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.


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, as well as the effect 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.
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 [Ristonen 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 [Sztitzel 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 Decryptphon program (http://www.decryptphon. fr/english/), which provides access to a computational grid as well as data storage and federation facilities, via the Decryptphon 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://decryptphon.ibgmc.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 sequence, structure, and evolution are processed for each disease-related protein. Two MACS are computed 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 [Jolander 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 [Caprioni et al., 2005]: a positive ΔΔG value implies a protein stabilization 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 SSSTSy 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).

![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) SSSTSy workbench construction; (4) mutant multidimensional characterization; (5) SM2PH-db data integration.](image)
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.

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 SStELsy 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.
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 location 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 Decryption 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 diffusing 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 Decryption 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 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 Decryption 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


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