universal Mutation Database (UMD) [Beroud et al., 2005], the Leiden Open (source) Variation Database (LOVD) [Fokkema et al., 2005], and MUTbase [Riikonen and Vihinen, 1999].

In this context, significant efforts have been devoted to providing links between human SNPs and their molecular effects, and have led to the development of novel systems combining data storage and analysis tools, both of which are indispensable for the characterization of the consequences of SNPs [Tavtigian et al., 2008]. Systems such as coliSNP [Kono et al., 2008], LS-SNP [Karchin et al., 2005], MutDB [Singh et al., 2008], SAAPdb [Hurst et al., 2009], SNPs3D [Yue et al., 2006], and topoSNP [Stitziel et al., 2004] all present structural information related to the mutated protein, which has been shown to be essential [Chasman and Adams, 2001; Wang and Moult, 2001]. Among these, however, LS-SNP is the only one that provides computationally generated comparative protein structure models when no experimentally determined structure is available. Concerning the phenotypic effects associated to the SNPs, only MutDB provides a direct access to the observed phenotype, although this is limited to the disease name. SNPs3D and LS-SNP give access to pathogenicity prediction such as the SIFT score [Ng and Henikoff, 2003], as well as in-house developed tools, but the absence of phenotypic descriptions clearly represents an obstacle in the interpretation of the molecular consequences of a nsSNP and the related disease mechanisms. Furthermore, the postgenomic era is characterized by a torrent of biological information flooding the databases: a major requirement for any newly developed database is hence to be based on a solid computing infrastructure and to ensure its regular and automated update [Philippi and Kohler, 2006]. ColiSNP and SAAPdb are the only systems that are regularly updated and consequently respond to these requirements.

To address these limitations, we have developed SM2PH-db, which stands for “from Structural Mutation to Pathology Phenotypes in Human database.” SM2PH-db provides access to a wide range of up-to-date interconnected information related to the relationship between genotype and phenotype, with particular attention being focused on structural information via 3D structure or comparative models. A detailed multidimensional (physicochemical, functional, structural, and evolutionary) characterization of the mutations based on a Sequence–Structure–Evolution Inference in Systems (SStEISy), as well as an interactive analysis platform, ensure the investigation of structural and functional impacts of missense mutations toward a better understanding of their potential molecular effects, with regard to their individual phenotypic effects.

SM2PH-db deals mainly with proteins involved in human monogenic diseases, also known as Mendelian diseases. These disorders are particularly studied, because the emergence of a pathological phenotype is linked to the alteration of a single gene. Phenotypic diversity in monogenic diseases primarily reflects mutation heterogeneity, although the action of gene modifiers and epigenetic and environmental factors must also be considered [Jirtle and Skinner, 2007; Weatherall, 1998]. Although these so-called genetic, epigenetic, and environmental backgrounds are not integrated in our system, the accurate multidimensional characterization of missense mutations will provide an initial insight into the molecular basis of monogenic diseases that could open the way in the future, to a significant improvement in our understanding of the molecular and genetic basis of common, complex diseases [Antonarakis and Beckmann, 2006].

The establishment of a suitable infrastructure for SM2PH-db required the development of automated procedures to allow the integration of information from heterogeneous sources as well as regular updates. This prompted us to perform our developments in the context of the Decrypthon program (http://www.decrypthon.fr/english/), which provides access to a computational grid as well as data storage and federation facilities, via the Decrypthon Data Center [Nguyen et al., 2008].

Currently (August 2009), SM2PH-db holds a total of 2,249 human proteins related to genetic diseases and is publicly accessible online at http://decrypthon.igbmc.fr/sm2ph.

SM2PH-db Content

The information associated with each disease-related protein can be classified into two main categories: data retrieved from existing databases and information produced for SStEISy purposes.

Data retrieved from Existing Databases

The retrieved information is organized into two classes:

1. General information such as the gene and protein names and synonyms, as well as the known splicing variants and

sequences, are extracted from the UniProtKB database [UniProt Consortium, 2008]. Cytogenetic band information is downloaded from the Genecards database [Safran et al., 2003], gene ontology annotations are extracted from the GO database [Harris et al., 2004] and the associated disease names are obtained from the Online Mendelian Inheritance in Man (OMIM) database [McKusick, 2007].

2. Missense mutations related to the disease proteins linked to their associated phenotypes are extracted from the index of "Human polymorphisms and disease mutations" from UniProtKB/Swissprot (http://www.uniprot.org/docs/humsavar.txt) and from two LSDBs (UMD-MTM1 [Biancalana et al., in preparation] and the Tissue Nonspecific Alkaline Phosphatase Gene Mutations Database, http://www.sesep.uvsq.fr/database_hypo/Mutation.html). Each disease-causing missense mutation is linked to the name of its related disease, supplemented by a severity description when available. Nonpathogenic missense mutations are associated with the "polymorphism" term, in accordance with the nomenclature used in the UniProtKB database.

At the time of this writing (August 2009), 27,884 missense mutations are recorded in SM2PH-db, among which 20,252 are considered as disease-causing and 7,632 as nonpathogenic.

Information Produced through SStEISy Approaches

To establish an appropriate SStEISy workbench, data related to sequence, structure, and evolution are processed for each disease-related protein and missense mutations are then characterized in this context. The information resulting from the processing of these data is listed below:

1. Evolutionary background information network: structural and functional annotations are linked to each disease-related protein thanks to MACSIMS (Multiple Alignment of Complete Sequences Information Management System) [Thompson et al., 2006], an information management system that combines knowledge-based methods with complementary ab initio sequence-based predictions. MACSIMS takes advantage of the multiple alignment ontology MAO [Thompson et al., 2005] to integrate several types of data in the framework of Multiple Alignments of Complete Sequences (MACS). Indeed, subfamily characterization based on MACS allows to highlight some discriminative aspects of sequence information, such as distinct conservation/variability patterns or domain organization between different phylogenetic levels [Lecompte et al., 2001]. For each disease-related protein, two MACS are computed with a modified version of the PipeAlign suite of programs [Plewniak et al., 2003], which integrates several steps ranging from homolog searches in protein sequence and structure databases to the definition of the hierarchical relationships between subfamilies. The first MACS is composed of the closest eukaryotic sequences and is used to identify evolutionary constraints at particular sequence positions that are characteristic of the protein family or subfamily. The second MACS is constructed with a sampling strategy to significantly reduce the number of aligned sequences, while at the same time maintaining the potential structural and functional information in the alignment [Friedrich et al., 2007].
2. Structural information network: the availability of a 3D structure or model of the protein is essential to gain insight into the structural impact of a mutation. The best source of protein structural information is the PDB [Berman et al., 2000], which stores almost all the experimentally resolved crystallographic structures. However, only 574 proteins out of our 2,249 human disease-related proteins are currently represented in the PDB. To enhance the available experimental data, 3D models of the wild-type proteins are automatically constructed by homology, using Modeller [Eswar et al., 2008]. The models are built by inferring the structure of a protein (the target) from the structure of another putatively homologous protein (i.e., a sequence sharing at least 30% identity) solved by experimental methods (the template). The selection of a suitable template is based on BLAST similarity searches in the PDB: templates covering the full protein are preferred, but shorter domains can be modeled when a full template is lacking, in which case several 3D models may be associated with a single protein. The pairwise alignment of the target and template proteins is extracted from the sampled MACS and is used as input to Modeller. Five homology models are constructed and the one with the best normalized DOPE score [Eramian et al., 2008] is integrated in SM2PH-db. Currently, 1,551 structures and 3D models related to wild-type proteins are stored in the database, concerning 1,370 different proteins.
3. Missense mutation information network: missense mutations are characterized according to 31 parameters (Supp. Table S1), which can be classified into three main levels of information: (a) physicochemical changes induced by the amino acid substitution: modifications in size, charge, polarity, and hydrophobicity are independently described [Taylor, 1986]. A global score reflecting the degree of modification induced by the substitution is also assigned. This score corresponds to the residues interdistance (Supp. Fig. S1) based on a vector representation of the amino acids [French and Robson, 1983] and normalized on a scale from 0 to 100, with larger distances implying less conservative substitutions; (b) functional and structural features related to the substituted position: these features include MACSIMS annotations, descriptions of the 3D context (e.g., residue relative accessibility, in contact with an annotated site, etc.) and a conservation ranking. This ranking is calculated using an in-house developed method based on a three-step process: (i) two independent conservation scores are computed for each column of the sampled MACS, namely, the free energy score [Lockless and Ranganathan, 1999] and a score based on the two-dimensional vector representation of the amino acids (Supp. Fig. S2); (ii) these scores are classified with a Dirichlet mixture algorithm [Sjolander et al., 1996] to define "groups of conservation;" (iii) the groups are then ranked. This process is reiterated for each MACS subfamily. Thus, two global conservation classes (rank 1 and rank 2) and one subfamily conservation class per subfamily are finally defined; (c) structural modifications induced by the amino acid substitution, based on the mutant 3D models. These are automatically constructed with Modeller for missense mutations that can be mapped onto a wild-type 3D model sharing more than 50% identity with its PDB template: at the time of this writing (August 2009), 9,435 3D mutant models are available in SM2PH-db (7,863 for disease-causing mutants and 1,572 for mutants considered as nonpathogenic). The change in protein relative stability upon single-site mutation (ΔΔG value) [Casadio et al., 1995] is predicted with I-Mutant2.0 [Capriotti et al., 2005]: a positive ΔΔG value implies a protein stability increase, whereas a negative ΔΔG value suggests a destabilizing mutation. The wild-type residue contacts, computed with the CSU software [Sobolev et al., 1999], are

compared to the mutated residue ones in the 3D model and any change induced by the substitution is stored.

SM2PH-db Data Generation and Update

A fully automated workflow has been developed for the generation and regular update of the entire database contents (Fig. 1), which can be divided into five main processes:

1. Protein entry list update, based on the OMIM database. The list of gene entries with disease-causing mutations is obtained as described in Amberger et al. [2009]. Based on this list, a file containing all the selected human entry sequences in FASTA format is created.
2. Mining of heterogeneous databases (OMIM, UniProtKB, Genecards, GO, several LSDBs) through the Decrypthon Data Center, based on the protein entry list. The entire set of retrieved data associated with each disease-related protein is constituted during this process.
3. Construction of the SStEISy workbench. First, structural templates are determined based on similarity searches in the PDB and two MACS are constructed and annotated, including the template sequences. Second, a new wild-type 3D model is generated if the structural template differs from the one used in the previous version of SM2PH-db or if a new 3D template is available.
4. Multidimensional characterization of the mutants. This process initiates with the generation of mutant 3D models not available in the previous version of SM2PH-db (e.g., for novel proteins, newly generated wild-type 3D models, or new mutations). Physicochemical changes and structural modifications induced by the substitution as well as functional and structural features related to the mutated position are then listed.
5. Finally, the entire content of the SM2PH-db is upgraded, by integrating all this information. Particular attention has been paid to the automation and optimization of this workflow. We have optimized communication and synchronization between the processes by analyzing the data dependencies. To speed up the entire workflow, processes 2 and 3 are launched in parallel. Moreover, as SM2PH-db includes about 2,250 entry proteins, the update of all the alignments and their annotation represented a potential bottleneck in terms of computing time and memory. To overcome these limitations, we have implemented PipeAlign and MACSIMS on the Decrypthon computation grid and the required databases have been integrated in the Decrypthon Data Center. The construction and annotation of all SM2PH-db alignments takes around two days using the Decrypthon infrastructure, which is constituted of 120 computational nodes, whereas the same tasks took about 1 month to run on our internal server (four processor SUN Enterprise V40z). Currently, the complete SM2PH-db update procedure is launched every 2 months, guaranteeing that the user is working with up-to-date information.

Software Implementation

SM2PH-db is implemented in a relational database management system (PostgreSQL), has a schema with 26 tables and runs on a Linux server. It takes advantage of a molecular model database management system called Modeome3D [Garnier et al., in preparation]. The generation of the Web interface is programmed in Python and uses the AJAX (Asynchronous JavaScript and XML) methodology for dynamic updating of the pages.

SM2PH-db Web Interface

Database Search

SM2PH-db can be queried using textual search forms (Fig. 2A), based on a combination of keywords (e.g., protein name, disease name) or via a BLAST search [Altschut et al., 1997] of the database entries.

The search results consist of a summary of the data related to each protein that matches the required criteria (Fig. 2B). *Protein*

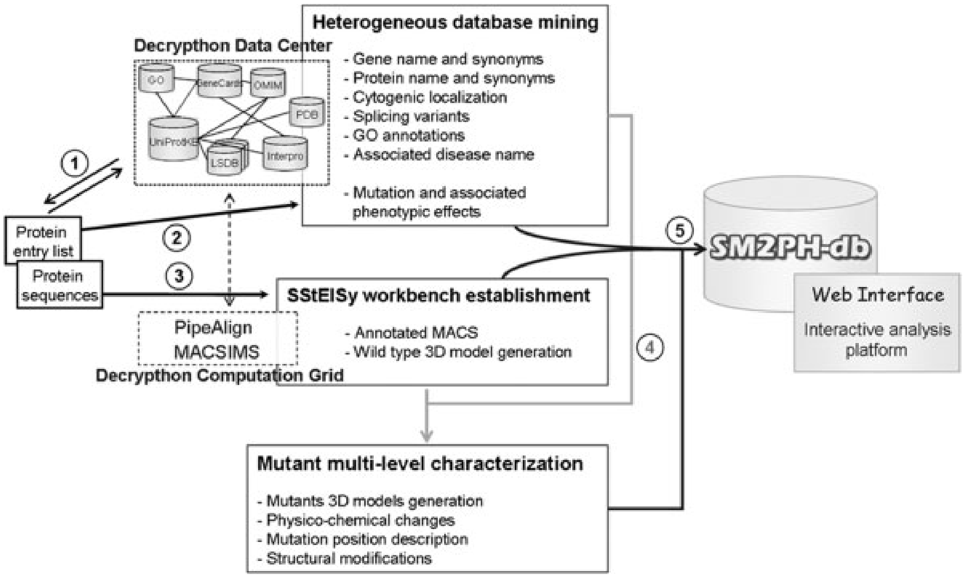
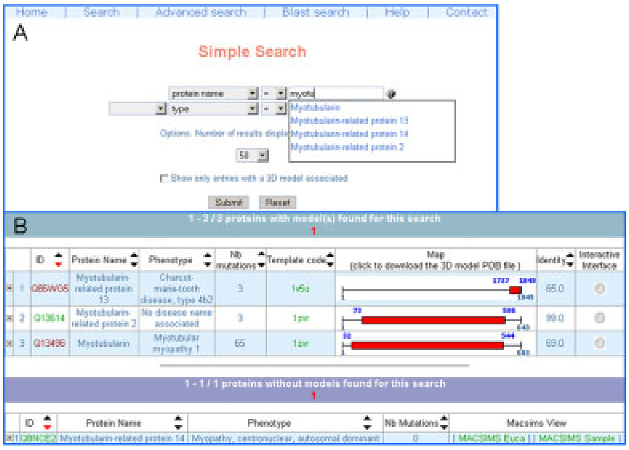

Heterogeneous database mining
Decrypthon Data Center
GO
GeneCards
OMIM
PDB
UniProtKB
LSDB
Interpro
- Gene name and synonyms
- Protein name and synonyms
- Cytogenic localization
- Splicing variants
- GO annotations
- Associated disease name
- Mutation and associated phenotypic effects
1
Protein entry list
2
Protein sequences
3
SStEISy workbench establishment
- Annotated MACS
- Wild type 3D model generation
PipeAlign
MACSIMS
Decrypthon Computation Grid
4
5
SM2PH-db
Web Interface
Interactive analysis platform
Mutant multi-level characterization
- Mutants 3D models generation
- Physico-chemical changes
- Mutation position description
- Structural modifications

Figure 1. Schematic representation of SM2PH-db automated workflow for data generation and integration: (1) protein entry list update; (2) heterogeneous data mining via the Decrypthon Data Center; (3) SStEISy workbench construction; (4) mutant multidimensional characterization; (5) SM2PH-db data integration.

details pages, that assemble all the retrieved data, can be accessed by clicking on the protein identifiers. When a 3D model is available, a diagram that schematizes the modeled region(s) of the sequence is presented and an interactive analysis interface can be accessed that allows the visualization and manipulation of the processed information.


Home | Search | Advanced search | Blast search | Help | Contact
A
B
Submit
Reset

Simple Search

protein name	=	myotu
type	=	

Myotubularin
Myotubularin-related protein 13
Myotubularin-related protein 14
Myotubularin-related protein 2

Options: Number of results disple 50

☐ Show only entries with a 3D model associated

1 - 3 / 3 proteins with model(s) found for this search
1

	ID	Protein Name	Phenotype	Nb mutations	Template code	Map (click to download the 3D model PDB file)	Identity	Interactive Interface
1	Q86WG5	Myotubularin-related protein 13	Charcot-marie-tooth disease, type 4b2	3	1z6a	1 1727 1848 1849	65.0	○
2	Q13614	Myotubularin-related protein 2	No disease name associated	3	1zsr	1 73 586 643	99.0	○
3	Q13496	Myotubularin	Myotubular myopathy 1	65	1zsr	1 32 544 603	69.0	○

1 - 1 / 1 proteins without models found for this search
1

	ID	Protein Name	Phenotype	Nb Mutations	Macsims View
1	Q8NCE2	Myotubularin-related protein 14	Myopathy, centronuclear, autosomal dominant	0	[MACSIMS Euca] [MACSIMS Sample]

Figure 2. SM2PH-db search interface. **A:** The textual search form allows querying of the database with a combination of keywords. A list of keywords matching the first letters entered by the user is automatically proposed to facilitate the search. **B:** The result page displays a table containing the protein identifier, its name, the associated disease name and information related to the 3D model if one has been constructed. Links to the entry protein details and to the interactive analysis interface are provided. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Protein Details Page

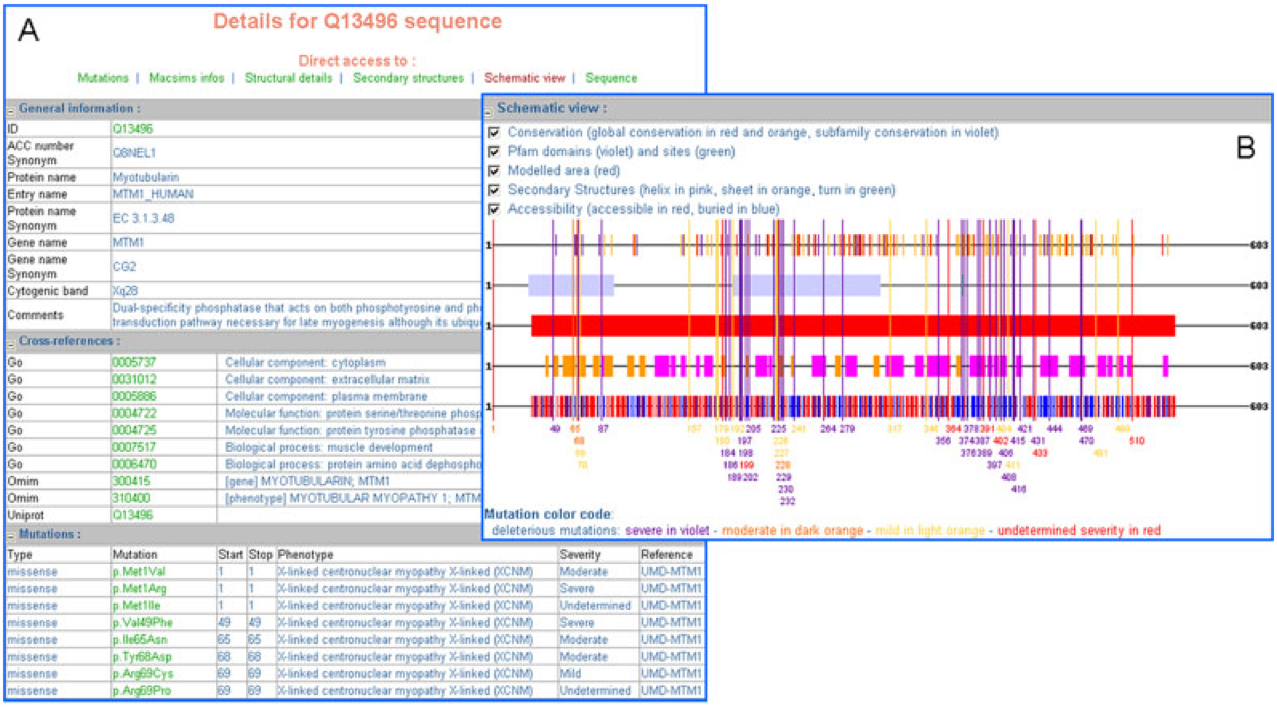
The *protein details* page can be considered to be a protein "identity card." It displays data retrieved from existing databases; links to external databases are provided and a visual summary of the processed information. The retrieved data are divided into three main sections: *General Information*, *Crossreferences*, and *Mutations* (Fig. 3A). The latter includes a list of missense mutations associated with a short description of their phenotypic consequences.

The processed information summary mainly concerns the functional domain annotations assigned to the protein by MACSIMS, details of the 3D model if one has been built and its secondary structure composition. The schematic view reveals all these features in a simple, linear representation (Fig. 3B).

Interactive Graphical Interface

The SM2PH-db interactive interface (Fig. 4A) is based on the MAGOS Web server [Garnier et al., 2006]. It has been upgraded to meet the SStEISy requirements and enhanced to allow integrative missense mutation analyses. This graphical interface is divided into three frames, each related to a produced information network:

1. Frame 1: the right-hand part of the interface shows the annotated MACS. The sequence of the human disease-related protein is always the first in the alignment and the modeled region is underlined. By default, one sequence of each subfamily of the MACS is displayed; subfamily sequences can be expanded with the "Uncollapse all" button. Functional and structural annotations such as active sites, PFAM domains, residues conservation during evolution, transmembrane regions, etc., are available through the *Features* menu: these are mapped on the MACS when selected.



Details for Q13496 sequence

Direct access to :
Mutations | Macsims infos | Structural details | Secondary structures | Schematic view | Sequence

General information :

ID	Q13496
ACC number Synonym	Q8NEL1
Protein name	Myotubularin
Entry name	MTM1_HUMAN
Protein name Synonym	EC 3.1.3.48
Gene name	MTM1
Gene name Synonym	CG2
Cytogenic band	Xq28
Comments	Dual-specificity phosphatase that acts on both phosphotyrosine and ph transduction pathway necessary for late myogenesis although its ubiqu

Cross-references :

Go	0005737	Cellular component: cytoplasm
Go	0031012	Cellular component: extracellular matrix
Go	0005886	Cellular component: plasma membrane
Go	0004722	Molecular function: protein serine/threonine phosp
Go	0004725	Molecular function: protein tyrosine phosphatase
Go	0007517	Biological process: muscle development
Go	0006470	Biological process: protein amino acid dephospho
Omim	300415	[gene] MYOTUBULARIN; MTM1
Omim	310400	[phenotype] MYOTUBULAR MYOPATHY 1; MTM
Uniprot	Q13496	

Mutations :

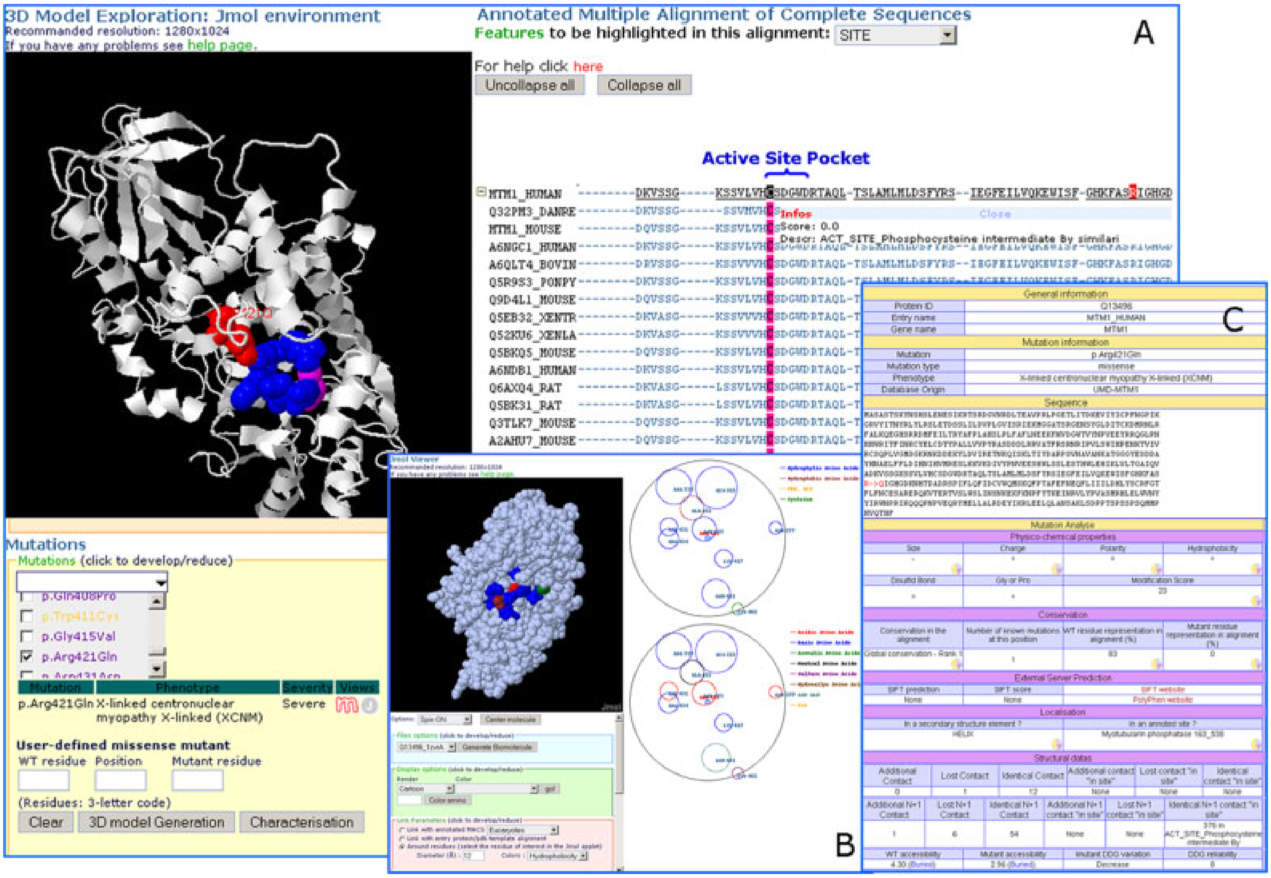
Type	Mutation	Start	Stop	Phenotype	Severity	Reference
missense	p.Met1Val	1	1	X-linked centronuclear myopathy X-linked (XCNM)	Moderate	UMD-MTM1
missense	p.Met1Arg	1	1	X-linked centronuclear myopathy X-linked (XCNM)	Severe	UMD-MTM1
missense	p.Met1Ile	1	1	X-linked centronuclear myopathy X-linked (XCNM)	Undetermined	UMD-MTM1
missense	p.Val49Phe	49	49	X-linked centronuclear myopathy X-linked (XCNM)	Severe	UMD-MTM1
missense	p.Ile65Asn	65	65	X-linked centronuclear myopathy X-linked (XCNM)	Moderate	UMD-MTM1
missense	p.Tyr68Asp	68	68	X-linked centronuclear myopathy X-linked (XCNM)	Moderate	UMD-MTM1
missense	p.Arg69Cys	69	69	X-linked centronuclear myopathy X-linked (XCNM)	Mild	UMD-MTM1
missense	p.Arg69Pro	69	69	X-linked centronuclear myopathy X-linked (XCNM)	Undetermined	UMD-MTM1

Schematic view :

☑ Conservation (global conservation in red and orange, subfamily conservation in violet)
☑ Pfam domains (violet) and sites (green)
☑ Modelled area (red)
☑ Secondary Structures (helix in pink, sheet in orange, turn in green)
☑ Accessibility (accessible in red, buried in blue)

Mutation color code:
deleterious mutations: severe in violet - moderate in dark orange - mild in light orange - undetermined severity in red

Figure 3. Screenshots of the Myotubularin details page. **A:** The retrieved data are displayed, divided into several sections. **B:** The schematic view associated with Myotubularin summarizes, in a linear representation, the main protein functional and structural features as well as the mutant positions. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]



3D Model Exploration: Jmol environment
Recommanded resolution: 1280x1024
If you have any problems see help page.

Annotated Multiple Alignment of Complete Sequences
Features to be highlighted in this alignment: SITE

For help click here

Uncollapse all | Collapse all

Active Site Pocket

```
MTM1_HUMAN  --------DKVSSG-----KSSVLVHCSDGWDRTAQL-TSLAMLMLDSFYRS--IEGFEILVQKEWISF-GHKFASRIGHGD
Q32PM3_DANRE --------DKVSSG------SSVMVHCS
MTM1_MOUSE  --------DQVSSG-----KSSVLVHCS
A6NGC1_HUMAN --------DKVSSG-----KSSVLVHC
A6QLT4_BOVIN --------DKVSSG-----KSSVVVHCSDGWDRTAQL-TSLAMLMLDSFYRS--IEGFEILVQKEWISF-GHKFASRIGHGD
Q5R9S3_PONPY --------DKVSSG-----KSSVLVHCSDGWDRTAQL-T
Q9D4L1_MOUSE --------DQVSSG-----KSSVLVHCSDGWDRTAQL-T
Q5EB32_XENTR --------DKVASG-----KSSVVVHCSDGWDRTAQL-T
Q52KU6_XENLA --------DKVASG-----KSSVVVHCSDGWDRTAQL-T
Q5BKQ5_MOUSE --------DQVSSG-----KSSVLVHCSDGWDRTAQL-T
A6NDB1_HUMAN --------DKVSSG-----KSSVLVHCSDGWDRTAQL-T
Q6AXQ4_RAT   --------DKVASG-----LSSVLVHCSDGWDRTAQL-T
Q5BK31_RAT   --------DKVASG-----LSSVLVHCSDGWDRTAQL-T
Q3TLK7_MOUSE --------DQVSSG-----KSSVLVHCSDGWDRTAQL-T
A2AHU7_MOUSE --------DQVSSG-----KSSVLVHCSDGWDRTAQL-T
```

Infos | Close
Score: 0.0
Descr: ACT_SITE_Phosphocysteine intermediate By similari

Mutations
Mutations (click to develop/reduce)

□ p.Gln408Pro
□ p.Trp411Cys
□ p.Gly415Val
☑ p.Arg421Gln
□ p.Arg421Asp

Mutation	Phenotype	Severity	Views
p.Arg421Gln	X-linked centronuclear myopathy X-linked (XCNM)	Severe	

User-defined missense mutant
WT residue: ____ Position: ____ Mutant residue: ____
(Residues: 3-letter code)

Clear | 3D model Generation | Characterisation

General information	
Protein ID	Q13496
Entry name	MTM1_HUMAN
Gene name	MTM1
Mutation information	
Mutation	p.Arg421Gln
Mutation type	missense
Phenotype	X-linked centronuclear myopathy X-linked (XCNM)
Database Origin	UMD-MTM1

Sequence

[illegible]

Mutation Analyse

Physico-chemical properties

Size	Charge	Polarity	Hydrophobicity
-	+	+	+

Disulfid Bond	Gly or Pro	Modification Score
+	+	23

Conservation

Conservation in the alignment	Number of known mutations at this position	WT residue representation in alignment (%)	Mutant residue representation in alignment (%)
Global conservation - Rank 1	1	83	0

External Server Prediction

SIFT prediction	SIFT score	SIFT website
None	None	PolyPhen website

Localisation

In a secondary structure element ?	In an annoted site ?
HELIX	Myotubularin phosphatase 163_538

Structural datas

Additional Contact	Lost Contact	Identical Contact	Additional contact "in site"	Lost contact "in site"	Identical contact "in site"
0	1	12	None	None	None

Additional N+1 Contact	Lost N+1 Contact	Identical N+1 Contact	Additional N+1 contact "in site"	Lost N+1 contact "in site"	Identical N+1 contact "in site"
1	6	54	None	None	376 in ACT_SITE_Phosphocysteine intermediate By

WT accessibility	Mutant accessibility	mutant DDG variation	DDG reliability
4.30 (Buried)	2.96 (Buried)	Decrease	8

Figure 4. SM2PH-db interactive graphical interface for Myotubularin. **A:** The dynamic Web page is divided into three interconnected frames: the right-hand frame gives access to the annotated MACS interconnected with the 3D model/structure displayed in the upper left frame. All the features mined and predicted by MACSIMS can be displayed both on the query sequence in the context of the MACS and on its structural representation. The lower left frame shows the missense mutations that can be positioned and analysed within the structural and functional context of the protein. **B:** Residues surrounding a position of interest can be highlighted with the *Around residues* option. The diameter of investigation can be modified by the user. **C:** The p.Arg421Gln mutation characterization page, which provides descriptions concerning the modifications induced by the substitution as well as information related to the conservation of the mutated residue, its position relative to functional features and within the 3D model, etc. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

2. Frame 2: the upper part of the left-hand side of the interface displays the 3D model in a Jmol environment (http://jmol.sourceforge.net/) and allows its in-depth exploration. The *Display options* section is dedicated to a basic manipulation of the 3D model by the intermediary of predefined rendering types (e.g., cartoon, spacefill, wireframe) and residue coloring (e.g., according to polarity, accessibility), whereas the *Console* section allows a high-level 3D model manipulation using the Jmol/Rasmol command line. The visualization of the residues positioned in a structural environment close to a residue of interest (i.e., within a distance defined by the user) can be performed with the *Around residues* option found in the *Link Parameters* section (Fig. 4B).
3. Frame 3: the *Mutations* section, in the lower part of the left-hand side of the page, is devoted to the missense mutation informational network. The missense mutations mapped on the 3D model can be accessed via the top menu of this section. A table is shown under mutation selection, in which the associated phenotypes and two links are furnished. The first link provides access to the mutation characterization page (Fig. 4C), that includes the 31 parameters describing mutations (i.e., physicochemical changes induced by the substitution, information related to the substituted position such as conservation, functional, and structural features). The second link allows the replacement of the wild-type 3D model by the mutant model in the Jmol window.

Although these frames represent suitable analysis tools on their own, the main advantage of SM2PH-db graphical interface relies on their interconnection, which allows high-level integrative missense mutation analysis to be performed. First, the disease-related protein sequence in the MACS is connected to its 3D model: the annotations assigned to the underlined part of the sequence (i.e., the modeled part) are simultaneously displayed on the 3D model using the same color code. The selection of a residue in the structure will also highlight this residue within the sequence and vice versa. Furthermore, the mutation frame is connected to the other two frames: the selection of a missense mutation leads to the simultaneously mapping of the wild-type residue within the wild-type 3D model and in the MACS. This allows the study of the mutated position in a SStEISy environment.

Thanks to this graphical interface, in addition to providing insights into the disease mechanisms, SM2PH-db represents an ideal workbench for an in-depth analysis of novel user's mutations through an interactive approach. Indeed, the user can study the substituted position in terms of structural/functional features, ask for the construction of a 3D model for his mutant of interest, and explore this model through the Jmol interface (*Generate* button in the *Mutations* section). Moreover, to support specific user investigations, the complete mutation characterization page, summarizing the main information related to this substitution and its position can be generated "on the fly."

SM2PH-db Statistics Page

General statistics related to the mutations stored in SM2PH-db can be accessed via the *Statistics* link, in the main menu. These statistics concern 10 parameters (out of the 31 computed) that characterize either the mutation substitutions or their substituted positions. For each of these parameters, two graphs representing the distribution of the associated values are provided: one for disease-causing mutations, the other for nonpathogenic mutations.

Each graph represents one single parameter and cannot be used in an independent manner to define a threshold to discriminate between disease-causing and nonpathogenic mutations. However, coupled with human expertise, these statistics could facilitate the interpretation of the molecular consequences of a given mutation. Therefore, these graphs can be viewed directly in a pop-up window from the mutation characterization pages, by clicking on the statistics icon located in the lower right corner of the concerned parameter cell.

Discussion

SM2PH-db has been organized to give the user an easy access to a wealth of information relevant to the study of genotype/phenotype relationships in the context of human genetic diseases. Our database regroups up-to-date heterogeneous interconnected information, ranging from sequence to structure and including evolutionary and functional features, thus providing a high-level SStEISy workbench. This workbench, in combination with the interactive analysis platform, allows the investigation of known missense mutation molecular impacts with regard to their phenotypic effects as well as the in-depth exploration of novel missense mutations to infer their potential molecular and phenotypic effects.

At time of writing (August 2009), a total of 27,884 missense mutations are recorded in the database, among which 20,252 (73%) are disease-causing and 7,632 (27%) are considered as nonpathogenic. 9,435 3D mutant models have been created and are available via the *Interactive graphical interface*, of which 7,863 are disease-causing and 1,572 nonpathogenic.

SM2PH-db: Insight into Mutation Effects

The molecular consequences of missense mutations are related to the functional and structural contexts of the affected position, as well as to the physicochemical characteristics of the substitution [Saunders and Baker, 2002; Terp et al., 2002]. All these types of information are represented in SM2PH-db for the stored missense mutations.

The integration capabilities of SM2PH-db can be illustrated by the analysis of the molecular consequences of a selected missense mutation. Here, we consider the p.Arg421Gln missense mutation that affects Myotubularin, which is associated with a severe deleterious phenotype.

Myotubularin can be searched by querying the *Protein name* field in the textual search form. This search indicates that four proteins whose names contain the term myotubularin are stored in the database, including three myotubularin-related proteins (Fig. 2). It should be noted that a 3D model of Myotubularin has been constructed with a template that shares 69% of identity, suggesting that this model is of good quality (Fig. 2B).

Before any further analysis, the protein details page should be visualized (Fig. 3). Here, the domain and active site localizations are of particular interest when trying to interpret mutation consequences. These can be viewed in the *Macsims infos* section, as well as in the linear schematic view of this entry, where structural and evolutionary information are also provided.

The interactive graphical interface can then be accessed via the *View in Jmol* link. The Arginine in position 421 can be mapped onto the 3D structure under selection in the *Mutations* section (Fig. 4A). A quick visual inspection shows that this residue is part of an alpha-helix and seems to be buried. The mutation characterization page (Fig. 4C) can then be accessed via the link provided in the table in the *Mutations* section. The physicochemical modifications linked to this substitution are not drastic: the modified score is 23 and the main change to be noted is the decrease of the residue size. The substituted position is located in the *Myotubularin phosphatase* domain and is well conserved during evolution: arginine represents more than 95% of this column in the alignment and has been classified as a rank 1 conserved position. In this example, an important feature can be observed, concerning the wild-type and mutated residue contacts: a residue that is in contact with our substituted residue is in direct contact with the Myotubularin active site, at position 375. This contact, although predicted as maintained, can obviously be modified by the substitution by a smaller residue. As a consequence, one might hypothesize that the p.Arg421Gln substitution destabilizes the active site pocket, which could explain the severe phenotype associated.

In conclusion, this example shows that the SM2PH-db infrastructure is suitable for in-depth investigations of mutations and can support the formulation of hypotheses related to the molecular consequences of known or newly discovered mutations.

Phenotypic Information

The availability of phenotypic information is central to SM2PH-db, because this information should help the scientist in the understanding of the disease pathogenesis. The retrieval of missense mutations from the UniProtKB/Swissprot database ensures the integration of phenotypic data for almost all the disease-related proteins, but our goal is to provide access to more precise information. Our system has consequently been organized to store information from LSDBs, excluding any patient data, although their integration has to be envisaged in a step-by-step manner because of the highly variable data formats. In the future, this task should become simpler, thanks to current standardization efforts in the context of the Gen2Phen European initiative (http://www.gen2phen.org) or the Human Variome Project [Ring et al., 2006], devised to collect and curate all genetic variation, its phenotypes and associated diseases, that will probably speed up the homogenization process. However, apart from the format aspects, another limiting factor for the integration of LSDB data is the reluctance of some clinicians to provide access to allelic variations and related information until after publication. To address this problem, we provide restricted access to private data on demand, based on the use of personal logins and passwords.

The Decrypthon Data Center is already capable of managing the UMD and LOVD formats and SM2PH-db stores data mined for example, from the UMD–MTM1 database. This database is dedicated to the Myotubularin protein (UniProt/Swiss-Prot: Q13496), involved in Myotubular Myopathy (MIM# 310400), and includes data related to 68 different missense variants, of which 61 have a corresponding three-level degree of severity.

We are currently contacting LSDB curators, seeking their consent to access and possibly diffuse their nonconfidential data, to provide users with precise genotypic/phenotypic information for a large number of disease-related proteins.

Conclusions and Perspectives

SM2PH-db represents an initial effort toward an effective and automated integration of mutation and phenotypic data involved in human genetic diseases. The sustainability of the database is guaranteed by the robust Decrypthon infrastructure on which it is based. Moreover, this latter also ensures that the biological data in SM2PH-db is always up to date.

In its current state, SM2PH-db and the numerous features offered by this novel system facilitates efficient study and interpretation of the molecular consequences of a given mutation.

In the future, a tool dedicated to the automated discrimination of disease-causing from nonpathogenic mutations will be integrated in our server. These predictions will be presented as propositions and complemented with all individual characterizing parameters, to give the user access to all the information required to judge the prediction accuracy. This complete system will hopefully contribute to the elucidation of the chain of events leading from a molecular defect to the pathology.

To achieve this, we intend to further enhance the available data by including, not only more detailed genotypic and phenotypic information, but also interactomic data such as functional and physical interactions mined from the STRING [Jensen et al., 2009] and IntAct [Kerrien et al., 2007] databases, as well as structural surface topology descriptions and interacting interface predictions [Albou et al., 2009]. We also plan to create *variant description* pages “on the fly” for missense variants of interest to the user. This will allow us to provide a simple access to customized information, which will aid the interpretation of the potential molecular effects of the mutation on the user’s protein.

Acknowledgments

The authors are grateful to Raymond Ripp for his assistance during this work and Julie Thompson for a critical reading of the manuscript. This work was funded by the French Association against Myopathy via the Decrypthon program (AFM12727-13836), EVI-GENORET LSHG-CT-2005-512036, EvolHHuPro ANR-07-BLAN-0054-02, as well as institute funds from the Institut National de la Santé et de la Recherche Médicale, the Centre National de la Recherche Scientifique, and the Université de Strasbourg.

References

Albou LP, Schwarz B, Poch O, Wurtz JM, Moras D. 2009. Defining and characterizing protein surface using alpha shapes. Proteins 76:1–12.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402.

Amberger J, Bocchini CA, Scott AF, Hamosh A. 2009. McKusick’s Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res 37:D793–D796.

Antonarakis SE, Beckmann JS. 2006. Mendelian disorders deserve more attention. Nat Rev Genet 7:277–282.

Antonarakis SE, Krawczak M, Cooper DN. 2000. Disease-causing mutations in the human genome. Eur J Pediatr 159(Suppl 3):S173–S178.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. Nucleic Acids Res 28:235–242.

Beroud C, Hamroun D, Collod-Beroud G, Boileau C, Soussi T, Claustres M. 2005. UMD (Universal Mutation Database): 2005 update. Hum Mutat 26:184–191.

Capriotti E, Fariselli P, Casadio R. 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 33:W306–W310.

Casadio R, Compiani M, Fariselli P, Vivarelli F. 1995. Predicting free energy contributions to the conformational stability of folded proteins from the residue sequence with radial basis function networks. Proc Int Conf Intell Syst Mol Biol 3:81–88.

Chasman D, Adams RM. 2001. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol 307:683–706.

Dayhoff MO, Eck RV, Park CM. 1972. A model of evolutionary change in proteins. In: Foundation NBR, editor. Atlas of protein sequence and structure. Washington, DC. p 89–99.

Eramian D, Eswar N, Shen MY, Sali A. 2008. How well can the accuracy of comparative protein structure models be predicted? Protein Sci 17:1881–1893.

Eswar N, Eramian D, Webb B, Shen MY, Sali A. 2008. Protein structure modeling with MODELLER. Methods Mol Biol 426:145–159.

Fokkema IF, den Dunnen JT, Taschner PE. 2005. LOVD: easy creation of a locus-specific sequence variation database using an “LSDB-in-a-box” approach. Hum Mutat 26:63–68.

French S, Robson B. 1983. What is a conservative substitution? J Mol Evolut 19:171–175.

Friedrich A, Ripp R, Garnier N, Bettler E, Deleage G, Poch O, Moulinier L. 2007. Blast sampling for structural and functional analyses. BMC Bioinformatics 8:62.

Galperin MY, Cochrane GR. 2009. Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009. Nucleic Acids Res 37:D1–D4.

Garnier N, Friedrich A, Bolze R, Bettler E, Moulinier L, Geourjon C, Thompson JD, Deleage G, Poch O. 2006. MAGOS: multiple alignment and modelling server. Bioinformatics 22:2164–2165.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, de la Cruz N, Tonellato P, Jaiswal P, Seigfried T, White R, Gene Ontology Consortium. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32:D258–D261.

Horaitis O, Cotton RG. 1999. 6th International HUGO Mutation Database Meeting, March 27, 1999, Brisbane, Australia. Hum Mutat 14:183–185.

Horaitis O, Talbot Jr CC, Phommarinh M, Phillips KM, Cotton RG. 2007. A database of locus-specific databases. Nat Genet 39:425.

Hurst JM, McMillan LE, Porter CT, Allen J, Fakorede A, Martin AC. 2009. The SAAPdb web resource: a large-scale structural analysis of mutant proteins. Hum Mutat 30:616–624.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature 431:931–945.

Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M, Bork P, von Mering C. 2009. STRING 8—a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37: D412–D416.

Jirtle RL, Skinner MK. 2007. Environmental epigenomics and disease susceptibility. Nat Rev Genet 8:253–262.

Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A. 2005. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics 21:2814–2820.

Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H. 2007. IntAct—open source resource for molecular interaction data. Nucleic Acids Res 35: D561–D565.

Kim BC, Kim WY, Park D, Chung WH, Shin KS, Bhak J. 2008. SNP@Promoter: a database of human SNPs (single nucleotide polymorphisms) within the putative promoter regions. BMC Bioinformatics 9(Suppl 1):S2.

Kono H, Yuasa T, Nishiue S, Yura K. 2008. coliSNP database server mapping nsSNPs on protein structures. Nucleic Acids Res 36:D409–D413.

Krawczak M, Thomas NS, Hundrieser B, Mort M, Wittig M, Hampe J, Cooper DN. 2007. Single base-pair substitutions in exon–intron junctions of human genes:

nature, distribution, and consequences for mRNA splicing. Hum Mutat 28:150–158.

Lecompte O, Thompson JD, Plewniak F, Thierry J, Poch O. 2001. Multiple alignment of complete sequences (MACS) in the post-genomic era. Gene 270:17–30.

Lockless SW, Ranganathan R. 1999. Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286:295–299.

McKusick VA. 2007. Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet 80:588–604.

Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31:3812–3814.

Nguyen H, Friedrich A, Berthommier G, Poidevin L, Moulinier L, Ripp R, Poch O. 2008. Introduction du nouveau Centre de Données Biomedicales Décrypthon. Tregastel: CORIA.

Philippi S, Kohler J. 2006. Addressing the problems with life-science databases for traditional uses and systems biology. Nat Rev Genet 7:482–488.

Plewniak F, Bianchetti L, Brelivet Y, Carles A, Chalmel F, Lecompte O, Mochel T, Moulinier L, Muller A, Muller J, Prigent V, Ripp R, Thierry JC, Thompson JD, Wicker N, Poch O. 2003. PipeAlign: a new toolkit for protein family analysis. Nucleic Acids Res 31:3829–3832.

Riikonen P, Vihinen M. 1999. MUTbase: maintenance and analysis of distributed mutation databases. Bioinformatics 15:852–859.

Ring HZ, Kwok PY, Cotton RG. 2006. Human Variome Project: an international collaboration to catalogue human genetic variation. Pharmacogenomics 7:969–972.

Safran M, Chalifa-Caspi V, Shmueli O, Olender T, Lapidot M, Rosen N, Shmoish M, Peter Y, Glusman G, Feldmesser E, Adato A, Peter I, Khen M, Atarot T, Groner Y, Lancet D. 2003. Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. Nucleic Acids Res 31:142–146.

Saunders CT, Baker D. 2002. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. J Mol Biol 322:891–901.

Singh A, Olowoyeye A, Baenziger PH, Dantzer J, Kann MG, Radivojac P, Heiland R, Mooney SD. 2008. MutDB: update on development of tools for the biochemical analysis of genetic variation. Nucleic Acids Res 36:D815–D819.

Sjolander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D. 1996. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. Comput Appl Biosci 12:327–345.

Sobolev V, Sorokine A, Prilusky J, Abola EE, Edelman M. 1999. Automated analysis of interatomic contacts in proteins. Bioinformatics 15:327–332.

Stenson PD, Ball E, Howells K, Phillips A, Mort M, Cooper DN. 2008. Human Gene Mutation Database: towards a comprehensive central mutation database. J Med Genet 45:124–126.

Stitziel NO, Binkowski TA, Tseng YY, Kasif S, 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.

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, Tyler-Smith C, Carter N, Scherer SW, Tavaré S, Deloukas P, Hurles ME, Dermitzakis ET. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315:848–853.

Tavtigian SV, Greenblatt MS, Lesueur F, Byrnes GB. 2008. In silico analysis of missense substitutions using sequence-alignment based methods. Hum Mutat 29:1327–1336.

Taylor WR. 1986. The classification of amino acid conservation. J Theor Biol 119:205–218.

Terp BN, Cooper DN, Christensen IT, Jorgensen FS, Bross P, Gregersen N, Krawczak M. 2002. Assessing the relative importance of the biophysical properties of amino acid substitutions associated with human genetic disease. Hum Mutat 20:98–109.

The International HapMap Consortium. 2003. The International HapMap Project. Nature 426:789–796.

Thompson JD, Holbrook SR, Katoh K, Koehl P, Moras D, Westhof E, Poch O. 2005. MAO: a Multiple Alignment Ontology for nucleic acid and protein sequences. Nucleic Acids Res 33:4164–4171.

Thompson JD, Muller A, Waterhouse A, Procter J, Barton GJ, Plewniak F, Poch O. 2006. MACSIMS: multiple alignment of complete sequences information management system. BMC Bioinformatics 7:318.

UniProt Consortium. 2008. The universal protein resource (UniProt). Nucleic Acids Res 36:D190–D195.

Wang Z, Moult J. 2001. SNPs, protein structure, and disease. Hum Mutat 17: 263–270.

Weatherall DJ. 1998. The phenotypic diversity of monogenic disease: lessons from the thalassemias. Harvey Lect 94:1–20.

Yip YL, Famiglietti M, Gos A, Duek PD, David FP, Gateau A, Bairoch A. 2008. Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. Hum Mutat 29:361–366.

Yue P, Melamud E, Moult J. 2006. SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics 7:166.