J Mol Model (2015) 21: 251
DOI 10.1007/s00894-015-2794-y

REVIEW

Computational approaches to study the effects of small genomic variations

Kamil Khafizov[1] · **Maxim V. Ivanov**[1] · **Olga V. Glazova**[1] · **Sergei P. Kovalenko**[1,2,3]

Received: 8 April 2015 / Accepted: 23 August 2015 / Published online: 8 September 2015

Abstract Advances in DNA sequencing technologies have led to an avalanche-like increase in the number of gene sequences deposited in public databases over the last decade as well as the detection of an enormous number of previously unseen nucleotide variants therein. Given the size and complex nature of the genome-wide sequence variation data, as well as the rate of data generation, experimental characterization of the disease association of each of these variations or their effects on protein structure/function would be costly, laborious, time-consuming, and essentially impossible. Thus, in silico methods to predict the functional effects of sequence variations are constantly being developed. In this review, we summarize the major computational approaches and tools that are aimed at the prediction of the functional effect of mutations, and describe the state-of-the-art databases that can be used to obtain information about mutation significance. We also discuss future directions in this highly competitive field.

Keywords Bioinformatics · Computational methods · Single-nucleotide polymorphism · Coding SNP · Mutation databases

✉ Kamil Khafizov
kkhafizov@gmail.com

[1] Moscow Institute of Physics and Technology, Dolgoprudny, Moscow Region, Russian Federation

[2] The Institute of Molecular Biology and Biophysics, Novosibirsk, Russian Federation

[3] Novosibirsk State University, Novosibirsk, Russian Federation

Abbreviations

TCGA	The Cancer Genome Atlas
ICGC	International Cancer Genome Consortium
SNP	Single nucleotide polymorphism
HGMD	Human Gene Mutation Database
sSNP	Nonsynonymous SNP
OMIM	Online Mendelian Inheritance in Man
HGV	Human Genome Variation
PMD	Protein Mutant Database
EVS	Exome Variant Server
COSMIC	Collection of somatic mutations in cancer
NCBI	National Center for Biotechnology Information
dbSNP	SNP Database
LSDB	Large number of locus-specific databases
HGVS	Human Genome Variation Society
MAF	Minor allele frequency
MSA	Multiple sequence alignment
PDB	Protein Data Bank
SS	Secondary structure
CAGI	Critical Assessment of Genome Interpretation

Introduction

The revolution in DNA sequencing technologies has resulted in an enormous increase in the number of publicly available gene sequences over the last several years, and these technologies continue to evolve [1, 2]. Next-generation sequencing projects not only generate millions of new gene sequences from a large variety of organisms but also huge numbers of previously unseen single nucleotide variants found therein [3]. Large ongoing initiatives, such as The Cancer Genome Atlas (TCGA [4]; Cancer Genome Atlas Network, http://cancergenome.nih.gov/), the International Cancer Genome Consortium (ICGC [5]; http://dcc.icgc.org), and the 1000

Genomes Project Consortium ([6]; http://www.1000genomes.org/), as well as the sequencing of genomes of patients affected by rare diseases [7–10], have produced comprehensive catalogs of mutations in the human genome. Single-nucleotide polymorphisms (SNPs) are by far the most common group of genetic variations found in humans. The International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/) has documented approximately 10 million common SNPs, defined as single base-pair substitutions with a minor allele frequency >1 % in (one or more) human populations. The 1000 Genomes Project Consortium has estimated that the difference between the genome of a randomly selected individual and the reference genome is at about 10,000 nonsynonymous (that lead to an amino acid substitution) sites, and about the same number of synonymous sites. In the Human Gene Mutation Database (HGMD, www.hgmd.org [11]), SNPs constitute more than half of all the disease-associated variations. Small indels (insertions and deletions) up to 20 base pairs (bp) long are the second largest class of mutations often associated with genetic diseases [11]. In the HGMD release 2015.2, deletions and insertions that cause inherited diseases account for 24,454 (27 %) and 10,617 (6 %), respectively, of all the mutations in the database, and combinations of short insertions and deletions add an extra 2,436 (1.4 %) variations. In protein coding regions, indels can be frameshifting (FS) or non-frameshifting (NFS) depending whether an indel inserts or deletes a multiple of three nucleotides. NFS indels thus do not alter the coding region but lead to an insertion or deletion of extra amino acids, which can affect the protein's function and, as a consequence, may also cause serious Mendelian diseases [12–16]. On the other hand, FS indels, having a length indivisible by three, shift the reading frame and alter the coding sequence downstream of the indel site, which also often leads to the introduction of a stop codon.

Given the size and complex nature of the genome-wide sequence variation data as well as the rate of data generation, experimentally characterizing the disease association of each of these variations and/or their effects on protein structure/function would be costly, laborious, time-consuming, and practically impossible. Thus, computational tools to predict the functional effects of sequence variations are constantly being developed. These tools have been used to investigate the impact of nonsynonymous SNPs (nsSNPs) and indels on protein function and to identify variations that are likely to be deleterious and those that are neutral. The ability to computationally discriminate between pathogenic and benign variants can significantly aid in targeting disease-causing mutations by helping in the selection and prioritization of the "best" candidates. Identifying and interpreting genomic variants associated with genetic diseases is an extremely important step which could pave the way for personalized medicine and diagnostics. For example, in silico mutation assessment can facilitate in vitro studies in reverse genetic projects in which mutations are introduced randomly into the genome of an experimental organism and in projects based on the random mutagenesis of genes of interest [17–19]. Prioritizing mutations before in vitro mutagenesis can increase the effectiveness of novel deleterious mutation searches. The tools that have been developed to predict mutation effects are quite general and can therefore be applied to any organisms, as was done, for example, in a large-scale TILLING reverse-genetics project that aimed to study mutations in *Arabidopsis*, zebrafish, maize, *Drosophila*, and *Lotus* [20]. Nonsynonymous variants are much better understood in terms of their biochemical and biophysical impact on the gene product than synonymous SNPs (i.e., SNPs that do not lead to changes in the protein sequences) and variations that occur in noncoding genomic regions; therefore, it is not surprising that most prediction tools have focused on nonsynonymous variants. However, some research (albeit to a significantly smaller extent) was also performed to investigate the impact of insertions and deletions on protein function. In this review, we summarize the major computational approaches and tools that aim to predict the functional effects of mutations. We also discuss future directions in this highly competitive field.

Databases

The diagnostics of Mendelian disorders consists mainly of the identification of genomic variations and their annotations. There are numerous publicly available mutation databases that can be used for this purpose [21]. Human genome variation databases also can serve as starting points for predicting the functional effects of mutations because they can provide both training and testing subsets for many of the in silico tools. The accuracy of prediction tools is highly dependent on the training sets used, which ideally should contain the disease-causing and neutral variations; therefore, the preparation of training datasets is a crucial step in the whole prediction process. Moreover, estimation of the accuracy of a prediction depends on the mutation sets used to test the tools. Further, the efficiency of a prediction tool is also influenced by the number of deleterious and benign mutations that can be collected from publically available databases. Deleterious nsSNPs are often collected from the Online Mendelian Inheritance in Man (OMIM) database [22], the HGMD [11], the dbSNP database [23, 24], and the Human Genome Variation (HGV) database [25]. In addition, mutations at the protein level can be collected from UniProtKB/Swiss-Prot [26, 27] and the Protein Mutant Database (PMD; [28]), though the PMD was last updated in 2007. Other mutation data sources are the 1000 Genome Project and the 6,500 Exomes Project

(Exome Variant Server; EVS; Seattle, WA; http://evs.gs.washington.edu/EVS/), with the former often being used to estimate mutation frequencies and serving as a source of neutral mutations, which are usually defined as mutations that have relatively high (>5 %) minor allele frequencies [29]. The major sources of somatic mutations related to cancers are TCGA (http://cancergenome.nih.gov/), the Catalogue of Somatic Mutations in Cancer (COSMIC; [30]; http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/), and the ICGC (https://dcc.icgc.org/), but these databases are not suitable for creating the training and testing sets because these mutations are of unknown significance. Although somatic mutation databases do not contain phenotype data, they can be used to estimate the efficiency of mutation effect prediction tools because the recurrence of somatic mutations has been correlated with their phenotypic effects [31]. Here we discuss the available mutation databases in more detail.

The 1000 Genomes Project (http://www.ncbi.nlm.nih.gov/variation/tools/1000genomes/) is an international research effort that was launched in 2008 to establish a detailed catalogue of human genetic variations. In 2012, the sequencing of 1,092 genomes from 14 populations was published [6]. The database contains almost 40 million genomic variants, which have been mapped to the human genome reference genome (version GRCh37/hg19). The database does not contain any phenotypic information but it can still be used to obtain important clues about which variants cause or are associated with genetic diseases; for instance, the observed polymorphism frequencies can be used to deduce whether a variation might be pathogenic or neutral (i.e., variations that indicate relatively high frequencies are not expected to cause any significant effect, otherwise they would be excluded by evolution).

The Exome Variant Server is another very popular source of genome variation frequency data. It focuses on rare genetic variations that may be associated with heart, lung, and blood disorders by sequencing the protein coding regions of the human genome across diverse populations. The EVS currently contains sequence information for the exomes of 6,503 individuals, and allele frequencies are provided for European-American and African-American populations. While individual genotype and phenotype data are not available on the EVS, these data can be accessed through the National Center for Biotechnology Information (NCBI) dbGAP database of genotypes and phenotypes [32]. An important feature of EVS is the inclusion of samples from patients with rare Mendelian disorders. Even researchers who are not interested specifically in heart, lung and blood disorders can use the database as a frequency filter in much the same way as the 1000 Genomes Project data are exploited.

The SNP Database (dbSNP; [23]) is a free public depository for genetic variations within and across different organisms which is developed and maintained by the NCBI in collaboration with the National Human Genome Research Institute. The dbSNP release 142 (16 October 2014) contains more than 88 million validated human refSNP clusters. The dbSNP accepts submissions for any organism from many sources including research laboratories, collaborative polymorphism discovery efforts, large-scale genome sequencing consortia, and other databases (e.g., the SNP consortium and HapMap [33]), and thereby serves as a major repository for genomic variations. The database does not, however, contain information about variant type, frequency, or causation, and variants present only in tumors can be included. In addition, the quality of the data in the dbSNP has been questioned and the database is known to contain a significant portion of false positives [34–36]. It is important to keep in mind that an entry in the dbSNP does not necessarily indicate disease causation, although some refSNP clusters include predicted clinical significance and associated scientific citations; therefore, the dbSNP has begun to offer limited phenotype information. Nevertheless, the dbSNP still serves as an important source of genome variation data and contains information from the 1000 Genomes and ESP projects.

One of the most notable attempts to gather together the disease-causing or disease-associated genome variations was the creation of the HGMD [37, 11], which collects published data from the scientific literature and has both free (public) and paid access options. HGMD is often used by researchers to annotate genomic variations found in sequencing experiments. The free public version of HGMD (http://www.hgmd.org) allows limited access for academic institutions and nonprofit organizations, while a subscription to HGMD Professional is required for full access. HGMD Professional provides pathogenicity predictions generated using several methods such as SIFT [38], PolyPhen-2 [39, 40], and MutPred [41]. The HGMD contains only published variants and often provides only the primary citation; thus, other mutation databases may provide more extensive reference lists.

A large number of locus-specific databases (LSDBs) have been developed and are typically curated by experts in the specific field. The Human Genome Variation Society (HGVS) maintains a list of available LSDBs (around 1600) on their website (http://www.hgvs.org/dblist/glsdb.html). General recommendations for the creation and curation of such databases have been proposed [42], and rules for the nomenclature of mutations have been discussed by den Dunnen and Antonarakis [43]. To improve uniformity between the databases, the HGVS has developed the Leiden Open Variation Database (LOVD; http://www.lovd.nl/3.0/home), a free tool for LSDB development that provides a consistent user interface [44]. Variant data can be quite extensive and include information about phenotype, functional data, family information, reported frequency, and references. One of the common problems with LSDBs is that

they are only as good as the information provided by submitters and curators, which varies.

Another popular resource is the Online Mendelian Inheritance in Man (OMIM [22]; http://www.omim.org), which was started in the 1960s by Victor A. McKusick as a physical catalog of Mendelian traits and disorders with or without known molecular etiology. Curation of the database is currently carried out at the Johns Hopkins University School of Medicine and it was transitioned to an online version in 1985. Variant information follows the textual summary and typically does not include precise HGVS nomenclature or genomic coordinates. For this reason, searching OMIM for specific variants can be challenging, and OMIM is often most useful for understanding the connection of a specific gene to a disease.

ClinVar ([45]; http://www.ncbi.nlm.nih.gov/clinvar/) was started in 2012 as a free public archive to store the links between genomic variations and phenotypes. The database was populated initially using variants from OMIM, GeneReviews (http://www.ncbi.nlm.nih.gov/books/NBK1116/), various LSDBs, and local databases provided by some scientific laboratories. Nevertheless, ClinVar now serves as a central repository for predictions of causation, with five standard categories that range from benign to pathogenic and additional categories that cover pharmacogenomics and complex traits. At the beginning of 2015, ClinVar contained more than 150,000 accessioned submissions. In addition to individual variant entries, ClinVar, like HGMD, provides gene information from external sources and links it to a wealth of related information in the database.

The UniProtKB/Swiss-Prot database (The UniProt Consortium; http://www.uniprot.org/ [46, 27]) contains high-quality manually curated protein sequence and functional information. Both manual curation and automatic annotation are used to add new information to the database entries, and cross-referencing to more than 120 external databases provides links to additional relevant information. Thus, the database contains a significant amount of information about the biological functions of proteins derived from the scientific literature. Human polymorphisms and disease-associated mutations from UniProt are collected into a single text file, Humsavar (http://www.uniprot.org/docs/humsavar), which contained about 25,000 disease-associated and about 38,000 common polymorphisms in the 4 February 2015 release. Based on information from the UniProtKB database, two training datasets (HumDiv and HumVar), both of which contain deleterious and benign mutations, were created by developers of the popular PolyPhen-2 program [40, 39]. HumDiv was compiled from all of the damaging alleles with known effects on molecular function (usually this means that they are associated with diseases) present in the UniProtKB database, plus likely neutral substitutions obtained from multiple sequence alignments of human proteins with closely related mammalian homologs. HumVar contains all human disease-causing mutations from UniProtKB, together with common human nsSNPs (MAF (minor allele frequency) >1 %) that have not been annotated as involved in disease and are therefore treated as nondamaging. Because of their simplicity of use, these datasets have been used to train and estimate the efficiency of mutation effect in a number of prediction tools [40, 47–49].

Methods

The molecular mechanisms that cause the dysfunction of mutagenic proteins vary significantly, and numerous methods to predict the possible effects of mutations have been developed. These methods can be classified into three main categories: (*i*) methods based on evolutionary (sequence conservation) information, (*ii*) methods based on the protein 3D structure, and (*iii*) methods based on various properties calculated directly from the amino acid sequence. The final goal of the majority of prediction tools is a binary (yes/no) classification of the variation as either pathogenic or benign. Most of these methods (see Fig. 1) first calculate a number of parameters related to changes caused by missense mutations. The resulting vector of the parameters is then used to obtain a "decision" score. Finally, a threshold is chosen such that possibly deleterious and benign mutations are separated in an optimal manner. Methods to transform the vector into a single score vary (see Table 1) and include machine learning techniques, such as neural networks (SNAP; [50]), random forests (MutPred, nsSNPAnalyzer [51, 41]), and support vector machines (PhD-SNP, SNPs&GO, SNPs&GO^{3d} [47, 52]), empirical rules (PolyPhen [53], MutationAssessor [54]), Bayesian methods (PolyPhen-2 [39]), and others (SIFT, Panther, MAPP; [55, 38, 56]). Machine learning has been shown to be a convenient tool for protein function predictions as it uses several criteria to infer output without explicitly assuming a pre-determined model [57]. However, supervised methods are highly dependent on the training datasets that are used, and to increase their reliability, large and accurate datasets are normally required. In many cases, the datasets contain mutations annotated as "pathogenic" and "neutral," but this information often conflicts with manual annotations made by experts. Here, we discuss the major approaches and methods used by prediction tools.

Sequence conservation methods

One of the approaches most commonly used to estimate the effects of mutations is sequence comparisons of many homologous proteins. These approaches employ the concept of natural selection, which states that organisms with malfunctioning proteins do not survive; therefore, mutations

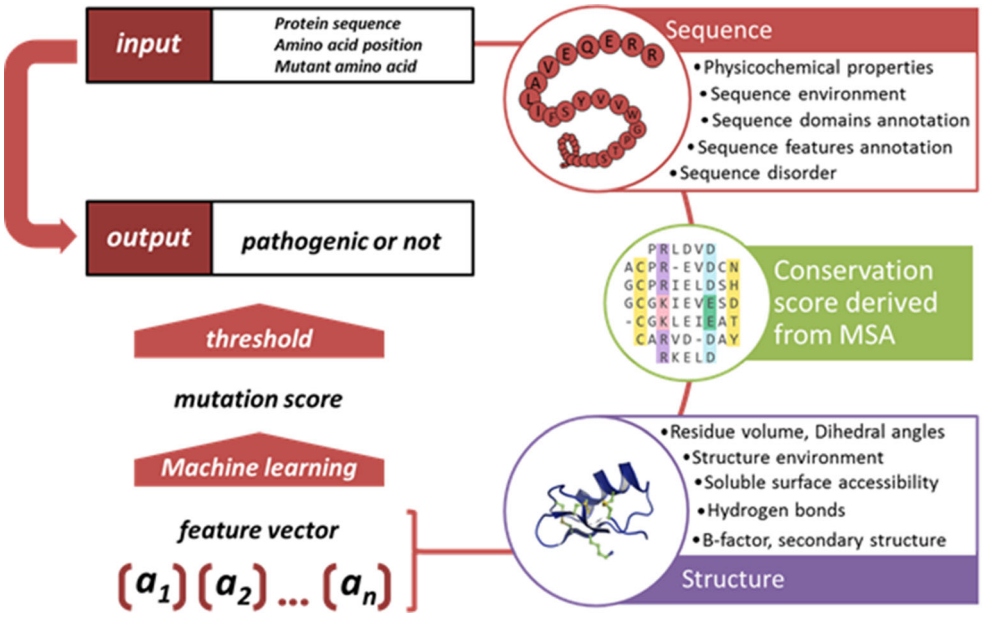

Fig. 1 General workflow employed by mutation effect prediction tools

that have a significant adverse impact are not normally detected in a healthy population. Thus, amino acid substitutions in highly conserved positions in the sequences are expected to be deleterious with a high probability, while amino acid substitutions in variable positions in the sequences are usually not so critical. Ideally, only true orthologs at high evolutionary distances should be compared to ensure that the selected homologous sequences have accumulated enough "neutral" mutations. Most of the tools that employ sequence conservation methods include paralogs in the total set of homologous sequences, which may confound the predictions of conserved residues only among the orthologs [56, 58]. Sequence comparisons are performed using multiple sequence alignments (MSAs) of homologous proteins. Numerous methods for MSA are available, but the alignment of distantly related proteins remains a challenging computational task. The commonly used progressive alignment methods do not guarantee an optimal solution and have had limited success in aligning divergent sequences [59]. While it is recommended [60] that many homologs should be included to build a reasonable MSA (usually 100 or more sequences), it is not always possible to collect enough high sequence identity orthologs, and the inclusion of sequences with low similarities introduces noise that lowers the confidence of predictions based on MSAs. Paralogous sequences are sometimes added to capture a sufficiently diverse sample of evolutionary variation; however, because paralogs may have different functions, this approach can reduce the sensitivity of the method. The selection of homologs is still a double-edged sword, and the way that a method handles the problem can influence its efficiency. Some tools take an MSA as input and play no part in selecting the homologs (MAPP, Align-GVGD [56, 61]), while others can either generate an MSA automatically or take one provided by the user (PolyPhen-2, PMut, SIFT [62, 38, 39]). To generate an MSA, basic automated tools search protein databases such as UniProt or NCBI nr [63, 64] using BLAST [65] programs with predefined search options (E-value, sequence identity) or employ information from the Pfam database of protein families [66]. The most widely used programs for MSAs include ClustalW, Clustal Omega [67], DIALIGN-T [68], MAFFT [69], MUSCLE [70], PROBCONS [71], and T-COFFEE [72]. Some tools combine several methods to generate a single output, such as the web-based, user-friendly M-Coffee [73], which allows users to select the alignment methods to combine. It has been reported that MSA tools are prone to errors, which can lead to incorrect phylogenies [74–76]. Significant efforts have been made to characterize the accuracy of the various MSA methods [59, 77–81], but no one program is ideal; therefore, it is advisable to use more than one method to obtain a reliable MSA, especially for the distantly related proteins.

Having constructed an MSA, heuristic models can be used to calculate positional conservation and estimate the possible effect of mutations. Most of the tools use amino acid substitution matrices such as BLOSUM, PAM, or PHAT combined with statistical models to perform this task. These matrices reflect the observed frequencies of amino acid substitutions in MSAs [82, 83]]. BLOSUM matrices (e.g., BLOSUM62 or BLOSUM45) are derived using different evolutionary distances in an MSA. BLOSUM62, for example, is built using amino acid sequences with similarities of <62 %. The average BLOSUM62 score is lower for pathogenic substitutions than

Table 1 Popular mutation effect prediction tools and the methods employed. [a] *AS* alignment score; [b] *NN* neural networks; [c] *HMM* hidden Markov models; [d] *RF* random forests; [e] *SVM* support vector machine; [f] *BC* Bayesian classification; [g] *DT* decision tree; [h] *ILP* inductive logic programming; [i] *nsSNP* nonsynonymous SNP; [j] *NFS indels* non-frameshifting insertions and deletions; [k] *FS indels* frameshifting insertions and deletions

Method	Aimed at	Based on	Conservation Analysis	Structure Attributes	Sequence Attributes
SIFT	nsSNP[i]	AS[a]			
PMut	nsSNP	NN[b]			
Panther	nsSNP	HMM[c]			
nsSNPAnalyzer	nsSNP	RF[d]			
MAPP	nsSNP	AS			
Align-GVGD	nsSNP	AS			
PhD-SNP	nsSNP	SVM[e]			
SNAP	nsSNP	NN			
Parepro	nsSNP	SVM			
SNP&GO	nsSNP	SVM			
MutPred	nsSNP	RF			
PolyPhen-2	nsSNP	BC[f]			
MutationTaster	nsSNP	BC			
SNP&GO 3d	nsSNP	SVM			
Mutation Assessor	nsSNP	AS			
Hansa	nsSNP	SVM			
PROVEAN	nsSNP, NFS indels[j]	AS			
FATHMM	nsSNP	HMM			
VEST3	nsSNP	RF			
CADD	nsSNP, NFS indels, FS indels[k]	SVM			
SIFT indel	FS indels	DT[g]			
KD4i	NFS indels	ILP[h]			
DDig-in	NFS indels FS indels	SVM			
PaPI	nsSNP, NFS indels, FS indels	RF			
HMMvar	NFS indels, FS indels[k]	HMM			

for neutral substitutions [84, 85]. For PAM matrices, the higher the matrix number, the more diverse the proteins used to build it (e.g., PAM-*X* means that *X* mutations per 100 amino acids may have happened). Many of the existing tools do not

calculate the conservation scores but use SIFT [51, 50] or Panther [86] scores instead, which generate reliable classifications based on the sequence conservation alone.

Structure-based methods

The 3D structure of a protein is the basis of its function. Numerous mutations involved in genetic diseases are known to affect the stability of protein structure [87–89]. Therefore, a number of protein structure stability analysis tools have been developed (reviewed in detail in [90]). Structural changes can alter protein–protein, protein–ligand, and protein–DNA interactions, which can also impact on the function of a protein. However, not much attention has been paid to such complexes, mainly because of the lack of an adequate amount of experimental data. The majority of structure stability prediction methods are based on supervised learning and are therefore highly dependent on the training datasets used. Databases with experimental thermodynamic parameters for both wild-type and mutant proteins, such as ProTherm [91], ProNIT [92], and more recently SKEMPI [93], which describes protein–protein complexes, have helped overcome the problem of data availability.

Protein structure stability prediction tools have implemented both statistical and physics-based free-energy potentials. A general equation that has been used to calculate protein free-energy changes as the result of a mutation is:

$$G = \alpha(\Delta V_{\mathrm{vdW}}) + \beta(\Delta V_{\mathrm{elec}}) + \gamma(\Delta SA) + \delta,$$

where α is the contribution of the van der Waals energy, β is the contribution of the electrostatic energy, and γ is the contribution of the nonpolar component of the solvation energy, which are calculated for both the wild-type and mutant protein structures. This formula can be extended to include contributions from hydrophobic groups, hydrogen bonds, steric overlaps between atoms in the structure, and entropy changes, as has been done in FoldX [94]. Other packages that use free-energy potentials to measure protein destabilization caused by a mutation include Eris [95], EGAD [96], TINKER [97], and Concoord [98]. Apart from statistical or biophysical methods to calculate free-energy potentials, an alternative approach is to represent the structure as a graph based on amino acid interactions [99, 100]. For instance, the Bongo method [99] predicts the structural effects of nsSNPs by evaluating graph-theoretic metrics and identifying key residues using a vertex cover algorithm. Distance patterns can be extracted from these graphs and summarized in a structural signature that can then be used as evidence to train predictive models. Da Silveira et al. [101] first reported the use of inter-residue distance patterns or signatures to define protein contacts and demonstrated that they were conserved across protein folds. The Cutoff Scanning Matrix is a protein structural signature [102] that has been used successfully in large-scale protein function predictions and structural classification tasks. Pires et al. [103] extended the inter-residue signature to the atomic level and successfully applied it in large-scale receptor-based ligand predictions.

Despite significant advances in methods to estimate protein stability changes caused by a mutation, they have rarely been implemented in missense variant classification tools. Although information about stability changes upon mutation could be useful in predicting the effect of a mutation on protein function, the degree of a stability change is likely to vary significantly among different proteins. Moreover, these algorithms return structural information that has to be processed by the user to make an informed assessment of pathogenicity. Thus, an in-depth understanding of these parameters is required for proper interpretation of the structural information. In addition, protein stability prediction tools have demonstrated a low correlation with high-quality experimental data [104]. Thus, methods based on the 3D-structural stability of proteins require further development before they can be widely used to predict the possible effect of protein structural variations on protein function.

PolyPhen-2 [40] is a widely used tool that predicts the impact of an amino acid change on the structure and function of a protein using simple geometric features, such as Ramachandran plots, changes in solvent-accessible surface areas, numbers of (side chain)–(side chain) and (main chain)–(side chain) H-bonds formed, distances to the closest atoms of other Protein Data Bank (PDB; www.rcsb.org; [105]) chains, and distances to important sites. Another popular tool that also employs protein 3D structure is SNP&GO^{3d} [47], which uses an amino acid environment derived from the protein structure, with further support vector machine implementation. Protein structures are usually obtained from the PDB, and homology modeling is used to predict the 3D structures of proteins that are not in the PDB. PolyPhen-2 can readily map the corresponding positions in an amino acid sequence to a selected PDB structure, which is a nontrivial task when dealing with significant amounts of sequence and structure data. Predictors based on structural features alone have been outcompeted [106]. Although the numbers of experimentally determined 3D structures have increased significantly over the last several years [2], one protein structure is often not enough to accurately determineprotein function. Posttranslational modifications are vital features of a protein that may be affected by sequence changes [107] but cannot be detected by structure methods. Nevertheless, the combination of protein structure predictions with amino acid sequence analysis can significantly increase the accuracy of structure–function prediction methods [108].

Sequence-based methods

Amino acid sequences can provide diverse information about protein properties. Physicochemical parameters of the residues in a sequence, including hydrophobicity [109–115], charge, polarity, dissociation constants for COOH– and NH_3– groups, molecular weight, and volume, are the most commonly used. By calculating changes in the physicochemical properties upon a mutation, it is possible to predict the potential effect of missense substitutions. The Grantham chemical difference matrix [116], which calculates differences in amino acid polarity, volume, and side-chain atomic composition (defined as the atomic weight ratio of hetero (non-carbon) elements in end groups or rings to carbons in the side chains), has been implemented in mutation effect prediction tools [61]. The average Grantham difference for pathogenic substitutions has been found to be higher than the average Grantham difference for neutral substitutions [117, 85, 118]. Apart from calculating physicochemical properties of amino acids, tools have been developed to predict secondary structure (SS), protein disorder segments, and transmembrane and signal domains based on an amino acid sequence. Predictions of the SS of proteins improved substantially in the 1990s and 2000s as evolutionary information from the diverse proteins belonging to the same structural family was included in the prediction tools. However, with the exponential growth in the numbers of X-ray protein structures, SS prediction tools were overshadowed by direct protein structure analysis programs, although the prediction of SS is still widely used because of its simplicity. Other less commonly used structural features that can be predicted based on an amino acid sequence alone include stability change [119, 120], solvent accessibility [121], coiled-coil structures [122], and B-factors [123]. Transmembrane and signal domain prediction tools are of particular interest because amino acid changes in these domains may imply a more severe restriction on differences in physicochemical properties between the wild-type and mutated amino acid. For example, changes in the hydrophilic or charged amino acids in a transmembrane domain may impact on the protein segments embedded in a membrane. Sequences are assumed to be disordered when they lack an ordered SS element. Disordered proteins, also known as "intrinsically disordered proteins" [124] or intrinsically unstructured proteins [125, 126], may be involved in protein–DNA binding, posttranslational modifications, phosphorylation, signaling, and regulation [127–129]. By analyzing the distribution of mutations within ordered and disordered segments of a sequence, Pajkos et al. [130] showed that cancer-associated mutations were more likely to occur in ordered regions, while neutral polymorphisms were more likely to occur in disordered segments. This finding clearly demonstrates the benefit of taking disorder levels into account when predicting the effects of amino acid mutations. Moreover, mutations themselves may introduce disorder into an ordered wild-type protein, causing incorrect folding. A large number of disorder predictors have been developed [131] that are based on protein amino acid composition, amino acid energy profiles and physicochemical properties, specific sequence patterns, as well as protein X-ray structures, and they are implemented in the protein MSA, SS prediction, and function annotation tools.

Additionally, specialized databases that focus on the annotation of the protein domain architecture, such as the popular Pfam database, have also been used for mutation effect assessment, since variations in functionally important regions have a higher probability of being pathogenic. Protein databases can also be used to derive posttranslational modifications, disulfide bonds, active sites, protein–protein interactions, and other sites that may be more susceptible to severe mutations because of their direct control of a protein's function. Since these features are defined by the protein sequences alone, there are methods to predict them based on amino acid composition [132–134, 123]. Some of these methods have been implemented in mutation effect prediction tools.

Understanding the effect of short insertions and deletions

Another important aspect to consider is short insertions and deletions. One significant limitation of the majority of in silico methods is that they usually aim at the single amino acid substitutions and do not deal with sequence variations such as small indels, even though these are known to account for almost a quarter of existing Mendelian disorders [37]. The majority of them are FS indels, the impact of which is expected to be pathological, and only two methods [135, 136] have been developed for them. At the same time, significantly more attention has been paid to studying the effect of NFS indels. However, development of the tools needed to accurately predict the effects of NFS indels is complicated due to a small number of exonic indels associated with human diseases (approximately 2,500 variants in the HGMD), especially the putatively neutral ones. The latter are usually obtained from the 1000 Genomes Project and account for approximately 1,500 variants [137]. Assuming the number of descriptors usually taken into account for machine learning procedures, training sets are usually not sufficient, as in the case of nsSNP. Nevertheless, several attempts to take into account this important class of genomic variations have been implemented over the last few years [60, 136–142]. These are also based on the descriptor collections and applications of classifiers as well as single-score MSA-based functions. For example, PROVEAN [60] introduced an alignment-based score as a measure to estimate the damaging effects of indels. Zhao et al. [137] developed a support-vector-machine-based method, DDIG-in, for predicting disease-causing NFS indels. Zhang et al. [143] developed a new classification method to distinguish deleterious and neutral NFS indels by combining

network and traditional sequential features of the environment surrounding the indels. Limongelli et al. [142] employed PolyPhen-2 and SIFT to predict the impact of indels via pseudo amino acid composition. In contrast to nsSNP effect prediction tools, these basically use two sequences as input instead of the sequence and point mutation. This makes it possible to judge not only the effects of indels but also those of the mutations, which can be useful in certain cases since multiple mutations may occur in the protein and predicting the effect of each of them can be confusing [140].

Limitations and future directions

Automated predictions of the possible impact of amino acid substitutions on protein function are of interest to researchers and medical doctors because of their importance in the annotation of the large genetic variation datasets that are being produced by modern sequencing technologies, as well as their application in identifying variants that cause rare Mendelian disorders [144] and for profiling the spectrum of variations uncovered by deep sequencing of large groups of individuals [145]. Although mutation effect prediction tools often can distinguish Mendelian-disorder-associated mutations from neutral ones, this is usually achieved by assigning a special score that should reflect the degree of function change. These approaches use datasets to train the methods. OMIM, HGMD, and Swiss-Prot are among the most widely used sources of deleterious mutations, and these databases have largely been created by utilizing data from the scientific literature. If a mutation is identified in a patient suffering from an inherited disease, it is often assumed that the mutation is associated with the disease and, unless other suspicious mutations are found, the disease is considered to be caused by a single genome sequence mutation. As a consequence of these assumptions, misinterpretations of the significance of a prediction often occur. First, even though a mutation may reduce the activity of a protein, it may be fixed in a certain population because of possible advantageous effects. The E6V substitution in human β-globin is a good example of this. The E6V mutation causes the hemoglobin aggregation that underlies sickle anemia and should be defined as deleterious [146]. However, the predictions for this mutation diverge among the tools, with the majority of them incorrectly predicting that this substitution is benign. The effect of this mutation is complicated by the fact that heterozygous carriers show resistance to malaria [147]. Second, apart from the case of overdominant selection, changes in protein function caused by environmental factors and complex diseases such as diabetes and hypertension are not generally considered in mutation effect prediction tools. Complex diseases are the consequences of numerous genetic variations as well as lifestyle and environmental factors and, importantly, it is often the case that not every mutation that contributes to a complex disease leads to the loss of protein activity. Thus, employing mutation effect prediction tools in complex disease applications is senseless.

Another limitation is that most of the existing methods for predicting the functional effects of mutations usually divide the variants into two categories: neutral (non-disease-associated) and deleterious (disease-associated). Applications in cancer genomics have revealed that two categories are not enough to functionally characterize single mutations [148]. Reva et. al. [54] have successfully realized the idea of categorizing mutations into five types: (*i*) "gain of function" mutations that convert genes into oncogenes, (*ii*) "loss of function" mutations that refer to pathogenic mutations, (*iii*) "drug resistant" mutations that overcome drug effects, (*iv*) "switch of function" mutations, which are intermediate between (*i*) and (*ii*), and (*v*) "neutral" mutations. Such categorization is supposed to be more realistic than the "two types" model because different variants usually have different impacts on a protein's function. Nevertheless, this idea of increasing the number of categories is yet to find wide application in Mendelian disease diagnostics.

Several possible directions have been explored to develop tools that allow more accurate predictions of the functional significance of mutations. These strategies include (*i*) the use of more relevant features that better describe the properties of substitutions (e.g., sequence or sequence-derived features, structures, annotations from databases, physicochemical properties, computed scores, families, and networks) [149]; (*ii*) selecting effective algorithms such as ensemble/multi-task-learning approaches to train classifiers with improved performance; (*iii*) collecting more insightful information from different sources and manually curating high-quality benchmark datasets; (*iv*) developing prediction methods for in-frame indels; (*v*) introducing new quantitative confidence scores that evaluate the quality of predictions; (*vi*) developing user-friendly tools or web servers that allow both single and batch queries; and (*vii*) creating pipelines that combine different tools that use different approaches (as has been done, for instance, in Condel [31], Meta-SNP [150], PredictSNP [151], and PON-P [152]).

As a consequence of codon degeneracy, not every DNA mutation results in an amino acid change. Such synonymous mutations are usually omitted in studies of inherited disease causing mutations. Although their significance and roles in trait development are unclear, there are several assumptions regarding the possible ways in which synonymous SNPs can initiate disease development. These include (*i*) alternative splicing [153], (*ii*) microRNA binding efficiency [154], and (*iii*) mRNA secondary structure differences resulting in a decrease in translation efficiency [155, 156] or protein misfolding caused by an altered translation speed [157]. Moreover, it was shown that some synonymous sites are highly conserved [158], indicating that conservation analysis at the

DNA level may also lead to new possibilities for mutation effect predictions.

It is also important to mention that the first Critical Assessment of Genome Interpretation (CAGI, http://genomeinterpretation.org/) experiment was organized in 2010. CAGI is an international effort to assess in silico tools for predicting the phenotypic impacts of genomic variations and to inform future research directions. In this experiment, the organizers provided participants with datasets of genetic variants to use for predictions of their functional impact. The predictions were evaluated against experimental characterizations by independent assessors. Each CAGI experiment culminates with a community workshop and publications to disseminate results. Three rounds of predictions (2010, 2011, 2012–2013) have now been carried out and have proven to be a great success [159]. Several available predictors performed well for disease and functional predictions. It is clear that CAGI will play a very important role in directing future efforts aimed at improving the quality of mutation effect prediction methods.

Acknowledgments This study was partially supported by RFBR, research project no. 15-04-04730, and grant no. RFMEFI60714X0098.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Levitt M (2009) Nature of the protein universe. Proc Natl Acad Sci USA 106(27):11079–11084. doi:10.1073/pnas.0905029106
2. Khafizov K, Madrid-Aliste C, Almo SC, Fiser A (2014) Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. Proc Natl Acad Sci USA 111(10): 3733–3738. doi:10.1073/pnas.1321614111
3. Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12(5):363–376. doi:10.1038/nrg2958
4. Giordano TJ (2014) The Cancer Genome Atlas research network: a sight to behold. Endocr Pathol 25(4):362–365. doi:10.1007/s12022-014-9345-4
5. The International Cancer Genome Consortium, Hudson T et al (2010) International network of cancer genome projects. Nature 464(7291):993–998. doi:10.1038/nature08987
6. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65. doi:10.1038/nature11632
7. Ng SB, Nickerson DA, Bamshad MJ, Shendure J (2010) Massively parallel sequencing and rare disease. Hum Mol Genet 19(R2):R119–R124. doi:10.1093/hmg/ddq390
8. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson DA, Shendure J, Bamshad MJ (2010) Exome sequencing identifies the cause of a Mendelian disorder. Nat Genet 42(1):30–35. doi:10.1038/ng.499
9. Thomas PD, Kejariwal A (2004) Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. Proc Natl Acad Sci USA 101(43):15398–15403. doi:10.1073/pnas.0404380101
10. Boycott KM, Vanstone MR, Bulman DE, MacKenzie AE (2013) Rare-disease genetics in the era of next-generation sequencing: discovery to translation. Nat Rev Genet 14(10):681–691. doi:10.1038/nrg3555
11. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN (2014) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 133(1):1–9. doi:10.1007/s00439-013-1358-4
12. Bi XH, Lu CM, Liu Q, Zhang ZX, Zhao HL, Yu J, Zhang JW (2012) A 14 bp indel variation in the NCX1 gene modulates the age at onset in late-onset Alzheimer's disease. J Neural Transm 119(3):383–386. doi:10.1007/s00702-011-0696-4
13. Dong B, Chen J, Zhang X, Pan Z, Bai F, Li Y (2013) Two novel PRP31 premessenger ribonucleic acid processing factor 31 homolog mutations including a complex insertion-deletion identified in Chinese families with retinitis pigmentosa. Mol Vis 19:2426–2435
14. Yu Q, Zhou C, Wang J, Chen L, Zheng S, Zhang J (2013) A functional insertion/deletion polymorphism in the promoter of PDCD6IP is associated with the susceptibility of hepatocellular carcinoma in a Chinese population. DNA Cell Biol 32(8):451–457. doi:10.1089/dna.2013.2061
15. Glanzmann B, Lombard D, Carr J, Bardien S (2014) Screening of two indel polymorphisms in the 5′UTR of the DJ-1 gene in South African Parkinson's disease patients. J Neural Transm 121(2): 135–138. doi:10.1007/s00702-013-1094-x
16. Ross JS, Wang K, Al-Rohil RN, Nazeer T, Sheehan CE, Otto GA, He J, Palmer G, Yelensky R, Lipson D, Ali S, Balasubramanian S, Curran JA, Garcia L, Mahoney K, Downing SR, Hawryluk M, Miller VA, Stephens PJ (2014) Advanced urothelial carcinoma: next-generation sequencing reveals diverse genomic alterations and targets of therapy. Mod Pathol: Off J US Can Acad Pathol Inc 27(2):271–280. doi:10.1038/modpathol.2013.135
17. Wrobel JA, Chao SF, Conrad MJ, Merker JD, Swanstrom R, Pielak GJ, Hutchison CA 3rd (1998) A genetic approach for identifying critical residues in the fingers and palm subdomains of HIV-1 reverse transcriptase. Proc Natl Acad Sci USA 95(2): 638–645
18. Zwick ME, Cutler DJ, Chakravarti A (2000) Patterns of genetic variation in Mendelian and complex traits. Annu Rev Genomics Hum Genet 1:387–407. doi:10.1146/annurev.genom.1.1.387
19. Hainaut P, Hernandez T, Robinson A, Rodriguez-Tome P, Flores T, Hollstein M, Harris CC, Montesano R (1998) IARC database of p53 gene mutations in human tumors and cell lines: updated compilation, revised formats and new visualisation tools. Nucleic Acids Res 26(1):205–213
20. Henikoff S, Comai L (2003) Single-nucleotide mutations for plant functional genomics. Annu Rev Plant Biol 54:375–401. doi:10.1146/annurev.arplant.54.031902.135009
21. Johnston JJ, Biesecker LG (2013) Databases of genomic variation and phenotypes: existing resources and future needs. Hum Mol Genet 22(R1):R27–R31. doi:10.1093/hmg/ddt384
22. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33:D514–D517. doi:10.1093/nar/gki033
23. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29(1):308–311

24. Smigielski EM, Sirotkin K, Ward M, Sherry ST (2000) dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Res 28(1):352–355
25. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW (2014) The Database of Genomic Variants: a curated collection of structural variation in the human genome. Nucleic Acids Res 42:D986–D992. doi:10.1093/nar/gkt958
26. UniProt Consortium (2008) The Universal Protein Resource (UniProt). Nucleic Acids Res 36:D190–D195. doi:10.1093/nar/gkm895
27. UniProt Consortium (2015) UniProt: a hub for protein information. Nucleic Acids Res 43:D204–D212. doi:10.1093/nar/gku989
28. Kawabata T, Ota M, Nishikawa K (1999) The Protein Mutant Database. Nucleic Acids Res 27(1):355–357
29. Thusberg J, Olatubosun A, Vihinen M (2011) Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat 32(4):358–368. doi:10.1002/humu.21445
30. Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S, Kok CY, Jia M, De T, Teague JW, Stratton MR, McDermott U, Campbell PJ (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 43:D805–D811. doi:10.1093/nar/gku1075
31. Gonzalez-Perez A, Lopez-Bigas N (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. Am J Hum Genet 88(4):440–449. doi:10.1016/j.ajhg.2011.03.004
32. Tryka KA, Hao L, Sturcke A, Jin Y, Wang ZY, Ziyabari L, Lee M, Popova N, Sharopova N, Kimura M, Feolo M (2014) NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Res 42:D975–D979. doi:10.1093/nar/gkt1211
33. International HapMap Consortium, Frazer KA et al (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449(7164):851–861. doi:10.1038/nature06258
34. Reich DE, Gabriel SB, Altshuler D (2003) Quality and completeness of SNP databases. Nat Genet 33(4):457–458. doi:10.1038/ng1133
35. Mitchell AA, Zwick ME, Chakravarti A, Cutler DJ (2004) Discrepancies in dbSNP confirmation rates and allele frequency distributions from varying genotyping error rates and patterns. Bioinformatics 20(7):1022–1032. doi:10.1093/bioinformatics/bth034
36. Musumeci L, Arthur JW, Cheung FS, Hoque A, Lippman S, Reichardt JK (2010) Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. Hum Mutat 31(1):67–73. doi:10.1002/humu.21137
37. Stenson PD, Ball EV, Mort M, Phillips AD, Shaw K, Cooper DN (2012) The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. Curr Protoc Bioinformatics Chapter 1:Unit 1.13. doi:10.1002/0471250953.bi0113s39
38. Ng PC, Henikoff S (2003) SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 31(13):3812–3814
39. Adzhubei I, Jordan DM, Sunyaev SR (2013) Predicting functional effect of human missense mutations using PolyPhen-2. Curr Protoc Hum Genet Chapter 7:Unit 7.20. doi:10.1002/0471142905.hg0720s76
40. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010) A method and server for predicting damaging missense mutations. Nat Methods 7(4):248–249. doi:10.1038/nmeth0410-248
41. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P (2009) Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25(21):2744–2750. doi:10.1093/bioinformatics/btp528
42. Cotton RG, Auerbach AD, Beckmann JS, Blumenfeld OO, Brookes AJ, Brown AF, Carrera P, Cox DW, Gottlieb B, Greenblatt MS, Hilbert P, Lehvaslaiho H, Liang P, Marsh S, Nebert DW, Povey S, Rossetti S, Scriver CR, Summar M, Tolan DR, Verma IC, Vihinen M, den Dunnen JT (2008) Recommendations for locus-specific databases and their curation. Hum Mutat 29(1):2–5. doi:10.1002/humu.20650
43. den Dunnen JT, Antonarakis SE (2000) Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat 15(1):7–12. doi:10.1002/(SICI)1098-1004(200001)15:1<7::AID-HUMU4>3.0.CO;2-N
44. Fokkema IF, Taschner PE, Schaafsma GC, Celli J, Laros JF, den Dunnen JT (2011) LOVD v. 2.0: the next generation in gene variant databases. Hum Mutat 32(5):557–563. doi:10.1002/humu.21438
45. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR (2014) ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42:D980–D985. doi:10.1093/nar/gkt1113
46. 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(3):361–366. doi:10.1002/humu.20671
47. Capriotti E, Calabrese R, Casadio R (2006) Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 22(22):2729–2734. doi:10.1093/bioinformatics/btl423
48. Tian J, Wu N, Guo X, Guo J, Zhang J, Fan Y (2007) Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines. BMC Bioinformatics 8:450. doi:10.1186/1471-2105-8-450
49. Hicks S, Wheeler DA, Plon SE, Kimmel M (2011) Prediction of missense mutation functionality depends on both the algorithm and sequence alignment employed. Hum Mutat 32(6):661–668. doi:10.1002/humu.21490
50. Bromberg Y, Rost B (2007) SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 35(11):3823–3835. doi:10.1093/nar/gkm238
51. Bao L, Zhou M, Cui Y (2005) nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res 33:W480–W482. doi:10.1093/nar/gki372
52. Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 30(8):1237–1244. doi:10.1002/humu.21047
53. Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30(17):3894–3900
54. Reva B, Antipin Y, Sander C (2011) Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 39:e118. doi:10.1093/nar/gkr407
55. Mi H, Guo N, Kejariwal A, Thomas PD (2007) PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. Nucleic Acids Res 35:D247–D252. doi:10.1093/nar/gkl869
56. Stone EA, Sidow A (2005) Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res 15(7):978–986. doi:10.1101/gr.3804205
57. Larranaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armananzas R, Santafe G, Perez A, Robles V (2006) Machine learning in bioinformatics. Brief Bioinform 7(1):86–112
58. Ng PC, Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function. Annu Rev Genomics Hum Genet 7:61–80. doi:10.1146/annurev.genom.7.080505.115630

59. Pervez MT, Babar ME, Nadeem A, Aslam M, Awan AR, Aslam N, Hussain T, Naveed N, Qadri S, Waheed U, Shoaib M (2014) Evaluating the accuracy and efficiency of multiple sequence alignment methods. Evol Bioinformatics Online 10:205–217. doi:10.4137/EBO.S19199
60. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP (2012) Predicting the functional effect of amino acid substitutions and indels. PLoS One 7:e46688. doi:10.1371/journal.pone.0046688
61. Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, Samollow PB, de Silva D, Zharkikh A, Thomas A (2006) Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. J Med Genet 43(4):295–305. doi:10.1136/jmg.2005.033878
62. Ferrer-Costa C, Gelpi JL, Zamakola L, Parraga I, de la Cruz X, Orozco M (2005) PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics 21(14):3176–3178. doi:10.1093/bioinformatics/bti486
63. Pruitt KD, Tatusova T, Maglott DR (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 33: D501–D504. doi:10.1093/nar/gki025
64. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35:D61–D65. doi:10.1093/nar/gkl842
65. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410. doi:10.1016/S0022-2836(05)80360-2
66. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer EL, Eddy SR, Bateman A, Finn RD (2012) The Pfam protein families database. Nucleic Acids Res 40:D290–D301. doi:10.1093/nar/gkr1065
67. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7: 539. doi:10.1038/msb.2011.75
68. Subramanian AR, Weyer-Menkhoff J, Kaufmann M, Morgenstern B (2005) DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. BMC Bioinformatics 6:66. doi:10.1186/1471-2105-6-66
69. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30(4):772–780. doi:10.1093/molbev/mst010
70. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5): 1792–1797. doi:10.1093/nar/gkh340
71. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res 15(2):330–340. doi:10.1101/gr.2821705
72. Notredame C, Higgins DG, Heringa J (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 302(1):205–217. doi:10.1006/jmbi.2000.4042
73. Wallace IM, O'Sullivan O, Higgins DG, Notredame C (2006) M-Coffee: combining multiple sequence alignment methods with T-Coffee. Nucleic Acids Res 34(6):1692–1699. doi:10.1093/nar/gkl091
74. Kim J, Ma J (2011) PSAR: measuring multiple sequence alignment reliability by probabilistic sampling. Nucleic Acids Res 39(15):6359–6368. doi:10.1093/nar/gkr334
75. Martin W, Roettger M, Lockhart PJ (2007) A reality check for alignments and trees. Trends Genet 23(10):478–480. doi:10.1016/j.tig.2007.08.007
76. Loytynoja A, Goldman N (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320(5883):1632–1635. doi:10.1126/science.1158395
77. Pais FS, Ruy Pde C, Oliveira G, Coimbra RS (2014) Assessing the efficiency of multiple sequence alignment programs. Algorithms Mol Biol 9(1):4. doi:10.1186/1748-7188-9-4
78. Ahola V, Aittokallio T, Vihinen M, Uusipaikka E (2006) A statistical score for assessing the quality of multiple sequence alignments. BMC Bioinformatics 7:484. doi:10.1186/1471-2105-7-484
79. Golubchik T, Wise MJ, Easteal S, Jermiin LS (2007) Mind the gaps: evidence of bias in estimates of multiple sequence alignments. Mol Biol Evol 24(11):2433–2442. doi:10.1093/molbev/msm176
80. Nuin PA, Wang Z, Tillier ER (2006) The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics 7:471. doi:10.1186/1471-2105-7-471
81. Raghava GP, Searle SM, Audley PC, Barber JD, Barton GJ (2003) OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. BMC Bioinformatics 4:47. doi:10.1186/1471-2105-4-47
82. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89(22):10915–10919
83. Dayhoff MOSRM (1978) A model of evolutionary change in proteins. Atlas Protein Seq Structure 5:345–351
84. Ferrer-Costa C, Orozco M, de la Cruz X (2002) Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. J Mol Biol 315(4):771–786. doi:10.1006/jmbi.2001.5255
85. Balasubramanian S, Xia Y, Freinkman E, Gerstein M (2005) Sequence variation in G-protein-coupled receptors: analysis of single nucleotide polymorphisms. Nucleic Acids Res 33(5): 1710–1721. doi:10.1093/nar/gki311
86. Brunham LR, Singaraja RR, Pape TD, Kejariwal A, Thomas PD, Hayden MR (2005) Accurate prediction of the functional significance of single nucleotide polymorphisms and mutations in the ABCA1 gene. PLoS Genet 1(6):e83. doi:10.1371/journal.pgen.0010083
87. Bross P, Corydon TJ, Andresen BS, Jorgensen MM, Bolund L, Gregersen N (1999) Protein misfolding and degradation in genetic diseases. Hum Mutat 14(3):186–198. doi:10.1002/(SICI)1098-1004(1999)14:3<186::AID-HUMU2>3.0.CO;2-J
88. Wang Z, Moult J (2001) SNPs, protein structure, and disease. Hum Mutat 17(4):263–270. doi:10.1002/humu.22
89. Yue P, Melamud E, Moult J (2006) SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics 7:166. doi:10.1186/1471-2105-7-166
90. Kucukkal TG, Yang Y, Chapman SC, Cao W, Alexov E (2014) Computational and experimental approaches to reveal the effects of single nucleotide polymorphisms with respect to disease diagnostics. Int J Mol Sci 15(6):9670–9717. doi:10.3390/ijms15069670
91. Gromiha MM, Uedaira H, An J, Selvaraj S, Prabakaran P, Sarai A (2002) ProTherm, thermodynamic database for proteins and mutants: developments in version 3.0. Nucleic Acids Res 30(1):301–302
92. Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, Uedaira H, Sarai A (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein–nucleic acid interactions. Nucleic Acids Res 34:D204–D206. doi:10.1093/nar/gkj103
93. Moal IH, Fernandez-Recio J (2012) SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. Bioinformatics 28(20):2600–2607. doi:10.1093/bioinformatics/bts489

94. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L (2005) The FoldX web server: an online force field. Nucleic Acids Res 33:W382–W388. doi:10.1093/nar/gki387
95. Yin S, Ding F, Dokholyan NV (2007) Eris: an automated estimator of protein stability. Nat Methods 4(6):466–467. doi:10.1038/nmeth0607-466
96. Pokala N, Handel TM (2005) Energy functions for protein design: adjustment with protein-protein complex affinities, models for the unfolded state, and negative design of solubility and specificity. J Mol Biol 347(1):203–227. doi:10.1016/j.jmb.2004.12.019
97. Pappu RV, Hart RK, Ponder JW (1998) Analysis and application of potential energy smoothing and search methods for global optimization. J Phys Chem B 102(48):9725–9742. doi:10.1021/Jp982255t
98. deGroot BL, vanAalten DMF, Scheek RM, Amadei A, Vriend G, Berendsen HJC (1997) Prediction of protein conformational freedom from distance constraints. Proteins 29(2):240–251. doi:10.1002/(Sici)1097-0134(199710)29:2<240::Aid-Prot11>3.0.Co;2-O
99. Cheng TMK, Lu YE, Vendruscolo M, Lio P, Blundell TL (2008) Prediction by graph theoretic measures of structural effects in proteins arising from non-synonymous single nucleotide polymorphisms. PLoS Comp Biol 4(7):e1000135. doi:10.1371/journal.pcbi.1000135
100. Pires DEV, Ascher DB, Blundell TL (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 30(3):335–342. doi:10.1093/bioinformatics/btt691
101. da Silveira CH, Pires DEV, Minardi RC, Ribeiro C, Veloso CJM, Lopes JCD, Meira W, Neshich G, Ramos CHI, Habesch R, Santoro MM (2009) Protein cutoff scanning: a comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. Proteins 74(3):727–743. doi:10.1002/Prot.22187
102. Pires DE, de Melo-Minardi RC, dos Santos MA, da Silveira CH, Santoro MM, Meira W Jr (2011) Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. BMC Genomics 12(Suppl 4):S12. doi:10.1186/1471-2164-12-S4-S12
103. Pires DE, de Melo-Minardi RC, da Silveira CH, Campos FF, Meira W Jr (2013) aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. Bioinformatics 29(7):855–861. doi:10.1093/bioinformatics/btt058
104. Potapov V, Cohen M, Schreiber G (2009) Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. Protein Eng Des Sel 22(9):553–560. doi:10.1093/protein/gzp030
105. 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(1):235–242
106. Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z (2013) Assessment of computational methods for predicting the effects of missense mutations in human cancers. BMC Genomics 14(Suppl 3):S7. doi:10.1186/1471-2164-14-S3-S7
107. Gnad F, Ren S, Choudhary C, Cox J, Mann M (2010) Predicting post-translational lysine acetylation using support vector machines. Bioinformatics 26(13):1666–1668. doi:10.1093/bioinformatics/btq260
108. Saunders CT, Baker D (2002) Evaluation of structural and evolutionary contributions to deleterious mutation prediction. J Mol Biol 322(4):891–901
109. Eisenberg D, Weiss RM, Terwilliger TC (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. Proc Natl Acad Sci USA 81(1):140–144
110. Engelman DM, Steitz TA, Goldman A (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu Rev Biophys Biophys Chem 15:321–353. doi:10.1146/annurev.bb.15.060186.001541
111. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157(1):105–132
112. Wimley WC, White SH (1996) Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat Struct Biol 3(10):842–848
113. Hessa T, Kim H, Bihlmaier K, Lundin C, Boekel J, Andersson H, Nilsson I, White SH, von Heijne G (2005) Recognition of transmembrane helices by the endoplasmic reticulum translocon. Nature 433(7024):377–381. doi:10.1038/nature03216
114. Hopp TP, Woods KR (1981) Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci USA 78(6):3824–3828
115. Stamm M, Staritzbichler R, Khafizov K, Forrest LR (2014) AlignMe—a membrane protein sequence alignment web server. Nucleic Acids Res 42:W246–W251. doi:10.1093/nar/gku291
116. Grantham R (1974) Amino acid difference formula to help explain protein evolution. Science 185(4154):862–864
117. Abkevich V, Zharkikh A, Deffenbaugh AM, Frank D, Chen Y, Shattuck D, Skolnick MH, Gutin A, Tavtigian SV (2004) Analysis of missense variation in human BRCA1 in the context of interspecific sequence variation. J Med Genet 41(7):492–507
118. Miller MP, Kumar S (2001) Understanding human disease mutations through the use of interspecific genetic variation. Hum Mol Genet 10(21):2319–2328
119. 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. doi:10.1093/nar/gki375
120. Capriotti E, Fariselli P, Rossi I, Casadio R (2008) A three-state prediction of single point mutations on protein stability changes. BMC Bioinformatics 9(Suppl 2):S6. doi:10.1186/1471-2105-9-S2-S6
121. Rost B (1996) PHD: predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol 266:525–539
122. Delorenzi M, Speed T (2002) An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. Bioinformatics 18(4):617–625
123. Radivojac P, Obradovic Z, Smith DK, Zhu G, Vucetic S, Brown CJ, Lawson JD, Dunker AK (2004) Protein flexibility and intrinsic disorder. Protein Sci 13(1):71–80. doi:10.1110/ps.03128904
124. Melamud E, Moult J (2003) Evaluation of disorder predictions in CASP5. Proteins 53(Suppl 6):561–565. doi:10.1002/prot.10533
125. Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure–function paradigm. J Mol Biol 293(2):321–331. doi:10.1006/jmbi.1999.3110
126. Tompa P (2002) Intrinsically unstructured proteins. Trends Biochem Sci 27(10):527–533
127. Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 6(3):197–208. doi:10.1038/nrm1589
128. Dunker AK, Brown CJ, Obradovic Z (2002) Identification and functions of usefully disordered proteins. Adv Protein Chem 62:25–49
129. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol 323(3):573–584
130. Pajkos M, Meszaros B, Simon I, Dosztanyi Z (2012) Is there a biological cost of protein disorder? Analysis of cancer-associated mutations. Mol BioSyst 8(1):296–307. doi:10.1039/c1mb05246b
131. He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK (2009) Predicting intrinsic disorder in proteins: an overview. Cell Res 19(8):929–949. doi:10.1038/cr.2009.87

132. Radivojac P, Vucetic S, O'Connor TR, Uversky VN, Obradovic Z, Dunker AK (2006) Calmodulin signaling: analysis and prediction of a disorder-dependent molecular recognition. Proteins 63(2): 398–410. doi:10.1002/prot.20873
133. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK (2004) The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res 32(3):1037–1049. doi:10.1093/nar/gkh253
134. Daily MD, Masica D, Sivasubramanian A, Somarouthu S, Gray JJ (2005) CAPRI rounds 3–5 reveal promising successes and future challenges for RosettaDock. Proteins 60(2):181–186. doi:10.1002/prot.20555
135. Folkman L, Yang Y, Li Z, Stantic B, Sattar A, Mort M, Cooper DN, Liu Y, Zhou Y (2015) DDIG-in: detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations employing sequence and structural properties at nucleotide and protein levels. Bioinformatics 31(10):1599–1606. doi:10.1093/bioinformatics/btu862
136. Hu J, Ng PC (2013) SIFT Indel: predictions for the functional effects of amino acid insertions/deletions in proteins. PLoS One 8(10):e77940. doi:10.1371/journal.pone.0077940
137. Zhao HY, Yang YD, Lin H, Zhang XJ, Mort M, Cooper DN, Liu YL, Zhou YQ (2013) DDIG-in: discriminating between disease-associated and neutral non-frameshifting micro-indels. Genome Biol 14(3):R23. doi:10.1186/Gb-2013-14-3-R23
138. Zia A, Moses AM (2011) Ranking insertion, deletion and nonsense mutations based on their effect on genetic information. BMC Bioinformatics 12:299. doi:10.1186/1471-2105-12-299
139. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J (2014) A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46(3):310–315. doi:10.1038/ng.2892
140. Liu M, Watson LT, Zhang L (2014) Quantitative prediction of the effect of genetic variation using hidden Markov models. BMC Bioinformatics 15:5. doi:10.1186/1471-2105-15-5
141. Bermejo-Das-Neves C, Nguyen HN, Poch O, Thompson JD (2014) A comprehensive study of small non-frameshift insertions/deletions in proteins and prediction of their phenotypic effects by a machine learning method (KD4i). BMC Bioinformatics 15:111. doi:10.1186/1471-2105-15-111
142. Limongelli I, Marini S, Bellazzi R (2015) PaPI: pseudo amino acid composition to score human protein-coding variants. BMC Bioinformatics 16:123. doi:10.1186/s12859-015-0554-8
143. Zhang N, Huang T, Cai YD (2015) Discriminating between deleterious and neutral non-frameshifting indels based on protein interaction networks and hybrid properties. Mol Genet Genomics 290(1):343–352. doi:10.1007/s00438-014-0922-5
144. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, Shendure J (2011) Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet 12(11): 745–755. doi:10.1038/nrg3031
145. Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, Bustamante CD, Bamshad MJ, Akey JM, Broad GO, Seattle GO, Project NES (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337(6090):64–69. doi:10.1126/science.1219240
146. Alper SL (2013) Harnessing red cell membrane pathophysiology towards point-of-care diagnosis for sickle cell disease. J Physiol 591(Pt 6):1403–1404. doi:10.1113/jphysiol.2013.252429
147. Aidoo M, Terlouw DJ, Kolczak M, McElroy PD, ter Kuile FO, Kariuki S, Nahlen BL, Lal AA, Udhayakumar V (2002) Protective effects of the sickle cell gene against malaria morbidity and mortality. Lancet 359(9314):1311–1312. doi:10.1016/S0140-6736(02)08273-9
148. Gong S, Blundell TL (2010) Structural and functional restraints on the occurrence of single amino acid variations in human proteins. PLoS One 5(2):e9186. doi:10.1371/journal.pone.0009186
149. Wang MJ, Sun ZW, Akutsu T, Song JM (2013) Recent advances in predicting functional impact of single amino acid polymorphisms: a review of useful features, computational methods and available tools. Curr Bioinform 8(2):161–176
150. Capriotti E, Altman RB, Bromberg Y (2013) Collective judgment predicts disease-associated single nucleotide variants. BMC Genomics 14(Suppl 3):S2. doi:10.1186/1471-2164-14-S3-S2
151. Bendl J, Stourac J, Salanda O, Pavelka A, Wieben ED, Zendulka J, Brezovsky J, Damborsky J (2014) PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. PLoS Comput Biol 10(1):e1003440. doi:10.1371/journal.pcbi.1003440
152. Olatubosun A, Valiaho J, Harkonen J, Thusberg J, Vihinen M (2012) PON-P: integrated predictor for pathogenicity of missense variants. Hum Mutat 33(8):1166–1174. doi:10.1002/humu.22102
153. Faa V, Coiana A, Incani F, Costantino L, Cao A, Rosatelli MC (2010) A synonymous mutation in the CFTR gene causes aberrant splicing in an Italian patient affected by a mild form of cystic fibrosis. J Mol Diagn 12(3):380–383. doi:10.2353/jmoldx.2010.090126
154. Brest P, Lapaquette P, Souidi M, Lebrigand K, Cesaro A, Vouret-Craviari V, Mari B, Barbry P, Mosnier JF, Hebuterne X, Harel-Bellan A, Mograbi B, Darfeuille-Michaud A, Hofman P (2011) A synonymous variant in IRGM alters a binding site for miR-196 and causes deregulation of IRGM-dependent xenophagy in Crohn's disease. Nat Genet 43(3):242–245. doi:10.1038/ng.762
155. Wang DX, Sadee W (2006) Searching for polymorphisms that affect gene expression and mRNA processing: example ABCB1 (MDR1). AAPS J 8(3):E515–E520. doi:10.1208/Aapsj080361
156. Nackley AG, Shabalina SA, Tchivileva IE, Satterfield K, Korchynskyi O, Makarov SS, Maixner W, Diatchenko L (2006) Human catechol-*O*-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. Science 314(5807):1930–1933. doi:10.1126/science.1131262
157. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, Gottesman MM (2007) A "silent" polymorphism in the MDR1 gene changes substrate specificity. Science 315(5811):525–528. doi:10.1126/science.1135308
158. Katsnelson A (2011) Breaking the silence. Nat Med 17(12):1536–1538. doi:10.1038/Nm1211-1536
159. Fernald GH, Capriotti E, Daneshjou R, Karczewski KJ, Altman RB (2011) Bioinformatics challenges for personalized medicine. Bioinformatics 27(13):1741–1748. doi:10.1093/bioinformatics/btr295