Chapter 12

Bioinformatics Approaches to the Functional Profiling of Genetic Variants

Biao Li, Predrag Radivojac and Sean Mooney

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/45900

1. Introduction

In the search for genetic mutations susceptible to human diseases, researchers take either genome-wide approaches or candidate gene approaches [1]. Traditional techniques in both approaches, such as chromosomal scan on the pedigree data and case-control design for a small number of genes of interest, however, have limitations in either achieving high resolution to identify specific genes, or obtaining whole genome coverage. Discoveries from pedigree linkage usually pointed to one or a few chromosomal regions related to the phenotype of interest, and these regions generally harbor many (perhaps hundreds) of genes, which rendered pinpointing actual genetic causes a daunting task. On the other hand, association studies typically focused on a couple of genes, some of which may participate in the same pathway, and the number of interrogated variants was always experimentally manageable. However, technical advances have brought high-throughput approaches within the reach of more and more scientists, increasing the volume of variants that researchers can interrogate by genotyping array and next-generation sequencing techniques at an exponential pace. A recent dbSNP build (build 135), a large public-domain database of single-nucleotide polymorphisms (SNPs), hosts more than 41.7 million validated human mutations, and with ongoing large-scale efforts such as the 1000 Genomes Project [2], that number is poised to grow significantly larger.

Of all genomic variants, those occurring in the protein-coding genes and resulting in amino acid substitutions hold special interest, as we have more knowledge about coding genes and their products than other genomic elements. Amino acid substitutions, or nonsynonymous SNPs (nsSNPs), not only change primary protein sequence but also have the potential for altering protein structure and disrupting or creating functional sites. These consequences can be tested experimentally, although doing so is costly and time-consuming.

Currently, about 1.2 million nsSNPs have been mapped to NCBI RefSeq proteins (2012/06), but we only have knowledge for a small fraction of them. The Human Gene Mutation

Database (HGMD; [3]) logs roughly 69,000 nsSNPs that are associated with diseases or traits; UniProt documents 37,000 nsSNPs as being neutral. For every six nsSNPs deposited in the public databases, five will have no disease or phenotype association. This gap will even grow larger as the emerging personal genome projects (www.personalgenomes.org) and whole-exome sequencing [4, 5] discover more rare variants.

Accompanying the compilation of a myriad of variants, a natural question arises about interpreting them in the context of human health. More specifically, how do we assess the disease risk for individual variants based on available biomedical information? Population studies, such as genome-wide association studies, have in recent years provided estimates of an odds ratio by comparing the frequencies of hundreds of thousands of genomic variants between disease/trait patients and healthy controls. One centralized resource, namely the Catalog of Published Genome-Wide Association Studies from the National Human Genome Research Institute [6], has collected published association studies involving at least 100,000 variants from 2008. The latest version (2012/06) records 8,063 significant mutation-trait associations from 1,287 studies. Most of these associations present a modest effect size with a median odds ratio (OR) of 1.36 (interquartile range [IQR]: 1.19–2.02). One clear observation from these studies is that the majority of variants occur in non-coding regions where the two most frequent locations are intergenic regions (43 percent) and introns (40 percent). In sharp contrast, only 368 nsSNPs associated with 177 diseases/traits were reported, with a slightly stronger effect size: a median OR of 1.52 (IQR: 1.21–3.33). This examination makes clear that the number of cohort studies will not keep pace with the increase in nsSNP data generation, suggesting that computational approaches may provide an important aid to our understanding of mutation-disease relationships.

Among all genome-level characteristics, scientists have collected the most knowledge about protein-coding genes, and they have published many investigations into the impacts of missense variants. Through mapping disease-associated nsSNPs and amino acid changes without disease annotations to the multispecies sequence alignment, researchers have observed that mutations related to monogenic diseases occurred significantly more frequently at slow-evolving positions, while neutral nsSNPs were enriched at fast-evolving positions [7, 8]. This observation therefore suggests that evolutionary rate could act as an indicator for discriminating diseases from neutral mutations. Also, the availability of crystal structure for numerous proteins provides us an opportunity to examine nsSNP consequences in the steric context. For example, p53, a well-studied tumor suppressor protein, is involved in many critical cell processes, such as DNA repair and cell-cycle regulation; p53 is inactive in half of all cancers [9]. Six mutation hot spots, such as R175H, R273H, and R282W, have been mapped to the p53 DNA-binding core domain that is critical to its activation, and most of them destabilize protein structure, leading to the degradation of p53 [10]. Intriguingly, certain mutations introduced to the mutant p53 could counteract this reduced stability and potentially rescue its functionality [11]. For example, nsSNP N268D in mutant p53 results in a hydrogen bond which bridges two strands and ultimately leads to an increase in thermodynamic stability. Finally, nsSNPs could influence a broad array of functional sites, including protein- and ligand-binding sites, catalytic residues, and numerous post-translational modification (PTM) sites. N-linked glycosylation, one type of PTM, is essential for the folding of some proteins. Proteins subjected to N-linked glycosylation contain an `NX[ST]` motif recognized by enzymes.

For example, amino acid substitution T183A, identified in the prion protein (PRNP), can cause spongiform encephalopathy by disrupting the consensus sequence `NX[ST]` through the loss of the threonine [12].

Many computational tools aiming to establish that nsSNPs cause disease are based on evolutionary characteristics, structural consequences, or functional impact, alone or in combination. One early and established method, SIFT (sort intolerant from tolerant substitutions; [13]), estimates the predisposition to disease for mutation solely by exploiting conservation information from sequence homology. Another well-known tool, PolyPhen-2 [14], uses predicted physicochemical features based on protein sequence in a naive Bayes classifier, in addition to sequence alignment.

In this chapter, we discuss the structural and functional impact of nsSNPs on the underlying proteins. We will provide concrete examples of both aspects, showing mechanisms through which amino acid substitutions affect proteins and contribute to disease phenotypes. We describe algorithms for predicting stability changes and for assigning probabilities to putative phosphorylation sites. We then apply these concepts/tools to the problem of distinguishing deleterious mutations from neutral ones. Finally, we will present another nsSNP prediction approach, MutPred, and apply it to a subset of dbSNP. Through these efforts, we aim to characterize a variety of computational approaches to the problem of inferring disease consequences for genetic variants, and demonstrate that these approaches are fruitful.

2. Structural impact of mutations

A classic disease that results from protein structural change via amino acid substitution is sickle cell anemia [15]. Replacement of a hydrophilic glutamic acid residue with a strong hydrophobic valine on the sixth amino acid of hemoglobin subunit beta causes the protein to aggregate and form rigid molecules, which in turn reshape the red blood cells as sickle-like [16]. The sickle cells die prematurely and thus result in anemia. Other possible structural abnormalities that nsSNPs can induce include changes of secondary structure, gain or loss of protein stability, and other physicochemical property alterations. In this section, we will illustrate two mutations on a cancer-related gene, BRCA1, and then describe an algorithm for predicting protein stability; finally, we will discuss its application to discriminating neutral and deleterious mutations.

BRCA1 is a well-known suppressor of breast and ovarian cancer tumors. Two C-terminal sequence repeats (BRCT) are essential for BRCA1's function, since mutations of stop codon and missense substitutions on these regions were observed in breast cancer patients [17, 18]. The crystal structure of the BRCT segments [19] shows that these two domains pack to each other in a tandem manner where one helix on the N-terminal domain and two helices on the C-terminal domain form an inner-domain interaction surface (Figure 1).

Two amino acid substitutions occur on this interface at A1708E, located near the end of the α1 helix, and at M1775R, located near the beginning of the α2 helix. At position 1708, the mutant glutamic acid is much larger than the original alanine (having a molecular weight of 147 versus 89) and introduces negative charge. Because M1708 lies near the center of the interaction surface, the compact core cannot accommodate this mutation sterically. Thus,

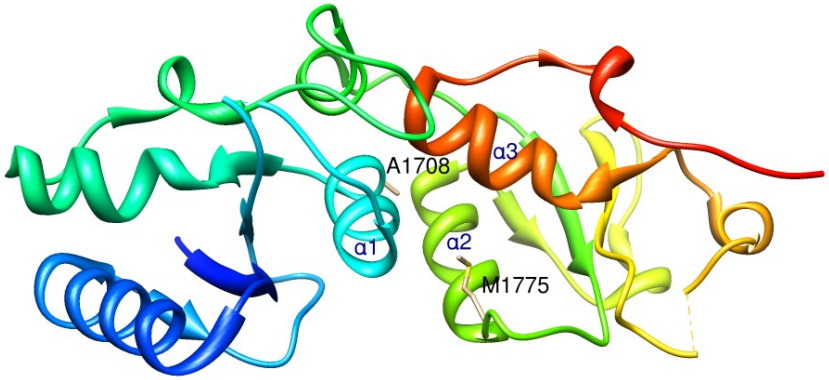

Figure 1. The crystal structure of human BRCT domains (PDB ID: 1JNX). The N-terminus is shown in blue; the C-terminus, in red. Residues A1708 and M1775 are depicted as ball and stick models. Three helices, α1 from the N-terminus and both α2 and α3 from the C-terminus, pack into a hydrophobic core that is important to the folding of BRCT domains.

A1708E would destabilize the BRCT interaction. On the other hand, although R1775 could be placed on the edge of the BRCT interface spatially, it positions a positive charge against the nearby R1835. Thus, both mutations would destabilize the BRCT core through either sterical incompatibility or disruption of electrostatic interactions [19]. This explanation found support from a mutation sensitivity assay that measures the stability of the inner domain interaction subject to proteolytic degradation. The wild-type protein resists the digestion by trypsin, elastase, and chymotrypsin, whereas the mutant with M1775R was partially degraded and A1708E was almost completely degraded [19]. The BRCT structure and in vitro experiments suggest that the genetic variants A1708E and M1775R cause the BRCA1 defect by destabilizing its inner-domain interaction.

From this example, we can see that crystal structure can be a powerful tool in interpreting possible consequences of nsSNPs by physicochemical principles. However, we cannot reasonably expect every protein and its mutants to have high-resolution three-dimensional (3D) structures or homology models available, either because of difficulties in structural determination, such as for membrane proteins, or because some proteins are intrinsically disordered [20].

To overcome this severe limitation, many computational tools aiming to predict structural properties use sequence information as input, either by direct use of sequence or through derived features such as amino acid composition and sequence motifs. Here, we describe a stability prediction method proposed by [21], namely MUpro, which was based on a sophisticated machine learning technique–Support Vector Machine (SVM)–and which achieved good performance.

In traditional molecular dynamics simulation, potential functions from a force field were usually calculated to obtain $\Delta\Delta G$, which was mainly influenced by interactions between nonlocal amino acids [22]. Although it is generally difficult, if not completely impossible, to infer protein structural architecture accurately based solely on amino acid sequence, pioneering work from [23, 24] showed that protein sequence was effective in the prediction

of secondary structure and solvent accessibility. MUpro fit a set of features derived from protein sequence to an experimental stability data by nonlinear transformation through SVM. The ProTherm database [25] collects from the literature a range of experimentally measured thermodynamic parameters, such as Gibbs free energy changes for wild-type and mutant proteins, with experimental conditions, including pH and temperature. From ProTherm MUpro used protein sequences and mutations for training and test purposes, along with numeric energy changes.

MUpro adopted a standard binary classification scheme in feature generation by selecting a window centered on a mutant position and then encoding each amino acid in the window as a vector of 20 elements. In this kind of vector, each element corresponds to one of 20 standard amino acids and takes a value of 1 if the corresponding amino acid is identical to the one observed or else 0. MUpro considered a window of seven amino acids for each mutation, thereby representing the feature set by a 140-element vector. The first 20-element vector records information about wild-type and mutant amino acids at the mutant position, and the final six vectors document the six flanking amino acids.

In a two-dimensional space, linear classifiers are designed to separate two classes of data points by a straight line. As illustrated in Figure 2 (left plot), any lines passing through the space between two parallel lines can separate the blue points (one class) from the orange (the other class) perfectly, and thus would be a good choice for linear classification. However, SVM algorithms [26] would select the dashed line, which distances two lines equally, as the class boundary. In other words SVMs optimize a margin separator that maximizes its distance to data points. Figure 2 shows the margin m between two classes, which is the optimization object in SVMs algorithm. Mathematically, larger m is expected to provide the classifier greater generalization, which measures how well the classifier performs on new, unseen data points.

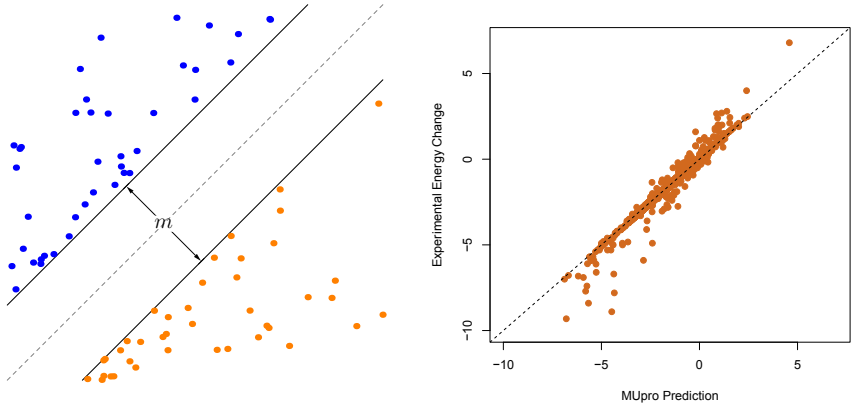

Figure 2. The left plot illustrates a linear classification on separable data with two classes (blue and orange). The class boundary (dashed line) is the middle line between two parallel lines. The right plot shows MUpro predictions against experimental values for 1,008 nsSNPs; points on the diagonal represent exact predictions.

When data sets overlap, SVMs still try to optimize a new objective function that considers both m and penalties from misclassification. Regardless of the separability of the data, m depends only on points located on the parallel lines (completely separable) or points located between them (partially separable). These points are called support vectors.

Besides data classification, SVMs can perform regression for data points with continuous response values, where the objective function measures the difference between prediction and actual values. But unlike typical linear regression, SVM regressions do not penalize differences falling within a predefined range.

The abilities of SVMs, however, go beyond linear classification and regression. By projecting the original data points into higher dimensional spaces, SVMs actually create additional, and usually more complex, features from the input points. By using the same linear settings as described above in these newly high-dimensional spaces, SVMs can effectively capture highly nonlinear relationships among data which otherwise would be missed.

MUpro applied a popular SVM implementation, SVM^{light} [27], to carry out energy change sign classification and regression. In 1,008 training mutations, MUpro performed rather well against true energy changes, with a root-mean-square deviation (RMSD) of 0.39 (Figure 2, right plot). Moreover, it made more accurate predictions with less dramatic actual stability changes between wild-type and mutant amino acids. Generally, MUpro tended to underestimate larger energy changes.

In one early comprehensive examination of the effects of nsSNPs on protein function, [28] catalogued nsSNP effects according to structural and sequence changes caused by the introduction of mutant amino acids. That study extracted 262 disease-causing missense variants from the HGMD and 42 neutral variants from hypertension-associated genes. Proteins harboring these variants either had 3D structures deposited in the Protein Data Bank (PDB) or they could find homologous ones with a sequence similarity of at least 40 percent. They then modeled both wild-type and mutant protein structures based on available 3D structures. By examining a broad range of physicochemical parameters from built models, including loss of hydrogen bonds, loss of a salt bridge, over-packing, and disruption of binding, Wang *et al.* could compare distributions of effects observed in disease-causing and neutral variants (Table 1). Their results clearly demonstrated that loss of stability accounts for many more disease-causing variants than neutral variants (83 versus 26 percent) and that 70 percent of neutral variants cause no measurable effects on the protein structure.

Effect	Disease	Neutral
Stability	83	26
Ligand binding	5	2
Other	2	2
No effect	10	70

Table 1. Percentage of effects from missense variants on protein function (adapted from Figure 2 in [28])

This survey suggests that nsSNPs giving rise to stability changes will more likely be disease-related than not, and this property might be useful in distinguishing disease-causing from neutral nsSNPs. Moreover, computational tools like MUpro capable of predicting

stability greatly facilitate this task by applying to virtually any protein with sequences available.

3. Functional impact of mutations

Besides structural consequences, variants can disrupt molecular functional sites, such as catalytic residues and DNA/protein binding sites, which are usually position-specific or share consensus motifs. Those disruptions, however, do not necessarily involve disruption of structure. A prominent class of sites that variants would affect consists of diverse PTM sites, of which some of the most frequent types are phosphorylation, glycosylation, acetylation, methylation, and ubiquitination. PTMs play an important role in cellular signal transduction and regulation, and activating and inactivating certain key proteins rely on precise modulation of PTMs in cell activities. For instance, without environmental stress, p53 is suppressed through ubiquitination catalyzed by E3 ubiquitin ligases, while in the presence of stress, such as DNA damage, p53 is activated by a variety of PTM enzymes, including acetylation and phosphorylation on its flexible DNA-binding domain [29]. PTM sites and flanking residues generally form consensus sequences with a high degree of variety, and therefore variants within these enzyme-specific motifs could abolish known functionalities or create new ones. This section starts by detailing two concrete examples of functional changes due to variants, followed by a description of DisPhos (Disorder-enhanced Phosphorylation sites predictor), an established phosphorylation predictor, and then explain how the concepts of gain and loss of phosphorylation can be used to analyze a cancer data.

FGFR2 (fibroblast growth factor receptor 2), one of four members of FGFR family of receptor tyrosine kinases, plays an important role in transmembrane signal transduction. Recent research identified one missense mutation, A628T, as being involved in LADD syndrome through severely impairing the kinase activity of FGFR2 [30]. Residue A628 is in the center of the catalytic pocket in the tyrosine kinase domain of FGFR2. A mutant structure, A628T-FGFR2 [31], reveals that the substitution of the smaller amino acid alanine at position 628 with the larger, polar threonine pushes one of the key residues, R630, out of the catalytic pocket; that movement disrupts the hydrogen bond between D626 and R630 existed in the wild-type structure (Figure 3, left). Although the position of D626 remains almost unchanged, R630 is too far away from the catalytic pocket and fails to stabilize the interaction with substrates, which consequently greatly compromises the catalytic ability of FGFR2. Compared with wild-type FGFR2, the A628T-FGFR2 mutant has roughly the same structure but highly reduced kinase activity.

It has been observed that amino acid substitutions occurred on non-PTM-sites could spread their influence to neighboring PTM sites on the same protein. One of such examples is PTPS, human PTP (protein tyrosine phosphatase) synthase, which catalyzes triphosphate elimination. PTPS participates in the biosynthetic pathway for tetrahydrobiopterin (BH4). Lack of PTPS catalytic activity causes a deficiency of BH4, which in turn leads to hyperphenylalaninemia (HPA), an autosomal recessive disorder. Missense mutation R16C was associated with HPA and resulted in reduced activity of PTPS [32]. Moreover, phosphorylation of S19 on PTPS is required for maximal enzyme activity [33]. So how does R16C affect phosphorylation on S19? There are multiple potential explanations. One is that the structure of PTPS shows the exposure of both R16 and S19 on the surface of the protein (Figure

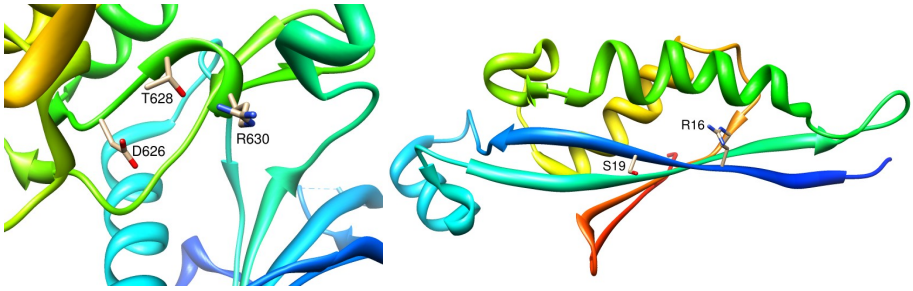

Figure 3. The crystal structure of the catalytic pocket of the A628T-FGFR2 mutant (left, PDB ID: 3B2T) and ribbon view of human PTPS structure (right, PDB ID: 3I2B). In both cases, the N-terminus is colored in blue and the C-terminus in red. Residues of interest are depicted as ball and stick models.

3, right; [34]) that forms the consensus sequence $R_{16}XXS_{19}$ for cGMP protein kinase II. The substitution C16 disrupts this kinase-recognizable motif and thus hinders phosphorylation, which ultimately leads to the inactivation of PTPS. Another explanation is that a removal of R16 prevents a salt bridge between it and a phosphate group when attached, which in turn results the loss of stability of the modified protein.

As with the stability prediction tool MUpro, described in the previous section, experimental difficulties have promoted the development of computational approaches to estimating many common PTM sites based on protein sequence. For the prediction of phosphorylation, DisPhos differs from other available methods like NetPhos [35] and ScanSite [36], since its model explicitly includes a range of characteristic features from the predicted disorder region around the phosphorylation site [37].

In some cases, researchers have found phosphorylation sites located on intrinsically disordered regions or have observed disorder-to-order or order-to-disorder conformational changes upon phosphorylation [38]. DisPhos exploited such observations by integrating predicted disorder information with the motif profile to improve its predictive performance.

Because phosphorylation occurs on residues S, T, and Y (S/T/Y), DisPhos assembled three pairs of positive-negative data sets, with each pair corresponding to one residue-specific predictor. First, it extracted proteins with phosphorylation annotations from UniProt (Universal Protein Resource); it then combined this data with data from Phospho.ELM [39]. DisPhos placed a 25-residue segment centered on each annotated S/T/Y into a positive set, while placing the same length segment around every non-annotated S/T/Y on the same protein into a negative set. To reduce the sequence bias caused by homologs or duplications, DisPhos only kept entries with a pairwise sequence similarity of less than 30 percent, which means that it allowed up to seven matches from alignment without gap. Due to the small size of experimentally verified phosphorylation sites, the filtered data sets were highly unbalanced (Table 2).

DisPhos used a broad range of features to discriminate positive from negative sites (Table 3).

To cope with the highly dimensional, yet sparse feature space, DisPhos performed feature selection by applying a permutation test to binary features and applying principal component

Residue	Positive Sites (P)	Negative Sites (N)	N/P Ratio
S	613	10,798	17.6
T	140	9,051	64.7
Y	136	5,103	37.5

Table 2. Data sets used in DisPhos (adapted from Table 1 in [37])

Type	Features	Dimension
Amino acid composition	Binary coding	480
Amino acid frequency	Binary coding	20
Disorder	VLXT, VL2, VLV, VLC, VLS	5
Secondary structure	Helix, loop and sheet	7
Sequence property	Complexity and flexibility	2
Residue property	Net charge, aromatic content, Hydrophobic moment, Hydrophobicity, exposed/buried	5

Table 3. Descriptive and predicted features used in DisPhos training.

analysis (PCA) to continuous features and then fitted logistic regression models to the transformed data sets.

Generally, binary classifiers work best in settings of balanced or close to balanced data sets in terms of accuracy, sensitivity, and specificity. For a classification in which the class boundary is determined by a solution that maximizes accuracy–the default configuration for many popular classifiers–training on highly unbalanced data sets inevitably results in extreme values for sensitivity or specificity, ultimately leading to poor generalization. DisPhos adopted an ensemble strategy to correct this issue in the S/T/Y data sets.

The combination of data filtering, feature selection, and sophisticated training and test configurations enabled DisPhos to achieve accuracy ranges between 70 and 80 percent, an improvement over the accuracy of other similar predictors. Moreover, the features derived from disorder predictions improved the accuracy by two percent on average, and these improvements showed the usefulness of disorder features in the prediction of phosphorylation sites.

DisPhos represents outcomes as probabilities, which quantitatively measure the likelihood that the underlying residues are phosphorylation sites. This characteristic facilitated the definition of gain and loss of phosphorylation for a specific site [40], and since these concepts can be interpreted readily, they may help provide insight into the underlying molecular mechanisms of mutations associated with diseases. Actually, the definitions of gain and loss are not limited to phosphorylation sites and can apply just as well to many other functional and structural properties.

Using bioinformatics tools that predict functional and structural attributes on both wild-type and mutant protein sequences provides us with two probabilistic estimates for a property p: $P(p = 1 \text{ at } s_i^w)$ and $P(p = 1 \text{ at } s_i^m)$ at site s_i, with s_i^w denoting a wild type site and s_i^m denoting a mutant site. Then, conceptually, we have

$$P(\text{loss of property } p \text{ at site } s_i) = P(p = 1 \text{ at } s_i^w \text{ AND } p = 0 \text{ at } s_i^m). \quad (1)$$

Given that s^w and s^m are actually different molecules, we consider that $P(p = 1 \text{ at } s_i^w)$ and $P(p = 0 \text{ at } s_i^m)$ are not dependent because of any underlying process. Therefore, we can expand the right hand of equation (1) as a product:

$$\begin{aligned} P(p = 1 \text{ at } s_i^w \text{ AND } p = 0 \text{ at } s_i^m) &= P(p = 1 \text{ at } s_i^w) \cdot P(p = 0 \text{ at } s_i^m) \\ &= P(p = 1 \text{ at } s_i^w) \cdot [1 - P(p = 1 \text{ at } s_i^m)] \end{aligned} \tag{2}$$

By substituting equation (1) with equation (2), we get

$$P(\text{loss of property } p \text{ at site } s_i) = P(p = 1 \text{ at } s_i^w) \cdot [1 - P(p = 1 \text{ at } s_i^m)] \tag{3}$$

Likewise, we can define gain of a property as

$$P(\text{gain of property } p \text{ at site } s_i) = [1 - P(p = 1 \text{ at } s_i^w)] \cdot P(p = 1 \text{ at } s_i^m) \tag{4}$$

Figure 4 shows the contour of gain of a property. Note that we can still compute gain/loss even if the predictions for the property are the same for wild-type and mutant sequences. The value of gain/loss varies from 0 to 0.25 when both predictions take a value of 0 through 0.5.

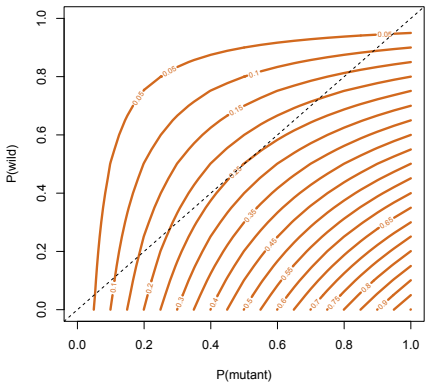

Figure 4. The contour of gain of property with respect to probability on mutant sequence–x-axis, $P(\text{mutant})$–and wild-type sequence–y-axis, $P(\text{wild})$). The dashed line denotes sites with equal probabilities for the two types of sequences.

[40] showed one application of gain and loss of phosphorylation. An experiment in their study collected 1,099 breast and colorectal cancer nsSNPs occurring on 847 proteins from a large-scale cancer-tumor-sequencing project [41]. Radivojac *et al.* then paired control and mutation data by randomly mutating on the same set of 847 wild-type proteins at the codon level. Their study then calculated gain and loss of phosphorylation for each mutation in both data sets, and found that disease-associated nsSNPs were significantly more likely to be involved in adding new phosphorylation sites (Table 4).

Phosphorylaiton change	Disease nsSNPs	Control nsSNPs	*P*-value
Gain	1.91	0.86	0.014
Loss	1.70	1.50	0.59

Table 4. Percentage of mutations predicted to have undergone gain or loss of phosphorylation. *P*-values were computed by *t*-test.

This survey showed how the concepts of gain and loss of phosphorylation could distinguish cancer-associated from neutral somatic mutations; it also suggested that they could serve as useful features for discriminating between general disease-related nsSNPs and neutral ones.

4. Mutation prediction: MutPred

In light of the above observations on the wide variety of consequences of a single mutation, we developed a large range of features for each variant and employed a popular machine learning technique, random forest, to distinguish disease-associated mutations from neutral ones. We called the model MutPred [42].

In a supervised learning scenario, we collected two sets of disease-associated mutations. One set came from the HGMD [3], in which 95 percent of mutations were annotated to monogenic diseases. We extracted the other set from a cancer-sequencing project [41]. Also, we created two corresponding control data sets (Table 5). For the HGMD data, we took a set of variants from UniProt that were annotated as polymorphisms to serve as controls (SPP). We identified all neutral mutations that occurred on the same proteins observed in the cancer data set and used them as the cancer controls. On average, HGMD proteins harbored 7.3 times as many variants as SPP proteins, while we observed a much less dramatic difference between cancer data set and its controls.

Data set	Mutations	Proteins	Type
HGMD	39,218	1,879	Disease
SPP	26,439	9,305	Neutral
Cancer	653	519	Disease
Cancer control	1,016	312	Neutral

Table 5. Summary of disease and neutral data sets.

We generated a total of 130 numeric attributes based on protein sequences for each mutation and utilized them as the input into a random forest classifier. These attributes can be divided into three major types (Table 6). Other evolutionary attributes include position-specific scoring matrix (PSSM) generated by PSI-BLAST, Pfam domain profile, and transition frequency from SNAP [43].

As the PTPS example shows, the influence of nsSNPs could spread to neighboring PTM sites. Accordingly, we expanded the definitions for gain/loss of structural and functional properties to pick up the largest gain/loss changes within an 11-residue window centered on the mutant position.

Random forest is an ensemble learning technique based on a population of binary decision trees, each of which is grown on a proportion of randomly chosen features and bootstrapped samples [54]. For classification, the outcome is the majority voting of individual trees.

Type	Property	Software
Functional properties	DNA-binding residues	DBS-PRED [44]
	Catalytic residues	†
	MoRFs	[45]
	Phosphorylation sites	DisPhos [37]
	Methylation sites	[46]
	Glycosylation sites	†
	Ubiquitination sites	[47]
Structure and dynamics	Secondary structure	PHD/Prof [48]
	Solvent accessibility	PHD/Prof [48]
	Stability	MUpro [21]
	Intrinsic disorder	DISPROT [49]
	B-factor	[50]
	Transmembrane helix	HMMTOP [51]
	Coiled-coil structure	marcoil [52]
Evolutionary information	Sequence Conservation	SIFT [13]
		Conservation index‡[53]

Table 6. Major attributes used in MutPred. † unpublished in-house program. ‡ used in latest version of MutPred.

Compared to a normal single decision tree, each subtree within a random forest uses only partial features and samples, which results in small correlations among subtrees and effectively reduces the overall variance of the model. Moreover, random forests inherit some attractive properties from decision trees, such as robustness to outliers and ease of interpretation.

In our model, we specified 1,000 trees to build the classifier between disease and neutral mutations. The HGMD achieved better accuracy than the somatic cancer data, suggesting that monogenic disease-related mutations are more suited to MutPred than somatic cancer mutations (Table 7). This is likely due to the large number of passenger variants (not causative) in tissue cancer sequencing data sets. Also, in terms of area under the curve (AUC) MutPred observed 0.86 in HGMD and 0.69 in cancer data sets (Figure 5, left).

Data set	Sensitivity	Specificity	Accuracy
HGMD	76.8	79.0	77.7
Cancer	60.9	68.4	65.5

Table 7. Percentage of classification performance measurement for HGMD and cancer data sets.

MutPred can provide not only comparable predictions for a mutation's predisposition to cause diseases [55], but it also allows the estimation of the significance level for individual gain/loss of properties (Figure 5, right). It is reasonable to assume that the distribution of property p in the neutral data set provides an unbiased approximation of the true null distribution, given the fact that UniProt provided the largest available set of curated neutral variants. Therefore, we could generate hypotheses about the molecular mechanism underlying variants at three different confidence levels: (1) actionable hypotheses: $0.78 \geq$ MutPred score > 0.5 AND property score < 0.05; (2) confident hypotheses: MutPred score > 0.78 AND $0.01 \leq$ property

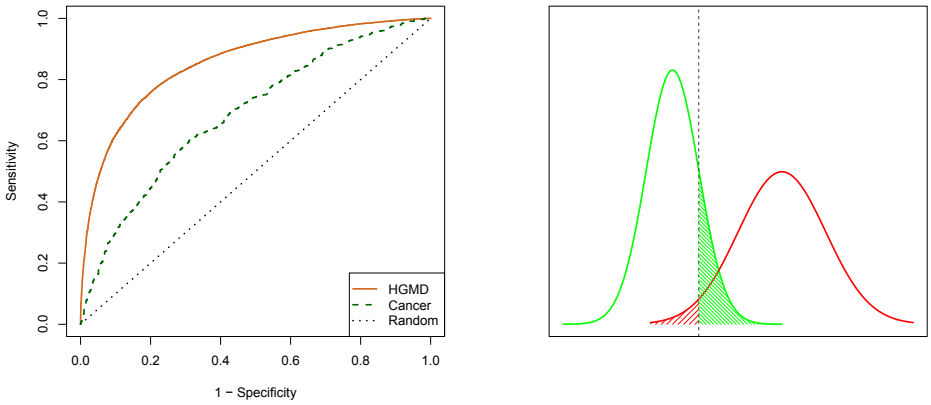

Figure 5. The Receiver Operating Characteristic (ROC) curves for HGMD and cancer data sets (left), and example distributions of gain/loss property p in neutral and disease sets (green and red, respectively; right). An empirical distribution of the putatively neutral substitutions can be used to define a threshold r on the false positive rate that, in turn, can be used to accept/reject the null hypothesis on new substitutions. The area shaded in green represents the P-value threshold (corresponding to the score r) that is used by MutPred to hypothesize molecular cause of disease. A particular area under the right tail of the neutral distribution is referred to as the property score.

score < 0.05; (3) very confident hypotheses: MutPred score > 0.78 AND property score < 0.01, where 0.78 corresponds to specificity 0.95 in HGMD data set.

We applied MutPred to 203,899 nsSNPs deposited in the dbSNP (build 135) and examined the score distribution and frequent hypotheses behind predicted deleterious mutations. In general, 35 percent of mutations were predicted with scores higher than 0.5; thus, we classified them as disease-associated (Figure 6). Of these deleterious mutations, 19.6 percent got at least one functional or structural hypothesis of possible molecular mechanism. The top three hypotheses all pointed to structural changes: gain of disorder (9.7 percent), loss of stability (8.5 percent), and loss of disorder (6.2 percent). This result agrees with [28]–at least in the sense that these changes are the most frequently seen. On the other hand, common functional alterations involved in disease included loss of MoRF binding (6.0 percent), gain of methylation (5.9 percent), and gain of catalytic residue (5.6 percent).

5. Conclusion

Understanding mutation data generated in biomedical research stimulates the development of computational methods. Previous studies have revealed structural and functional impacts on underlying proteins from variants, and research has proven that these impacts can differentiate between disease-associated and neutral mutations. Most current prediction tools have taken advantage of these characteristics, along with evolutionary information readily available from sequence alignment. Such tools have demonstrated impressive classification

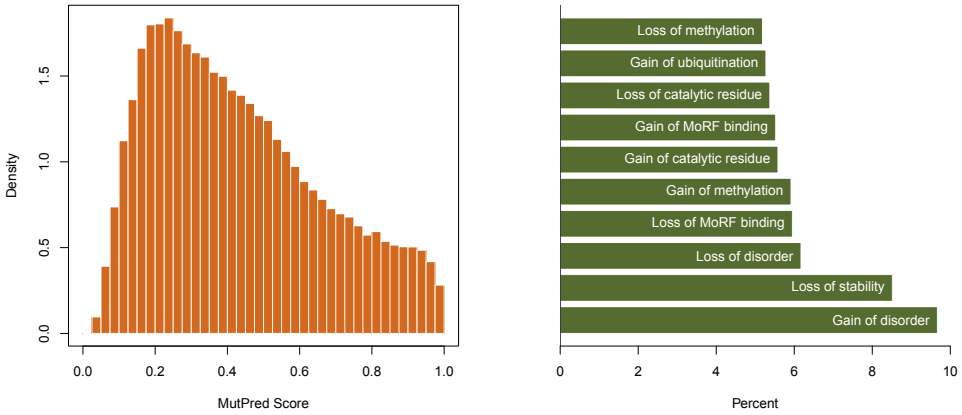

Figure 6. The distribution of MutPred scores for nsSNPs from dbSNP (left), and the top ten hypotheses for disease-associated mutations (right). The density on the left is a normalized frequency to ensure a total area in the bar plot equals one.

accuracy in monogenic disease-associated mutations but have performed less well for cancer somatic mutations. One explanation from an evolutionary perspective for this descrepency is that cancers usually arise late in life, so they are subjected to less purifying selection. This makes conservation information in cancers less useful than in monogenic diseases [56]. This field faces two immediate challenges: (1) How can we improve these tools to improve performance with somatic mutations? If the consensus opinion holds that tools depending on evolutionary knowledge are less effective than when applied to monogenic-disease-related mutations, it seems that research should explore other avenues. Inclusion of the mutation context in the model–e.g., pathways containing disease proteins–might offer a starting point for new directions. (2) How can we more accurately elucidate the molecular mechanisms for predicted deleterious mutations? MutPred has demonstrated this concept through definitions of gain/loss of individual properties. Similar features should be considered once they prove capable of reliably discriminating between disease-associated and neutral mutations. By continuously improving our computational tools, we can obtain better and more accurate understandings of biology and human health.

Author details

Biao Li
The Buck Institute for Research on Aging, Novato, CA 94945, USA

Predrag Radivojac
Indiana University, Bloomington, IN 47405, USA

Sean Mooney
The Buck Institute for Research on Aging, Novato, CA 94945, USA

6. References

[1] David Altshuler, Mark J. Daly, and Eric S. Lander. Genetic mapping in human disease. *Science*, 322(5903):881–888, 2008.

[2] 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, 2010.

[3] Peter D Stenson, Matthew Mort, Edward V Ball, Katy Howells, Andrew D Phillips, Nick St Thomas, and David N Cooper. The human gene mutation database: 2008 update. *Genome Med*, 1(1):13, 2009.

[4] Jamie K Teer and James C Mullikin. Exome sequencing: the sweet spot before whole genomes. *Hum Mol Genet*, 19(R2):R145–51, 2010.

[5] Jens G. Lohr, Petar Stojanov, Michael S. Lawrence, Daniel Auclair, Bjoern Chapuy, Carrie Sougnez, Peter Cruz-Gordillo, Birgit Knoechel, Yan W. Asmann, Susan L. Slager, Anne J. Novak, Ahmet Dogan, Stephen M. Ansell, Brian K. Link, Lihua Zou, Joshua Gould, Gordon Saksena, Nicolas Stransky, Claudia Rangel-Escareño, Juan Carlos Fernandez-Lopez, Alfredo Hidalgo-Miranda, Jorge Melendez-Zajgla, Enrique Hernández-Lemus, Angela Schwarz-Cruz y Celis, Ivan Imaz-Rosshandler, Akinyemi I. Ojesina, Joonil Jung, Chandra S. Pedamallu, Eric S. Lander, Thomas M. Habermann, James R. Cerhan, Margaret A. Shipp, Gad Getz, and Todd R. Golub. Discovery and prioritization of somatic mutations in diffuse large b-cell lymphoma (dlbcl) by whole-exome sequencing. *Proceedings of the National Academy of Sciences*, 109(10):3879–3884, 2012.

[6] Lucia A. Hindorff, Praveen Sethupathy, Heather A. Junkins, Erin M. Ramos, Jayashri P. Mehta, Francis S. Collins, and Teri A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

[7] C D Bottema, R P Ketterling, S Ii, H S Yoon, J A Phillips, 3rd, and S S Sommer. Missense mutations and evolutionary conservation of amino acids: evidence that many of the amino acids in factor ix function as "spacer" elements. *Am J Hum Genet*, 49(4):820–38, Oct 1991.

[8] M. P. Miller and S. Kumar. Understanding human disease mutations through the use of interspecific genetic variation. *Hum Mol Genet*, 10(21):2319–28, 2001.

[9] C Prives. How loops, beta sheets, and alpha helices help us to understand p53. *Cell*, 78(4):543–6, 1994.

[10] Y Cho, S Gorina, PD Jeffrey, and NP Pavletich. Crystal structure of a p53 tumor suppressor-dna complex: understanding tumorigenic mutations. *Science*, 265(5170):346–355, 1994.

[11] Andreas C. Joerger, Mark D. Allen, and Alan R. Fersht. Crystal structure of a superstable mutant of human p53 core domain. *Journal of Biological Chemistry*, 279(2):1291–1296, 2004.

[12] E Grasbon-Frodl, Holger Lorenz, U Mann, R M Nitsch, Otto Windl, and H A Kretzschmar. Loss of glycosylation associated with the t183a mutation in human prion disease. *Acta Neuropathol*, 108(6):476–84, Dec 2004.

[13] P C Ng and S Henikoff. Predicting deleterious amino acid substitutions. *Genome Res*, 11(5):863–874, 2001.

[14] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, Vasily E Ramensky, Anna Gerasimova, Peer Bork, Alexey S Kondrashov, and Shamil R Sunyaev. A method and server for predicting damaging missense mutations. *Nat Methods*, 7(4):248–9, 2010.

[15] L Pauling and H A Itano. Sickle cell anemia a molecular disease. *Science*, 110(2865):543–8, 1949.

[16] B C Wishner, K B Ward, E E Lattman, and W E Love. Crystal structure of sickle-cell deoxyhemoglobin at 5 a resolution. *J Mol Biol*, 98(1):179–94, 1975.

[17] Y Miki, J Swensen, D Shattuck-Eidens, P A Futreal, K Harshman, S Tavtigian, Q Liu, C Cochran, L M Bennett, and W Ding. A strong candidate for the breast and ovarian cancer susceptibility gene brca1. *Science*, 266(5182):66–71, 1994.

[18] L S Friedman, E A Ostermeyer, C I Szabo, P Dowd, E D Lynch, S E Rowell, and M C King. Confirmation of brca1 by analysis of germline mutations linked to breast and ovarian cancer in ten families. *Nat Genet*, 8(4):399–404, 1994.

[19] R S Williams, R Green, and J N Glover. Crystal structure of the brct repeat region from the breast cancer-associated protein brca1. *Nat Struct Biol*, 8(10):838–42, 2001.

[20] A K Dunker, J D Lawson, C J Brown, R M Williams, P Romero, J S Oh, C J Oldfield, A M Campen, C M Ratliff, K W Hipps, J Ausio, M S Nissen, R Reeves, C Kang, C R Kissinger, R W Bailey, M D Griswold, W Chiu, E C Garner, and Z Obradovic. Intrinsically disordered protein. *J Mol Graph Model*, 19(1):26–59, 2001.

[21] Jianlin Cheng, Arlo Randall, and Pierre Baldi. Prediction of protein stability changes for single-site mutations using support vector machines. *Proteins*, 62(4):1125–1132, 2006.

[22] D Gilis and M Rooman. Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. *J Mol Biol*, 272(2):276–90, 1997.

[23] P Y Chou and G D Fasman. Prediction of protein conformation. *Biochemistry*, 13(2):222–45, Jan 1974.

[24] N Qian and T J Sejnowski. Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol*, 202(4):865–84, Aug 1988.

[25] M D Shaji Kumar, K Abdulla Bava, M Michael Gromiha, Ponraj Prabakaran, Koji Kitajima, Hatsuho Uedaira, and Akinori Sarai. Protherm and pronit: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Res*, 34(Database issue):D204–6, 2006.

[26] Trevor Hastie, Robert Tibshirani, and J. H Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer series in statistics. Springer, New York, NY, 2nd edition, 2009.

[27] Thorsten Joachims. *Learning to classify text using support vector machines*, volume SECS 668. Kluwer Academic Publishers, Boston, 2002.

[28] Z Wang and J Moult. Snps, protein structure, and disease. *Hum Mutat*, 17(4):263–270, 2001.

[29] Christopher L Brooks and Wei Gu. p53 ubiquitination: Mdm2 and beyond. *Mol Cell*, 21(3):307–15, 2006.

[30] Imad Shams, Edyta Rohmann, Veraragavan P Eswarakumar, Erin D Lew, Satoru Yuzawa, Bernd Wollnik, Joseph Schlessinger, and Irit Lax. Lacrimo-auriculo-dento-digital syndrome is caused by reduced activity of the fibroblast growth factor 10 (fgf10)-fgf receptor 2 signaling pathway. *Mol Cell Biol*, 27(19):6903–12, 2007.

[31] Erin D Lew, Jae Hyun Bae, Edyta Rohmann, Bernd Wollnik, and Joseph Schlessinger. Structural basis for reduced fgfr2 activity in ladd syndrome: Implications for fgfr autoinhibition and activation. *Proc Natl Acad Sci U S A*, 104(50):19802–7, 2007.

[32] B Thöny, W Leimbacher, N Blau, A Harvie, and C W Heizmann. Hyperphenylalaninemia due to defects in tetrahydrobiopterin metabolism: molecular characterization of mutations in 6-pyruvoyl-tetrahydropterin synthase. *Am J Hum Genet*, 54(5):782–92, 1994.

[33] T Scherer-Oppliger, W Leimbacher, N Blau, and B Thöny. Serine 19 of human 6-pyruvoyltetrahydropterin synthase is phosphorylated by cgmp protein kinase ii. *J Biol Chem*, 274(44):31341–8, 1999.

[34] T Oppliger, B Thöny, H Nar, D Bürgisser, R Huber, C W Heizmann, and N Blau. Structural and functional consequences of mutations in 6-pyruvoyltetrahydropterin synthase causing hyperphenylalaninemia in humans. phosphorylation is a requirement for in vivo activity. *J Biol Chem*, 270(49):29498–506, 1995.

[35] N Blom, S Gammeltoft, and S Brunak. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J Mol Biol*, 294(5):1351–62, 1999.

[36] M B Yaffe, G G Leparc, J Lai, T Obata, S Volinia, and L C Cantley. A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nat Biotechnol*, 19(4):348–53, 2001.

[37] Lilia M Iakoucheva, Predrag Radivojac, Celeste J Brown, Timothy R O'Connor, Jason G Sikes, Zoran Obradovic, and A Keith Dunker. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res*, 32(3):1037–1049, 2004.

[38] D P Teufel, M Bycroft, and A R Fersht. Regulation by phosphorylation of the relative affinities of the n-terminal transactivation domains of p53 for p300 domains and mdm2. *Oncogene*, 28(20):2112–8, 2009.

[39] Holger Dinkel, Claudia Chica, Allegra Via, Cathryn M Gould, Lars J Jensen, Toby J Gibson, and Francesca Diella. Phospho.elm: a database of phosphorylation sites–update 2011. *Nucleic Acids Res*, 39(Database issue):D261–7, 2011.

[40] Predrag Radivojac, Peter H Baenziger, Maricel G Kann, Matthew E Mort, Matthew W Hahn, and Sean D Mooney. Gain and loss of phosphorylation sites in human cancer. *Bioinformatics*, 24(16):i241–7, 2008.

[41] Tobias Sjöblom, Siân Jones, Laura D Wood, D Williams Parsons, Jimmy Lin, Thomas D Barber, Diana Mandelker, Rebecca J Leary, Janine Ptak, Natalie Silliman, Steve Szabo, Phillip Buckhaults, Christopher Farrell, Paul Meeh, Sanford D Markowitz, Joseph Willis, Dawn Dawson, James K V Willson, Adi F Gazdar, James Hartigan, Leo Wu, Changsheng Liu, Giovanni Parmigiani, Ben Ho Park, Kurtis E Bachman, Nickolas Papadopoulos, Bert Vogelstein, Kenneth W Kinzler, and Victor E Velculescu. The consensus coding sequences of human breast and colorectal cancers. *Science*, 314(5797):268–74, 2006.

[42] Biao Li, Vidhya G Krishnan, Matthew E Mort, Fuxiao Xin, Kishore K Kamati, David N Cooper, Sean D Mooney, and Predrag Radivojac. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics*, 25(21):2744–50, 2009.

[43] Yana Bromberg and Burkhard Rost. Snap: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res*, 35(11):3823–35, 2007.

[44] Shandar Ahmad, M Michael Gromiha, and Akinori Sarai. Analysis and prediction of dna-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, 20(4):477–86, 2004.

[45] Predrag Radivojac, Slobodan Vucetic, Timothy R O'Connor, Vladimir N Uversky, Zoran Obradovic, and A Keith Dunker. Calmodulin signaling: analysis and prediction of a disorder-dependent molecular recognition. *Proteins*, 63(2):398–410, 2006.

[46] Kenneth M. Daily, Predrag Radivojac, and A. Keith Dunker. Intrinsic disorder and protein modifications: building an svm predictor for methylation. In *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2005*, pages 475–481, 2005.

[47] Predrag Radivojac, Vladimir Vacic, Chad Haynes, Ross R Cocklin, Amrita Mohan, Joshua W Heyen, Mark G Goebl, and Lilia M Iakoucheva. Identification, analysis, and prediction of protein ubiquitination sites. *Proteins*, 78(2):365–80, 2010.

[48] B Rost. Phd: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol*, 266:525–39, 1996.

[49] Kang Peng, Predrag Radivojac, Slobodan Vucetic, A Keith Dunker, and Zoran Obradovic. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*, 7:208, 2006.

[50] Predrag Radivojac, Zoran Obradovic, David K Smith, Guang Zhu, Slobodan Vucetic, Celeste J Brown, J David Lawson, and A Keith Dunker. Protein flexibility and intrinsic disorder. *Protein Sci*, 13(1):71–80, 2004.

[51] A Krogh, B Larsson, G von Heijne, and E L Sonnhammer. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J Mol Biol*, 305(3):567–80, 2001.

[52] Mauro Delorenzi and Terry Speed. An hmm model for coiled-coil domains and a comparison with pssm-based predictions. *Bioinformatics*, 18(4):617–25, 2002.

[53] J Pei and N V Grishin. Al2co: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, 17(8):700–12, 2001.

[54] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[55] Janita Thusberg, Ayodeji Olatubosun, and Mauno Vihinen. Performance of mutation pathogenicity prediction methods on missense variants. *Hum Mutat*, 32(4):358–68, 2011.

[56] Sudhir Kumar, Joel T Dudley, Alan Filipski, and Li Liu. Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations. *Trends Genet*, 27(9):377–86, 2011.