J Comput Aided Mol Des (2016) 30:401–412
DOI 10.1007/s10822-016-9914-3

Computational approaches for predicting mutant protein stability

Shweta Kulshreshtha[1,3] · Vigi Chaudhary[1] · Girish K. Goswami[2] · Nidhi Mathur[1]

Received: 9 February 2016 / Accepted: 2 May 2016 / Published online: 9 May 2016

Abstract Mutations in the protein affect not only the structure of protein, but also its function and stability. Prediction of mutant protein stability with accuracy is desired for uncovering the molecular aspects of diseases and design of novel proteins. Many advanced computational approaches have been developed over the years, to predict the stability and function of a mutated protein. These approaches based on structure, sequence features and combined features (both structure and sequence features) provide reasonably accurate estimation of the impact of amino acid substitution on stability and function of protein. Recently, consensus tools have been developed by incorporating many tools together, which provide single window results for comparison purpose. In this review, a useful guide for the selection of tools that can be employed in predicting mutated proteins' stability and disease causing capability is provided.

Keywords Mutated protein · Protein stability · Computational tools · Protein function · Databases

✉ Shweta Kulshreshtha
shweta_kulshreshtha@rediffmail.com

[1] Amity Institute of Biotechnology, Amity University Rajasthan, 14-Gopal Bari, Ajmer Road, Jaipur 302006, India

[2] C U Shah Institute of Life Sciences, C U Shah University, Wadhwan City, Surendranagar, Gujarat 363030, India

[3] A-200 Vaishali Nagar, Jaipur 302021, India

Introduction

All kinds of biological functions including enzymatic reactions, biochemical reactions and immunological reactions, co-ordination of nerve impulses, transport and storage are carried out by different types of proteins. Proteins are present in primary, secondary, tertiary and quaternary structures that are maintained by bonds and interactions among its amino acids. The folding of protein depends upon the sequence of amino acids. Proteins carry out biological functions in vivo and in vitro conditions. Protein structure is an important key to understand its role in various biological functions. Proteins are encoded by genes which determine the sequence of amino acids in them.

The sequential process of primary assembly of amino acid leads to proper interaction of the amino acids forming the secondary structure which finally minimizes the energy within the protein, as folding of the protein into a stable conformation produces a stable 3D protein. This 3D structure provides stability to protein for it to be functionally active. All the forces which act in accordance with each other as the hydrogen bonding, hydrophobic interaction, and the negative and positive forces along with conformational entropy decide the protein stability. The correct orientation due to proper interaction between the specific amino acid in form of appropriate amino acid sequence will produce a stable conformation and functionally active protein.

The function of protein such as ligand or substrate binding, catalysis of any reaction and post-translational modification is affected by addition or deletion of nucleotide(s). However, to which extent this affect the structure and function of a protein, varies. According to Tokuriki and co-workers [1] mutations can be classified into two types: (1) New function mutations and (2) Other mutations.

*New function mutations give the mutated protein with new activities or selectivity while o*ther *mutations produce mutant protein without any functional change*.

Mutations of amino acids in protein may be present either in the core or on the surface in buried or exposed forms. Mutations occurring at the surface of a domain are usually considered as neutral, but may affect the binding affinities while those in the core may alter the stability of the domain fold. However, mutations at the interface of protein complexes are called *hot spot* which are associated with diseases [2]. The mutations which are in the residues at active site or at any other residue at the distant site affect the structure of enzyme, substrate and cofactor binding site. Non-synonymous Single Nucleotide Polymorphisms (nsSNPs) and mutations can produce a variety of adverse effect on proteins like changing genotype and phenotype of any protein which may cause diseases like cancer. In contrast to this negative effect [1, 3, 4], many mutations exert a positive effect on the stability, function and structure of proteins. Some mutations are silent and do not exert any effect on protein function, but increase the stability of proteins. The changes in the protein stability are associated with single or multiple amino acid substitution(s). *The detailed knowledge of the effect of mutation can be employed for constructing proteins of interest by protein engineering techniques and in disease treatment/management*. Mutations that confer stability can be used for identifying a wide spectrum of drug resistance mechanisms. Aforesaid description illustrates that mutations and SNPs produce a variety of effects that may be advantageous or disadvantageous for the living organism. Therefore, it is utmost important to study the relationship between protein function and stability in order to understand evolutionary dynamics of protein. Besides this, an insight of nsSNPs and mutations can be helpful in diagnosing diseases at an early stage, prognosis, prevention and treatment of diseases. It is also *helpful in understanding the aspects of protein engineering for designing novel protein with enhanced stability and activity* in the laboratory.

Mutations on another aspect not just provide us an understanding of the novel protein designing with enhanced stability and activity, but it also gives us a platform where we can design mutants with a strong understanding about the importance of a particular amino acid at specific location, having a stability effect. The stability in this context leads to developing a thermo-tolerant and pH stable protein. The mutation of the protein is thus targeted as per our need; designing of high temperature tolerant enzymes, alkali stable or stable in acidic environment. All this can be performed by recognizing the most mutation tolerant residues in the protein, without affecting the overall conformation and 3D structure of the designed protein.

In this review, light is shed on the effect of mutation on protein stability for its functionality and enhancing thermal stability and pH tolerance. Protein stability is defined as the active conformation of the protein which is functional and performs its specific role as enzyme, substrate or receptor. The computational approaches and their basis of prediction are also discussed in this paper.

Stability of protein: a major concern

Stability of protein is a major concern because only stable proteins can perform their function efficiently. Stability of protein or enzyme depends on the organization of residues at active site which mainly involves two types of residues: (1) functional residues and (2) key catalytic residues. Generally, *functional residues are polar or charged, embedded in hydrophobic cleft however, key catalytic residues possess unfavorable backbone angles* [1].

Many researchers have reported significant loss of protein stability when mutations are related to gain of function [1]. Mutation in key positions of protein leads to the evolution of new protein function at the expense of stability. In contrast to this, mutations may sometimes evolve a more stable protein with loss of function compared to the native one [5]. Since most mutations lead to have destabilizing effect on proteins conferring new or altered protein function, these mutations must be studied. *The study of impact of mutation on protein stability provides the details of mechanism of protein folding and identifies the role of specific residues in function and stability* [6, 7]. This information can be used for designing new proteins with desired characteristics such as specific levels of enzymatic activities and stability by introducing point mutations using site directed mutagenesis and random mutagenesis. Introducing point mutations by these methods in protein *requires financial investments, time, resources and labor*. Second important aspect of study of stability, induced by point mutation is—to find out the new mutated stable protein, which has a wide spectrum of drug resistance among the unstable proteins [8, 9].

Analysis of stability

Protein stability is maintained by various non-covalent interactions such as hydrophobic, electrostatic, van der Waals and hydrogen bonds [10–12]. These interactions are of incalculable value for the analysis and prediction of protein stability. Several methods have been developed for assessing and predicting the factors affecting the stability of protein upon mutation or SNPs. These methods also

provide a basis for discriminating mutated stable proteins from native unstable proteins, disease causing mutations from non-disease causing mutations and developing novel enzymes with improved function and stability.

Several methods have been developed over the years for predicting the factors influencing the stability of mutant proteins even upon single amino acid substitution [13–20]. Xencor developed techniques for high throughput generation of myriad sequence variants, coupled with computational protein design automation, for cytokine and growth factor protein therapeutic, and later antibody, protein stability improvement [2]. *Traditional methods have several limitations which are overcome by computational approaches based on sequence, structure and energy features.* These methods can provide better prediction of stability if used in combined form rather than used alone.

Stability predicting features of computational tools

Computational tools work on algorithms (set of rules) which are based on following predictive features:

(a) Structural features: Hydrophobic area, packing and folding of protein, backbone angles, electrostatic interactions are some of the important features that can be used for stability prediction [5].
(b) Sequence features: This is based on conserved sequences and amino acid position. The impact on protein viability can be assessed. However, it provides no direct insight into the underlying mechanism [5].
(c) Combination of structural and sequence features: In this approach, all the above mentioned features can be used together to predict stability.
(d) Energy features: Energy features are important for assessing the stability of protein. The energy of unfolding of the target protein is the sum of various energies such as Van der Waals interaction, solvation energy, extra—stabilizing free energy, etc. [5].
(e) Molecular features: Solvent accessible surface area of the interface and hydrophobic and hydrophilic area is used for stability prediction.

Machine learning approaches are based on the study of structure from data. All these features were used for the development of machine learning approaches such as *support vector machines* (SVM), neural network and decision tree, by the incorporation of functional effects. These methods are designed to predict a change in single amino acid substitution using secondary structure, surface accessibility and sequence attributes. The objective of these methods is—to identify and use non-redundant features that are required for accurate classification [21].

Steps for prediction of mutant protein stability by computational approaches

There are three main steps for predicting the mutant protein stability mentioned below:

(i) Development of a database for proteins and mutants: Various stability predicting tools require different types of *databases*. These databases provide a template whose structure and all details are known. This template is used to compare the structure and stability of query sequence. These databases are given in Table 1.
(ii) Understanding the factors influencing protein mutant stability: The comparison of various structural and sequence features provide better understanding of the factors affecting stability of mutant protein. The detailed insight of these factors can give direction to solve problems related to protein stability.
(iii) Prediction of protein stability upon mutation: This can be done by the help of different computational approaches and tools. Many tools which are available as web based tools or standalone tools are mentioned in Table 2.

The information gathered by these tools can be used for stability prediction and further for designing of new proteins and diseases caused by non-synonymous mutations.

Computational approaches

In the last decade, many tools have been developed to predict the effect of Single Nucleotide Polymorphisms (SNPs) and mutation on genomic location (coding, non-coding and regulatory sequences) and on translated protein (synonymous and non-synonymous SNPs) which was considered as unsolved problems earlier. SNPs are of three types, i.e., neutral, fully disruptive and partially disruptive. The *prediction of neutral, fully disruptive SNPs is relatively easy. However, prediction of SNPs that produce intermediate phenotypic effects is a great challenge.* To overcome this problem, computational tools such as PolyPhen, HOPE, SNPeffect and many other tools have been developed [41, 42]. All SNPs and mutations are not associated with the origin of disease. It is of utmost importance to discriminate disease associated mutations with non-disease mutations. This can also be done by computational tools such as PoPMuSiC-2.0, Site Directed Mutator, Mutation assessor, PhD-SNP and PANTHER etc. All the computational tools can be categorized into four broad categories: (1) Structural features, (2) Sequence

Table 1 Different databases provide information related to mutant protein

S. no.	Database	Website	Remarks	References
1.	PMD (protein mutant database)	http://pmd.ddbj.nig.ac.jp/~pmd/pmd.html	A literature based database for protein mutants	[22]
2.	dbSNP (single nucleotide polymorphism database)	http://www.ncbi.nlm.nih.gov/SNP/	A catalog of short variations in nucleotide sequences from a wide range of organisms	[23]
3.	HGVbase (human genome variation database)	https://gwas.biosciencedbc.jp/cgi-bin/hvdb/hv_top.cgi	A repository system for human mutation data	[24]
4.	Protherm	http://www.abren.net/protherm/	A collection of numerical data of thermodynamic parameters for wild type and mutant proteins	[25]
5.	F-SNP	http://compbio.cs.queensu.ca/F-SNP/	A collection of functional SNPs, specifically prioritized for disease association studies	[26]
6.	HGMDHuman gene mutation database	http://www.hgmd.cf.ac.uk/ac/index.php	It represents an attempt to collate known gene lesions responsible for human inherited disease	[27]
7.	dbNSFP (database for nonsynonymous SNPs' functional predictions)	https://sites.google.com/site/jpopgen/dbNSFP	A database for functional prediction and annotation of all potential non-synonymous single-nucleotide variants (nsSNVs) in the human genome	[28]
8.	COSMIC	http://cancer.sanger.ac.uk/cosmic	Stores and displays information related to human cancers and somatic mutations	[29]
9.	UniProt/SWISS-PROT	http://www.uniprot.org/	Human protein variant dataset	[30]
10.	OMIM (online Mendelian inheritance in man)	http://www.omim.org/	A catalog of human genes and genetic conditions	[31]

features, (3) Energy parameters and (4) Combined features. In this review, we have also focused on the use of consensus tools for predicting stability.

Structure based approaches/tools

These tools predict the stability changes by observing structural properties such as secondary structure and accessible surface area of mutated residue [20]. PolyPhen tool was developed to assess intermediate phenotypic effects of point mutation. Similar to HOPE, this tool predicts the protein structure by statistical analysis, but unable to provide any information on the amount of free energy changes on point mutation and therefore, cannot be used for the correct prediction of stability. This limitation is tackled by the development of SNPeffect tool. It is a structure based tool that uses FoldX; force field for quantitative estimation of free energy and thus, gives accurate information about the protein stability. However, it has some limitations related to the quality of structure without which protein structure cannot be modelled with more than 90 % sequence identity to that of template structure. *This tool also helps in finding out the protein homeostasis landscape i.e. the amount of proteins which must be present in various cellular compartments of cell* [54]. Further modification in this tool has led to the development of latest version *i.e.* SNPeffect 4.0. It is integrated meta-analysis tool that is designed by the integration of FoldX for predicting protein misfold, TANGO and WALTZ for protein aggregation, LIMBO for chaperon interaction. It enables the study of large scale data mining and graphical representation of data. It provides detailed information on functional sites, structural features and post-translational modification of protein [55]. It can also be applied for molecular characterization and presentation of disease linked polymorphism in humans owing to the database for phenotypes of human single nucleotide polymorphisms (SNPs).

To analyze the effects of nsSNPs on phenotypic characteristics nsSNP Analyzer, a web based tool, was developed which provides detailed information related to the effect of SNP on structure of protein, surface accessibility, environment and multiple sequence alignment. This tool facilitates the identification of disease-associated nsSNPs from neutral nsSNPs. This software uses a machine learning method called Random Forest for prediction and requires structural and evolutionary information from a query nsSNP [33].

CUPSAT (Cologne University Protein Stability Analysis Tool) was developed to predict changes in stability of protein upon point mutation with good efficiency [8]. *It provides information about the site of mutation and structural attributes such as solvent accessibility, secondary structure and torsion angles affected by mutation.*

Table 2 This table depicts different tools with their salient features and respective websites for accession

S. no.	Tool	Website	Remarks	References
1.	I-Mutant	http://folding.biofold.org/i-mutant/i-mutant2.0.html	Prediction of protein stability change upon single point mutation	[32]
2.	nsSNPAnalyzer	http://snpanalyzer.uthsc.edu/	Predict whether a nonsynonymous single nucleotide polymorphism (nsSNP) has a phenotypic effect	[33]
3.	PANTHER	http://www.pantherdb.org/tools/	Protein Analysis THrough Evolutionary Relationships classifies proteins (and their genes) in order to facilitate high-throughput analysis	[34]
4.	MUpro	http://mupro.proteomics.ics.uci.edu/	Prediction of Protein Stability Changes for Single-Site Mutations from Sequences	[17]
5.	PhD-SNP	http://snps.biofold.org/phd-snp/phd-snp.html	Predictor of human deleterious SNPs	[35]
6.	SNPs3D	http://www.snps3d.org/	Assigns molecular functional effects of nsSNPs based on structure and sequence analysis	[36]
7.	Allign GVGD	http://agvgd.iarc.fr/index.php	Predicts where missense substitutions in genes of interest fall in a range from deleterious to neutral	[37]
8.	CUPSAT	http://cupsat.tu-bs.de/	Predicts changes in protein stability upon point mutations	[13]
9.	iPTREE-STAB	http://210.60.98.19/IPTREEr/iptree.htm	Interpretable decision tree based method for predicting protein stability changes upon mutations	[38]
10.	Eris	http://troll.med.unc.edu/eris/login.php	Predicts change in protein stability induced by mutation	[39]
11.	SIFT	http://sift.bii.a-star.edu.sg/	Predicts whether an amino acid substitution affects protein function	[40]
12.	HOPE	http://www.cmbi.ru.nl/hope/input;jsessionid=3a676f9318c73d61a7295af3360d?0	Analyze the effect of certain mutation on protein structure	[41]
13.	PolyPhen-2	http://genetics.bwh.harvard.edu/pph2/	Predicts change in structure and function of protein upon single amino acid substitution	[42]
14.	AUTO-MUTE	http://proteins.gmu.edu/automute/	AUTOmated server for predicting functional consequences of amino acid MUTations in proteins	[43, 44]
15.	MutationAssessor	http://www.ngrl.org.uk/Manchester/page/mutation-assessor	Predicts the functional impact of amino-acid substitutions in proteins	[45]
16.	ProMaya	http://bental.tau.ac.il/ProMaya/	Predicts protein stability free energy difference arising from a single amino acid substitution compared to the wild type	[46]
17.	CONDEL	http://bg.upf.edu/fannsdb/	Assess the outcome of nonsynonymous SNVs using a consensus deleteriousness score	[47]
18.	PoPMuSiC	http://dezyme.com/	Tool for computer aided design of mutant proteins with controlled stability properties	[48]
19.	PROVEAN	http://provean.jcvi.org/index.php	Predicts whether an amino acid substitution has an impact on the biological function of a protein	[49]
20.	I-Stable	http://predictor.nchu.edu.tw/istable/index.php	Integrated predictor for protein stability change upon single mutation	[50]
21.	MetaSNP	http://snps.biofold.org/meta-snp/	Detection of disease-associated nsSNVs by integrating four existing methods: PANTHER, PhD-SNP, SIFT and SNAP	[51]
22.	WS-SNPs&GO	http://snps.biofold.org/snps-and-go/	Predicts human disease-related single point protein mutations	[51]
23.	DUET	http://bleoberis.bioc.cam.ac.uk/duet/stability	Server for predicting effects of nsSNP mutations on protein stability	[52]
24.	ZEBRA	http://biokinet.belozersky.msu.ru/zebra	Identify the subfamily-specific positions (SSPs) which can be used as hotspots for directed evolution or rational design experiments to enhance protein function and create novel biocatalysts	[53]

The stability is predicted by calculating difference in free energy of unfolding between wild type and mutated protein. As a rule this tool requires protein structure in Protein Data Bank (PDB) format and the location of the residue to be mutated.

Auto-Mute tool is also a computational approach. Like SDM, it uses knowledge based potential and machine learning approach to predict the function and stability of mutant proteins according to the score generated. However, SDM was proved to be better than Auto-Mute. Like CUPSAT, this tool also requires the data in PDB format [44].

Structure-based tools have a *drawback.* These tools cannot be used if structure of protein in crystallized, 3-dimentional and high resolution form is unavailable. All methods need structure only in PDB format. However, mutation may occur at genome (instead of proteome) level, which can change the structure and other properties of mutated protein. This limitation can be resolved by using sequence based prediction tools. Nowadays, huge amount of data is generated from various genome sequencing projects which is utilized for making libraries or databases and further, applied for comparison of query sequence with the target sequence.

Sequence based approaches/tools

Sequences, which code for functional region of protein, are considered as conserved regions and not useful for predicting the structure of mutated protein. Therefore, sequence based methods rely on evolutionary conservation of homologous protein sequences [56].

The most simple, easy to handle tool is HOPE (Homotopy optimization method) because the *results are depicted in the form of animated videos and figures* which provide an easy basis to discriminate between neutral and deleterious mutations by either trained or new user. It is used for the prediction of intermediate phenotypic effects caused by single amino acid change. It is a sequence based tool to facilitate calculation of potential energy associated with a protein model. Any user of online web server can submit a sequence and mutation. It can also collect information about the protein from various sources (3D protein structure, UniProt Database and DAS servers). It requires a known protein with minimum energy conformation which acts as a template and used to predict the structure of query sequence on statistical basis. The comparison of query sequence with the template provides information about the effect of mutation on protein structure, function and stability [41]. Further, *it builds a model of mutated protein if the amount of identity between query sequence and PDB file exceeds the threshold value*.

Another tool for the prediction of SNPs or mutation due to single amino acid substitution is SIFT (Sorting Intolerant from Tolerant) which was developed by Ng and Henikoff [57]. It is a tool based solely on sequence homology which discriminate neutral or deleterious SNPs using normalized probability score (score < 0.05: deleterious and score > 0.05: neutral) [57]. It collects the sequences from PSI-BLAST and consider only closely related sequences. The basis of closely related sequences is the conservation in the conserved region. This tool is *very sensitive and has been applied to human variant databases* in order to find out deleterious mutations (disease causing) among neutral SNPs [58]. Like PolyPhen, this is used to predict loss of function mutations. It is also applicable to predict mutations induced by in silico procedures. The most important advantage of this tool is that it can predict a large number of substitutions which provides higher number of affected phenotypes. Moreover, application of this tool during the in silico analysis can reduce the number of assays required for screening of desired mutants. The main limitation of this technique is—*it gives generalized probability score rather than specific for the mutant protein of interest*. It has very low specificity and therefore, results must be interpreted carefully [59]. This tool will not provide molecular mechanism or details to understand the cause of disease [2].

PhD-SNP (Predictor of human Deleterious Single Nucleotide Polymorphism) tool is also a sequence based approach for the prediction of mutant protein stability caused by single nucleotide polymorphism. It is basically a Support Vector Machine (SVM) based classifier, specifically designed for human dataset associated with disease causing mutations. It collects the data from Swiss-Prot database for input and results in the generation of protein sequences and profile information about the mutant protein [35]. PhD-SNP server provides datasets for Parepro, a computational tool, based on SVM. P*arepro requires evolutionary information instead of structural information* to predict the effect of nsSNP with higher accuracy than any other method [60].

Another sequence based tool is MUpro that can accurately predict the stability of protein using primary sequence information only by machine learning approaches. Enhanced accuracy of this tool is reported if structure is also provided, although it is not necessary [61].

Mutation Assessor also calculates stability score that predicts the impact of mutation on protein. It relies on evolutionary conservation approach and uses multiple sequence alignment (MSA) to generate conservation score, which classifies the mutation into three categories; neutral, low, high or medium. All these mutations affect the functionality of protein and their stability. This tool was also used for SNPs associated with diseases such as cancer. It

has its own database for the prediction of diseases; however, it can retrieve data from other databases such as COSMIC, UniProt and Pfam. This is available only as a web based tool.

Like Mutation Assessor, PANTHER (Protein Analysis Through Evolutionary Relationships) is also one of the tools that are based on position specific evolutionary conservation score to find out the relation of any mutation with disease and its impact on protein function and stability. The score decides the type of mutation; if score range falls in between 0 (neutral) to 10 (deleterious). It applies *hidden Markov models (HMMs) for aligning protein sequences* and provides an idea about the variants in protein families and subfamilies. This tool can also be used for designing of novel stable proteins.

I-Mutant (old version of I-Mutant 2.0) implements a neural network algorithm for predicting the stability change caused by single nucleotide substitution due to point mutation. It is a support vector machine based tool which only requires the input of sequence and therefore, can be used even if crystal structure of protein is not known [16].

A web based tool PoPMuSiC (Prediction of Protein Mutant Stability Changes) can also be used for predicting all mutations irrespective of their effect on stability. It is a neural network based approach, which predicts the mutant protein stability on the basis of sequence. Compared to the other methods, it is a very fast technique which requires only few minutes of time for predicting stability changes of any protein possessing single-site mutations. It is user friendly and easy to exploit [62, 63] for computer-aided designing of mutant proteins.

Similar to PoPMuSiC, SNAP (Screening for Non-Acceptable Polymorphisms) is also a neural-network based tool which collects information solely from sequence to study effect of mutation on protein structure and function. *This computational approach can predict secondary structure, solvent accessibility and other information related to protein structure in addition to evolutionary information*. This enables the discrimination of gain of function mutations from loss of function mutations. It can be used in designing of protein. This tool predicts effects of mutation by score/reliability index for each substitution with improved accuracy. It can incorporate mutations at the position of user's choice [64]. However, the most advanced version of PoPMuSiC i.e., PoPMuSiC 2.1 can be used to find out interesting sites for introducing mutation. Therefore, it can be suitably used for protein designing. Unlike the previous version, PoPMuSiC 2.1 can predict only those region or sequences that correspond to structural and functional weakness [65]. This property of tool is based on the use of evolutionary information. It requires only sequence input, uses machine learning approaches for processing data, and predicts the direction and value of stability.

The entire above mentioned tools can predict single amino acid substitution in the sequences only. There is only one computational tool PROVEAN (PROtein Variation Effect ANalyser) that not only predicts single/multiple amino acid substitutions, but also predicts single/multiple insertions and deletions [49]. This tool applies a new metric measure *i.e., alignment-based score to predict the change caused by mutation*. At first, this tool collects all possible supporting sequence using search tool BLAST, which corresponds to input protein sequence. Then, delta alignment scores are computed with respect to each sequence variation which is further used for calculating PROVEAN score (50). It provides *more accurate results for mutant proteins* compared to Mutation assessor, SIFT and PolyPhen-2 [49]. Castellana and Mazza [66] used this tool for the classification of nsSNPs by using SNP-associated chromosomal positions as an input protein sequence.

Annotation of SNPs is difficult in the species lacking a reference genome. This problem is solved by a newly developed tool known as SNPMeta. This tool collects information about SNPs by comparing sequences with the sequences present in GenBank databases. Results, obtained from this tool, are incomparable with that obtained from a reference genome [67].

Energy based approaches/tools

Different methods have been developed and implemented in order to predict the stability of protein with.respect to wild type protein. Some computational tools are based on energy functions to compute the free energy changes related to stability of proteins. The *physical and statistical energy based approaches provide good results qualitatively, however, do not provide precise values and cannot be applied to large datasets*. In contrast to this, empirical potential approaches provide rapid results with precise values to evaluate the contribution of an amino acid substitution to the stability of protein [68].

Approaches based on the binding free energy cannot be applied to predict core mutations due to biophysical nature, polar and electrostatic attractions of protein–protein interfaces. Therefore, only dissociation rate of protein instead of association rate is used to predict difference in energy of mutant protein and native protein. This was done by alanine scanning method that is based on empirical energy functions. The main advantage of this method is—its applicability to predict the effect and molecular effect of multiple amino acid substitution. It provides more accurate prediction compared to sequence based tools.

PoPMuSiC is a sequence based tool. However, PoPMuSiC 2.0, a latest version of this tool, works on energy

based function and measures the change in protein upon single nucleotide substitution. It is statistical potential based method that has also been used to characterize in silico the effect of mutation on stability of protein. Besides this, it also predicts mutation responsible for hereditary diseases; acquired drug resistance and natural heterogeneity of a viral protein [65]. This tool can estimate stability changes in medium size protein within seconds and robustness of structure. It depicts highest linear correlation (0.63) between predicted and measured stability values.

After PoPMuSiC algorithm, Site Directed Mutator (SDM) has the highest correlation between predicted and measured stability values [65]. It is better than Auto-Mute tool. SDM is also knowledge based approach that predicts the mutant protein stability associated with disease development and engineering protein. It is developed by Topham and colleagues in 1997 [69]. It works on the statistical potential energy function *to predict the stability score* (negative score for destabilizing mutation and positive score for stabilizing mutations). This score is not only useful in disease prediction, but also in protein engineering. The main problem associated with this tool is its least bias in predicting stabilizing and destabilizing mutation. The performance of SDM is therefore, improved only when highly stabilizing and destabilizing mutations are considered.

Combined features based tools

In the last decade, many in silico approaches based on computational tools have been developed to predict mutant protein stability on the basis of structure, sequence and energy based features. As discussed earlier, structural based approaches can be applied only when structure is known otherwise sequence based approaches are suggested. However, prediction accuracy of sequence based approaches is lower than the structure based approaches [38]. None of them is proven to be accurate and provide complete mutant protein analysis. The prediction accuracy can be increased by using the right combination of features [70]. Therefore, combined and consensus approaches are developed with increased accuracy and efficiency and for use in all situations [52]. *Mutations in the core are difficult to predict by using one method and therefore, sequence, structural and energetic features based methods can be used in a suitable combination by machine learning methods for prediction*.

I-Mutant tool (mentioned earlier) is modified and developed into its latest version, known as I-Mutant 2.0, to be used either with sequence or structure. It is based on SVM and support vector regression [71] automatic prediction of stability upon single nucleotide substitution. The accuracy of prediction depends upon the sequence and structural information provided to it. Besides predicting value and direction of stability, it can be used for predicting point mutation associated diseases in humans. It can also be useful in protein engineering. Unlike the other tools, *it requires the input of data in raw format which is a unique feature of the tool* [71].

Another tool based on sequence, structural and phylogenetic features is PolyPhen-2. It is automatic tool and an advanced version of PolyPhen used to predict the impact of amino-acid substitution on the structure and function of human protein by machine learning approaches. It requires data from human protein database (UniProt KB/swiss-Prot), known 3D structure or homologous proteins (if known structure is unavailable) to predict amino acid replacement in the core of protein corresponding to the known structure. Similar to SIFT, it can also accept FASTA protein sequences. It is a Bayesian classifier used to categorize pathogenicity [66]. It also provides information on functional impact of SNPs by using input from F-SNP database (mentioned in Table 1). *The best feature of this tool is its built-in support system for high performance computing, which makes it suitable for handling huge amount of data generated by next generation sequencing projects*.

EvoD tool [72] has better accuracy compared to PolyPhen-2 and Condel. In this tool, multiple sequences submitted as NCBI RefSeq. Protein IDs are used to predict the nucleotide substitution affected sites on the basis of biochemical and evolutionary properties. This tool *distinguishes neutral mutations from deleterious mutations*.

An integrated tool, I-Stable is also developed based on both structural and sequential information. It is SVM based tool that not only provides information regarding protein stability, but also provide accurate predictions for secondary structures, relative solvent accessibility and classification of protein into super families. It can also be used for the designing and engineering of protein.

Recently, Berliner and colleagues [2] developed a Stability Meta-Predictor for predicting core and domain–domain interface mutations by integrating sequence and structural features. This novel tool is known as ELASPIC (Ensemble Learning Approach for Stability Prediction of Interface and Core mutations). To increase the accuracy of prediction, *it uses Stochastic Gradient Boosting of Decision Trees* (*SGB-DT*) *algorithm which combines both sequence and structural features.* This tool not only predicts stability and affinity of mutant protein, but also reveals the molecular principles behind disease-causing mutations. It decorously discriminates disease-causing mutations from neutral mutations.

Consensus tools

Consensus tools are based on the integration of various tools that compares sequences of homologous protein using multiple alignment process. This comparison gives a consensus sequence which further compares with existing protein sequences in order to predict the differences generated by point mutations or nsSNPs. This offers the selection of best mutation related to increasing stability of proteins. In this way, consensus tools provide information, that need to design a novel protein with desired characteristic and stability.

PON-P is a machine learning-based method that requires the submissions of sequences in multi-FASTA protein sequences. It collects input data from SIFT, PhD-SNP, PolyPhen-2.0 and SNAP tools for analysis. Therefore, it delivers the results predicted by all the aforesaid programs due to which these outputs can be compared effortlessly [66].

Condel is also *a consensus tool* which is developed by integrating SIFT, PolyPhen-2 and MutationAssessor. It collects, assembles and present the results obtained by these tools. This can be used as a web server or standalone tool (run on computer after downloading) [47].

Another consensus tool PredictSNP collects the input data from six tools such as MutPred, Polyphen-1, Polyphen-2, SNAP, MAPP, PhD-SNP, SIFT, SNPs&GO for predicting the effect of single amino acid substitution. Most of these tools are based on machine learning methods, especially designed to *classify neutral and deleterious mutations on the basis of physicochemical, sequence and structural parameters*. PredictSNP provides a consensus prediction with improved accuracy and efficiency over individual integrated tools [73].

DUET is a *valuable integrated tool* which uses sequence based approaches. DUET integrates mCSM and SDM in a consensus tool and analyzes the results using SVM. It is used to predict mutant protein stability, which can be applied for protein engineering approaches and anticipation of disease. It predicts the structure by multiple sequence alignment, which can be done automatically by this tool or manually by user [74]. The availability of sequences may affect prediction accuracy of tool. It can be applied for the prediction of nsSNP of human and non-human genomes. Besides, it helps in engineering novel protein with enhanced stability [52].

These consensus tools improve accuracy and efficiency of the tools and predict the results better than any other tool. These tools provide a platform for comparing the results obtained by different tools in a common place and hence, reduce the efforts required to study the data by using different tools individually.

Tools for study of buried residues

Aforementioned tools can be used to study the change in the core, especially to provide information about the interaction of exposed residues only. The amino acid changes occurring in the buried residues of core are difficult to predict. Recently, NeEMO *tool is developed for the study of stability changes on the basis of residue interaction network (RINs).* RINs are very comprehensive data structures that help the management of heterogeneous data, such as evolutionary and topological data. RINs are used to describe interaction of mutant amino acid with its structural environment. It is very effective and provides an accurate prediction of the buried residues that are difficult to predict by any other method. Unlike other methods, *it does not require the modelling of mutant protein structure and therefore, not only avoid the errors introduction but also makes it computationally economic.* The NeEMO web server can be freely accessed from URL: http://protein.bio.unipd.it/neemo/. It requires PDB files of PoPMuSiC 2.0 dataset as an input. It can be used for protein engineering and study of diseases caused by unstable mutant protein [75].

Conclusion

Stability of proteins is extremely valuable for the study of diseases caused by unstable proteins and for engineering protein with desired characteristics. Stability is an important parameter to be considered when the effect of mutations is observed. It provides fundamental knowledge based on sequence, structure and evolutionary relationship. In the last decade, various tools have been developed for the prediction of stability. These tools provide details on nsSNP induced effects on protein structure and their molecular basis. However, these tools have their own limitations which restrict their ability of predictions. The restricted ability is due to the availability parameters which need to be understood, the tools should be used after understanding the prediction parameters which hold importance for a specific type of enhancement. Recently, some consensus tools have been developed by integrating various tools. These tools provide more reliable and accurate prediction due to compelling and comparing data from various tools. NeEMO, a novel tool, is also developed for predicting the effect of un-annotated (buried) SNPs or mutations on stability. These tools also have some limitations which restrict their use with a particular dataset. Research is going on for the development of complete tool, fit for all types of data. In future, these tools will provide more reliable and accurate assessment as the submission of

new data in the databases is rapidly increasing. However, till then, for reliability of results obtained from these tools, one can further use the concept of molecular dynamics and simulation to substitute with a more confident approach for stability. This not only provides an insight in understanding stability, folding process and interactions among residues, but also depicts any specific slight changes occurring due to mutations. Tools such as MD Analysis, Desmond of Schrodinger, X-PLORE, MDTraj, WORDOM, MDWeb etc., are available to carry out molecular dynamic analysis. Mutation analysis along with molecular dynamics simulation studies do support the design and analysis of naturally occurring as well as novel designed mutants with good stability.

Compliance with ethical standards

Conflict of interest There is no conflict of interest with this manuscript.

References

1. Tokuriki N, Stricher F, Serrano L, Tawfik DS (2008) How protein stability and new functions trade off. PLoSComput Biol 4:e1000002
2. Luo P, Hayes RJ, Chan C, Stark DM, Hwang MY, Jacinto JM, Juvvadi P, Chung HS, Kundu A, Ary ML, Bassil I (2002) Dahiyat development of a cytokine analog with enhanced stability using computational ultrahigh throughput screening. Protein Sci 11:1218–1226
3. Hartl FU, Bracher A, Hayer-Hartl M (2011) Molecular chaperones in protein folding and proteostasis. Nature 475:324–332
4. Tokuriki N, Tawfik DS (2009) Chaperonin overexpression promotes genetic variation and enzyme evolution. Nature 459:668–673
5. Yue P, Li Z, Moult J (2005) Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol 353:459–473
6. Lehmann M, Wyss M (2001) Engineering proteins for thermostability: the use of sequence alignments versus rational design and directed evolution. Curr Opin Biotechnol 12:371–375
7. Yang DF, Wei YT, Huang RB (2007) Computer-aided design of the stability of pyruvate formate-lyase from *Escherichia coli* by site-directed mutagenesis. Biosci Biotechnol Biochem 71:746–753
8. Parthiban V, Gromiha MM, Schomburg D (2006) CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res 34((Web Server issue)):W239–W242
9. Foot E, Kleyn D, Foster PE (2010) Pharmacogenetics-pivotal to the future of the biopharmaceutical industry. Drug Discov Today 15:325–327
10. Dill KA (1990) Dominant forces in protein folding. Biochemistry 29:7133–7155
11. Pace CN (1990) Conformational stability of globular proteins. Trends Biochem Sci 15:14–17
12. Ponnuswamy PK, Gromiha MM (1994) On the conformational stability of folded proteins. J Theor Biol 166:63–74
13. Parthiban V, Gromiha MM, Hoppe C, Schomburg D (2007) Structural analysis and prediction of protein mutant stability using distance and torsion potentials: role of secondary structure and solvent accessibility. Proteins 66:41–52
14. Bordner AJ, Abagyan RA (2004) Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. Proteins 57:400–413
15. Khatun J, Khare SD, Dokholyan NV (2004) Can contact potentials reliably predict stability of proteins? J Mol Biol 336:1223–1238
16. Capriotti E, Fariselli P, Casadio R (2004) A neural-network-based method for predicting protein stability changes upon single point mutations. Bioinformatics 20:I63–I68
17. Cheng JL, Randall A, Baldi P (2006) Prediction of protein stability changes for single-site mutations using support vector machines. Proteins 62:1125–1132
18. Huang LT, Saraboji K, Ho SY, Hwang SF, Ponnuswamy MN, Gromiha MM (2006) Prediction of protein mutant stability using classification and regression tool. Biophys Chem 125:462–470
19. Saraboji K, Gromiha MM, Ponnuswamy MN (2005) Relative importance of secondary structure and solvent accessibility to the stability of protein mutants: a case study with amino acid properties and energetics on T4 and human lysozymes. Comput Biol Chem 29:25–35
20. Saraboji K, Gromiha MM, Ponnuswamy MN (2006) Average assignment method for predicting the stability of protein mutants. Biopolymers 82:80–92
21. Kamath U, De Jong K, Shehu A (2014) Effective automated feature construction and selection for classification of biological sequences. PLoS One 9:e99982
22. Kawabata T, Ota M, Nishikawa K (1999) The protein mutant database. Nucleic Acids Res 27:355–357
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 29:308–311
24. Fredman D, Siegfried M, Yuan YP, Bork P, Lehväslaiho H, Brookes AJ (2002) HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. Nucleic Acids Res 30:387–391
25. Kumar MD, Bava KA, Gromiha MM, Parabakaran P, Kitajima K, Uedaira H, Sarai A (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nuleic Acids Res 34:D204–D206
26. Lee PH, Shatkay H (2008) F-SNP: computationally predicted functional SNPs for disease association studies. Nucleic Acids Res 36:D820–D824
27. Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H (2013) Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum Genet 132:1077–1130
28. Liu X, Jian X, Boerwinkle E (2013) dbNSFP v2.0: a database of human non-synonymous SNPs and their functional predictions and annotations. Hum Mutat 34:E2393–E2402
29. 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 J, Teague JW, Stratton MR, McDermott U, Campbell PJ (2014) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 43:D805–D811
30. UniProt Consortium (2015) UniProt: a hub for protein information. Nucl. Acids Res 43((Database issue)):D204–D212
31. Shaw CA, Campbell IM (2015) Variant interpretation through Bayesian fusion of frequency and genomic knowledge. Genome Med 7:4
32. Bava KA, Gromiha MM, Uedaira H, Kitajima K, Sarai A (2004) ProTherm, version 4.0: thermodynamic database for proteins and mutants. Nucleic Acids Res 32:D120–D121
33. Bao L, Zhou M, Cui Y (2005) nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res 33:W480–W482

34. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, Rabkin S, Guo N, Muruganujan A, Doremieux O, Campbell MJ, Kitano H, Thomas PD (2005) The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res 33:D284–D288
35. 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 22:2729–2734
36. Yue P, Melamud E, Moult J (2006) SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinform 7:166
37. Mathe E, Olivier M, Kato S, Ishioka C, Hainaut P, Tavtigian SV (2006) Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. Nucleic Acids Res 34:1317–1325
38. Huang LT, Gromiha MM, Ho SY, Ho SY (2007) Sequence analysis and rule development of predicting protein stability change upon mutation using decision tree model. J Mol Model 13:879–890
39. Yin S, Ding F, Dokholyan NV (2007) Eris: an automated estimator of protein stability. Nat Methods 4:466–467
40. Kumar P, Henikoff S, Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4:1073–1081
41. Venselaar H, te BeekG TAH, Kuipers RKP, Hekkelman ML, Vriend G (2010) Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. BMC Bioinform 11:548
42. 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 7:248–249
43. Masso M, Vaisman II (2011) A structure-based computational mutagenesis elucidates the spectrum of stability-activity relationships in proteins. Conf Proc IEEE Eng Med Biol Soc 2011:3225–3228
44. Masso M, Vaisman II (2014) AUTO-MUTE 2.0: a portable framework with enhanced capabilities for predicting protein functional consequences upon mutation. Adv Bioinform http://dx.doi.org/10.1155/2014/278385
45. Reva B, Antipin Y, Sander C (2011) Predicting the functional impact of protein mutations: applications to cancer genomics. Nucleic Acids Res 39:e118
46. Wainreb G, Wolf L, Ashkenazy H, Dehouck Y, Ben-Tal N (2011) Protein stability: a single recorded mutation aids in predicting the effects of other mutations in the same amino acid site. Bioinformatics 27:3286–3292
47. González-Pérez A, López-Bigas N (2011) Improving the assessment of the outcome of nonsynonymous SNPs with a consensus deleteriousness score, Condel. Am J Hum Genet 88:440–449
48. Gonnelli G, Rooman M, Dehouck Y (2012) Structure-based mutant stability predictions on proteins of unknown structure. J Biotechnol 161:287–293
49. 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
50. Chen CW, Lin J, Chu YW (2013) iStable: off-the-shelf predictor integration for predicting protein stability changes. BMC Bioinform 14(Suppl 2):S5
51. Capriotti E, Altman RB, Bromberg Y (2013) Collective judgment predicts disease-associated single nucleotide variants, mutations in proteins. BMC Genom 14(suppl 3):S2
52. Pires DEV, Ascher DV, Blundell TL (2014) DUET: a server for predicting effects of mutations on protein stability via an integrated computational approach. Nucleic Acids Res 42(W1):W314–W319
53. Suplatov D, Shalaeva D, Kirilin E, Arzhanik V, Švedas V (2014) Bioinformatic analysis of protein families for identification of variable amino acid residues responsible for functional diversity. J Biomol Struct Dyn 32:75–87
54. Powers ET, Morimoto RI, Dillin A, Kelly JW, Balch WE (2009) Biological and chemical approaches to diseases of proteostasis deficiency. Annu Rev Biochem 78:959–991
55. Baets GD, Durme JV, Reumers J, Maurer-Stroh S, Vanhee P, Dopazo J, Schymkowitz J, Rousseau F (2012) SNPeffect 4.0: online prediction of molecular and structural effects of protein-coding variants. Nucleic Acids Res 40((Database issue)):D935–D939
56. Han JH, Batey S, Nickson AA, Teichmann SA, Clarke J (2007) The folding and evolution of multidomain proteins. Nat Rev Mol Cell Biol 8:319–330
57. Ng PC, Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function. Annu Rev Genom Hum Genet 7:61–80
58. Ng PC, Henikoff S (2002) Accounting for human polymorphisms predicted to affect protein function. Genom Res 12:436–446
59. Flanagan SE, Patch AM, Ellard S (2010) Using SIFT and PolyPhen to predict loss-of-function and gain-of-function mutations. Genet Test Mol Biomark 14:533–537
60. 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 Bioinform 8:450
61. Cheng G, Qian B, Samudrala R, Baker D (2005) Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. Nucleic Acids Res 33:5861–5867
62. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics 25:2537–2543
63. Johnston MA, Sondergaard C, Nielsen JE (2011) Integrated prediction of the effect of mutations on multiple protein characteristics. Proteins 79:165–178
64. Bromberg Y, Rost B (2007) SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 35:3823–3835
65. Dehouck Y, Kwasigroch MJ, Gilis D, Rooman M (2011) PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinform 12:151
66. Castellana S, Mazza T (2013) Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools. Brief Bioinform 14:448–459
67. Kono TJY, Seth K, Poland JA, Morrell PL (2014) SNPMeta: SNP annotation and SNP metadata collection without a reference genome. Mol Ecol Resour 14:419–425
68. 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 22:553–560
69. Topham CM, Srinivasan N, Blundell TL (1997) Prediction of protein mutants based on structural environment-dependent amino acid substitution and propensity tables. Protein Eng 10:7–21
70. Folkman L, Stantic B, Sattar A (2013) Sequence-only evolutionary and predicted structural features for the prediction of stability changes in protein mutants. BMC Bioinform 14(Suppl 2):1
71. 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((Web Server issue)):W306–W310

72. Kumar S, Sanderford M, Gray VE (2012) Evolutionary diagnosis method for variants in personal exomes. Nat Methods 9:855–856
73. 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:e1003440
74. Dunbrack RL Jr (2006) Sequence comparison and protein structure prediction. Curr Opin Struct Biol 16:374–384
75. Giollo M, Martin AJM, Walsh I, Ferrari C, Tosatto SCE (2014) NeEMO: a method using residue interaction networks to improve prediction of protein stability upon mutation. BMC Genom 15(Suppl 4):S7