Briefings in Bioinformatics, 22(1), 2021, 3–19

doi: 10.1093/bib/bbz146
Advance Access Publication Date: 8 December 2019
Review article

Molecular dynamics simulations for genetic interpretation in protein coding regions: where we are, where to go and when

Juan J. Galano-Frutos, Helena García-Cebollada and Javier Sancho

Corresponding author: Javier Sancho, Departamento de Bioquímica y Biología Molecular y Celular. Facultad de Ciencias. Universidad de Zaragoza. Pedro Cerbuna 12, Zaragoza 50009, Spain; Biocomputation and Complex Systems Physics Institute (BIFI), Joint Units BIFI-IQFR (CSIC) and GBs-CSIC, Zaragoza 50018, Spain; Aragon Health Research Institute (IIS Aragón), Zaragoza 50009, Spain; Tel.: (+34) 976 761286; Fax: (+34) 976 762123; E-mail: jsancho@unizar.es

Abstract

The increasing ease with which massive genetic information can be obtained from patients or healthy individuals has stimulated the development of interpretive bioinformatics tools as aids in clinical practice. Most such tools analyze evolutionary information and simple physical–chemical properties to predict whether replacement of one amino acid residue with another will be tolerated or cause disease. Those approaches achieve up to 80–85% accuracy as binary classifiers (neutral/pathogenic). As such accuracy is insufficient for medical decision to be based on, and it does not appear to be increasing, more precise methods, such as full-atom molecular dynamics (MD) simulations in explicit solvent, are also discussed. Then, to describe the goal of interpreting human genetic variations at large scale through MD simulations, we restrictively refer to all possible protein variants carrying single-amino-acid substitutions arising from single-nucleotide variations as the human variome. We calculate its size and develop a simple model that allows calculating the simulation time needed to have a 0.99 probability of observing unfolding events of any unstable variant. The knowledge of that time enables performing a binary classification of the variants (stable-potentially neutral/unstable-pathogenic). Our model indicates that the human variome cannot be simulated with present computing capabilities. However, if they continue to increase as per Moore's law, it could be simulated (at 65°C) spending only 3 years in the task if we started in 2031. The simulation of individual protein variomes is achievable in short times starting at present. International coordination seems appropriate to embark upon massive MD simulations of protein variants.

Key words: genetic interpretation; mutation-effect prediction tools; single amino acid variation; molecular dynamics simulation; protein stability; large-scale phenotype prediction

Introduction

Many diseases have a genetic component. The Online Mendelian Inheritance in Man (OMIM) database (https://www.omim.org/) reports so far (last update: 16 June 2019) 7129 phenotypes associated to genetic disorders in humans, which can be passed on from generation to generation. Over the last decades, an increasing number of disease-causing mutations have been identified thanks, in part, to efforts by international collaborative large-scale sequencing initiatives such as the Human Genome Project [1, 2], the Human Variome Project [3], the ENCODE Project [4], the HapMap Project [5] or the 1000 Genomes Project [6] among others, which have paved the way to more recent

Juan J. Galano-Frutos, PhD, is a post-doctoral researcher working on molecular modeling, computational drug discovery and bioinformatics at the 'Protein Folding and Molecular Design (ProtMol)' group at BIFI, University of Zaragoza.
Helena García-Cebollada is a PhD fellow with a Degree in Biotechnology, member of the BIFI, working on bioinformatics issues at the 'Protein Folding and Molecular Design (ProtMol)' group.
Javier Sancho is a Biochemistry Professor. His laboratory investigates protein stability using experimental and computational approaches in order to improve rational protein stabilization strategies and existing genetic interpretation tools.
Submitted: 23 July 2019; **Received (in revised form):** 22 September 2019

Table 1. Summary of up-to-date (16 June 2019) statistics for *Homo sapiens* provided by the main biological databases referred in this manuscript

Database (URL)	Feature	Entries or report
Online Mendelian Inheritance in Man (OMIM) (https://www.omim.org)	Genetic disorders	7129
NCBI's dbSNP Short Genetic Variations Database (https://www.ncbi.nlm.nih.gov/snp)	SNPs	605 048 595
RCSB Protein Data Bank (https://www.rcsb.org/)	3D protein structures	1559
Ensembl (Human GRCh38.p12) (https://www.ensembl.org/Homo_sapiens/Info/Annotation)	Protein-coding genes	20 465
Human Gene Mutation Database (HGMD Professional 2019.1) (http://www.hgmd.cf.ac.uk/ac/index.php)	Mutations	256 070

initiatives, e.g. the Cancer Genome [7] and the 100,000 Genomes [8] projects. According to the NCBI's dbSNP Short Genetic Variations database (build 151) [9], ~1 billion uniquely mapped (non-redundant) human genetic variations have been reported, including single- and multiple-base variations, insertions and deletions (indels) and short tandem repeats. Advances in high-throughput sequencing techniques [10–12] have enabled whole-exome sequencing (WES) [13] and genome-wide association studies (GWAS) [14] on humans [15–17] and a number of model species [18–21], contributing to a better understanding of how single-nucleotide variations (SNVs), the most frequent variations present in DNA [22], are related to diseases. An accurate interpretation of SNVs constitutes a major challenge in genetics and health.

Genetic variation, disease, protein structure and the proteome

Genes contain the information required to produce proteins, the main components of the cellular machinery, combining 20 essential building blocks called amino acids. Changes in the amino acid sequence of proteins arise from variations in protein-coding regions of the DNA, including synonymous and non-synonymous single-nucleotide variations (sSNVs and nsSNVs, respectively), as well as indels and multi-nucleotide variations (MNVs). Regarding their effects on individual phenotypes, frameshifts and non-sense SNVs are likely to produce null or totally dysfunctional proteins. The effects of missense nsSNVs, in-frame indel variations or MNVs are more difficult to assess. They may range from not affecting protein function to causing severe dysfunctionality of the encoded protein variants, but they can also enhance the original function of the protein [23, 24] or even drive the acquisition of a new one [25]. Missense and in-frame indel variations may affect not only protein folding and the stability of the native protein conformation [26–28] but also protein expression [29], post-translational modification [30, 31] or binding affinity [32–34]. Typically, missense variations are described in a binary way as being deleterious/pathogenic/damaging or tolerated/neutral/benign. In monogenic disorders, deleteriousness tends to be associated to loss of structural stability [35, 36] while impaired binding interactions with partners or cofactors [34, 37, 38] might be frequent in more complex diseases where disease-associated variations may not affect protein stability so often [39].

As the three-dimensional structure of a protein determines its function, the availability of protein structures solved at atomic level can greatly facilitate understanding which mutations impact protein function and why [40–42]. Knowledge of the structure allows to perform a variety of *in silico* analyses, such as the calculation of structure-based properties [43, 44], the study of protein dynamics using methodologies such as molecular dynamics (MD) simulations [45] or the combined application of molecular docking [46] and MD approaches [47] to uncover functional issues related to protein/cofactor or protein/protein interactions. Advances in resolution of 3D protein structures over the last decades, including the efforts by international consortia [41, 48, 49], have gathered a wealth of structural information that has laid the foundation for proteome-wide structural analyses. To date (16 June 2019), 149 518 structures including wild-type proteins, mutated variants, protein domains, protein/cofactor, protein/protein and protein/nucleic acids complexes are available in the Protein Data Bank (PDB) (https://www.rcsb.org/stats/growth/overall). Within this wealth of structural information, human proteins are the larger group. Still, those solved at atomic resolution (around 1559) constitute a small percentage of the estimated number of human genes (~20 400, see Table 1), which in turn represents a lower limit to the still unclear size of the human proteome [50, 51].

Human genetic variation space: the variome

The combination of technological advances with GWAS has allowed large-scale identification of human variations (~90% corresponding to SNV) (https://www.ncbi.nlm.nih.gov/snp/?term=human). A subset of them, responsible for human inherited disease, is collected in the Human Gene Mutation Database (http://www.hgmd.cf.ac.uk/ac/index.php) which, in its latest professional version as of June 2019, reports 256 070 entries. Over 57% of these variations correspond to SNVs (missense/nonsense). Most probably, they only account for a small fraction of all clinically relevant genetic variants present in human genomes. Although mutations associated with complex diseases seem often to arise in non-coding regions [52], most well-annotated genetic diseases are linked to coding variants [53], which—fortunately—are the more amenable to structural analysis. Knowing the size of the human genetic variation space may be useful. If the functional implications of all possible human nsSNVs could be accurately calculated, the interpretation of those already reported and of those being found in the future would be greatly speeded up. In this respect, the number of nsSNVs that can exist, and therefore be found sooner or later in a given gene, can be computed easily. In previous work [54], we showed that a small 37-residue-long repeat of the LDL receptor could give rise to 227 nsSNVs, including many that had been related to familial hypercholesterolemia. The number of hypothetical nsSNVs for a given gene product can be calculated from the number of possible nsSNVs associated to each codon ($nsSNVc_i$) and the absolute frequencies of codons in the secondary transcript (see Supporting Information Table S1). Likewise, the complete nsSNV space for all the human genes, human_nsSNVome (HumanV for short), can be accurately calculated from the $nsSNVc_i$ values together with the absolute frequencies of codons in coding regions, fc_i.

$$\text{HumanV} = \sum_{i=1}^{61} fc_i * nsSNVc_i \quad (1)$$

Those frequencies can be obtained from the HIVE online resource [55] (see Supporting Information Table S2) and the canonical protein sequences downloaded from UnitProt after a search filtered by 'reviewed:yes AND organism: "*Homo sapiens* (Human) [9606]"'). A total of 66474822 nsSNVs are estimated this way, which will be referred to as the theoretical human non-synonymous single nucleotide variome (t-human_nsSNVome; t-HumanV, for short) as opposed to the sub-space of it that has been already described, which will be referred to as the d-HumanV.

Protein interaction space: the interactome

Amino acid variations occurring at hotspots of protein binding interfaces can dramatically affect binding affinity. The STRING database [56, 57] includes 19 257 human proteins for which protein–protein network connections have been reported (out of the ~20 400 human genes reported so far). In humans, each protein participates in about 3–10 protein–protein interactions (PPIs) [58]. In this sense, in-depth characterization of protein–protein interaction networks (PPINs) is crucial to understanding cellular pathways and devising strategies to effectively treat human illnesses [59–61]. The complete set of PPI taking place in a defined biological context (organelle, cell, organism, etc.) constitutes its interactome. Large-scale PPI screening techniques, in particular the yeast two-hybrid method (Y2H) [62], have allowed to uncover complete interactomes in a number of species including humans [63]. Very recently, proteome co-evolutionary methods have also shown an impressive capacity of working with millions of protein pairs to systematically identify PPIs on the whole-proteome scale [64]. In this context, current estimations of the human interactome range from 130 000 [65] to ~650 000 PPI [66], excluding trans-organism PPI (relevant for infectious diseases).

A growing number of computational tools used for prediction of PPI and PPIN have been released over the last few years, e.g. iLoops [67] HOMCOS [68], COTH [69], InterPreTS [70] and PRISM [71]. Also, databases storing annotated and/or predicted PPI are currently available, e.g. DIP [72], BIND [73], PrePPI [74] and STRING [75]. One way to identify key residues involved in PPI, to then establish mutation/function relationships, is to focus on the identification of protein–protein interaction sites (PPISs). Prediction of PPIS is facilitated by recently developed computational resources such as predPPIS [76] and IntPred [77], as well as by many others previously released [78]. The algorithms implemented in PPI/PPIN and PPIS predictive tools utilize methods based on protein sequence data, structure data or a combination of these (hybrid approaches). Despite the analytical progress condensed in those applications, which use a variety of mathematical–statistical methods (e.g. support vector machine, random forest, neural network, Bayes, hidden Markov model), the accuracy achieved barely reach 80% in the best cases [67–71, 76–78]. This is likely due to the still very low number of structures of protein complexes available in the PDB. Therefore, the applicability of this type of predictive tools to clinical diagnosis and therapeutics is still limited.

Approaches in use for the interpretation of mutation effects at the protein level

Much of what we know about mutation-driven effects on protein stability, folding or protein–protein interactions has been revealed through mutational studies based on individual amino acid changes. Recently, deep mutational scanning (DMS), a high-scale DNA-sequencing-based method, is becoming an invaluable tool for experimental evaluation of missense variants [79–81]. DMS aims at testing in a single, multiplexed assay the effects of hundreds or thousands of variations by focusing on the presence of a target property (e.g. cell growth, presence of fluorescent reporter or ligand binding) in a large library of variants [79–81]. However, the proteome-wide DMS-based approach faces difficulties, such as the need for specific assays or the complexity of the equipment needed.

The interpretation of mutation effects at the protein level can also be approached in a predictive scenario using computational methods [22, 82]. Many computational tools have been developed over the last 20 years for the prediction of mutation effects (mutation-effect prediction tools, MEPT), some of which have become quite popular (e.g. SIFT [83–86], PolyPhen2 [87] or CADD [88], see Table 2).

As PPI/PPIN/PPIS prediction tools, most MEPT utilize methods based on evolutionary conservation of sequence (homology), structure and structure-derived data (structure-based) or a combination of the two (hybrid approaches). Some (meta-predictors) combine outputs from several MEPTs to provide consensus scores (Figure 1 and Table 2). Representing each of these groups, SIFT (homology) is based on 'the degree of conservation of amino acid residues in sequence alignments derived from closely related sequences, collected through PSI-BLAST' [85]; HoTMuSiC (structure-based) 'uses standard and temperature-dependent statistical potentials combined with an artificial neural network to predict the change in melting temperature ΔT_m upon point mutations' [89]; Polyphen2 (hybrid) 'uses eight sequence-based and three structure-based predictive features which were selected automatically by an iterative greedy algorithm' [87]; and Meta-SNP (meta-predictor) provides a consensus score based on SNAP, SIFT, PhD-SNP and PANTHER outputs [90–92]. MEPTs based on structural data often try to calculate $\Delta\Delta G$, the change in free energy (i.e. in conformational stability) brought about by a single-amino-acid variation. For a comprehensive review on protein stability and folding principles, or for an explanation of the way to determine experimentally $\Delta\Delta G$, see references [93] and [94], respectively.

Although most existing MEPTs only provide predictions for single-amino-acid variants (SAVs), some recently developed applications, e.g. PROVEAN [95], are already able to generate predictions for multiple-amino-acid substitutions, insertions and deletions. While this progress is encouraging, and most published MEPTs are freely available, the maximum accuracy achieved by these methodologies so far (~80–85%) [96] strongly limits their usage in clinical diagnosis. One of the major caveats of current MEPTs' development is the limited size of the datasets available for training and validating these tools. In this context, DMS studies [79–81] and novel developments [97] are expected to provide MEPTs' community with validation and training sets larger than those currently available, which will boost the next generation of MEPTs.

Protein dynamics in interpretation of mutations

Perhaps due to the limited number of human protein structures available at the beginning of the MEPT era, most of these applications rely heavily on protein sequence data analysis (Table 2). Even MEPTs using hybrid approaches commonly include more sequence-based prediction features than structure-based ones. On top of that, many of the structure-based features selected as

Table 2. Sample of Mutation-Effect Prediction Tools (MEPT) currently available

MEPT	Year	Input	Source/URL	Reference(s)	Citations
Structural					
FoldX	2005	PDB files or IDs	http://foldxsuite.crg.eu/	[131]	776
ANGDelMut	2013	Amino acid variation and attributes that distinguish deleterious mutations	http://bioschool.iitd.ernet.in/DelMut/	[117]	2
SDS	2014	DNA query or protein query	http://sds-p.cig.biology.gatech.edu/	[174]	8
ENCoM	2015	PDB file or ID	https://labworm.com/tool/encom	[112]	35
HoTMuSiC	2015	PDB ID and T_m when available	https://soft.dezyme.com/	[89]	17
MODICT	2016	PDB files from 3D protein models	https://github.com/IbrahimTanyalcin/MODICT	[175]	0
SDM2	2017	PDB file or ID	http://marid.bioc.cam.ac.uk/sdm2	[130]	19
mCSM–NA	2017	PDB file or ID and variation(s)	http://biosig.unimelb.edu.au/mcsm_na/	[176]	10
DynaMut	2018	PDB file or ID and variation(s)	http://biosig.unimelb.edu.au/dynamut/prediction	[118]	5
Homology					
SIFT	2001	NCBI GI number, SNP ID(s), or alignment	http://sift.jcvi.org/	[83–86]	3408
SNPs&GO	2009	Protein sequence	http://snps.biofold.org/snps-and-go/snps-and-go.html	[177, 178]	297
PANTHER	2003	Protein sequence, variation(s), organism	http://www.pantherdb.org/tools/csnpScoreForm.jsp	[179]	460
MAPP	2005	FASTA alignment and phylogenetic tree	http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html	[180]	189
PhD-SNP	2006	Swiss-Prot code or protein sequence, and variation	http://snps.biofold.org/phd-snp/phd-snp.html	[181]	346
Align-GVGD	2006	FASTA alignment and variation(s)	http://agvgd.hci.utah.edu/	[182]	372
Parepro	2007	FASTA alignment	http://www.mobioinfor.cn/parepro/	[183]	33
CHASM	2009	Passenger mutation rates and a list of variations	http://wiki.chasmsoftware.org/index.php/Main_Page	[184]	231
MutationTaster2	2010	Transcript, gene, or ORF	http://www.mutationtaster.org/	[185, 186]	1419
mutationassessor	2011	UniProt protein accession or a RefSeq protein ID and variation(s)	http://mutationassessor.org/	[187]	676
PROVEAN	2012	FASTA protein sequence and variation(s)	http://provean.jcvi.org/index.php	[95, 188, 189]	970
AnnTools	2012	VCF and pileup files, and user specified tabular formats	http://anntools.sourceforge.net/	[190]	25
FATHMM	2013	SwissProt/TrEMBL, RefSeq or Ensembl protein IDs, and variation(s)	http://fathmm.biocompute.org.uk/	[191]	338
Evolutionary Action	2014	UniProt protein accession or gene name and a variation	http://mammoth.bcm.tmc.edu/uea/	[192]	38
PON-P2	2015	Protein/gene ID(s) and variation(s)	http://structure.bmc.lu.se/PON-P2/	[193]	60
Hybrid					
SNPeffect	2005	FASTA protein sequence or PDB file (also PDB ID or UniProt ID), and variation(s)	http://snpeffect.switchlab.org/	[104]	105
LS-SNP/PDB	2005	PDB, gene, dbSNP rs, Kegg pathway or UniProt IDs, or genomic region	http://ls-snp.icm.jhu.edu/ls-snp-pdb/	[194, 195]	159

(Continued)

Table 2. Continued

MEPT	Year	Input	Source/URL	Reference(s)	Citations
nsSNPAnalyzer	2005	FASTA protein sequence, PDB file (optional) and variation(s)	http://snpanalyzer.uthsc.edu/	[99]	117
MUpro	2005	Protein sequence, PDB file (optional)	http://mupro.proteomics.ics.uci.edu/	[103]	329
Pmut	2005	FASTA, Ensembl, dbSNP, UniProt or PDB IDs, and variation(s)	http://mmb.irbbarcelona.org/PMut/	[196, 197]	347
SNPs3D	2006	SNP rs ID, sequence accession number, or gene ontology (GO)	http://www.snps3d.org/	[198]	331
SNAP	2007	FASTA protein sequence	https://www.rostlab.org/services/SNAP/	[102, 199]	420
I-mutant3.0	2007	Protein sequence, mutation, PDB ID	http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi	[200]	141
AUTO-MUTE 2.0	2010	PDB ID(s) and variation(s)	http://binf.gmu.edu/automute/	[134, 135]	47
PolyPhen2	2010	Protein identifier or sequence and mutation query	http://genetics.bwh.harvard.edu/pph2/	[87]	5829
SNPs&GO^{3d}	2011	PDB ID	http://snps.biofold.org/snps-and-go/snps-and-go-3d.html	[201]	46
HANSA	2012	FASTA protein sequence or GI, GenBank, REFSEQ, SWISSPROT or PDB ID	http://www.cdfd.org.in/HANSA/	[98, 202]	33
CADD	2014	VCF file	http://cadd.gs.washington.edu/score	[88]	1660
VarMod	2014	FASTA protein sequence, UniProt ID (optional), and variation(s)	http://www.wasslab.org/varmod/	[105]	9
ELASPIC	2014	UniProt or PDB ID	http://elaspic.kimlab.org/	[101]	26
SuSPect	2014	UniProt accessions, chromosomal locations, VCF file, FASTA sequence or PDB file.	http://www.sbg.bio.ic.ac.uk/suspect/index.html	[107]	66
VIPUR	2016	FASTA protein sequence or PDB file, and variation(s)	https://osf.io/bd2h4/	[100]	10
MutPred2	2017	FASTA protein sequence or file, *P* value threshold (optional)	http://mutpred.mutdb.org/	[106]	19
Meta-predictor					
Condel	2011	List of variant(s) in a specific format	http://bbglab.irbbarcelona.org/fannsdb/	[91]	419
CanDrA	2013	TXT file with chr. number, gene coordinate, and ref. & mut. alleles	http://bioinformatics.mdanderson.org/main/CanDrA	[92]	31
Meta-SNP	2013	FASTA protein sequence and variation(s)	http://snps.biofold.org/meta-snp/	[90, 203]	332
DUET	2014	PDB file or ID and variation(s)	http://structure.bioc.cam.ac.uk/duet	[129]	133

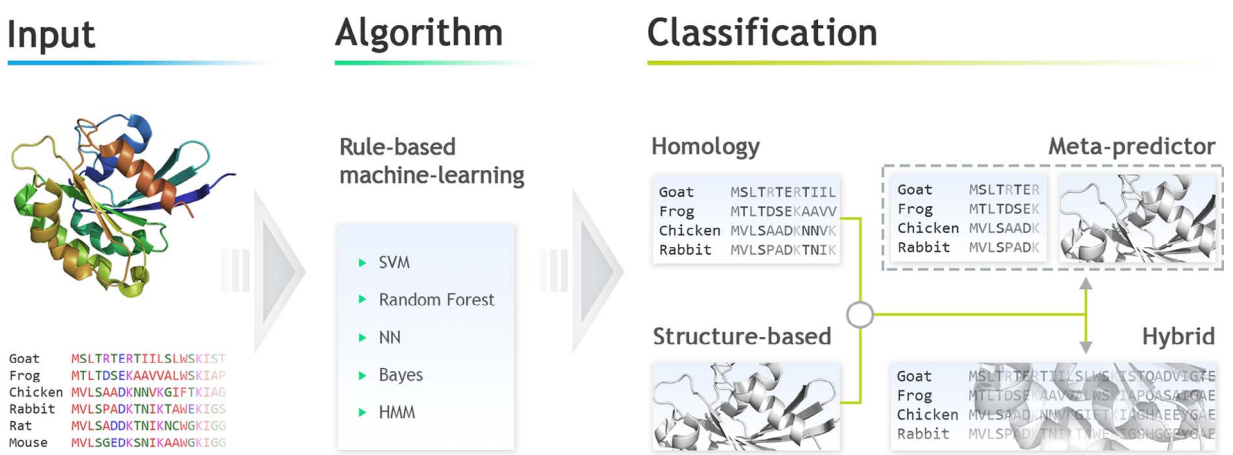

Figure 1. Overview of the methods underlying Mutation-Effects Prediction Tools (MEPTs). Homology methods rely on evolutionary conservation of protein sequence; structure-based ones utilize 3D structure or 3D-structure-derived data to train their algorithms (SVM: support vector machine; NN: neural networks; HMM: hidden Markov models); hybrid methods combine both sequence conservation and structural data; and meta-predictors obtain consensus scores based on outputs by individual MEPT.

predictors in MEPTs are obtained from either homology models or directly predicted from sequence rather than from true structural information [98–107]. This issue is being addressed in more recently developed MEPTs with the introduction of predictive features based on known 3D protein structures and on the usage of larger training sets including experimentally solved 3D protein structures. Nevertheless, such efforts have not led to significant increases in MEPTs' accuracies so far [108].

Looking for new approaches to enhance MEPTs' reliabilities, researchers have begun to exploit the availability of 3D protein structures to include dynamic aspects of proteins in the prediction of mutation effects [40, 47, 54, 109–120]. As protein function is defined by both structure and dynamics [121–123], dynamic patterns have begun to be recognized as descriptive of proteins [124, 125]. To our knowledge, DynaMut [118] and ENCoM [112] are the only available general (non protein-specific) MEPTs that include assessment of protein dynamics for predicting the impact of mutations on protein stability. Those two MEPTs apply normal mode analysis (NMA) simulations [126], a coarse-grained version in the case of ENCoM. According to the authors, considering destabilizing mutations alone, DynaMut comparably performs to methods such as mCSM-NA (the current version of mCSM [127]), I-mutant3.0 (the current version of I-mutant2.0 [128]) and DUET [129] and outperforms ENCoM [112] and SDM2 [130]. On the other hand, ENCoM's authors compared this tool to several existing methods, i.e. FoldX [131], HoTMusic (an improved version of PoPMusic [132] whose current last version is PoPMuSiC^{Sym} [133]), AUTO-MUTE [134, 135] and I-mutant2.0 [128], and concluded that 'ENCoM proved to be the most self-consistent and least biased method' but 'not the best overall predictive method when considering both stabilizing and destabilizing mutations together' [112]. Overall, even though many available MEPTs may show overestimated accuracies and bias, and comparisons between them by authors are at times misleading, DynaMut and ENCoM do not seem to overcome the maximum performances obtained by the best predictor with which they are compared. One reason for this may be that, although NMA-based methods add dynamics elements to static-structure-based methods, they do not sample sufficiently the conformational landscape. Also of note, the NMA simulations in those methods are performed in absence of water, disregarding fundamental solvation interactions.

MD simulation is a powerful, reliable tool used to study protein stability, dynamics or function [136, 137]. Very recently, its usefulness in accurately calculating protein folding energetics has been shown [138]. The possibility offered by MD simulation to explore the conformational energy landscape of proteins in very realistic settings, including explicit solvent molecules and specific solution conditions (e.g. temperature, pH, concentration, pressure), alongside the continuous progress made in related areas, such as the development of force fields [139], water models [140] and next-generation graphics processing units (GPUs) (high-performance computing), makes this tool a logical choice to address the analysis of SAV. In this context, two different situations can arise when MD is used to simulate protein mutation effects. In the infrequent case where both the native and mutant structures are known, unfolding simulations can be performed taking the corresponding structural models as the starting points. In the usual case where only the wild-type structure is available, a starting model of the mutant structure should be constructed by modeling the single-amino-acid substitution. Then, the evolution in time of the modeled mutant structure would be simulated by MD. We refer to this approach as relaxation molecular dynamics (rMD) simulations because they describe the relaxation of the protein native structure after the introduction of a mutation.

Relaxation molecular dynamics

Given the artifactual nature of the starting model, the pertinence of this approach may be arguable. Nevertheless, a number of rMD-based approaches have been used to study mutation effects on phenotypes in the last few years [47, 54, 110, 116, 117, 119]. Many of these works showed quite good correlations between their predictions and experimental data and, at the same time, allowed to extract meaningful insights underlying the mechanisms through which mutations impair protein function. In one of these works, a web-based tool, ANGDelmut, used for the prediction and analysis of functional loss mechanisms of deleterious mutations related to amyotrophic lateral sclerosis (ALS), has been made available (http://bioschool.iitd.ernet.in/DelMut/) [117]. This is—as far as we know—the only currently available tool for prediction of SAV effects relying on a classical atomistic MD approach. However, ANGDelmut is implemented only for the

analysis of angiogenin, the target protein in ALS. Simulations performed in [117] use an implicit-solvent MD approach that is faster than explicit-solvent ones but may not describe in full the important solvation effects associated to the protein dynamics.

The promising results obtained in works where classical atomistic MD simulations with explicit water were performed [47, 54, 110, 116, 119] reinforce the idea that using MD-based methods under realistic conditions allows to accurately model and reveal the complexity of the structural changes associated to point mutations in proteins. In previous work, we followed this approach to explore the complete t-HumanV of the low-density lipoprotein receptor (LDL-r) LA5 domain by performing all-atom rMD simulations in explicit solvent. The LA5 domain is a 37-residue repeat that plays a key role in the uptake of LDL particles from the blood plasma and in their release in the endosome. The LA5 domain concentrates the highest rate of mutations reported as disease-causing for familial hypercholesterolemia (FH) [54]. A stability analysis based on principal component analysis (PCA) of MD trajectories, combined with PPI information of the binding site [141, 142], allowed to satisfactorily predict the pathogenicity of 49 out of the 50 FH mutations known by the time and to obtain a higher true positive rate than that provided by PMUT, Condel and PolyPhen2. The reliability of this approach raises hopes that using MD-based methodologies to address the prediction of SAV's deleteriousness will greatly contribute to obtain higher accuracies than those shown by currently available MEPT. Likewise, it demonstrates that performing integrative approaches capable of exploiting PPI information is a way to address the prediction of mutation effects in a reliable manner, which is required if prediction is to be applied to diagnosis. In the above example [54], short MD simulations (20 ns productive phase) sufficed to capture structural differences between wild-type-like and pathogenic mutant conformational ensembles and so to unveil the (un)compatibility of the different variants with the native conformation of the small, 37-residue LA5 domain. However, it is quite likely that the simulation time span required to observe structural disruption in rMD trajectories of larger proteins will be larger or may vary considerably depending on protein size [143], on the intrinsic dynamics determined by protein folds or on specific folding/unfolding mechanisms [125, 143–145]. Furthermore, the PCA-based analysis method used in the study of the LA5 domain may or may not be adequate for larger proteins.

Problems and challenges to address

The unavailability of MEPT integrating MD simulations onto SAV-predictive approaches is clearly linked to the high computational cost of the task. The still insufficient computing power and the lack of novel algorithms that could further accelerate full-atom MD simulations constitute a challenge for the expansion and generalization of their use for massive prediction of SAV's deleteriousness. Many approaches have been suggested or implemented in order to speed up atomistic MD simulations. Temperature or pressure raising [146–148], the inclusion of molecules of denaturing agents such as urea [149, 150] and the application of force [151, 152] are options often chosen in simulations thought for unfolding. From a more general perspective, the use of higher or mixed time-step algorithms [153], split GPU–CPU algorithms [154], force fields and MD programs optimized for GPU architectures [155], special-purpose hardware designed specifically for MD simulations (a prominent case being the Anton platform [156]), implicit solvent [157] or the incorporation of multiscale modeling algorithms [158–160] may also help. Such advances have not been systematically implemented in SAV-predictive approaches, and at present, the computational cost associated to using rMD to provide accuracy to genetic diagnostic in a variome-wide scale continues being too high. Yet, the increasing number of SNVs being identified and of proteins with known 3D structure available argues for the consideration of that goal as a worthy endeavor claiming for new concerted efforts.

There is also a challenge in performing quantitative analysis of MD trajectories in simulations of protein folding/unfolding events [161], specially in large-scale simulation projects. At present, a broad set of analyses intended for detecting protein conformational changes in MD trajectories are available, which often focus on basic root mean square deviations (RMSDs) and root mean square fluctuations (RMSFs) of atomic positions, gyration radius, secondary structure content, template modeling score (TM score), principal component analysis (PCA), percentage of native contacts, solvent-accessible surface area (SASA), hydrogen bonds number, phi and psi backbone angles, distances, etc. [162–164]. Albeit promising results have been obtained using these analysis methods, as in the aforementioned study on the LA5 domain [54], it is difficult to establish robust metrics to quantitatively assess conformational changes from trajectory analyses and, therefore, to accurately determine the impact of mutations on protein stability. In this context, there is a developing research field relying on the so-called higher-order statistics methods wherein the semi-automated analysis of large sets of data, e.g. from long MD simulations, is intended to efficiently and accurately detect conformational changes in molecular ensembles [165]. As this area is still in an early stage of development, a close collaboration between structural biologists, programmers and developers of MD tools is needed.

While protein structure prediction methods continue to improve [166], a high-resolution 3D all-atom protein structure remains the starting point of choice required to reliably perform full-atom MD simulations. In this sense, the current availability of 3D structures of human proteins still constitutes a limitation to the expansion of MD usage in SAV-predictive approaches. Related to this is the question of how many proteins constitute the human proteome and, therefore, how many simulations need to be done to analyze the complete t-HumanV once their structures are solved. The answer is not trivial. Although the Human Genome Project revealed a lower-than-anticipated number of genes, ~20 400, a much higher number of protein species (proteoforms) can be produced in an individual from alternatively spliced RNA transcripts and from post-translational modifications [51, 167]. Besides, a variety of additional chemical entities including, for instance, cofactors or ions, as well as acetylated, phosphorylated and/or methylated residues, may appear in the scene, which increases the complexity of the task.

Estimation of computational resources needed to perform rMD simulations on the complete HumanV at 25°C

Performing rMD for the whole t-HumanV, as it has been defined in a previous section, is a formidable task, and the time required with present and foreseeable computational capabilities should be estimated in order to decide when embarking upon the calculation would be appropriate. As explained above, the t-HumanV comprises 66 474 822 nsSNV. Thus, the real time (T) required to simulate all the variants arising in the 20 410 different proteins (canonical transcripts) and being able to provide an accurate

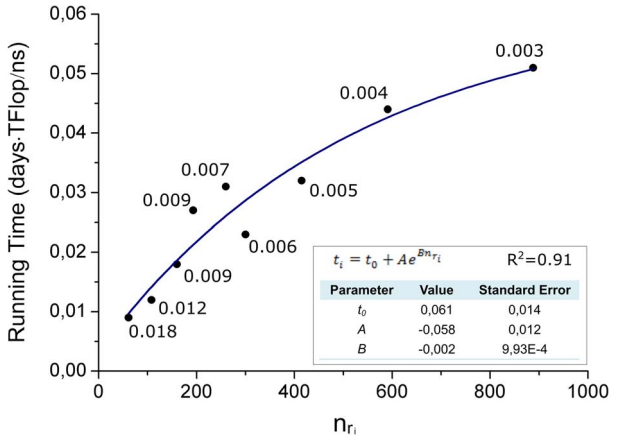

Figure 2. Running times (in days · TFlop/ns) versus number of residues (n_{ri}) fitting for nine proteins (see Table S3) simulated (three replicas of each protein) using the CHARMM27 force field, the TIP3P water model, an octahedron solvation box (with 1 nm minimum distance between protein surface and the end of the solvation box) and 48 Intel Xeon E5-2680v3 2.5GHz cores. Values indicated next to the points were calculated by dividing the radius of gyration of the protein by its number of residues (a structural sparseness measure, Table S3). These values decreasing with protein length leads to a lower number of water molecules per residue needed for the solvation box of larger proteins. The inset displays the fitting equation, the fitted parameters, their standard errors and the squared Pearson coefficient.

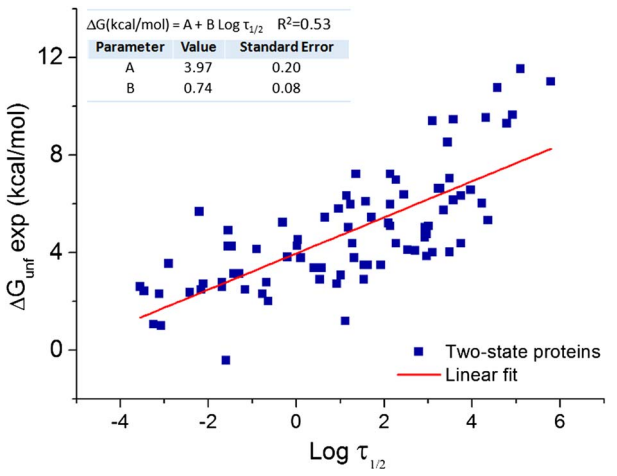

Figure 3. Correlation plot between protein conformational stability and the logarithm of unfolding half-lives showing a linear fit and the fitting parameters. The protein conformational stabilities have been calculated from folding and unfolding rate constants ($\Delta G_{unf} = -RT \times \ln (k_u/k_f)$) of 89 two-state proteins (Manavalan *et al.* dataset) normalized at 25.0°C. [168]. Half-lives have been calculated as $\tau_{0.5} = -\ln 2/k_u$. The inset displays the fitting equation, the fitted parameters, their standard errors and the squared Pearson coefficient.

binary prediction can be expressed as

$$T = \sum_{i=1}^{20\ 410} \sum_{j=1}^{n_{r_i}} nsSNVc_j \times \frac{t_i}{TFlops} \times \tau_{x_{unf}} \quad (2)$$

where is the number of amino acid residues of protein i, t_i is the running time required to simulate a unitary time span (e.g. 1 ns) of protein *i* using 1 TFlop of computing power, TFlops is the number of TFlops used in the calculation and $\tau_{x_{unf}}$ is the simulation time span (in ns) required to arrive to a given molar fraction (x_{unf}) of unfolded molecules (or to have a probability x_{unf} of observing unfolding in a single simulation, see below).

To obtain an estimation of the term t_i in Equation 2, we have simulated nine roughly spherical globular proteins, containing from 60 to 900 residues (see Supporting Information Table S3). The simulations (300 ns long; three replicas of each protein) have been done using octahedral solvation boxes affording for a 1 nm minimum distance from protein surface atoms to the end of the solvation box, using the CHARMM27 force field, TIP3P waters and 48 Intel Xeon E5-2680v3 2.5GHz cores. Values of t_i are processor-type-independent as they are calculated for 1 TFlop (the number of TFlops is explicitly introduced in Equation 2). These values as a function of protein length (Supporting Information Table S3) are shown in Figure 2, fitted to an exponential function (Equation 3):

$$t_i = t_0 + A \times e^{B \times n_{r_i}} \quad (3)$$

where t_0, *A* and *B* are fitting constants. The reasonably good fit of the experimental data to this simple function allows to estimate the time needed to simulate 1 ns for the different proteins of the proteome as: $t_i = 0.061 - 0.058e^{-0.002n_{r_i}}$, although proteins with non-spherical global shapes may require longer times due to the increase in the number of water molecules conforming the solvation shell in the rMD simulations.

When it comes to the term $\tau_{x_{unf}}$ in Equation 2, we have used normalized experimental data of folding (k_f) and unfolding (k_u) rate constants of 89 two-state proteins [168] to calculate their conformational stabilities (ΔG_{unf}) from the $\Delta G_{unf} = -RT \times \ln (k_u/k_f)$ relationship. Then, we have obtained an approximately linear correlation between the conformational stabilities of the protein and the logarithm of their half-lives of unfolding ($\tau_{0.5}$) (Equation 4):

$$\Delta G_{unf} = 3.97 + 0.74 \times \log \tau_{0.5} \quad (4)$$

indicating that more stable proteins take longer to unfold (see Figure 3). If conformational stabilities directly determined by chemical denaturation using the linear extrapolation method are used instead for calculating the correlation, a similar equation yielding slightly shorter unfolding times can be obtained (Supporting Information Figure S1).

Furthermore, the time required ($\tau_{x_{unf}}$) to obtain a given fraction (x_{unf}) of unfolded molecules starting from a population of fully folded proteins can be described as

$$\tau_{x_{unf}} = -\tau_{0.5} \times \ln (1 - x_{unf}) / \ln 2 \quad (5)$$

The conformational stability of a protein determines the percentage of folded molecules at equilibrium. For stable proteins, the percentage is close to 100%, and it decreases as the protein becomes less and less stable. The stability of most functional folded proteins studied ranges from 3 to 15 kcal/mol [169]. A protein with a low stability of 3 kcal/mol still has, at 25°C, 99.4% of the molecules folded at any time. According to Equation 4, the average $\tau_{0.5}$ for a protein in the high stability range (15 kcal/mol) is of 25 ± 7 million years (Myr), while that for one in the low stability range (3 kcal/mol) is of only 50 ± 14 ms. A protein that, due to a SNV, becomes destabilized to the point of having its function compromised will likely exhibit a lower conformational stability [170]. If we take 2 kcal/mol as the stability threshold below which a protein is not stable enough to perform its cellular function (3% of its molecules will be unfolded at 25°C at any moment), the time needed to observe the unfolding of individual

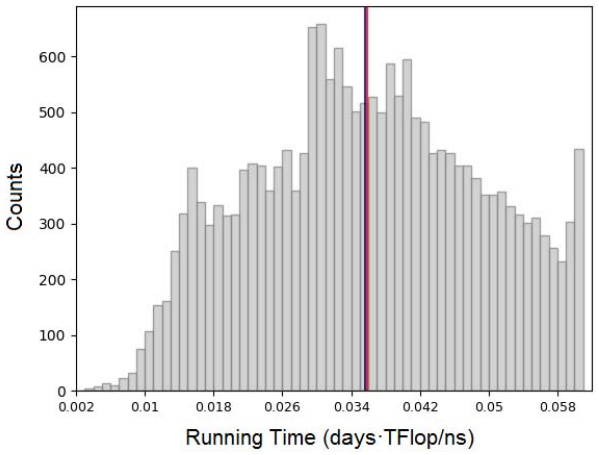

Figure 4. Histogram of estimated running times (in days·TFlop/ns) for all the proteins encoded in the human genome. The red line is the mean of the distribution (0.0357), and the navy blue line is the median (0.0358). The running time values were calculated with Equation 3 (see Figure 2).

protein molecules in experiment or simulation can be similarly calculated from Equation 4. As it turns out, half of the molecules of protein variants with stabilities of 2 kcal/mol or less will have experienced unfolding events in 2 ± 0.6 ms. Thus, if 2 ms of a single molecule of such destabilized protein variant are simulated at 25°C, the probability of observing unfolding will be of just 0.5. Equations 4 and 5 enable to calculate the average time needed for a given fraction of a protein of a given stability to become unfolded. Following with the example, 9 out of 10 molecules of that protein will be observed to unfold in 7 ± 2 ms and, in 14 ± 4 ms, 99% of the molecules will have unfolded. Therefore, if a single copy of the protein is simulated for 14 ms, the probability of observing its unfolding will be of 0.99. This 2 kcal/mol stability threshold is certainly arbitrary and should be only taken as a reasonable average value. This is so because protein variants causing amyloid-related diseases may become deleterious even if they display higher stabilities, as a small fraction of denatured molecules may initiate and drive the aggregation. On the other hand, proteins causing disease by a loss-of-function mechanism may provide some functionality to the cell even having stabilities below that level. It should be also borne in mind that some proteins display non-cooperative, i.e. non two-state, unfolding equilibria [94]. For those proteins, which may abound among the large ones, it is the unfolding of the weakest energetic domain [171] rather than that of the entire protein that should be monitored in the simulations. The weakest domain will correspond in most cases to the structural domain containing the deleterious variation, and it is the domain to which the 2 kcal/mol stability threshold applies.

Combining Equations 1, 2 and 3 and taking into account the length of the 20 410 different proteins of the human proteome (which can be obtained from the canonical sequences annotated in UniProt, see Supporting Information Figure S2), the average time needed to simulate a single protein for 1 ns is $\bar{t}_i$ = 0.0358 days·TFlop (the distribution of t_i values for the different proteins is shown in Figure 4). Thus, the approximate time needed to simulate 1 ns of the entire t-HumanV is 731 ± 306 days·TFlop. This means that, in order to have a 0.99 probability of detecting any destabilized SAV with conformational stability below 2 kcal/mol, 29 Myr·TFlop would be required.

Differently to other MEPTs (e.g. FoldX, Rosetta or I-Mutant), the approach proposed here is based on the capabilities of MD-trajectory analysis to reliably detect conformational changes associated to very low conformational stability, Therefore, it does not allow, nor it tries, to compute ΔΔG values associated to mutations. However, recent developments in higher-order statistics methods to fine tune semi-automated analysis of long MD simulations [165] together with improved force fields [138] might allow in the future the quantitative analysis of conformational changes in MD trajectories, bringing an extra added value to MD-based predictive methods.

When should we embark upon this endeavor?

The very long time (29 Myr·TFlop) calculated above, which is beyond present computing capacity, is expected to decrease in the coming years both due to increased computing power of the CPUs and to improved efficiency of the MD-simulating algorithms. Taking a conservative approach, we may consider that the first factor will be the more relevant one and may accordingly disregard the improving contribution of the second factor. Then, assuming that the evolution of the computing power of CPUs will follow Moore's law [172], the time needed in the foreseeable future (at year X) to compute 1 ns of the t-HumanV becomes

$$t^{@\text{year } X} = \frac{t^{@2019}}{2^{(X-2019)/2}} \tag{6}$$

Combining Equations 2, 4 and 6 yields Equation 7 which allows to estimate the time required (T^X) to simulate the complete human variome using an *a priori* defined computer power and beginning the simulations at year *X*.

$$T^X = \sum_{i=1}^{20410} \sum_{j=1}^{n_{r_i}} \frac{\text{nsSNVc}_j \times \left(0.061 - 0.058e^{-0.002n_{r_i}}\right)}{\text{TFlops} \times 2^{(X-2019)/2}} \times \tau_{x_{unf}} \tag{7}$$

If, as explained above, $\tau_{x_{unf}}$ is set as 14 ms (to have a 0.99 probability of observing the unfolding of protein variants with stability below 2 kcal/mol in any rMD simulation), we may estimate how long it would take to simulate the complete HumanV to identify any severely destabilizing mutation. Equation 7 indicates that beginning at present, 2019, and using the total TFlops allocated in the top 10 supercomputers in the TOP500 ranking (https://www.top500.org/) plus those accounted in the project of distributed computing Folding@Home (606 826 TFlops, see Table 3), it would take around 193 000 years to complete the simulations. However, if we began the simulation in 2051, only 2.9 ± 0.5 years would be required using a similar number of supercomputers/projects of the time (see the blue curve in Figure 5A, and Table 3). Year 2051 would be the optimal moment to begin an efficient international effort to simulate t-HumanV as it affords the earliest date of completion of the task (Figure 5A).

For the more modest task of calculating the entire variome of individual human proteins using the same stability threshold of 2 kcal/mol, Figure 6 indicates that, starting at present, 2019, the mean time required to calculate a single protein variome is of 9.4 years, with half of the individual protein variomes needing less than 5.5 years.

When should we begin to simulate the human variome if we perform the simulations at 45°C or at 65°C?

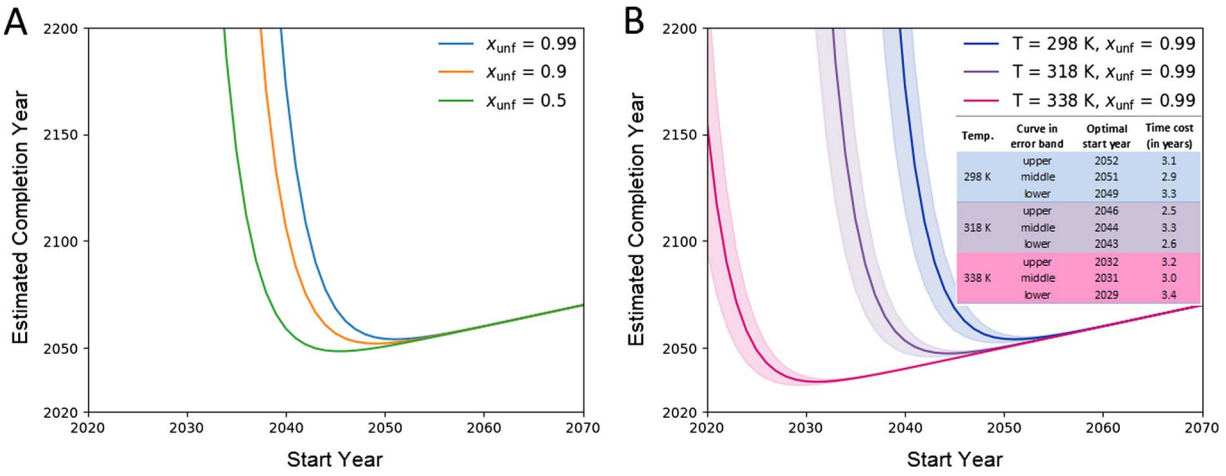

Figure 5. Estimated completion year versus start year profiles for the simulation of the entire t-HumanV. (A) Profiles calculated using different values of the fraction of unfolded molecules observed in the simulations (i.e. different probabilities of observing unfolding events in a single rMD simulation): x_{unf}, a protein stability threshold of 2 kcal/mol, and a simulation temperature of 25°C (298 K). (B) Profiles calculated at three different temperatures by setting $x_{unf} = 0.99$ and a protein stability threshold of 2 kcal/mol. Depicted error bands have been calculated by propagating the fitting errors obtained for the different parameters (see the standard errors of the fittings in Figures 2 and 3) through Equation 7. The inserted table summarizes the optimal start years and time costs estimated for performing the required rMD simulations of the t-HumanV at the indicated temperatures. Profiles in both panels were calculated taking TFlops in Equation 7 as the number of TFlops allocated in the top 10 supercomputers worldwide plus those in the project of distributed computing Folding@Home (606 826, see Table 3) and their future equivalence according to Moore's law.

Table 3. Estimated time required to complete the simulation of the t-HumanV using the TFlops allocated in the Top Ten supercomputers worldwide plus those in the distributed computing project Folding@Home, or the expected TFlops in equivalent computing infrastructures in future years

Supercomputer ranking	Supercomputer/project	Rmax (TFlop/s)	Start year[a]	Time required (years)[a]
1	Summit, DOE/SC/Oak Ridge National Laboratory, USA	143 500	2019 2055[b]	814 757 3.1
2	Sierra, DOE/NNSA/LLNL, USA	94 640	2019 2056[b]	1 235 393 3.3
3	Sunway TaihuLight, National Supercomputing Center in Wuxi, China	93 015	2019 2056[b]	1 902 801 3.4
4	Tianhe-2A, National Super Computer Center in Guangzhou, China	61 445	2019 2058[b]	5 507 189 2.6
5	Piz Daint, Swiss National Supercomputing Centre (CSCS), Switzerland	21 230	2019 2061[b]	5 799 773 2.6
6	Trinity, DOE/NNSA/LANL/SNL, USA	20 159	2019 2061[b]	5 881 167 2.8
7	AI Bridging Cloud Infrastructure, AIST, Japan	19 880	2019 2061[b]	6 002 856 2.8
8	SuperMUC-NG, Leibniz Rechenzentrum, Germany	19 477	2019 2061[b]	6 646 823 2.9
9	Titan, DOE/SC/Oak Ridge National Laboratory, USA	17 590	2019 2061[b]	6 646 823 3.2
10	Sequoia, DOE/NNSA/LLNL, USA	17 173	2019 2061[b]	6 808 223 3.2
Distributed computing project	Folding@Home, The Pande Lab, Stanford University and Stanford University Medical Center	98 747	2019 2056[b]	1 184 012 3.2
-	Top 10 + Folding@Home	606 826	2019 2051[b]	192 671 2.9

[a]Time required to observe unfolding—0.99 probability—in single simulations at 25°C of single molecules of unstable protein variants (stability below 2 kcal/mol).
[b]Optimal moment—providing the earliest completion date—to begin the simulation of the complete human variome (see also Figure 5) using the corresponding TFlops calculated at that time as per Moore's law.

Rate constants of protein (un)folding are temperature-dependent. [173] The Eyring–Kramers equation

$$k = \frac{k_B T}{h} \times e^{-\left(\frac{\Delta G^{\ddagger}}{RT}\right)} \quad (8)$$

(where k_B and h are Boltzmann's and Planck's constant, respectively) indicates that such dependency is modulated by the activation free-energy ($\Delta G^{\ddagger}$), which can be determined experimentally. Analysis of experimental data on the variation of the unfolding rate constants of 11 proteins (ranging from 48 to 118

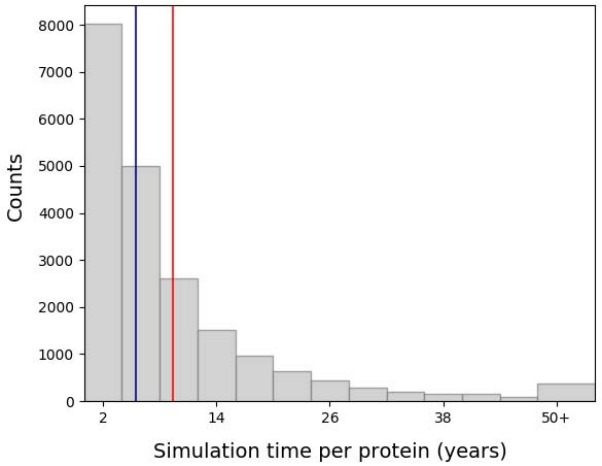

Figure 6. Histogram of the real simulation times (in years) required to simulate at present full individual protein variomes, calculated using Equation 7. The calculation was done by setting a protein stability threshold of 2 kcal/mol and a value of 0.99 for the probability of detecting that a particular SNV has lowered the stability of the wild-type protein below that threshold (i.e. setting $\tau x_{unf} = 14$ ms), as well as a number of TFlops equivalent to those allocated together in the top 10 supercomputers and the distributed computing project Folding@Home at present (2019) (see Table 3). The red line represents the mean of the distribution (9.4), and the navy blue one, the median (5.5).

amino acid residues) as a function of temperature (Supporting Information Table S4) shows no correlation between the activation free energies of those proteins and their sizes (Supporting Information Figure S3). Therefore, we have calculated from that set of proteins an average activation free energy of unfolding of 82.5 ± 12.1 kJ/mol ($\Delta G^{\ddagger} \pm$ SD). Taking this value as an invariant, Equation 9 indicates that kinetic constants of protein unfolding (k_u) become 8.7 times larger if the temperature is raised from 298 to 318 K, and 58.4 times larger if temperature is raised from 298 to 338 K.

$$\frac{k_1}{k_2} = \frac{T_1}{T_2} \times e^{\left(\frac{\Delta G^{\ddagger}}{RT}\right)\left(\frac{1}{T_2} - \frac{1}{T_1}\right)} \quad (9)$$

As a consequence, the time needed to have a 0.99 probability of observing unfolding in a given rMD simulation performed at 298 K, as described by Equations 4 and 5, is reduced at 318 K by a factor of 0.12 compared to that at 298 K and, at 338 K, by a factor of 0.017. As it seems, modest increases of simulation temperature can greatly reduce the computer time needed and bring the optimal moment when the simulation of the entire variome should be initiated significantly closer to present. The table inserted in Figure 5B summarizes the optimal start year for the simulation of the HumanV and the time cost at that point for the three simulation temperatures here discussed. While simulations at 25°C (298 K) should begin in 2051 and last for 2.9 years, those at 45°C (318 K) should begin in 2044 (lasting for 3.3 years), and those at 65°C (338 K) should begin in 2031 and could be completed in 3.0 years.

As indicated above, the cost of simulating a single protein variome is much lower and, starting at present, 2019, the mean time required to simulate a single protein variome at 45°C would be of 1.2 years and, at 65°C, of just 2 months. The simulation of individual protein variomes at temperatures close to the physiological one is already feasible, and that of the entire human variome will become feasible in few years. Having a reliable assessment of the structural impact of any single-amino-acid variation arising in the human proteome ready may greatly contribute to increase the accuracy of genetic interpretation of human SNV.

Conclusions

The continuous improvement of current MEPT based on evolutionary information and simple assessment of physical–chemical properties of amino acid residues appears not to be reaching the level of accuracy required for their generalized use in medical diagnosis. The exploration of alternative methods that fully analyze the impact of variations on protein structure, stability and binding, such as those using MD simulations, may provide the more accurate predictive tools in need. We have developed a simple model that allows to estimate the time needed to perform a predictive MD analysis of the entire human variome (here defined as all possible protein variants carrying a single-amino-acid replacement arising from a SNV in the corresponding coding sequence) with existing computing capabilities. Our model indicates that the structural impact of all human SAVs should begin to be assessed around 2031 and they could be completed by 2034 using explicit-solvent full-atom rMD simulations performed at a moderate temperature (65°C). On the other hand, full variomes of individual proteins can be already analyzed at present, even at a lower temperature, closer to the physiological one.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Author Contributions

J.S. conceived and directed the investigation. J.J.G-F. and J.S. wrote the manuscript. J.J.G-F. and H.G-C. revised the literature, analyzed the data and made the figures and tables.

Key Points

- Obtaining genetic information describing variations in patients and healthy individuals has become easy and cheap, yet its accurate interpretation is still an unattained goal.
- Bioinformatics tools for genetic interpretation appear to have reached a maximum accuracy of around 85% in the binary interpretation of nsSNV in protein coding regions as deleterious or neutral, which is not sufficient for them to be trusted by clinicians.
- Full-atom molecular dynamics (MD) simulation of proteins carrying SAV offers good prospects of providing the extra accuracy needed, but this technique is slow, and its potential for massive interpretation of SAV at proteome scale needs to be appraised.
- We calculate here the size of the human nsSNV space (human Variome) at 66 474 822 protein variations and develop a physical model to determine the MD simulation time span required to have a high probability of observing unfolding of the unstable variants, a key datum enabling the use of MD simulations as binary predictors of protein stability.
- According to our model, an international effort initiated around 2031 could complete the MD simulation of the human Variome in 3 years, providing an accurate binary

classification (destabilizing/not destabilizing) of all possible nsSNVs. For a single human protein of average size, it could be done at present in few months.

Acknowledgements

We thank the Biocomputation and Complex Systems Physics Institute (BIFI) of the University of Zaragoza for computing facilities granted to perform molecular dynamics simulations. We thank Alfonso López for his help in the creation of some figures.

Funding

This work was supported by the Ministerio de Economía y Competitividad, Spain (BFU2016-78232-P); Gobierno de Aragón, Spain (E45_17R); and ERDF-InterregV-A POCTEFA (PIREPRED-EFA086/15). The FPU16/04232 doctoral contract was conceded by the Ministerio de Educación, Cultura y Deporte, Spain, to H.G-C.

References

1. Shapiro LJ. Human genome project. *West J Med* 1993;**158**:181.
2. Hood L, Rowen L. The Human Genome Project: big science transforms biology and medicine. *Genome Med* 2013;**5**:79–9.
3. Cotton RG, Auerbach AD, Axton M, *et al*. The Human Variome Project. *Science* 2008;**322**:861–2.
4. Management EPS. The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science* 2004;**306**:636–40.
5. Consortium IH. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007;**449**:851–61.
6. Consortium GP, Abecasis GR, Altshuler D, *et al*. A map of human genome variation from population-scale sequencing. *Nature* 2010;**467**:1061–73.
7. Consortium TICG. International network of cancer genome projects. *Nature* 2010;**464**:993–8.
8. Peplow M. The 100 000 Genomes Project. *BMJ* 2016;**353**: i1757.
9. Kitts A, Phan L, Ward M, *et al*. *The Database of Short Genetic Variation (dbSNP) The NCBI Handbook [Internet]*, 2nd edn. Bethesda (MD): National Center for Biotechnology Information (US), 2013, Updated 2014 Apr 3.
10. Ambardar S, Gupta R, Trakroo D, *et al*. High throughput sequencing: an overview of sequencing chemistry. *Indian J Microbiol* 2016;**56**:394–404.
11. Park SJ, Saito-Adachi M, Komiyama Y, *et al*. Advances, practice, and clinical perspectives in high-throughput sequencing. *Oral Dis* 2016;**22**:353–64.
12. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. *Mol Cell* 2015;**58**:586–97.
13. Bick D, Dimmock D. Whole exome and whole genome sequencing. *Curr Opin Pediatr* 2011;**23**:594–600.
14. Witte JS. Genome-wide association studies and beyond. *Annu Rev Public Health* 2010;**31**:9–20.
15. Consortium WTCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007;**447**:661–78.
16. Gonzaga-Jauregui C, Lupski JR, Gibbs RA. Human genome sequencing in health and disease. *Annu Rev Med* 2012;**63**:35–61.
17. Yang Y, Muzny DM, Reid JG, *et al*. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *N Engl J Med* 2013;**369**:1502–11.
18. Brachi B, Morris GP, Borevitz JO. Genome-wide association studies in plants: the missing heritability is in the field. *Genome Biol* 2011;**12**:232.
19. Horton MW, Bodenhausen N, Beilsmith K, *et al*. Genome-wide association study of Arabidopsis thaliana leaf microbial community. *Nat Commun* 2014;**5**:5320.
20. Lee BY, Lee KN, Lee T, *et al*. Bovine genome-wide association study for genetic elements to resist the infection of foot-and-mouth disease in the field. *Asian-Australas J Anim Sci* 2015;**28**:166–70.
21. Wang K, Liu D, Hernandez-Sanchez J, *et al*. Genome wide association analysis reveals new production trait genes in a male Duroc population. *PLos One* 2015;**10**: e0139207.
22. Katsonis P, Koire A, Wilson SJ, *et al*. Single nucleotide variations: biological impact and theoretical interpretation. *Protein Sci* 2014;**23**:1650–66.
23. Blanco EH, Peinado JR, Martín MG, *et al*. Biochemical and cell biological properties of the human prohormone convertase 1/3 Ser357Gly mutation: a PC1/3 hypermorph. *Endocrinology* 2014;**155**:3434–47.
24. Isrie M, Breuss M, Tian G, *et al*. Mutations in either TUBB or MAPRE2 cause circumferential skin creases Kunze type. *Am J Hum Genet* 2015;**97**:790–800.
25. Tokuriki N, Stricher F, Serrano L, *et al*. How protein stability and new functions trade off. *PLoS Comput Biol* 2008; **4**:e1000002.
26. Gaboriau DC, Rowling PJ, Morrison CG, *et al*. Protein stability versus function: effects of destabilizing missense mutations on BRCA1 DNA repair activity. *Biochem J* 2015; **466**:613–24.
27. Sergouniotis PI, Barton SJ, Waller S, *et al*. The role of small in-frame insertions/deletions in inherited eye disorders and how structural modelling can help estimate their pathogenicity. *Orphanet J Rare Dis* 2016;**11**:125.
28. Shi Z, Sellers J, Moult J. Protein stability and in vivo concentration of missense mutations in phenylalanine hydroxylase. *Proteins* 2012;**80**:61–70.
29. Haraksingh RR, Snyder MP. Impacts of variation in the human genome on gene regulation. *J Mol Biol* 2013;**425**: 3970–7.
30. Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol Syst Biol* 2013;**9**:637.
31. Yates CM, Sternberg MJ. The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein-protein interactions. *J Mol Biol* 2013;**425**:3949–63.
32. Duning K, Wennmann DO, Bokemeyer A, *et al*. Common exonic missense variants in the C2 domain of the human KIBRA protein modify lipid binding and cognitive performance. *J Translat Psych* 2013;**3**:e272.
33. Feinberg H, Rowntree TJ, Tan SL, *et al*. Common polymorphisms in human langerin change specificity for glycan ligands. *J Biol Chem* 2013;**288**:36762–71.
34. Jubb HC, Pandurangan AP, Turner MA, *et al*. Mutations at protein-protein interfaces: small changes over big surfaces have large impacts on human health. *Prog Biophys Mol Biol* 2017;**128**:3–13.
35. Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol* 2005;**353**:459–73.

36. Wang Z, Moult J. SNPs, protein structure, and disease. *Hum Mutat* 2001;**17**:263–70.
37. David A, Razali R, Wass MN, *et al*. Protein-protein interaction sites are hot spots for disease-associated nonsynonymous SNPs. *Hum Mutat* 2012;**33**:359–63.
38. Nishi H, Tyagi M, Teng S, *et al*. Cancer missense mutations alter binding properties of proteins and their interaction networks. *PLos One* 2013;**8**:e66273.
39. Sahni N, Yi S, Taipale M, *et al*. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* 2015;**161**:647–60.
40. Fleming N, Kinsella B, Ing C. Predicting protein thermostability upon mutation using molecular dynamics timeseries data. *bioRxiv* 2016;**078246**.
41. Montelione GT. The protein structure initiative: achievements and visions for the future. *F1000 Biol Rep* 2012;**4**:7.
42. Azia A, Uversky VN, Horovitz A, *et al*. The effects of mutations on protein function: a comparative study of three databases of mutations in humans. *Israel Journal of Chemistry* 2013;**53**:217–26.
43. Chwastyk M, Jaskolski M, Cieplak M. Structure-based analysis of thermodynamic and mechanical properties of cavity-containing proteins–case study of plant pathogenesis-related proteins of class 10. *FEBS J* 2014;**281**: 416–29.
44. Estrada J, Bernado P, Blackledge M, *et al*. ProtSA: a web application for calculating sequence specific protein solvent accessibilities in the unfolded ensemble. *BMC Bioinformatics* 2009;**10**:104.
45. Karplus M, Kuriyan J. Molecular dynamics and protein function. *PNAS* 2005;**102**:6679–85.
46. Barradas-Bautista D, Fernández-Recio J. Docking-based modeling of protein-protein interfaces for extensive structural and functional characterization of missense mutations. *PLos One* 2017;**12**:e0183643.
47. Priya Doss CG, Chakraborty C, Chen L, *et al*. Integrating in silico prediction methods, molecular docking, and molecular dynamics simulation to predict the impact of ALK missense mutations in structural perspective. *Biomed Res Int* 2014;895831.
48. Jones MM, Castle-Clarke S, Brooker D, *et al*. The structural genomics consortium: a knowledge platform for drug discovery: a summary. *Rand Health Quarterly* 2014;**4**:19.
49. Terwilliger TC, Stuart D, Yokoyama S. Lessons from structural genomics. *Annu Rev Biophys* 2009;**38**:371–83.
50. Kim MS, Pinto SM, Getnet D, *et al*. A draft map of the human proteome. *Nature* 2014;**509**:575–81.
51. Ponomarenko EA, Poverennaya EV, Ilgisonis EV, *et al*. The size of the human proteome: the width and depth. *Int J Anal Chem* 2016;**2016**:7436849.
52. 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.
53. Ma M, Ru Y, Chuang L, *et al*. Disease-associated variants in different categories of disease located in distinct regulatory elements. *BMC Genomics* 2015;**16**:S3.
54. Angarica VE, Orozco M, Sancho J. Exploring the complete mutational space of the LDL receptor LA5 domain using molecular dynamics: linking SNPs with disease phenotypes in familial hypercholesterolemia. *Hum Mol Genet* 2016; **25**:1233–46.
55. Athey J, Alexaki A, Osipova E, *et al*. A new and updated resource for codon usage tables. *BMC Bioinformatics* 2017; **18**:391.
56. Szklarczyk D, Franceschini A, Wyder S, *et al*. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 2015;**43**:D447–52.
57. Szklarczyk D, Morris JH, Cook H, *et al*. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* 2017;**45**:D362–8.
58. Futschik ME, Chaurasia G, Herzel H. Comparison of human protein–protein interaction maps. *Bioinformatics* 2007; **5**:605–11.
59. Acuner Ozbabacan SE, Engin HB, Gursoy A, *et al*. Transient protein–protein interactions. *Protein Eng Des Sel* 2011; **24**:635–48.
60. Ngounou Wetie AG, Sokolowska I, Woods AG, *et al*. Investigation of stable and transient protein-protein interactions: past, present, and future. *Proteomics* 2013;**13**:538–57.
61. Snider J, Kotlyar M, Saraon P, *et al*. Fundamentals of protein interaction network mapping. *Mol Syst Biol* 2015;**11**:848.
62. Brückner A, Polge C, Lentze N, *et al*. Yeast two-hybrid, a powerful tool for systems biology. *Int J Mol Sci* 2009;**10**:2763–88.
63. Rual JF, Venkatesan K, Hao T, *et al*. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 2005;**437**:1173–8.
64. Cong Q, Anishchenko I, Ovchinnikov S, *et al*. Protein interaction networks revealed by proteome coevolution. *Science* 2019;**365**:185.
65. Venkatesan K, Rual JF, Vazquez A, *et al*. An empirical framework for binary interactome mapping. *Nat Methods* 2009;**6**:83–90.
66. Stumpf MP, Thorne T, de Silva E, *et al*. Estimating the size of the human interactome. *Proc Natl Acad Sci USA* 2008; **105**:6959–64.
67. Planas-Iglesias J, Marin-Lopez MA, Bonet J, *et al*. iLoops: a protein-protein interaction prediction server based on structural features. *Bioinformatics* 2013;**29**:60–2.
68. Fukuhara N, Kawabata T. HOMCOS: a server to predict interacting protein pairs and interacting sites by homology modeling of complex structures. *Nucleic Acids Res* 2008; **36**:W185–9.
69. Mukherjee S, Zhang Y. Protein-protein complex structure predictions by multimeric threading and template recombination. *Structure* 2011;**19**:955–66.
70. Aloy P, Russell RB. InterPreTS: protein interaction prediction through tertiary structure. *Bioinformatics* 2003;**19**:161–2.
71. Keskin O, Nussinov R, Gursoy A. PRISM: protein-protein interaction prediction by structural matching. *Methods Mol Biol* 2008;**484**:505–21.
72. Xenarios I, Rice DW, Salwinski L, *et al*. DIP: the database of interacting proteins. *Nucleic Acids Res* 2000;**28**:289–91.
73. Bader GD, Betel D, Hogue CW. BIND: the biomolecular interaction network database. *Nucleic Acids Res* 2003;**31**:248–50.
74. Zhang QC, Petrey D, Garzón JI, *et al*. PrePPI: a structure-informed database of protein-protein interactions. *Nucleic Acids Res* 2013;**41**:D828–33.
75. von Mering C, Huynen M, Jaeggi D, *et al*. STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* 2003;**31**:258–61.
76. Tzu-Hao K, Kuo-Bin L. Predicting protein–protein interaction sites using sequence descriptors and site propensity of neighboring amino acids. *Int J Mol Sci* 2016;**17**: 1788.
77. Northey TC, Barešić A, Martin ACR. IntPred: a structure-based predictor of protein–protein interaction sites. *Bioinformatics* 2018;**34**:223–9.

78. Esmaielbeiki R, Krawczyk K, Knapp B, *et al*. Progress and challenges in predicting protein interfaces. *Brief Bioinform* 2016;**17**:117–31.
79. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. *Nat Methods* 2014;**11**:801–7.
80. Stein A, Fowler DM, Hartmann-Petersen R, *et al*. Biophysical and mechanistic models for disease-causing protein variants. *Trends Biochem Sci* 2019;**44**:575–88.
81. Fowler DM, Stephany JJ, Fields S. Measuring the activity of protein variants on a large scale using deep mutational scanning. *Nat Protoc* 2014;**9**:2267–84.
82. Potapov V, Cohen M, Schreiber G. Assessing computational methods for predicting protein stability upon mutation: good on average but not in the details. *Protein Eng Des Sel* 2009;**22**:553–60.
83. Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc* 2009;**4**:1073–81.
84. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res* 2001;**11**:863–74.
85. Ng PC, Henikoff SSIFT. Predicting amino acid changes that affect protein function. *Nucleic Acids Res* 2003;**31**:3812–4.
86. Ng PC, Henikoff S. Predicting the effects of amino acid substitutions on protein function. *Annu Rev Genomics Hum Genet* 2006;**7**:61–80.
87. Adzhubei IA, Schmidt S, Peshkin L, *et al*. A method and server for predicting damaging missense mutations. *Nat Methods* 2010;**7**:248–9.
88. Kircher M, Witten DM, Jain P, *et al*. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 2014;**43**:310–5.
89. Pucci F, Bourgeas R, Rooman M. Predicting protein thermal stability changes upon point mutations using statistical potentials: introducing HoTMuSiC. *Sci Rep* 2016;**6**: 23257.
90. Capriotti E, Altman RB, Bromberg Y. Collective judgment predicts disease-associated single nucleotide variants. Mutations in proteins. *BMC Genomics* 2013;**14**:S2.
91. González-Pérez A, López-Bigas N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am J Hum Genet* 2011; **88**:440–9.
92. Mao Y, Chen H, Liang H, *et al*. CanDrA: cancer-specific driver missense mutation annotation with optimized features. *PLos One* 2013;**8**:e77945.
93. Dill KA, MacCallum JL. The protein-folding problem, 50 years on. *Science* 2012;**338**:1042.
94. Sancho J. The stability of 2-state, 3-state and more-state proteins from simple spectroscopic techniques… plus the structure of the equilibrium intermediates at the same time. *Arch Biochem Biophys* 2013;**531**:4–13.
95. Choi Y, Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. *Bioinformatics* 2015;**31**:2745–7.
96. Riera C, Padilla N, de la Cruz X. The complementarity between protein-specific and general pathogenicity predictors for amino acid substitutions. *Hum Mutat* 2016;**37**: 1013–24.
97. Nisthal A, Wang CY, Ary ML, *et al*. Protein stability engineering insights revealed by domain-wide comprehensive mutagenesis. *Proc Natl Acad Sci* 2019;**116**:16367.
98. Acharya V, Nagarajaram HA. Hansa: an automated method for discriminating disease and neutral human nsSNPs. *Hum Mutat* 2012;**33**:332–7.
99. Bao L, Zhou M, Cui Y. nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res* 2005;**33**:W480–2.
100. Baugh EH, Simmons-Edler R, Müller CL, *et al*. Robust classification of protein variation using structural modelling and large-scale data integration. *Nucleic Acids Res* 2016; **44**:2501–13.
101. Berliner N, Teyra J, Colak R, *et al*. Combining structural modeling with ensemble machine learning to accurately predict protein fold stability and binding affinity effects upon mutation. *PLos One* 2014;**9**:e107353.
102. Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res* 2007;**35**:3823–35.
103. Cheng J, Randall A, Baldi P. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins* 2006;**62**:1125–32.
104. De Baets G, Van Durme J, Reumers J, *et al*. SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Res* 2012;**40**:D935–9.
105. Pappalardo M, Wass MN. VarMod: modelling the functional effects of non-synonymous variants. *Nucleic Acids Res* 2014;**42**:W331–6.
106. Pejaver V, Urresti J, Lugo-Martinez J, *et al*. MutPred2: inferring the molecular and phenotypic impact of amino acid variants. *bioRxiv* 2017;134981.
107. Yates CM, Filippis I, Kelley LA, *et al*. SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. *J Mol Biol* 2014;**426**:2692–701.
108. Grimm DG, Azencott CA, Aicheler F, *et al*. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum Mutat* 2015;**36**:513–23.
109. Dewan S, McCabe K, Regnier M, *et al*. Molecular effects of cardiac troponin DCM mutations on calcium sensitivity and Myofilament activation - an integrated multi-scale modeling study. *Biophys J* 2017;**112**:322a.
110. Elmore DE, Dougherty DA. Molecular dynamics simulations of wild-type and mutant forms of the mycobacterium tuberculosis MscL channel. *Biophys J* 2001;**81**:1345–59.
111. Feng Z, Alqarni MH, Yang P, *et al*. Modeling, molecular dynamics simulation, and mutation validation for structure of cannabinoid receptor 2 based on known crystal structures of GPCRs. *J Chem Inf Model* 2014;**54**: 2483–99.
112. Frappier V, Chartier M, Najmanovich RF. ENCoM server: exploring protein conformational space and the effect of mutations on protein function and stability. *Nucleic Acids Res* 2015;**43**:W395–400.
113. Gapsys V, Michielssens S, Seeliger D, *et al*. Accurate and rigorous prediction of the changes in protein free energies in a large-scale mutation scan. *Angew Chem Int Ed* 2016; **55**:7364–8.
114. Jordan EJ, Radhakrishnan R. We can predict the effects of kinase domain mutations using molecular dynamics and machine learning. *Biophys J* 2017;**112**:322a.
115. Koukos PI, Glykos NM. Folding molecular dynamics simulations accurately predict the effect of mutations on the stability and structure of a vammin-derived peptide. *J Phys Chem* 2014;**118**:10076–84.
116. Kumar A, Purohit R. Use of long term molecular dynamics simulation in predicting cancer associated SNPs. *PLoS Comput Biol* 2014;**10**:e1003318.
117. Padhi AK, Vasaikar SV, Jayaram B, *et al*. ANGDelMut – a web-based tool for predicting and analyzing functional loss

mechanisms of deleterious angiogenin mutations causing amyotrophic lateral sclerosis. *F1000Res* 2013;**2**.
118. Rodrigues CHM, Pires DEV, Ascher DB. DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability. *Nucleic Acids Res* 2018; gky300.
119. Schadzek P, Schlingmann B, Schaarschmidt F, *et al*. Data of the molecular dynamics simulations of mutations in the human connexin46 docking interface. *Data Brief* 2016;**7**:93–9.
120. Zimmermann MT, Urrutia R, Oliver GR, *et al*. Molecular modeling and molecular dynamic simulation of the effects of variants in the TGFBR2 kinase domain as a paradigm for interpretation of variants obtained by next generation sequencing. *PLos One* 2017;**12**:e0170822.
121. Hensen U, Meyer T, Haas J, *et al*. Exploring protein dynamics space: the dynasome as the missing link between protein structure and function. *PLos One* 2012;**7**:e33931.
122. Henzler-Wildman KA, Thai V, Lei M, *et al*. Intrinsic motions along an enzymatic reaction trajectory. *Nature* 2007;**450**:838–44.
123. Stein A, Rueda M, Panjkovich A, *et al*. A systematic study of the energetics involved in structural changes upon association and connectivity in protein-protein interaction networks. *Structure* 2011;**19**:881–9.
124. Gerstein M, Krebs W. A database of macromolecular motions. *Nucleic Acids Res* 1998;**26**:4280–90.
125. Orozco M. The dynamic view of proteins: comment on "comparing proteins to their internal dynamics: exploring structure–function relationships beyond static structural alignments". *Phys Life Rev* 2013;**10**:29–30.
126. Case DA. Normal mode analysis of protein dynamics. *Curr Opin Struct Biol* 1994;**4**:285–90.
127. Pires DEV, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics* 2014;**30**:335–42.
128. 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.
129. Pires DEV, Ascher DB, Blundell TL. DUET: a server for predicting effects of mutations on protein stability using an integrated computational approach. *Nucleic Acids Res* 2014;**42**:W314–9.
130. Pandurangan AP, Ochoa-Montaño B, Ascher DB, *et al*. SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Res* 2017;**45**:W229–35.
131. Schymkowitz J, Borg J, Stricher F, *et al*. The FoldX web server: an online force field. *Nucleic Acids Res* 2005;**33**:W382–8.
132. Dehouck Y, Grosfils A, Folch B, *et al*. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics* 2009;**25**:2537–43.
133. Pucci F, Bernaerts KV, Kwasigroch JM, *et al*. Quantification of biases in predictions of protein stability changes upon mutations. *Bioinformatics* 2018;**34**:3659–65.
134. Masso M, Vaisman I. AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements. *Protein Eng Des Sel* 2010;**23**:683–7.
135. Masso M, Vaisman II. AUTO-MUTE 2.0: a portable framework with enhanced capabilities for predicting protein functional consequences upon mutation. *Advances in bioinformatics* 2014;**2014**:278385–5.
136. Hospital A, Goñi JR, Orozco M, *et al*. Molecular dynamics simulations: advances and applications. *Adv Appl Bioinform Chem* 2015;**8**:37–47.
137. Kumari I, Sandhu P, Ahmed M, *et al*. Molecular dynamics simulations, challenges and opportunities: a Biologist's prospective. *Curr Protein Pept Sci* 2017;**18**:1163–79.
138. Galano-Frutos JJ, Sancho J. Accurate calculation of Barnase and SNase folding energetics using short MD simulations and an atomistic model of the unfolded ensemble. Evaluation of force fields and water models. *J Chem Inf Model* 2019In Press. doi: 10.1021/acs.jcim.9b00430.
139. Nerenberg PS, Head-Gordon T. New developments in force fields for biomolecular simulations. *Curr Opin Struct Biol* 2018;**49**:129–38.
140. Ouyang JF, Bettens RP. Modelling water: a lifetime enigma. *Chimia* 2015;**69**:104–11.
141. Martínez-Oliván J, Arias-Moreno X, Velazquez-Campoy A, *et al*. LDL receptor/lipoprotein recognition: endosomal weakening of ApoB and ApoE binding to the convex face of the LR5 repeat. *FEBS J* 2014;**281**:1534–46.
142. Rudenko G, Henry L, Henderson K, *et al*. Structure of the LDL receptor extracellular domain at endosomal pH. *Science* 2002;**298**:2353–8.
143. García-Fandiño R, Bernadó P, Ayuso-Tejedor S, *et al*. Defining the nature of thermal intermediate in 3 state folding proteins: apoflavodoxin, a study case. *PLoS Comput Biol* 2012;**8**:e1002647.
144. Micheletti C. Comparing proteins by their internal dynamics: exploring structure-function relationships beyond static structural alignments. *Phys Life Rev* 2013;**10**:1–26.
145. Velázquez-Muriel RM, Cuesta I, *et al*. Comparison of molecular dynamics and superfamily spaces of protein domain deformation. *BMC Struct Biol* 2009;**9**:6.
146. Ghosh T, García AE, Garde S. Molecular dynamics simulations of pressure effects on hydrophobic interactions. *J Am Chem Soc* 2001;**123**:10997–1003.
147. Okumura H. Temperature and pressure denaturation of chignolin: folding and unfolding simulation by multibaric-multithermal molecular dynamics method. *Proteins* 2012;**80**:2397–416.
148. Day R, Bennion BJ, Ham S, *et al*. Increasing temperature accelerates protein unfolding without changing the pathway of unfolding. *J Mol Biol* 2002;**322**:189–203.
149. Bennion BJ, Daggett V. The molecular basis for the chemical denaturation of proteins by urea. *Proc Natl Acad Sci U S A* 2003;**100**:5142–7.
150. Camilloni C, Guerini Rocco A, Eberini I, *et al*. Urea and guanidinium chloride denature protein L in different ways in molecular dynamics simulations. *Biophys J* 2008;**94**:4654–61.
151. Gao M, Wilmanns M, Schulten K. Steered molecular dynamics studies of Titin I1 domain unfolding. *Biophys J* 2002;**83**:3435–45.
152. Lu H, Schulten K. Steered molecular dynamics simulations of force-induced protein domain unfolding. *Proteins* 1999; **35**:453–63.
153. Krieger E, Vriend G. New ways to boost molecular dynamics simulations. *J Comput Chem* 2015;**36**:996–1007.
154. Biedermann J, Ullrich A, Schöneberg J, *et al*. ReaDDyMM: fast interacting particle reaction-diffusion simulations using graphical processing units. *Biophys J* 2015;**108**:457–61.
155. Barney L. Speeding up molecular dynamics: modified GROMACS code improves optimization, parallelization. *Scientific Computing Advantage Business Marketing* 2016.
156. Shaw DE, Dror RO, Salmon JK, *et al*. Millisecond-scale molecular dynamics simulations on Anton. In: *SC'09: Proceedings of the Conference on High Performance Computing Networking, Storage and Analysis*. New York, NY, 2009, 1–11.

157. Kleinjung J, Fraternali F. Design and application of implicit solvent models in biomolecular simulations. 2014;**25**, **100**: 126–34.
158. Nielsen SO, Bulo RE, Moore PB, *et al*. Recent progress in adaptive multiscale molecular dynamics simulations of soft matter. *Phys Chem Chem Phys* 2010;**12**:12401–14.
159. Pronk S, Larsson P, Pouya I, *et al*. Copernicus: a new paradigm for parallel adaptive molecular dynamics. In: *SC '11: Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*. New York, NY: ACM, 2011.
160. Riniker S, van Gunsteren WF. Mixing coarse-grained and fine-grained water in molecular dynamics simulations of a single system. *J Chem Phys* 2012;**137**:044120.
161. Freddolino PL, Harrison CB, Liu Y, *et al*. Challenges in protein folding simulations: timescale, representation, and analysis. *Nat Phys* 2010;**6**:751–8.
162. Lindahl E, Hess B, van der Spoel D. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J Mol Model* 2001;**7**:306.
163. Michaud-Agrawal N, Denning EJ, Woolf TB, *et al*. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J Comput Chem* 2011;**32**:2319–27.
164. Roe DR, Cheatham IIITE. PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data. *J Chem Theory Comput* 2013;**9**:3084–95.
165. Papaleo E, Saladino G, Lambrughi M, *et al*. The role of protein loops and linkers in conformational dynamics and allostery. *Chem Rev* 2016;**116**:6391–423.
166. Kryshtafovych A, Monastyrskyy B, Fidelis K, *et al*. Evaluation of the template-based modeling in CASP12. *Proteins* 2018;**86**:321–4.
167. Smith LM, Kelleher NL. Proteoform: a single term describing protein complexity. *Nat Methods* 2013;**10**:186–7.
168. Manavalan B, Kuwajima K, Lee J. PFDB: a standardized protein folding database with temperature correction. *Sci Rep* 2019;**9**:1588.
169. Zeldovich KB, Chen P, Shakhnovich EI. Protein stability imposes limits on organism complexity and speed of molecular evolution. *Proc Natl Acad Sci USA* 2007;**104**: 16152–7.
170. Tokuriki N, Tawfik DS. Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol* 2009;**19**: 596–604.
171. Campos LA, Garcia-Mira MM, Godoy-Ruiz R, *et al*. Do proteins always benefit from a stability increase? Relevant and residual stabilisation in a three-state protein by charge optimisation. *J Mol Biol* 2004;**344**:223–37.
172. Moore GE. Progress in digital integrated electronics [Technical literature, Copyright 1975 IEEE. Reprinted, with permission. Technical Digest. International Electron Devices Meeting, IEEE, 1975, pp. 11-13]. *IEEE Solid-State Circuits Society Newsletter* 2006;**11**:36–7.
173. Bilsel O, Matthews CR. Barriers in protein folding reactions. *Adv Protein Chem* 2000;**53**:153–207.
174. Preeprem T, Gibson G. SDS, a structural disruption score for assessment of missense variant deleteriousness. *Front Genet* 2014;**5**:82.
175. Tanyalcin I, Stouffs K, Daneels D, *et al*. Convert your favorite protein modeling program into a mutation predictor: "MODICT". *BMC Bioinformatics* 2016;**17**:425–5.
176. Pires DEV, Ascher DB. mCSM–NA: predicting the effects of mutations on protein–nucleic acids interactions. *Nucleic Acids Res* 2017;**45**:W241–6.
177. Calabrese R, Capriotti E, Fariselli P, *et al*. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum Mutat* 2009;**30**:1237–44.
178. Capriotti E, Calabrese R, Fariselli P, *et al*. WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. *BMC Genomics* 2013;**14**:S6.
179. Mi H, Huang X, Muruganujan A, *et al*. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res* 2017;**45**:D183–9.
180. Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res* 2005;**15**: 978–86.
181. 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.
182. Tavtigian SV, Deffenbaugh AM, Yin L, *et al*. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J Med Genet* 2006;**43**:295–305.
183. Tian J, Wu N, Guo X, *et al*. Predicting the phenotypic effects of non-synonymous single nucleotide polymorphisms based on support vector machines. *BMC Bioinformatics* 2007;**8**:450.
184. Carter H, Chen S, Isik L, *et al*. Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res* 2009;**69**:6660–7.
185. Schwarz JM, Cooper DN, Schuelke M, *et al*. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* 2014;**11**:361–2.
186. Schwarz JM, Rödelsperger C, Schuelke M, *et al*. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* 2010;**7**:575–6.
187. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res* 2011;**39**:e118.
188. Choi Y. A fast computation of pairwise sequence alignment scores between a protein and a set of single-locus variants of another protein. In: *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine (BCB '12)*. New York, NY, 2012, 414–7.
189. Choi Y, Sims GE, Murphy S, *et al*. Predicting the functional effect of amino acid substitutions and indels. *PLos One* 2012;**7**:e46688.
190. Makarov V, O'Grady T, Cai G, *et al*. AnnTools: a comprehensive and versatile annotation toolkit for genomic variants. *Bioinformatics* 2012;**28**:724–5.
191. 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.
192. Katsonis P, Lichtarge O. A formal perturbation equation between genotype and phenotype determines the evolutionary action of protein-coding variations on fitness. *Genome Res* 2014;**24**:2050–8.
193. Niroula A, Urolagin S, Vihinen M. PON-P2: prediction method for fast and reliable identification of harmful variants. *PLos One* 2015;**10**:e0117380.
194. Karchin R, Diekhans M, Kelly L, *et al*. LS-SNP: large-scale annotation of coding non-synonymous SNPs based

on multiple information sources. *Bioinformatics* 2005;**21**: 2814–20.
195. Ryan M, Diekhans M, Lien S, *et al*. LS-SNP/PDB: annotated non-synonymous SNPs mapped to Protein Data Bank structures. *Bioinformatics* 2009;**25**:1431–2.
196. Ferrer-Costa C, Parraga I, Gelpí JL, *et al*. PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics* 2005;**21**:3176–8.
197. López-Ferrando V, Gazzo A, de la Cruz X, *et al*. PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. *Nucleic Acids Res* 2017;**45**:W222–8.
198. Yue P, Melamud E, Moult J. SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics* 2006;**7**:166.
199. Hecht M, Bromberg Y, Rost B. Better prediction of functional effects for sequence variants. *BMC Genomics* 2015;**16**:S1.
200. Capriotti E, Fariselli P, Rossi I, *et al*. A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics* 2008;**9**:S6.
201. Capriotti E, Altman RB. Improving the prediction of disease-related variants using protein three-dimensional structure. *BMC Bioinformatics* 2011;**12**(S3).
202. Acharya V, Nagarajaram HA. Response to: statistical analysis of missense mutation classifiers. *Hum Mutat* 2013; **34**:407.
203. Pei J, Grishin NV. AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics* 2001;**17**:700–12.