Role of Structural Bioinformatics in Drug Discovery by Computational SNP Analysis

Analyzing Variation at the Protein Level

David K. Brown, Özlem Tastan Bishop

Grahamstown, South Africa

ABSTRACT

With the completion of the human genome project at the beginning of the 21st century, the biological sciences entered an unprecedented age of data generation, and made its first steps toward an era of personalized medicine. This abundance of sequence data has led to the proliferation of numerous sequence-based techniques for associating variation with disease, such as genome-wide association studies and candidate gene association studies. However, these statistical methods do not provide an understanding of the functional effects of variation. Structure-based drug discovery and design is increasingly incorporating structural bioinformatics techniques to model and analyze protein targets, perform large scale virtual screening to identify hit to lead compounds, and simulate molecular interactions. These techniques are fast, cost-effective, and complement existing experimental techniques such as high throughput sequencing. In this paper, we discuss the contributions of structural bioinformatics to drug discovery, focusing particularly on the analysis of nonsynonymous single nucleotide polymorphisms. We conclude by suggesting a protocol for future analyses of the structural effects of nonsynonymous single nucleotide polymorphisms on proteins and protein complexes.

With the completion of the human genome project in 2003, biological science entered the genomic era. Since then, the rate of data generation has been increasing at an unprecedented rate. Improved technologies have given rise to next-generation sequencing capabilities that are able to sequence genomes faster and at a fraction of the cost of technologies that came before. These advances have made previously unfeasible undertakings, such as the 1000 Genomes Project [1] and the International HapMap Project [2], possible.

More recently, the Human Heredity and Health in Africa (H3Africa) Initiative was founded to facilitate genomic studies and to build research capacity on the African continent [3]. As part of this project, thousands of genomes from various populations around Africa are being sequenced and massive amounts of new data are being generated. One of the goals of the project is to identify and understand single nucleotide polymorphisms (SNPs) linked to disease. In order to identify SNPs associated with disease, various sequence-level techniques can be employed, including genome-wide association studies (GWAS) and candidate gene association studies (CGAS). These techniques associate SNPs with diseases by comparing the genomes/genes of healthy individuals with those of unhealthy individuals to determine which SNPs mostly occur in disease-affected patients. SNPs that occur at a statistically significant higher rate in the unhealthy individuals are said to be associated with disease.

Where techniques such as GWAS and CGAS are used to analyze variation at the DNA level, structural bioinformatics techniques provide a means for the downstream analysis of variation (i.e., the analysis of variation at the protein level). These techniques include methods such as homology modeling, molecular docking, molecular dynamics, and residue interaction network (RIN) analysis, and let researchers form hypotheses on what effects SNPs have on protein structure, stability, and inter- and intra-protein interactions. Unfortunately, structural bioinformatics techniques can be extremely computationally expensive. As such, even the filtered data sets provided by GWAS and CGAS can be too large. In this paper, we discuss the importance of structural bioinformatics in SNP analysis and drug discovery, and provide a suggested approach for analyzing variation at the protein level.

RETRIEVING AND FILTERING SNPs FOR USE IN STRUCTURAL STUDIES

There are roughly 100 million validated human variants in dbSNP build 147 [4]. It is simply not feasible to study each and every one of these variants in detail. Techniques such as GWAS and CGAS are applied at the sequence level and provide a quick means of filtering out SNPs that are likely not important for a disease. Additionally, tools that predict the effects of SNPs on protein function and stability can be used to further filter these datasets. This does not mean that the

The authors report no relationships that could be construed as a conflict of interest.
H3ABioNet is supported by the National Institutes of Health Common Fund (grant number U41HG006941).
The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.
From the Research Unit in Bioinformatics (RUBi), Department of Biochemistry and Microbiology, Rhodes University, Grahamstown, South Africa.
Correspondence: Ö. Tastan Bishop (o.tastanbishop@ru.ac.za).

GLOBAL HEART

VOL. 12, NO. 2, 2017
ISSN 2211-8160/$36.00.
http://dx.doi.org/10.1016/j.gheart.2017.01.009

TABLE 1. Variation databases

Database	Description	Link	Reference
COSMIC	Cancer-associated mutations	http://cancer.sanger.ac.uk/cosmic	[11]
ClinVar	Clinical significance of variation	http://www.ncbi.nlm.nih.gov/clinvar/	[8]
dbGaP	Database of genotypes and phenotypes	http://www.ncbi.nlm.nih.gov/gap/	[7]
dbNSFP	Functional predictions and annotations of nonsynonymous SNPs	https://sites.google.com/site/jpopgen/dbNSFP	[96-98]
dbSNP	Short variation	http://www.ncbi.nlm.nih.gov/projects/SNP/	[5]
dbVAR	Structural variation	http://www.ncbi.nlm.nih.gov/dbvar/	[6]
Database of Genomic Variants archive (DGVa)	Structural variation	http://www.ebi.ac.uk/dgva	[6]
European Genome-phenome Archive (EGA)	Private variation archive	https://www.ebi.ac.uk/ega/home	[9]
European Variation Archive (EVA)	Public variation archive	http://www.ebi.ac.uk/eva/	—
Ensembl	Comprehensive biological database including variation	http://www.ensembl.org/	[15]
HGMD	Disease-related gene lesions	http://www.hgmd.cf.ac.uk/	[99]
HGVD	Japanese genetic variation	http://www.genome.med.kyoto-u.ac.jp/SnpDB/	[100]
Human Mutation Analysis (HUMA)	Comprehensive biological database including variation	https://huma.rubi.ru.ac.za	—
LS-SNP/PDB	Nonsynonymous SNPs likely to affect biological function	http://ls-snp.icm.jhu.edu/ls-snp-pdb/	[18]
National Human Genome Research Institute-European Bioinformatics Institute (NHGRI-EBI) catalog	Manually curated database of published genome-wide association studies	http://www.ebi.ac.uk/gwas/home	[10]
Online Mendelian In Man (OMIM)	Human genes and genetic disorders	http://www.omim.org/	[13]
PinSnps	Protein-protein interaction networks	http://fraternalilab.kcl.ac.uk/PinSnps/	[17]
SNPeffect	Characterization and annotation of SNPs	http://snpeffect.switchlab.org/	[101]
SNPs3D	Functional effects of nonsynonymous SNPs	http://www.snps3d.org/	[102]
The Cancer Genome Atlas (TCGA)	Cancer-associated mutations	http://cancergenome.nih.gov/	[12]
Uniprot	Protein database including nonsynonymous SNPs	http://www.uniprot.org/	[14]
VnD	Variation and drugs	http://vnd.kobic.re.kr/	[51]

SNP, single nucleotide polymorphism.

remaining SNPs are important, however. Further studies are required to confirm their importance as well as to understand their role, if any, in the disease. It is at this point that structural bioinformatics techniques can be employed.

Variation databases

One of the challenges of bioinformatics is storing the enormous amounts of data being generated by next-generation sequencing projects. In line with this, various databases have been developed to store variation identified via these projects (Table 1). The most well-known of these databases is probably dbSNP [5], a database created and managed by the National Center for Biotechnology Information as a central repository for all known short variation. The dbSNP database incorporates data from projects such as 1000 Genomes and HapMap, as well as many others.

The National Center for Biotechnology Information also has various other variation databases, including dbVAR [6], dbGaP [7], and ClinVar [8]. Where dbSNP focuses on short variation, dbVAR stores structural variation such as insertions and deletions. On the other hand, dbGaP and ClinVar are focused on the relationship between genotype and phenotype and the clinical significance of variation, respectively.

The European Bioinformatics Institute (EBI) also hosts various variation databases including the European Variation Archive (EVA), the Database of Genomic Variants archive (DGVa) [6], and the European Genome-phenome Archive (EGA) [9]. EVA is a public variation archive, which stores all types of variation. DGVa, on the other hand, is EBI's version of dbVAR (i.e., a database for structural variation). Variation in EVA, DGVa, dbVAR, and dbSNP is exchanged on a regular basis, meaning that these databases generally mirror each other. EVA also stores data from ClinVar, making it a rich source for variation data.

The EGA stores complete data sets from genomic studies, allowing users to browse various aspects of the data. Unlike EVA, EGA is not a public data archive. Data sets are stored privately and researchers must be granted access by the specified Data Access Committee to view the data.

The EBI, along with the National Human Genome Research Institute, have also produced the National Human Genome Research Institute—EBI GWAS Catalog [10], a high-quality, manually curated collection of published GWAS. The GWAS Catalog stores SNP and SNP-trait associations for over 11,000 SNPs and from over 1,700 publications.

Some variation databases focus of variation related to a disease or groups of diseases. Examples of this include COSMIC [11] and the Cancer Genome Atlas [12], which focus on variation related to cancer. Other databases, such as the Online Mendelian In Man (OMIM) [13] database link variation to phenotypes. Uniprot [14], a database focused on proteins, maps nonsynonymous SNPs to these proteins.

One of the most comprehensive biological databases is hosted by Ensembl [15]. The Ensembl database stores various biological data including genes, transcripts, proteins, exons, and more. To this data, it links phenotypes and variation. Ensembl incorporates variation from numerous sources including dbSNP, ClinVar, COSMIC, dbGaP, DGVa, EGA, OMIM, and Uniprot. All this data is stored within a single, relational database and can be queried using BioMart [16], a powerful tool that provides simple and uniform access to various data sources.

The previously mentioned databases all focus on the analysis of SNPs at the sequence level. PinSnps [17] is a database where variation is mapped to protein structures. Variation data are collected from various sources including OMIM and COSMIC. Users of the PinSnps web server are then able to select their SNPs of interest and visualize them in the protein structure. PinSnps also links SNPs to protein interaction networks.

LS-SNP/PDB [18] is another variation database where SNPs are pre-mapped to protein structures. As with PinSnps, users can query the database for a protein or SNP of interest and then visualize SNPs in the structure of the protein.

Tools and databases, such as PinSnps and LS-SNP/PDB, that focus on the structural impacts of variation are, unfortunately, few and far between. Additionally, these databases tend to neglect the sequence level data. We have developed the Human Mutation Analysis (HUMA) web server and database, which focuses on the analysis of variation in humans both at the sequence and structural level. The HUMA database stores genes, proteins, proteins structures, diseases, and variants. Variation is pre-mapped to gene and protein sequences based on chromosome coordinates. Variants are also mapped to protein structures based on alignments between the protein sequences and sequences extracted from the PDB files for the respective proteins. Additional information about the protein structures, such as the ligands that were solved with the structure and the resolution at which the structure was solved are also stored. Proteins, genes, and variation are all linked to disease via data obtained from ClinVar and Uniprot. As part of the pipeline for mapping variation to protein sequences, HUMA also stores the coding sequences, coding DNA, and exons for proteins. As such, HUMA provides a resource for querying variation both at the sequence and structural level.

Predicting disease associated/deleterious mutations

The main challenge of computational SNP analysis at the sequence level is determining whether a SNP is associated with, or likely to be associated with, disease. As previously discussed, GWAS and CGAS are useful techniques for associating variants with disease. Association via these techniques is no guarantee that mutation is disease-related, however. Additionally, these techniques can miss variation that is important. As such, other methods are still required to further analyze the effects of variation.

At the protein level, numerous tools have been developed which predict the impact of nonsynonymous SNPs on protein function (Table 2). These tools usually fall into 1 of 2 categories. The first category is made up of tools that make predictions based solely on the sequence of a protein, while the second is made up of tools that incorporate structural information when making predictions [19].

Tools such as SIFT [20], PROVEAN [21], and PANTHER-PSEP [22] fall into the first category. These tools look at sequence conservation to determine whether mutations at a particular position will be deleterious. This is based on the theory that highly conserved regions of a sequence must be important to protein function. Mutations in these regions will therefore have detrimental effects. SIFT and PROVEAN look at the conservation of amino acids across homologs. While SIFT can predict the effects of SNPs, PROVEAN has the added advantage of being able to predict the effects of in-frame insertions and deletions. PANTHER-PSEP, on the other hand, looks at evolutionary conservation (i.e., the time since the last mutation occurred at a particular position in an amino acid sequence).

FATHMM [23] is also a sequence-based SNP analysis tool. As with the above tools, the FATHMM makes conservation-based predictions. However, FATHMM also includes a second, weighted algorithm. This algorithm essentially allows predictions to be adjusted based on the tolerance of the region of the protein to mutations.

Machine learning techniques have also been used to predict the functional effects of variation. PhD-SNP [24] and Parepro [25] are sequence-based support vector machine (SVM) methods for predicting the functional effects of SNPs. SVM methods are popular for handling biological data due to their ability to work with large data sets and to handle noise effectively.

PolyPhen-2 [26], Auto-Mute 2.0 [27], and SNAP [28] incorporate structural information when making predictions on the functional effects of mutations. As such, they fall into the second category of SNP analysis tools. PolyPhen-2 uses 3 structure-based predictive features as well as 8 sequence-based predictive features to classify variation. Predictions are made via a naive Bayes classifier.

TABLE 2. Tools for predicting the functional effects of nonsynonymous SNPs

Tool	Description	Link	Reference
Auto-Mute 2.0	Sequence and structure based	http://binf2.gmu.edu/automute/	[27]
FATHMM	Sequence based	http://fathmm.biocompute.org.uk/	[23]
MAPP	Sequence based	http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html	[103]
Meta-SNP	Consensus classifier	http://snps.biofold.org/meta-snp/	[30]
MuD	Sequence and structure based	http://mud.tau.ac.il/	[104]
MutPred	Sequence based	http://mutpred.mutdb.org/	[105]
PANTHER-PSEP	Sequence based	http://www.pantherdb.org/tools/csnpScoreForm.jsp	[22]
Parepro	Sequence-based	http://www.mobioinfor.cn/parepro/	[25]
PolyPhen-2	Sequence and structure based	http://genetics.bwh.harvard.edu/pph2/	[26]
PredictSNP	Consensus classifier	http://loschmidt.chemi.muni.cz/predictsnp/	[29]
Provean	Sequence and structure based	http://provean.jcvi.org/index.php	[21]
SIFT	Sequence-based	http://provean.jcvi.org/index.php	[20]
SNAP	Sequence-based	http://www.bio-sof.com/snap	[28]
SNPs&GO	Sequence and structure based	http://snps.biofold.org/snps-and-go/snps-and-go.html	[106]
Variant Analysis Portal (VAPOR)	Consensus classifier	https://huma.rubi.ru.ac.za/#vapor	—

Similarly, Auto-Mute 2.0 combines structural features with trained, machine-learning methods. SNAP, on the other hand, only requires sequence information as input, but structural and functional annotations help to improve predictions.

There are various other methods for predicting the functional effects of SNPs, which have not been discussed here. None of these methods are perfect, however. As such, it is a good idea to get a consensus from several different tools before deciding, which SNPs to select for further analysis. With this in mind, classifiers such as PredictSNP [29] and Meta-SNP [30] combine the predictions of various existing tools to gain a consensus on which SNPs are deleterious to protein function.

We have developed the Variant Analysis Portal (VAPOR), which has been incorporated into the HUMA web server. VAPOR is a workflow that accepts either a protein sequence or protein structure as input along with a list of SNPs. From here, it gets predictions from PROVEAN, PolyPhen-2, PhD-SNP, PANTHER-PSEP, and FATHMM and merges the results into a single table. Unlike PredictSNP and Meta-SNP, VAPOR does not generate a consensus score from these results. It remains as a useful tool for quickly getting results from multiple SNP analysis methods, however.

PREDICTING CHANGES IN PROTEIN STABILITY DUE TO MUTATIONS

Predicting the impact of SNPs on protein stability is another important area of SNP analysis. Nonsynonymous SNPs can result in changes of the internal energy of a protein as well as lead to changes in the structure of the protein. Calculating the change in Gibbs free energy between a wild type protein and the mutated form is a common measure of how much a mutation affects protein stability [31]. One thing to note when analyzing changes in protein stability is that increases and decreases in protein stability do not necessarily correspond to deleterious and beneficial effects, as increases in protein stability can also hamper protein function.

Various tools have been developed to predict changes in protein stability due to nonsynonymous SNPs (Table 3). The Auto-Mute 2.0 suite discussed earlier includes functionality for predicting stability changes. Additionally, I-Mutant2.0 [32] and MuPro [33] provide SVM based methods for predicting changes in stability. Both tools can be used, either to simply predict the sign of the change in stability, or to predict the actual size of the change. Both tools can also incorporate structural information when making predictions, but MuPro can achieve nearly the same accuracy when only the primary sequence is considered, making it a useful option when the tertiary structure of the protein is unknown.

NeEMO [34] is a machine learning method based on RINs. It incorporates information from RINs in a nonlinear neural network to improve prediction accuracy. RINs provide useful information regarding changes in residue interactions when a mutation is introduced as they implicitly incorporate detailed maps of chemical interactions within proteins.

The VAPOR workflow makes use of I-Mutant 2.0 and MuPro predictions to complement the functional predictions described in the previous section. Unfortunately, NeEMO is not available for download and, as such, could not be included as part of VAPOR. Including stability

TABLE 3. Tools for predicting changes in stability due to nonsynonymous SNPs

Tool	Description	Link	Reference
Auto-Mute 2.0	Sequence and structure based	http://binf2.gmu.edu/automute/	[27]
CUPSAT	Structure based	http://cupsat.tu-bs.de/	[107]
Eris	Structure based	http://troll.med.unc.edu/eris/login.php	[108]
I-Mutant2.0	Sequence and structure based	http://folding.biofold.org/i-mutant/i-mutant2.0.html	[32]
MuPro	Sequence and structure based	http://mupro.proteomics.ics.uci.edu/	[33]
NeEMO	Residue interaction networks	http://protein.bio.unipd.it/neemo/help.html	[34]
PoPMuSiC 2.1	Structure based	https://soft.dezyme.com/query/create/pop	[109]

prediction tools in VAPOR, however, adds an additional dimension to the workflow and differentiates it from similar tools.

ROLE OF STRUCTURAL BIOINFORMATICS: SNP ANALYSIS IN DRUG DISCOVERY

Structural bioinformatics is an area of bioinformatics focused on the structure, movement and interaction of biological macromolecules in 3-dimensional space. Structural bioinformatics techniques play an important role in drug discovery and can be used at every stage of the drug design process [35-39], where they can be used to complement, and sometimes replace more costly experimental techniques [40-42]. For example, protein structure prediction software provides alternatives to x-ray crystallography and nuclear magnetic resonance techniques, while virtual screening and molecular dynamics simulations can complement high throughput screening (HTS).

The use of computational techniques in drug discovery and design is often referred to as computer-aided drug design [53]. In this section, we will discuss the uses of structural bioinformatics as part of computer-aided drug design, specifically in the context of nonsynonymous SNP analysis.

Mutations have been associated with drug resistance in numerous diseases such as influenza, tuberculosis, HIV, and cancer [43-47]. Similarly, mutations can be linked to drug sensitivity in patients [48]. This opens the door to personalized medicines, where knowledge of drug resistant and drug sensitive SNPs allow treatments to be tailored to individual patients [49,50]. Understanding structural changes caused by nonsynonymous SNPs will enable the design of novel drugs to target these mutations and, thus, be key in advancing personalized medicine [51].

Protein structure prediction

In the post-genomic era, there is an abundance of available protein sequences. Unfortunately, solving the structures of these proteins is a slow and expensive process. As such, the gap between known protein sequences and solved protein structures is growing. To illustrate this, as of September 2016, the Protein Data Bank [52] contained a little over 120,000 protein structures, which pales in comparison to the 65 million sequences available in the Uniprot protein sequence database. Having the protein structure available lets researchers gain insight into the molecular function of the protein. An understanding of the structural and functional aspects of proteins opens up the door to drug design and discovery [38,53] and, as such, is of great interest to chemists as well as biologists. To counter the growing sequence-structure gap, various computational structure prediction methods have been developed. These methods can be categorized into 2 distinct groups, namely, template-based modeling, and ab initio (or de novo) techniques.

Ab initio modeling attempts to construct a model of a protein based solely on its amino acid sequence. This is a computationally intensive task that, despite ever increasing computational power, is currently only practical for small systems [54]. Additionally, according to the latest CASP results [55], ab initio methods have yet to catch up to template-based modeling techniques in terms of accuracy.

Template-based modeling is currently the most reliable method for protein structure prediction, producing decent quality models for roughly two-thirds of proteins with unsolved structures [55-57]. Template-based modeling can be divided into homology modeling and protein threading techniques.

Homology modeling is a structure prediction technique that relies on the observation that the structural conformation of a protein is more conserved than its amino acid sequence. As such, solved protein structures can be used as templates for predicting the tertiary structure of a target sequence, provided the sequence identity between the target and template sequences is high enough (roughly >30%) [38,58].

Protein threading is similar to homology modeling in that it uses the structures of previously solved proteins to predict the structure of a target sequence. Where homology modeling uses the structures of homologous proteins as templates, however, threading uses the structures of proteins, which are predicted to have the same folds. Threading is useful when there are no homologous proteins available that have solved structures [59].

Protein structure prediction can be used to introduce SNPs into a structure and determine the effects that these SNPs might have on the protein's function and stability. Once modeled, the wild type structure can be compared to the mutant structure in several ways. For example, the

RINs of the structures can be compared to see if introducing SNPs influences intra-protein communication. The structures can also be compared to see if new bonds have been introduced or existing bonds have been broken. In addition, the models can be further analyzed using molecular docking and molecular dynamics simulations, 2 important techniques for drug discovery.

Homology modeling has been used in various stages of drug discovery including the study of protein function and mechanisms [60], analysis of the effects of mutations in binding sites of receptor proteins [61], identification of druggable pockets [62], and various virtual screening studies [63-66].

Molecular docking and virtual screening

Molecular docking is a technique for predicting the bound conformations of a protein-ligand complex, and is used in structure-based drug design to study biomolecular interactions [67]. Docking is fast enough to allow libraries containing thousands of compounds to be docked against a receptor protein in a process called virtual screening. Virtual screening is used to scan a compound library for potential drug candidates [68-70]. As compounds are docked against the receptor, a score is calculated to determine the binding affinity of each compound to the receptor. Compounds with the highest binding affinity scores are selected for further study. Binding affinity scores are not infallible, and rankings based on these scores are, therefore, not necessarily reliable. Nevertheless, these binding affinity scores can distinguish likely from unlikely compounds, and can be used as potential hit compounds in the drug design process [68].

Molecular docking can also be used to assess the impact of SNPs on drug response. Mutations in the binding sites of receptor proteins can affect the binding affinity of drugs. This can lead to drug resistance or drug susceptibility. Molecular docking can be used in conjunction with protein structure prediction to predict the effect these mutations will have on drug response [61].

Virtual screening has become a routine procedure in drug discovery and can be used as a cheaper alternative to HTS [71]. Having access to a comprehensive compound library is an important part of virtual screening. As such, numerous compound libraries have been made available via online databases and portals such as ZINC [72], ChemSpider [73], the Traditional Chinese Medicine (TCM) Database@Taiwan [74], and SANCDB [75].

Molecular dynamics simulations

Protein structure prediction and molecular docking provide a snapshot in time of a protein structure and protein-ligand complex, respectively. Molecular dynamics, on the other hand, simulates the movements and trajectories of all the atoms in these structures over a period time. It can be used to check if a protein structure remains stable after the introduction of 1 or more SNPs. Similarly, it can be used to determine the stability of protein-ligand complexes after docking [76]. While molecular docking predicts how well a compound docks to a receptor, molecular dynamics can predict how stably bound the compound is and whether it will stay bound over a specified period.

Molecular dynamics results are usually analyzed via plots of their root mean square deviation (RMSD) and root mean square fluctuation. There first measurement, RMSD, measures the average movement in the structure's backbone over the course of the simulation. If, by the end of the simulation, it appears that the plot of the RMSD has leveled out, it can be assumed that the structure has stabilized.

Where RMSD measures the global movement of the protein, root mean square fluctuation, measures local movement (i.e., how much individual residues fluctuate over the course of the simulation). Spikes in this plot indicate residues that move a lot over the course of the simulation, while low values indicate residues that remain relatively fixed throughout.

Molecular docking simulations are often used in combination with homology modeling and virtual screening [76,77]. In terms of computational SNP analysis, molecular dynamics can be used to determine whether introducing a SNP will destabilize a protein or perhaps cause the protein to move or fold in a different way [78].

Inter- and intra-protein interactions

Inter- and intra-protein interactions play important roles in protein folding as well as in the stability and function of proteins and protein complexes. Due to protein folding, residues that are far apart in a protein's sequence can be right next to one another in 3-dimensional space. Interactions between these residues help the protein to adopt the correct structural conformation [79]. As such, disruptions to these interactions (e.g., residue substitutions) could cause instability and loss of protein function. It is, therefore, useful to understand, which residues are important in the structure and function of a protein. This can be done by analyzing the types of bonds (e.g., hydrogen bonds, disulfide bonds) that occur between residues.

RINs provide another means of analyzing protein structures. RINs have been analyzed using a branch of mathematics known as graph theory. In a RIN, each residue in the protein is a node in the network. An edge (or connection) between 2 nodes exists if there is an interaction between the 2 residues that they represent [80]. In RINs, interactions between residues exist if the residues are within a user-defined cutoff (usually around 6.5 to 7.5 Å) of each other [81].

Various network measures have been used to analyze RINs. Previously, the change in the average shortest path to each residue (ΔL) and the change in betweenness centrality of each residue (ΔBC) has been used to perform alanine scanning, where each residue is mutated to alanine to see its effect on the overall network [82].

The shortest path (L) between 2 nodes is the minimum number of edges that must be traversed to travel from one node to another. The average shortest path to a residue is

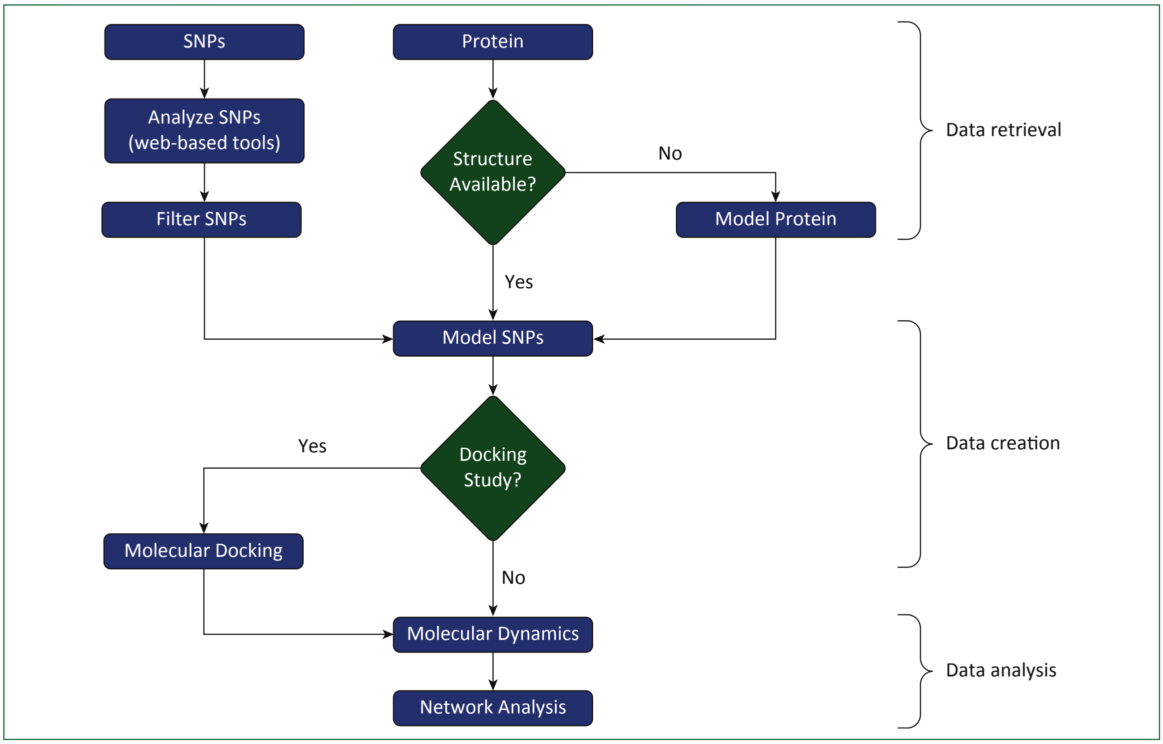

FIGURE 1. Protocol for analyzing nonsynonymous single nucleotide polymorphisms (SNPs). A flowchart depicting the steps required to analyze the effects of nonsynonymous SNPs using structural bioinformatics. The process can be divided into 3 phases: data retrieval, data creation, and data analysis.

calculated by summing the shortest path between a given residue and all other residues in the structure and dividing the result by N−1, where N is the number of residues in the structure. The result of this calculation is the average accessibility of the given residue from any other residue in the structure (i.e., selecting any other residue at random, what are the average number of edges that will need to be traversed to reach the given residue). When comparing a wild type protein to a mutant, ΔL can be calculated for each residue by subtracting the average shortest path to each residue in the mutant from the average shortest path to each respective residue in the wild type. The result describes whether the residue is more or less accessible in the mutated structure [82].

The betweenness centrality (BC) of a given node is a measurement of how often a shortest path between 2 nodes passes through the given node. As such, it measures the importance of the given node to efficient navigation of the network. A high BC means that the node occupies a central position in the network. When using this measure to perform an alanine scan, ΔBC for a residue is calculated by getting the difference between the BC for a residue in the mutant and wild type [82].

Network analysis techniques such as those describe above can be applied to both experimental and predicted PDB structures. In addition, network analysis can be carried out over the trajectory of a molecular dynamics simulation to monitor how the network changes over time [83]. Although L and BC have previously only been used to perform alanine scanning, we propose that these same techniques could be applied to SNP analysis.

PROTOCOL FOR ANALYZING SNPs USING STRUCTURAL BIOINFORMATICS

Structural bioinformatics is an important part of the drug discovery process. As discussed in previous sections, it can contribute to every stage of the drug design process. Here we propose a protocol for determining the effects of nonsynonymous SNPs on protein structure, function, and stability using structural bioinformatics techniques (Fig. 1).

The first requirement of any type of analysis is data. In our case, the required data to perform the analysis is the protein sequence and structure and the nonsynonymous SNPs that occur in the protein. As previously discussed, there are various public databases available that provide access to variation data (Table 1). For our purposes, the most useful of these databases are arguably Ensembl and HUMA. Both databases allow the user to search for their protein of interest and make both the sequence and all the known variation in that sequence available for download. Mutation data from these databases is linked to phenotypes, where possible. If there are experimentally determined structures available for the protein, these structures are also linked to. As such, Ensembl and HUMA provide convenient locations to access all of our required data.

If no protein structures are available, or if there are important missing residues in available structures, the structure of the protein must be modeled. Fortunately, various online structure prediction pipelines exist. Commonly used tools include HHPred [84], SWISS-MODEL [85], I-TASSER [86], and Phyre2 [87]. We have also developed PRIMO (PRotein Interactive MOdeling) [88], an interactive homology modeling platform that assists users through the modeling process.

As structural bioinformatics techniques tend to be computationally intensive, it is not possible to analyze every SNP in the protein in detail using these methods. As such, the SNP data set must be filtered before we move on to more computationally expensive techniques. Tools that predict the effects of SNPs on function (Table 2) and stability (Table 3) can be used to quickly analyze large SNP data sets. The results of this analysis, although not infallible, can be used to filter the data set to contain only SNPs that are likely to negatively affect function or stability. As a general rule of thumb, at least 4 or 5 of these tools should be run to gain a consensus as to the effect of the SNP.

To complement this analysis, the SNPs should be checked for known disease-associations in literature. Ensembl and HUMA link diseases to SNPs and, as such, provide useful resources for this purpose.

If a structure is available for the protein, or once the structure of the protein has been modeled, it may be useful to check, which residues in the structure are interacting. Interacting residues are likely to be important for protein function and stability and, as such, SNPs occurring at these locations may be important. Thus, protein inter- and intra-actions can be used to further filter the SNP data set. Various tools have been developed to calculate these interactions by determining the bonds, such as hydrogen bonds and disulfide bonds, that form between residues. These include web servers such as PIC [89], COCOMAPS [90], InterProSurf [91], PDBParam [92], and PDBSum [93].

Once the SNP data set has been filtered to a low enough level (dependent on available computational resources), the SNPs can be introduced into the protein structure via homology modeling. A model should be produced for every SNP (i.e., if there are 20 SNPs in the data set, 20 models should be produced, each containing one of the SNPs). Combinations of SNPs can also be modeled into the structure if, for example, it is known that the SNPs co-occur.

If the goal of the research is to determine whether SNPs will affect the binding affinity of a drug, it is at this point that molecular docking runs should be performed, both on the wild type structure and the mutants. Analyzing changes in the binding affinity of the drug between the wild type and the mutants will give an idea of whether drug responses may be affected in the mutants.

To improve the reliability of the docking results, or to analyze the stability of the wild type and protein models, molecular dynamics simulations should be run. Currently, the most popular molecular dynamics software available are arguably GROMACS [94] and NAMD [95]. These simulations will give insight into whether the docked drug will remain bound to the mutant proteins over a period of time. If the protein has been destabilized, this may not be the case. A destabilized protein may also have impaired function, which could indicate the involvement of the respective SNP in a disease phenotype.

RIN analysis can be performed after modeling or docking to determine how these methods have affected the network. Previous methods have minimized the protein structure before performing network analysis [82]. Another interesting option is to perform network analysis over the trajectory of the molecular dynamics simulation [83].

To predict whether a given SNP is associated with a disease, the networks of mutant models containing SNPs that are associated with the disease in literature (or in Ensembl and HUMA) can be compared with the network of the mutant model containing the given SNP. Similar changes in the network may indicate similar effects on protein function and stability.

SUMMARY

Structural bioinformatics techniques such as protein structure prediction, molecular docking, and molecular dynamics provide low cost alternatives to experimental techniques such as x-ray crystallography, nuclear magnetic resonance, and HTS. In this paper, we have discussed the use of these techniques in drug discovery, with a focus on the analysis of nonsynonymous SNPs. Mutations, such as SNPs, contribute to differences in drug response between individuals. Gaining further understanding of the reasons behind these differences will gives us insight into how we can take advantage of them and, thereby, usher in the age of personalized medicine.

REFERENCES

1. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature 2015;526:68–74.
2. The International HapMap Consortium. The International HapMap Project. Nature 2003;426:789–96.
3. The H3Africa Consortium. Research capacity. Enabling the genomic revolution in Africa. Science 2014;344:1346–8.
4. dbSNP 147 Data Summary. Available at: https://www.ncbi.nlm.nih.gov/projects/SNP/snp_summary.cgi?view+summary=view+summary&build_id=147. October 1, 2016.
5. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001;29:308–11.
6. Lappalainen I, Lopez J, Skipper L, et al. DbVar and DGVa: Public archives for genomic structural variation. Nucleic Acids Res 2013;41:D936–41.
7. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. Nat Genet 2007;39:1181–6.
8. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: Public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 2014;42:D980–5.
9. Lappalainen I, Almeida-King J, Kumanduri V, et al. The European Genome-phenome archive of human data consented for biomedical research. Nat Genet 2015;47:692–5.

10. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res 2014;42:D1001–6.
11. Bamford S, Dawson E, Forbes S, et al. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. Br J Cancer 2004;91:355–8.
12. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008;455:1061–8.
13. Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM). Hum Mutat 2000;15:57–61.
14. Apweiler R, Bairoch A, Wu CH, et al. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 2004;32:D115–9.
15. Hubbard T, Barker D, Birney E, et al. The Ensembl genome database project. Nucleic Acids Res 2002;30:38–41.
16. Smedley D, Haider S, Ballester B, et al. BioMart: biological queries made easy. BMC Genomics 2009;10:22.
17. Lu HC, Braga JH, Fraternali F. PinSnps: structural and functional analysis of SNPs in the context of protein interaction networks. Bioinformatics 2016;32:2534–6.
18. Ryan M, Diekhans M, Lien S, Liu Y, Karchin R. LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures. Bioinformatics 2009;25:1431–2.
19. Mah JTL, Low ESH, Lee E. In silico SNP analysis and bioinformatics tools: A review of the state of the art to aid drug discovery. Drug Discov Today 2011;16:800–9.
20. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31:3812–4.
21. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS One 2012;7:e46688.
22. Tang H, Thomas PD. PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation. Bioinformatics 2016:1–3.
23. Shihab HA, Gough J, Cooper DN, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat 2013;34:57–65.
24. Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 2006;22:2729–34.
25. Tian J, Wu N, Guo X, Guo J, Zhang J, Fan Y. Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines. BMC Bioinformatics 2007;8:450.
26. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–9.
27. Masso M, Vaisman II. AUTO-MUTE 2.0: A portable framework with enhanced capabilities for predicting protein functional consequences upon mutation. Adv Bioinformatics 2014;2014:278385.
28. Bromberg Y, Rost B. SNAP: Predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 2007;35:3823–35.
29. Bendl J, Stourac J, Salanda O, et al. PredictSNP: robust and accurate consensus classifier for prediction of disease-related mutations. PLoS One 2014;10:e1003440.
30. Capriotti E, Altman RB, Bromberg Y. Collective judgment predicts disease-associated single nucleotide variants. BMC Genomics 2013;14(Suppl 3):S2.
31. Thiltgen G, Goldstein RA. Assessing predictors of changes in protein stability upon mutation using self-consistency. PLoS One 2012;7:e46084.
32. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 2005;33:W306–10.
33. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. Proteins 2006;62:1125–32.
34. Giollo M, Martin AJ, Walsh I, Ferrari C, Tosatto SC. NeEMO: a method using residue interaction networks to improve prediction of protein stability upon mutation. BMC Genomics 2014;15:S7.
35. Chou K-C. Impacts of bioinformatics to medicinal chemistry. Med Chem 2015;11:218–34.
36. Blundell TL, Sibanda BL, Montalvão RW, et al. Structural biology and bioinformatics in drug design: opportunities and challenges for target identification and lead discovery. Philos Trans R Soc Lond B Biol Sci 2006;361:413–23.
37. Taboureau O, Baell JB, Fernández-Recio J, Villoutreix BO. Established and emerging trends in computational drug discovery in the structural genomics era. Chem Biol 2012;19:29–41.
38. Cavasotto CN, Phatak SS. Homology modeling in drug discovery: current trends and applications. Drug Discov Today 2009;14:676–83.
39. Kapetanovic IM. Computer-aided drug discovery and development (CADDD): In silico-chemico-biological approach. Chem Biol Interact 2008;171:165–76.
40. Scapin G. Structural biology and drug discovery. Curr Pharm Des 2006;12:2087–97.
41. Congreve M, Murray CW, Blundell TL. Structural biology and drug discovery. Drug Discov Today 2005;10:895–907.
42. Durrant JD, McCammon JA. Molecular dynamics simulations and drug discovery. BMC Biol 2011;9:71.
43. Sim S, Kacevska M, Ingelman-Sundberg M. Pharmacogenomics of drug-metabolizing enzymes: a recent update on clinical implications and endogenous effects. Pharmacogenomics J 2012;13:1–11.
44. Casali N, Nikolayevskyy V, Balabanova Y, et al. Evolution and transmission of drug-resistant tuberculosis in a Russian population. Nat Genet 2014;46:279–86.
45. Gottesman MM. Mechanisms of cancer drug resistance. Annu Rev Med 2002;53:615–27.
46. LI J, Linley L, Kline R, Ziebell R, Heneine W, Johnson JA. Sensitive sentinel mutation screening reveals differential underestimation of transmitted HIV drug resistance among demographic groups. AIDS 2016;30:1439–45.
47. Pielak RM, Schnell JR, Chou JJ. Mechanism of drug inhibition and drug resistance of influenza A M2 channel. Proc Natl Acad Sci U S A 2009;106:7379–84.
48. Garnett MJ, Edelman EJ, Heidorn SJ, et al. Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 2012;483:570–5.
49. Kumar RD, Chang LW, Ellis MJ, Bose R. Prioritizing potentially druggable mutations with dGene: an annotation tool for cancer genome sequencing data. PLoS One 2013;8:e67980.
50. Niu B, Scott AD, Sengupta S, et al. Protein-structure-guided discovery of functional mutations across 19 cancer types. Nat Genet 2016;48:827–37.
51. Yang JO, Oh S, Ko G, et al. VnD: a structure-centric database of disease-related SNPs and drugs. Nucleic Acids Res 2011;39:D939–44.
52. Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. Nucleic Acids Res 2000;28:235–42.
53. Kantardjieff K, Rupp B. Structural bioinformatic approaches to the discovery of new antimycobacterial drugs. Curr Pharm Des 2004;10:3195–211.
54. Chen M, Lin X, Zheng W, Onuchic JN, Wolynes PG. Protein folding and structure prediction from the ground up: the atomistic associative memory, water mediated, structure and energy model. J Phys Chem B 2016;120:8557–65.
55. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction: progress and new directions in round XI. Proteins 2016;82(Suppl 2):1–6.
56. Jacobson M, Sali A. Comparative protein structure modeling and its applications to drug discovery. Annu Rep Med Chem 2004;39:259–76.
57. Ma J, Wang S, Zhao F, Xu J. Protein threading using context-specific alignment potential. Bioinformatics 2013;29:i257–65.
58. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12:85–94.
59. Peng J, Xu J. Low-homology protein threading. Bioinformatics 2010;26:i294–300.

60. Petrey D, Chen TS, Deng L, et al. Template-based prediction of protein function. Curr Opin Struct Biol 2015;32:33–8.
61. Blair JMA, Bavro VN, Ricci V, et al. AcrB drug-binding pocket substitution confers clinically relevant resistance and altered substrate specificity. Proc Natl Acad Sci 2015;112:3511–6.
62. Vyas VK, Ghate M, Patel K, Qureshi G, Shah S. Homology modeling, binding site identification and docking study of human angiotensin II type I (Ang II-AT1) receptor. Biomed Pharmacother 2015;74:42–8.
63. Messaoudi A, Belguith H, Ben Hamida J. Homology modeling and virtual screening approaches to identify potent inhibitors of VEB-1 β-lactamase. Theor Biol Med Model 2013;10:22.
64. Ung PM-U, Song W, Cheng L, et al. Inhibitor discovery for the human GLUT1 from homology modeling and virtual screening. ACS Chem Biol 2016;11:1908–16.
65. Morya VK, Dung NH, Singh BK, Lee H-B, Kim E. Homology modelling and virtual screening of P-protein in a quest for novel antimelanogenic agent and In vitro assessments. Exp Dermatol 2014;23:838–42.
66. Fazi R, Tintori C, Brai A, et al. Homology model-based virtual screening for the identification of human helicase DDX3 inhibitors. J Chem Inf Model 2015;55:2443–54.
67. Forli S, Huey R, Pique ME, Sanner MF, Goodsell DS, Olson AJ. Computational protein-ligand docking and virtual drug screening with the AutoDock suite. Nat Protoc 2016;11:905–19.
68. Irwin JJ, Shoichet BK. Docking screens for novel ligands conferring new biology. J Med Chem 2016;59:4103–20.
69. Pyzer-Knapp EO, Suh C, Gómez-Bombarelli R, Aguilera-Iparraguirre J, Aspuru-Guzik A. What is high-throughput virtual screening? A perspective from organic materials discovery. Annu Rev Mater Res 2015;45:195–216.
70. Kumar V, Krishna S, Siddiqi MI. Virtual screening strategies: recent advances in the identification and design of anti-cancer agents. Methods 2015;71:64–70.
71. Lyne PD. Structure-based virtual screening: an overview. Drug Discov Today 2002;7:1047–55.
72. Irwin JJ, Shoichet BK. ZINC - A free database of commercially available compounds for virtual screening. J Chem Inf Model 2005;45:177–82.
73. Pence HE, Williams A. ChemSpider: an online chemical information resource. J Chem Educ 2010;87:1123–4.
74. Chen CY-C. TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico. PLoS One 2011;6:e15939.
75. Hatherley R, Brown DK, Musyoka TM, et al. SANCDB: A South African Natural Compound Database. J Cheminform 2015;7:29.
76. Musyoka TM, Kanzi AM, Lobb KA, Tastan Bishop Ö. Analysis of non-peptidic compounds as potential malarial inhibitors against Plasmodial cysteine proteases via integrated virtual screening workflow. J Biomol Struct Dyn 2016;34:2084–101.
77. Musyoka TM, Kanzi AM, Lobb KA, Tastan Bishop Ö. Structure based docking and molecular dynamic studies of plasmodial cysteine proteases against a South African natural compound and its analogs. Sci Rep 2016;6:23690.
78. Kumar A, Purohit R. Use of long term molecular dynamics simulation in predicting cancer associated SNPs. PLoS Comput Biol 2014;10:e1003318.
79. Gromiha MM, Selvaraj S. Inter-residue interactions in protein folding and stability. Prog Biophys Mol Biol 2004;86:235–77.
80. Grewal RK, Roy S. Modeling proteins as residue interaction networks. Protein Pept Lett 2015;22:923–33.
81. Atilgan AR, Turgut D, Atilgan C. Screened nonbonded interactions in native proteins manipulate optimal paths for robust residue communication. Biophys J 2007;92:3052–62.
82. Ozbaykal G, Rana Atilgan A, Atilgan C. In silico mutational studies of Hsp70 disclose sites with distinct functional attributes. Proteins 2015;83:2077–90.
83. Doshi U, Holliday MJ, Eisenmesser EZ, Hamelberg D. Dynamical network of residue—residue contacts reveals coupled allosteric effects in recognition, catalysis, and mutation. Proc Natl Acad Sci U S A 2016;113:4735–40.
84. Söding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 2005;33:W244–8.
85. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res 2003;31:3381–5.
86. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 2010;5:725–38.
87. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 2015;10:845–58.
88. Hatherley R, Brown DK, Glenister M, Tastan Bishop Ö. PRIMO: an Interactive homology modeling pipeline. PLoS One 2016;11:e0166698.
89. Tina KG, Bhadra R, Srinivasan N. PIC: protein interactions calculator. Nucleic Acids Res 2007;35:W473–6.
90. Vangone A, Spinelli R, Scarano V, Cavallo L, Oliva R. COCOMAPS: a web application to analyze and visualize contacts at the interface of biomolecular complexes. Bioinformatics 2011;27:2915–6.
91. Negi SS, Schein CH, Oezguen N, Power TD, Braun W. InterProSurf: a web server for predicting interacting sites on protein surfaces. Bioinformatics 2007;23:3397–9.
92. Nagarajan R, Archana A, Thangakani AM, Jemimah S, Velmurugan D, Gromiha MM. PDBparam: online resource for computing structural parameters of proteins. Bioinform Biol Insights 2016;10:73–80.
93. Laskowski RA. PDBsum: summaries and analyses of PDB structures. Nucleic Acids Res 2001;29:221–2.
94. Abraham MJ, Murtola T, Schulz R, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX 2015;1:19–25.
95. Phillips JC, Braun R, Wang W, et al. Scalable molecular dynamics with NAMD. J Comput Chem 2005;26:1781–802.
96. Liu X, Jian X, Boerwinkle E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 2011;32:894–9.
97. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat 2013;34:e2393–402.
98. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: a one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. Hum Mutat 2016;37:235–41.
99. Stenson PD, Mort M, Ball EV, Shaw K, Phillips AD, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum Genet 2014;133:1–9.
100. Higasa K, Miyake N, Yoshimura J, et al. Human genetic variation database, a reference database of genetic variations in the Japanese population. J Hum Genet 2016;61:547–53.
101. Reumers J, Schymkowitz J, Ferkinghoff-Borg J, Stricher F, Serrano L, Rousseau F. SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. Nucleic Acids Res 2005;33:D527–32.
102. Yue P, Melamud E, Moult J. SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics 2006;7:166.
103. Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res 2005;15:978–86.
104. Wainreb G, Ashkenazy H, Bromberg Y, et al. MuD: an interactive web server for the prediction of non-neutral substitutions using protein structural data. Nucleic Acids Res 2010;38:W523–8.
105. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 2009;25:2744–50.

106. Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 2009;30:1237–44.

107. Parthiban V, Gromiha MM, Schomburg D. CUPSAT: prediction of protein stability upon point mutations. Nucleic Acids Res 2006;34:W239–42.

108. Yin S, Ding F, Dokholyan NV. Eris: an automated estimator of protein stability. Nat Methods 2007;4:466–7.

109. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC Bioinformatics 2011;12:151.