FEBS Letters

FEBS PRESS
science publishing by scientists

REVIEW ARTICLE

Protein function in precision medicine: deep understanding with machine learning

Burkhard Rost[1], Predrag Radivojac[2] and Yana Bromberg[3]

1 Department of Informatics and Bioinformatics, Institute for Advanced Studies, Technical University of Munich, Garching, Germany
2 School of Informatics and Computing, Indiana University, Bloomington, IN, USA
3 Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ, USA

Correspondence
B. Rost, Technical University of Munich, Boltzmannstr. 3, 85748 Garching, Germany
Fax: +49 89 289 19414
Tel: +49 89 289 17811
E-mail: assistant@rostlab.org

(Received 27 June 2016, revised 12 July 2016, accepted 12 July 2016, available online 6 August 2016)

doi:10.1002/1873-3468.12307

Edited by Wilhelm Just

Precision medicine and personalized health efforts propose leveraging complex molecular, medical and family history, along with other types of personal data toward better life. We argue that this ambitious objective will require advanced and specialized machine learning solutions. Simply skimming some low-hanging results off the data wealth might have limited potential. Instead, we need to better understand all parts of the system to define medically relevant causes and effects: how do particular sequence variants affect particular proteins and pathways? How do these effects, in turn, cause the health or disease-related phenotype? Toward this end, deeper understanding will not simply diffuse from deeper machine learning, but from more explicit focus on understanding protein function, context-specific protein interaction networks, and impact of variation on both.

Keywords: computational prediction; molecular mechanism of disease; protein function; variant effect

To avoid problems with the next car you buy, you may consult the reliability statistics for every make and model that you are considering. This may reduce the odds of unpleasant experiences and may help avoid lemons (car jargon: an, often new, car that is found to be defective only after it has been bought). Once your specific car fails, however, such precautions no longer help. Then an expert has to cycle through a diagnostic protocol, excluding potential problems one at a time, until finding the actual cause of failure. In some ways, health and car problems are similar; failure of many disparate parts may ultimately lead to the same observable effect. In this analogy, we might argue that medicine has so far been often investing into mitigating the inconvenience with lemons and much less into improving and augmenting the protocols for finding the individual causes of problems.

In his recent State-of-the-Union address, the US President Barack Obama announced the Precision Medicine Initiative, making this challenge a national and international priority. Precision medicine refers to all attempts to merge complex data from molecular biology and medicine, e.g., from genome, proteome, metabolome, microbiome, imaging, electronic health records, and mobile devices, 'to tailor treatment and prevention strategies to people's unique characteristics' [1]. Human cognitive abilities of compressing the available data into a coherent and consistent diagnostic or prognostic set of conclusions are fairly limited [2]. On the other hand, machine learning, a subfield of artificial intelligence 'that gives computers the ability to learn without being explicitly programmed' [3], can be used to reasonably model the level of complexity of this precision-medicine relevant data for predictive understanding.

Abbreviations
AUC, area under the ROC curve; COSMIC, Catalogue of Somatic Mutations in Cancer; HGMD, Human Gene Mutation Database; OMIA, Online Mammalian Inheritance in Animals; OMIM, Online Mammalian Inheritance in Man; ROC, receiver operating characteristic.

Today, one major problem in the way of any health-related advances is the lack of data availability, owing to both the difficulties of acquiring high-resolution molecular information and to the regulatory restrictions governing access to such data. For over a decade, the expertise for the analysis of complex data in molecular biology and medicine cannot be encapsulated in transferable programs. Instead, experts in computational biology and biomedical informatics need to access the data in ways that do not limit their success. The 'completely open behind completely closed doors' model realized by Genomics England illustrates the complexity of the situation [4]. We propose that the available crucial (and expensive) biological data cannot any longer stay hidden from the most appropriate expert and/or optimal processing tools. Once this human-imposed obstacle is eliminated, some of the major challenges for precision medicine will reside in combining advanced machine learning with deep annotations of mechanistic molecular models that explain observed effects, to point out correlations and, more importantly, identify causes.

Here, we discuss an extraordinarily important subset of these challenges [5,6] – the need to broaden the coverage and deepen the annotation of (tissue-specific) protein functions, as well as to improve meaningful predictions of variant effects on protein stability and interactions with small chemicals and other macromolecules. We suggest that the efforts toward this end will necessarily include the development of new machine learning techniques. However, perhaps more importantly, these problems will require learning to correctly identify, quantify, and present the biological problems to existing and novel machine learning algorithms. For all higher eukaryotes, including humans, these approaches will contribute to substantially improving the efforts to unravel the dark proteome [7].

Machine learning: teaching machines to understand biology

Advanced machine learning implies understanding through prediction

Over the past few decades machine learning has influenced many aspects of research and everyday life. Recently, deep learning [8] has led to many advances, including publicity-effective breakthroughs such as a machine (AlphaGo) besting man in the board game Go [9], Google learning computer games from players [10], or even an algorithm predicting who will survive the next episode of the Game of Thrones series [11]. Its success in molecular biology, arguably, began with the leap resulting from a machine first learning to predict protein secondary structure [12–15]; its applications have become ubiquitous ever since. Machine learning often succeeds by matching the complexity of the solution to that of the problem. This complexity matching is generally overlooked, and frequently leads nonexperts and novices to confuse advanced methods for 'black boxes' that obfuscate rather than reveal the underlying mechanisms. When machine learning does succeed, it captures understanding, but on a level of complexity that may be difficult to grasp for its human designers and users.

Over almost three decades of experience with machine learning in biology, we learned that the major usability challenges lie in ascertaining the machine's abilities to understand the mechanistic principles underlying the data, instead of tripping over correlations. In school, we all encounter the difference between understanding and learning by heart. Turning machine learning into predictive understanding is a challenging endeavor that requires deep familiarity with the theoretical foundations of the field, dataset and model selection, performance estimation, and ways in which probabilistic models can be applied.

What are we trying to learn? Formalizing the prediction task and data selection

A major challenge in computational biology and medicine pertains to how representative of reality the available experimental data are. Ideally, the data used to train computational models should be a uniformly random sample from the data on which the predictor will be used in all future applications. This, however, is impossible to assure, e.g., predictors of effects of protein variants on molecular function are often used to judge mutation involvement in disease. Moreover, even when the future application is explicitly specified, the uniform sampling of biological reality in training is almost never true. Databases of experimentally determined protein structures [16,17] and annotations [18,19], are extremely biased, i.e., they are representative of the current state-of-the-art of experimental science, investigators' abilities and preferences, and concentrated exploration efforts (e.g., cancer initiatives or drug target searches). Hence, they are unrepresentative of reality [7, 20–22].

We illustrate some issues with problem formulation and data selection on an example of relating blood pressure to heart disease. For this task, we can obtain sets of blood pressure measurements from people with and without disease (Fig. 1A). These datasets may be of a similar size and intended to characterize the

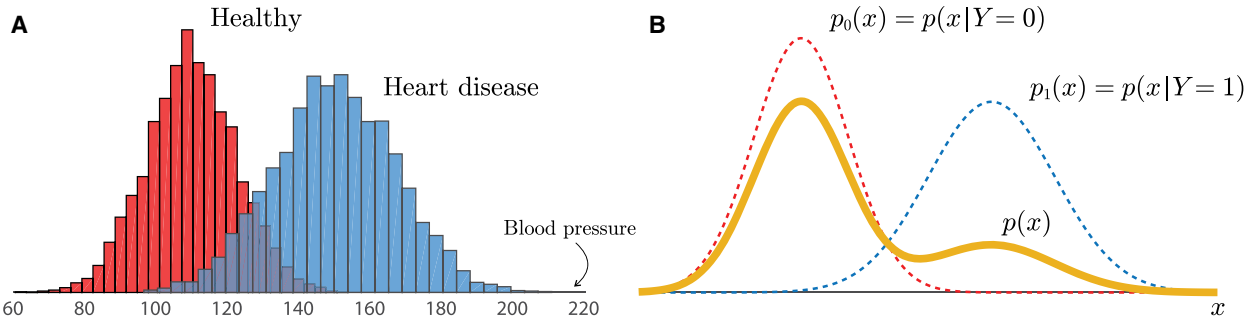

Fig. 1. Example for probability distributions. (A) Histograms of systolic blood pressure measurements from cohorts of subjects without (red) and with (blue) heart disease. (B) Probabilistic formulation where the cohort without heart disease (red, dashed) is represented by the distribution $p_0(x)$, the cohort with heart disease (blue, dashed) is represented by the distribution $p_1(x)$, and the underlying unlabeled data (yellow, solid) is represented by the distribution $p(x) = \alpha p_1(x) + (1-\alpha)p_0(x)$. The parameter $\alpha = P(Y=1)$ is referred to as class prior. In this example, we set $\alpha = 0.25$, i.e., we assume that a quarter of the population suffers from heart disease.

difference between the two populations. For example, a two-sample *t*-test applied to these data can provide a good measure of the difference between the average blood pressure of the two groups. The resulting *P*-value should, hopefully, be small enough to demonstrate statistical significance.

The problem with this type of setup in learning is that the collected samples may reflect a 'survivorship bias' [23]. That is, the odds of sampling 'healthy' patients from a population of those with family history of disease are generally higher than the odds of sampling truly healthy individuals who rarely consult physicians for this indication. Furthermore, both cohorts of individuals may contain health status errors, i.e., some seemingly healthy people may be sick, but below the threshold of diagnostic capabilities, and some of those with disease might have been misdiagnosed.

Although statistical significance of a difference between cohorts is often a useful metric, it is important to understand that it may be poorly related to the predictability of the phenomenon, i.e., the ability of the model to accurately assign a person's health status, given his/her blood pressure. The *P*-value from the example above reflects the probability that the observed difference between means of the two populations would be as large or larger if the sets of blood pressures were drawn from the same underlying distribution. Note that, in this definition, simply increasing the number of people in each cohort reduces the *P*-value until it reaches zero, if there is any discernible difference between the sample means. The accuracy of a health status prediction, on the other hand, is limited by the overlap between the two distributions and cannot be reduced below a certain quantity, regardless of the sample size.

We argue that, here, the goal of a machine learning approach is to approximate the posterior probability of the target concept, i.e., health status equals heart disease, given the evidence (here blood pressure). We write this probability as $P(Y=1|x)$, where Y is the random variable indicating health status (1 = disease; 0 = healthy) and x is the observation (blood pressure). In fact, we can even state that all that predictive machine learning needs to learn is the probability distribution; this abstract view is illustrated in Fig. 1B. For machine learning applications in computational biology and medicine, data have usually much higher complexity (dimensionality) than a simple blood pressure observation. Thus, this task is extremely complex, but entails the same concepts: training a classification model, say $f(x)$, that approximates as faithfully as possible the fixed but unknown underlying posterior probability (of heart disease, given the particular blood pressure) $P(Y=1|x)$, for every x. The simple binary decision then becomes 1 (heart disease) when $P(Y=1|x) > P(Y=0|x)$ and 0 (no heart disease) otherwise. Solid theory already supports using neural networks to approximate posterior distributions [24], while other models can be calibrated [25,26].

Why does the data need to be representative? To show this we use Bayes' rule to re-write the posterior probability as

$$P(Y=1|x) = \frac{P(x|Y=1)P(Y=1)}{P(x)} \qquad (1)$$

where we assume that x is discrete. We refer to $P(x|Y=1)$ as the class-conditional probability (of a particular blood pressure, given heart disease), $P(Y=1)$ (probability of heart disease) as the class prior, and $P(x) = P(x|Y=1)P(Y=1) + P(x|Y=0)P(Y=0)$ as the

probability of observing evidence x. If the distribution of $P(x|Y = 1)$ is not representative of the true distribution for the disease cohort, the classifier $f(x)$ will be biased and unable to approximate $P(Y = 1|x)$ correctly. Similar reasoning holds for the distribution of the values in the healthy cohort, which we can easily see by noting that $P(Y = 0|x) = 1-P(Y = 1|x)$.

The problem of biased data has been extensively studied in statistics and machine learning [27,28], offering a number of solutions supported by theory. However, none of these solutions fully apply to biology (much less to medicine) where the community expectation of the 'tool' is to recognize the concept regardless of any development issues or usability side-notes. Raising the awareness in the end user that the model relies on the data distribution $P(x)$ may raise the yield from these models. However, working to remove bias from training as much as possible would probably prove significantly more effective.

In molecular biology, some intriguing solutions to this problem were offered early on in the prediction of protein secondary structure or intrinsically disordered protein regions. There, data are clustered based on sequence identity thresholds [29], retaining a single representative sequence from each group. By eliminating 'redundant' sequences, these methods produce data distributions that are more uniform in the feature space as the similarity groups are unequal in size. This approach results in models that resemble having learned a concept, instead of a probability distribution. Well-defined theoretical support for this situation is an open problem that will formalize and improve understanding of this long-standing practice in computational biology.

How do we evaluate success? Effective performance estimation through cross-validation

In computer science terms, cross-validation implies data partitioning and (iteratively) using only some fraction of the available data to train/develop a model [30]. In any iteration, the remainder of the data points is used to establish whether the machine memorizes or understands. Here, for a true evaluation of performance, data points in each partition need to be unrelated in the defined feature space, to model as closely as possible appearance of new data. In standard out-of-the-box machine learning packages such as WEKA [31], cross-validation is applied to avoid getting stuck in local minima. However, data in biology are atypical for machine learning; it is too sparse and incomplete, too biased, and too noisy for the application of standard protocols. Furthermore, there often exist underlying relationships between data points, such as spatial proximity of two amino acids or two medical records of the members of the same family. As a result, most performance estimates remain short-lived and even experts often fail to distinguish fraudulently faulty estimates from the sustained sound ones. Data limitations make the task even more complicated, as illustrated by our development of a method that uses only protein sequence as input to predict whether two proteins physically interact [32]. Despite many levels of caution on top of standard cross-validation protocols, we initially ended up with a method predicting whether the two protein sequences originated from the PDB [16] or from UniProt [33], rather than whether their proteins interact [22]. This was a surprising result with interesting implications that created an irrelevant method. Proper cross-validation requires careful processing of the available experimental data, e.g., in this case, bias and overlap between training, cross-training, and testing sets had to be avoided through clustering. In fact, the application of machine learning methods in biology typically requires a detailed analysis of how to cluster the data and how to avoid overlap [22].

In a field marked by noisy and imbalanced data, the selection of proper performance measures is imperative. However, this is often an underappreciated aspect of applied machine learning – a place where theory meets practice. In binary classification, and in the absence of well-defined costs of (erroneous) classification, the field has converged to using a variety of measures such as the area under the receiver operating characteristic (ROC) curve (AUC), or F-measure [30]. However, each of these measures is flawed in some aspects, e.g., the AUC can provide deceivingly high numbers for imbalanced data, both ROC and AUC put the emphasis on performance in areas possibly not of interest for users, and the F-measure relies on unknown class priors (see Section 'How frequent are the concepts that we try to model? Estimation of class priors and posteriors'). We are yet to understand how to reason about score distributions of the prediction models on unlabeled data and how to best evaluate the outputs. In more complex classification tasks, such as the prediction of hierarchical structure in function prediction (e.g., Gene Ontology [34] or Human Phenotype Ontology [35]), the evaluation objectives are even less clear, as we are required to define similarities on ontological terms or subgraphs in the ontology [36–39]. For these reasons, characterizing the accuracy of a probabilistic model in computational biology and medicine approaches a form of art. While the community experience provides invaluable feedback about performance, it is necessary to further standardize evaluation protocols.

Overall, dataset selection and performance estimation are among the major challenges for turning machine learning into practically applied tools that enable us to fish for understanding in the ocean of big data. Unfortunately, most machine learning practitioners tend to spend a small fraction of their time and resources on dataset selection and performance evaluation and instead commit major effort to creating a new method. We argue that advanced machine learning applications in biology necessitate inverting this effort ratio.

How frequent are the concepts that we try to model? Estimation of class priors and posteriors

As we saw in Eqn (1), the representativeness of class-conditional distributions is necessary but not sufficient for the correct estimation of posterior probabilities; that is, not unless the positive and negative examples initially come in the right proportions. Therefore, one of the fundamental tasks in machine learning is proper estimation of prior probabilities, i.e., $P(Y = 1)$ in binary classification.

Semisupervised learning has been extensively researched. When both positive (a sample from $p_1(x)$; Fig. 1B) and negative (a sample from $p_0(x)$; Fig. 1B) data are available and representative of class-conditional distributions, the presence of representative unlabeled data (a sample from $p(x)$; Fig. 1B) guarantees a correct estimate of class priors [40]. However, in many cases, labeled data points are exclusively positive. For example, in training a model to recognize disease predisposition, it is easy to recruit positive samples as certain lab results can reliably establish that a person is sick. Negatives, on the other hand, are much harder to ascertain, as in many cases, it is nearly impossible to show definitively that the person is not currently sick (e.g., just below a diagnostic threshold), or will not get sick in the near future. In semisupervised learning, working in these scenarios is usually referred to as learning from positive and unlabeled data, or positive-unlabeled learning [41]. Here, class priors are not identifiable [42,43], i.e., there is no unique solution to the problem. Unless restrictive assumptions are made, we are only guaranteed to identify the upper bounds [43] of class frequency distributions. Several ideas have been recently explored in positive-unlabeled learning, from more [44–46] to less restrictive [43] approaches, giving reasonable estimates.

An extension of the positive-unlabeled framework should be further considered when the positive examples are contaminated with incorrectly labeled data points. Class priors in this situation remain unidentifiable [47,48] with the only available estimation algorithm of the upper bound just recently proposed [48]. A generalization of these solutions will be able to correct some forms of biased sampling in the estimation procedure. However, we are not aware of any algorithms that simultaneously handle biased class-conditional distributions.

Interestingly, Jain et al. [43,48] have also formulated the problem of learning the posterior probabilities from (noisy, high-dimensional) positive-unlabeled data. The authors propose that the optimal solution can be obtained by developing classifiers that distinguish between positive and unlabeled data. These positive-unlabeled classifiers can then be converted into classifiers that distinguish between positive and negative data using (nonlinear) deterministic transformations. It is important to keep the distinction between these two types of classifiers in mind because the unlabeled data may be 'contaminated' with a potentially large fraction of positives. That fraction equals $P(Y = 1)$, the positive class prior.

Precision medicine: proteins and disease mechanisms

Personalized to precision medicine: going molecular

While recent years have seen a push for personalized medicine, the term has been fiercely opposed by doctors who (accurately) noted that they 'practice personalized medicine' with every patient. To quote Sir William Osler, who was the first to require bedside clinical training for medical students so that they can better understand the difference between theory and practice [49]: 'If it were not for the great variability among individuals, medicine might well be a science and not an art' [50]. Unlike personalized medicine, precision medicine implies a more scientific approach to patient categorization, encompassing the idea that molecular and other individual-specific information could improve the precision with which the patients are diagnosed and treated [51]. For instance, integrating the patient hemoglobin levels, gender, age, tumor stage and location, and treatment dose could predict survival probability after a laryngeal carcinoma diagnosis [52]. Adding in gene expression profiling could then further contribute to accurate diagnosis, prognosis, and treatment assessment of this disease in individual patients [53]. Precision medicine is, thus, about increasing the level of resolution in medical practice and health choices. As such, it should integrate as many diverse types of high-resolution data as possible,

which could contribute to the proper patient categorization, selection of treatment, and informed lifestyle adjustments.

Hereafter we narrow down the focus of our review to the molecular aspects of precision medicine, particularly those related to the understanding of protein function and functional consequences of mutation. We argue that the interplay of recent research in basic life sciences and machine learning, together with technological advances, lays the foundation for successful translation of decades of biological, statistical, and computer science research into successful drug discovery and clinical practice.

How different are we? Understanding the effects of sequence variants

Sequence variation and structural changes that cause disease were first linked for sickle-cell anemia (or sickle-cell disease) through the work of Pauling and Perutz [54,55]. The sickle-cell disease is an autosomal recessive disorder caused by the amino acid substitution E6V in the β-chain of human hemoglobin [56], i.e., a replacement of glutamic acid (E) by a valine (V) at position 6 (not counting the first methionine, otherwise E7V, Fig. 2). Today, the Human Gene Mutation Database (HGMD [57]), Online Mammalian Inheritance in Man (OMIM [58]), Online Mammalian Inheritance in Animals (OMIA [59]), the Catalogue of Somatic Mutations in Cancer (COSMIC [60]), and other resources collect thousands of such single amino acid variants causative of or associated with disease, in addition to many other types of sequence variation.

The disease-causing variants from genetic databases are in sharp contrast to the numerous single amino acid differences between healthy individuals: any pair of (random) individuals differs by almost one variant in every other protein (roughly 10 000 in total) [63]. For most variants, even those experimentally implicated in disease, we have very little experimental information. In fact, while the sequence conservation of OMIM residues is higher than expected, for most of the known OMIM variants [64,65], their protein 3D structure is neither experimentally known, nor can be modeled [65,66]. Similarly underrepresented are functional annotations and there is a clear lack of experimental evaluations of variant effects [65]. Simply put, only *in silico* predictions can at this time bridge the gap between existing and desired information.

Many methods predict the effect of variants on protein structure or molecular function, as well as their involvement in disease [67–73]. However, due to the lack of comprehensive and unbiased experimental data for evaluation, their performance remains to be fully understood. The methods tend to perform well for existing data and to agree more with each other for annotated variants than for unknown variants. This might hint at a tendency of over-training [74]. However, what is even more alarming is that the goals of individual methods are often not well defined. For instance, a method trained to recognize 'disease' variants from variants between orthologous proteins may be labeled and used as predicting 'functional effect,' even in the absence of explicit evaluation of this particular ability. Furthermore, directional non-neutrality of variants, as compared to wild-type, is often missed, i.e., the class of variants termed 'deleterious' or 'damaging' does not represent 'advantageous' variants. Finally, destabilizing variants are not always 'deleterious' and variants that do not impact structure are not

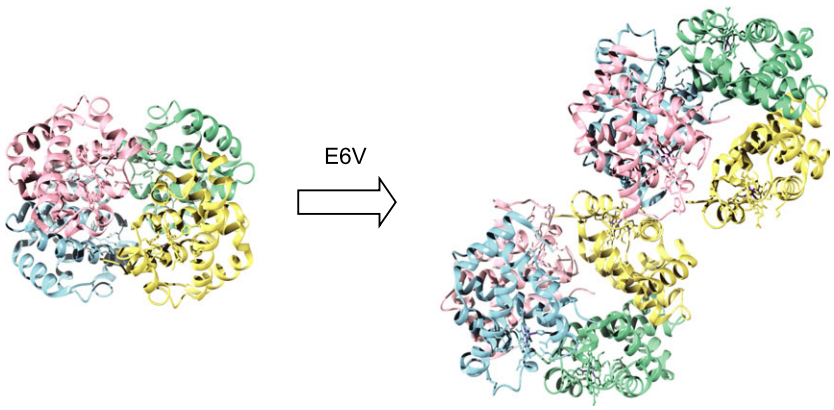

Fig. 2. The molecular mechanism underlying sickle-cell disease through an E6V mutation. (Left) The unmodified version of hemoglobin as a hetero tetramer consisting of two α and two β hemoglobins; PDB ID: 4hhb [61]. (Right) The introduction of valine recruits another molecule of hemoglobin likely through a hydrophobic interaction with F85 and L88 of the interacting molecule. The octameric complexes further polymerize, ultimately leading to fibrils that physically alter the shape of red blood cells. PDB ID: 2hbs [62].

necessarily 'neutral.' The various impacts of variants are correlated, but the concepts are not interchangeable. In fact, the effect of variants can be thought of in three categories – impact on evolutionary fitness, change in molecular functionality or protein structure/dynamics, and pathogenicity, i.e., induction of disease. A recent study proposed to expand the clinical utility of these models by focusing on the prediction of endophenotypes [75].

Directly disease-causing, often monogenic, variants will by definition alter protein function. Methods that predict molecular effects in turn identify variants from OMIM and OMIA as having very strong effects [64,76]. In fact, the predicted effects for these variants are even stronger than those of data points used in training a classifier to distinguish functionally significant variants [64,76]. A surprising behavior in machine learning, this result is most likely explained by the unusual experimental observation consistency, i.e., experimentally determined functional effects may vary between experiments and subjective evaluations, particularly for weak effects, but disease variants stand out objectively as unusually devastating to molecular function. This observation highlights another strength of machine learning in biology: consistent reproducibility of data tends to imply stronger effects. Prediction methods are then able to pick up strength of signal without ever learning to recognize these strengths on experimentally labeled data, i.e., we can predict protein–protein interaction hotspots [77] and the strength of variant effects [78,79] without using strength of effect labels in training. Here, the probabilistic outputs of machine learning tools correlate with severity of effect.

Arguably, for the purposes of elucidating disease mechanisms, one major challenge for both experimental and computational methods is to correctly identify variants with molecular and/or cellular effects. However, this knowledge of individual variant functional effects alone is not sufficient. As many as a quarter to nearly half of the variants between healthy individuals are predicted to impact function [80]. A significant fraction is predicted to have very strong effects. How can we pick the one that causes an OMIM-like disease in a sea of those that appear equally strong for each of us? And what about the others? These variants most certainly cannot all be pathogenic in the causative sense. However, they can (and likely do) create a disease-permissive background, which is likely involved in triggering manifestations of complex disease, providing support in favor of the so-called infinitesimal model of complex genetic disease [81]. Specifically, this can be interpreted as a vulnerability of an interaction pathway, whose multiple members are affected by individual-specific variation. In this light, many groups have been trying to move from the level of individual residues and proteins to that of protein interaction networks and the investigation of pathway-specific enrichments [82–85], and on to disease trajectories [86].

Why is mechanistic understanding important? Mapping sequence variants to molecular mechanisms through function prediction

The alteration of overall cellular activity often arises as a consequence of altered function of one or more individual proteins. The limiting factor in

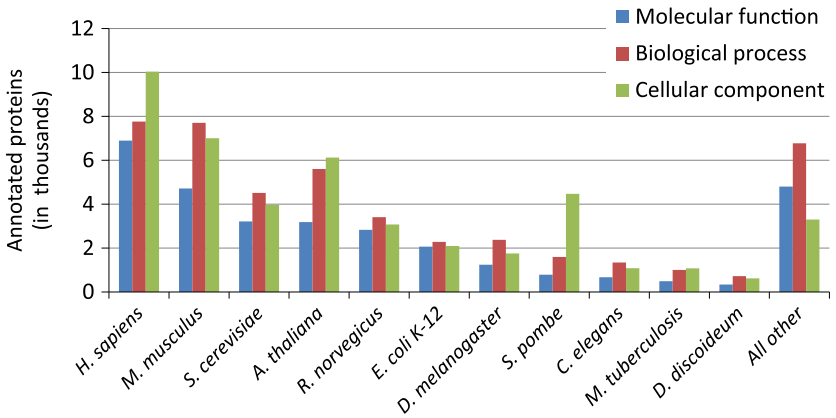

Fig. 3. The number of experimentally annotated (evidence codes: EXP, IPI, IMP, IGI, IDA, and IEP) proteins in the Swiss-Prot database (July 2016) shown separately for each of the three ontologies. The graph shows that the major annotation efforts are concentrated around a handful of model organisms, with more than 85% of annotated proteins being from these species. The total number of annotations equals 31 212 (from 1211 species) for Molecular Function, 45 075 (from 1303 species) for Biological Process, and 44 586 (from 665 species) for Cellular Component. Although some of the organisms appear well annotated by the GO terms, many of these annotations are shallow [89]. Evidence also suggests that some of the experimentally derived functional annotations may be incorrect [90].

understanding pathogenesis is the often missing knowledge of specific molecular events/functions that individual proteins are associated with Fig. 3. While we have some experimental data and computational inferences about the function of many human proteins, for a vast number of these macromolecules, the details remain experimentally unresolved [87,88]. In creating GO, the Gene Ontology, Michael Ashburner (University of Cambridge) and colleagues have recognized the complexity of defining function and proposed three views on every protein's (or gene's) function: (a) molecular function, largely describing biochemical activities, (b) role in a process, largely describing emergent biological functions, and (c) localization, describing cellular or extracellular locale where the protein is active [34]. Ultimately, however, biological function is better captured in a detailed narrative than in any one number. High-resolution human-interpretable function prediction is thus the next frontier in probabilistically inferring molecular mechanisms of disease.

In the meantime, prediction methods can help close the sequence-annotation gap [87,91], but with respect to deep annotations of function, *in silico* methods remain as limited as their experimental 'teachers' [39,92,93]. Machine learning plays critical roles in capturing protein function from the vast biomedical data resources [39,89]. However, if we are to compare among many different possible functions for a given molecule, it is critical to revisit the aforementioned importance of accurately identifying posterior probabilities (Section 'What are we trying to learn? Formalizing the prediction task and data selection'). Suppose that we were to computationally assign either tyrosine kinase or tyrosylprotein sulfotransferase activity to an unannotated human gene. The decision to choose one of these functions must be modulated by our knowledge (or expectation) of the number of proteins in either functional group that are present in the entire proteome. The lack of these *de facto* class priors confines the (proper) use of functional annotation to a single biological function, or a GO term, at a time and limits the extent of the probabilistic reasoning. Although the natural world is characterized by nonuniform distributions of functional categories, the hierarchical nature of biomedical ontologies further exacerbates the problem [39,89].

In addition to the whole-molecule-based view of protein function, it is often important to reason about the function at the level of individual residues. Is a particular residue involved in catalysis, does is bind DNA, is it important for stability and dynamics of the protein? We argue that a probabilistic approach to this problem offers an opportunity to move towards understanding molecular mechanisms of disease. Consider a function of the ith residue (s_i) in the protein sequence s and suppose our method can approximate the probability that s_i is functional, i.e., learn $P(Y_i = 1|s)$. Suppose now we are given another protein sequence s' that differs from s at a single position j (we can later extend this view). Naturally, an accurate predictor of function will be able to output a value similar to $P(Y_i = 1|s')$. These simple building blocks now lead us to a formal model for predicting loss of protein function in the presence of mutation as

$$P(\text{loss of function at } s_i|s, s') = P(Y_i = 1|s) \cdot (1 - P(Y_i = 1|s')).$$

where the first term on the right-hand side is the probability that the residue i is functional in the reference sequence s and the second term is the probability that the residue i is not functional in the mutant sequence s'. We can similarly model an increased propensity for particular function and loosely refer to it as gain of function, i.e.,

$$P(\text{gain of function at } s_i|s, s') = (1 - P(Y_i = 1|s)) \cdot P(Y_i = 1|s').$$

A comprehensive review of the methodologies for the computational function prediction at a residue level is beyond the scope of this text; we refer the reader to available reviews [87,94]. Our main concern is that these algorithms and tools, despite their breadth and availability, remain underexplored in understanding disease. Although some methods that predict specific molecular mechanisms of disease have been proposed either from sequence [95,96] or structure [97,98], the size of the pool of these tools suggests that it will soon be possible to probabilistically reason among many potential options. Of particular importance will be methods for intrinsically disordered proteins [99], i.e., proteins that cannot be characterized by a single dominant macro state due to their conformational dynamics. Disordered regions are known to exhibit conformational sensitivity to amino acid variants and environment [100], and have been implicated in disease upon mutation [101–103]. Better understanding of conformational dynamics and molecular recognition specific for disordered proteins will play significant roles in broadening the scope of molecular mechanisms involved in disease [104].

How does biological context define functionality? Functional pathways and tissue specificity

Regardless of data availability and our accuracy in prediction, understanding the function of individual proteins will not have the power to elucidate the causes of most diseases. However, grouping proteins and various other interactors into pathways, will. The field of pharmacogenomics can, arguably, serve as a training exercise for elucidating causative mechanisms of disease. Identifying all components of drug metabolism pathways is difficult [105]. However, the start and end points of these pathways are predefined and, once the specific sets of pathway components are identified, finding causes of individual differences in drug efficacy (including effectiveness and toxicity) is 'simply' mapping correlated genome variation to these genes/proteins. Thus, patients with mutations that cause functional deficiencies in drug-metabolizing enzymes will need lower doses of drugs (e.g., UGT1A1+ patients treated with irinotecan [106]). On the other hand, patients with a poorly expressed drug target (e.g., ER-negative breast cancer patients treated with tamoxifen [107]) or affected bottleneck genes in pathways downstream from the target (e.g., mutations in KRAS genes of the nonsquamous cell lung cancer patients treated with EGFR inhibitors [108]) will likely not respond to the drug. In these examples, potentially affected pathways could be first hypothesized and then the reasons for particular responses to a drug can be predicted and, eventually, validated. Pharmocogenomics has thus been at the forefront of melding together hypothesis-generating and hypothesis-driven approaches of drug metabolism pathway analysis [109].

In the case of disease, however, pathways are rarely known and need to be inferred from circumstantial evidence. As every person suffers from his/her individual disease manifestation, regardless of the overall disease label, the process of identifying molecular pathway culprits is often complicated by (a) the idiosyncrasies of individual genomic and epigenomic background, (b) pathway involvement in multiple diseases, whose symptoms may overlap or not, and (c) differences in environmental effects. Note that breaking down a disease into individual symptoms could potentially facilitate mapping of pathways to disease [110,111]. The specifics of the protein interactions within and outside the pathway, and the effects of genome variation on the required quantities and activities of each of the pathway components, define the type and severity of the resulting disease.

The definition of a molecular pathway itself presents a problem. Broadly, a molecular pathway is a set of interdependent molecular events that collectively produce some emergent functionality. Among many other activities, pathways lead to DNA, fat, and protein molecule assembly, signal changes in the environment, toggle genes on and off, and metabolize drugs. To define a specific pathway, one may need to establish the set of interacting genes/proteins and cofactors/metabolites leading to a particular endpoint. One way to visualize a pathway is as a specific route from state A to state B through a predefined map of all (possible) molecular interactions in a cell. Here, all molecular interactions can be described as a network, where individual nodes represent genes/proteins and edges are probabilistically weighed and assigned necessary interactors (e.g., enhancers for gene regulation or small molecules as co-factors). A disease pathway is then one where the endpoint is a disease phenotype that is expressed, perhaps, on the molecular level, as opposed to symptomatic manifestations. Thus, Type I diabetes could be more accurately described as a pathway leading to the autoimmune destruction of insulin-producing cells of islets of Langerhans, rather than a pathway leading to increased blood sugar levels [112].

One major challenge in systems biology is to infer pathways from currently available high-throughput and accumulating low-throughput data for further mapping to disease states. Many methods exist to determine the complete map of cellular interactions [113,114]. Similarly, a number of methods exist for establishing disease-associated sets of genes/proteins [115,116]. There are at least two ways of combining these pieces of information into a solid framework of disease-pathway elucidation [117]. First, one can check for enrichment of variation in known pathways extracted from available databases, e.g., KEGG [118], DIP [119], and TRANSPATH [120]. A slight variation on this theme is to only consider variants that are known or predicted to be in some way perturbing of the affected pathway component's molecular functionality (e.g., loss-of-function variants or variants that affect expression). Note that genes that participate in the same pathways are often involved in the same diseases [121,122]. Thus, from the perspective of predicting variant-induced pathogenicity, most (or all) functionally affected members of a disease-related pathway can be construed as disease genes [123,124]. A second approach is to start from the other direction and to first identify disease-associated variants (via GWAS and/or pathogenicity predictions) and then to look for enrichment of interactions between the affected pathway components using databases of interactions, e.g., STRING [125] and ConsensusPathDB [126]. Curiously, both of these approaches can benefit

from the results of multiple unrelated studies directly, without consideration for eliminating batch effects, i.e., as unrelated studies often pick up different disease-associated genes, collapsing the findings by pathway may alleviate population-specific considerations and contribute to the statistical enrichment and mechanistic understanding of disease [127]. However, it is also important to note that sets of pathways differ between tissues. Thus, disease-associated variation should somehow be representative of pathways that are active in tissues relevant to this disease [128]. Much of the logic described here for mapping pathways to disease has been encoded in a number of pathway analysis tools [116,117,129], both commercially and publicly available, using both expression [130,131] and GWAS [132,133] data.

Unfortunately, identification of disease pathways is plagued by many of the same issues as other biological big-data initiatives, including limited data sources, lack of standardization in protocols of data extraction and reporting, and bias in available data. Unless specifically tuned, machine learning approaches to finding disease pathways may be swayed by these biases, leading to an expensive and time-consuming phantom hunt for disease causes and potential drug targets. The future of mechanistic understanding of complex diseases is thus dependent on alleviation of these issues either via experimental means (which is unlikely) or via computational approaches that encompass proper dataset selection, correct class prior estimation, and fine-tuning of the algorithm parameters.

Conclusions

While we are not yet in the era of precision medicine, it seems that evidence-based and biomarker-led stratified medicine is already solidly in practice and moving forward [134]. Stratified medicine implies different treatments for different groups of people – an approach that is more effective than the earlier version of empirical medicine, which focused more on disease specifics than on the patient specifics. Arguably, however, the correlative stratified medicine cannot be made more precise without explicit understanding of disease mechanisms.

In the absence of experimental data that elucidates disease mechanisms, our best opportunity for understanding them is a proper and formal application of decades of research in statistics and machine learning. We have argued earlier that the key to these approaches in protein science and related tasks in precision medicine is to develop models that approximate posterior distributions. While this is an ideal outcome, we realize that such estimates are not always necessary. In many biological problems, as well as in many commercial applications, it often suffices to rank data points. An example of this is Google's presentation of the search engine results or, closer to computational biology, disease gene prioritization [135–138]. While useful, the results of such predictor outputs are less interpretable and do not naturally permit probabilistic reasoning about the general phenomena. Furthermore, our discussion of machine learning in previous sections focused on supervised and semisupervised learning as these frameworks are most thoroughly studied and widely applicable. We should, however, keep in mind the importance of unsupervised learning as well as the resurgence of reinforcement learning, particularly in time-dependant patient-centered applications.

Returning to the car parts and diagnostic protocols analogy, a major need in precision medicine is a comprehensive list that associates observable manifestations with particular mechanistic causes of failure. In biology, these might include sets of sequence variants in coding and noncoding regions, groups of phenotype-relevant proteins and tissue-specific pathways, cell/tissue specific genome structure alterations, as well as some annotations of the impacts of 'external' influences of the microbiome and the environment overall. An immense undertaking that, like every journey, will begin with the first steps – a catalog of phenotypes and their causes.

Acknowledgements

We thank friends and colleagues at the TUM, Rutgers, and Indiana University. We thank Jose Lugo-Martinez for producing Figure 2. We also thank Inga Weise for support on many levels. BR was supported by the Alexander von Humboldt foundation through the German Ministry for Research and Education (BMBF: Bundesministerium fuer Bildung und Forschung). PR was supported by the NIH grant R01 MH105524 and NSF grant DBI-1458477. YB was supported by the NIH U01 GM115486, U24 MH068457, and USDA-NIFA 1015:0228906 grants. Last, but not least, we thank all those who deposit their experimental data in public databases, and those who maintain these resources.

References

1 Obama B (2015) State of the union address, January 30, 2015.
2 Abernethy AP, Etheredge LM, Ganz PA, Wallace P, German RR, Neti C, Bach PB and Murphy SB (2010) Rapid-learning system for cancer care. *J Clin Oncol* **28**, 4268–4274.

3 Samuel AL (1959) Programming computers to learn from experience should eventually eliminate the need for much of this detailed programming effort. *IBM J Res Dev* **3**, 210.
4 Marx V (2015) The DNA of a nation. *Nature* **524**, 503–505.
5 Fernald GH, Capriotti E, Daneshjou R, Karczewski KJ and Altman RB (2011) Bioinformatics challenges for personalized medicine. *Bioinformatics* **27**, 1741–1748.
6 Bromberg Y (2013) Building a genome analysis pipeline to predict disease risk and prevent disease. *J Mol Biol* **425**, 3993–4005.
7 Perdigao N, Heinrich J, Stolte C, Sabir KS, Buckley MJ, Tabor B, Signal B, Gloss BS, Hammang CJ, Rost B *et al*. (2015) Unexpected features of the dark proteome. *Proc Natl Acad Sci USA* **112**, 15898–15903.
8 Schmidhuber J (2015) Deep learning in neural networks: an overview. *Neural Netw* **61**, 85–117.
9 Gibney E (2016) Google AI algorithm masters ancient game of Go. *Nature* **529**, 445–446.
10 Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G *et al*. (2015) Human-level control through deep reinforcement learning. *Nature* **518**, 529–533.
11 McCluskey M (2016) This computer algorithm predicted who will die next on Game of Thrones, April 19, 2016.
12 Qian N and Sejnowski TJ (1988) Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol* **202**, 865–884.
13 King RD and Sternberg MJ (1990) Machine learning approach for the prediction of protein secondary structure. *J Mol Biol* **216**, 441–457.
14 Bohr H, Bohr J, Brunak S, Cotterill RM, Fredholm H, Lautrup B and Petersen SB (1990) A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks. *FEBS Lett* **261**, 43–46.
15 Rost B and Sander C (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci USA* **90**, 7558–7562.
16 Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, and Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* **28**, 235–242.
17 Peng K, Obradovic Z and Vucetic S (2004) Exploring bias in the Protein Data Bank using contrast classifiers. *Pac Symp Biocomput* **2004**, 435–446.
18 Schneider M, Lane L, Boutet E, Lieberherr D, Tognolli M, Bougueleret L and Bairoch A (2009) The UniProtKB/Swiss-Prot knowledgebase and its plant proteome annotation program. *J Proteomics* **72**, 567–573.
19 Schnoes AM, Ream DC, Thorman AW, Babbitt PC and Friedberg I (2013) Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. *PLoS Comput Biol* **9**, e1003063.
20 Rost B (2002) Enzyme function less conserved than anticipated. *J Mol Biol* **318**, 595–608.
21 Schlessinger A, Schaefer C, Vicedo E, Schmidberger M, Punta M and Rost B (2011) Protein disorder – a breakthrough invention of evolution? *Curr Opin Struct Biol* **21,** 412–418.
22 Hamp T and Rost B (2015) More challenges for machine-learning protein interactions. *Bioinformatics* **31**, 1521–1525.
23 Wald A (1943) *A Method of Estimating Plane Vulnerability Based on Damage of Survivors*. Operations Evaluation Group, Center for Naval Analysis, Alexandria, VA, USA.
24 Rojas R (1996) A short proof of the posterior probability property of classifier neural networks. *Neural Comput* **8**, 41–43.
25 Platt JC (1999) Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In *Advances in Large Margin Classifiers* pp. 61–74. MIT Press, Cambridge, MA. Available at: https://books.google.com/books?id=gOXI3fO3VUwC&printsec=frontcove&source=rgbs_ge_summary_r&cad=0#v=onepage&q&f=false.
26 Niculescu-Mizil A and Caruana R (2005) Obtaining calibrated probabilities from boosting. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence*, UAI 2005, Edinburgh, UK, pp. 413–420. AUAI Press, Arlington, VA.
27 Heckman J (1979) Sample selection bias as a specification error. *Econometrica* **47**, 153–161.
28 Cortes C, Mohri M, Riley M and Rostamizadeh A (2008) Sample selection bias correction theory. In *Proceedings of the 19th International Conference on Algorithmic Learning Theory*, ALT 2008 Budapest, Hungary, pp. 38–53. Springer-Verlag, Berlin, Germany.
29 Rost B (1999) Twilight zone of protein sequence alignments. *Protein Eng* **12**, 85–94.
30 Hastie T, Tibshirani R and Friedman JH (2001) The elements of statistical learning: data mining, inference, and prediction. Springer Verlag, New York, NY.
31 Frank E, Hall M, Trigg L, Kirkby R, Schmidberger G, Ware M, Xu X, Bouckaert R, Wang Y, Inglis S *et al*. (2002) Weka 3 – Machine Learning Software in Java.
32 Hamp T and Rost B (2015) Evolutionary profiles improve protein-protein interaction prediction from sequence. *Bioinformatics* **31**, 1945–1950.

33 Consortium UniProt (2015) UniProt: a hub for protein information. *Nucleic Acids Res* **43**(Database issue), D204–D212.

34 Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT *et al*. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29.

35 Robinson PN and Mundlos S (2010) The human phenotype ontology. *Clin Genet* **77**, 525–534.

36 Lord PW, Stevens RD, Brass A and Goble CA (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* **19**, 1275–1283.

37 Verspoor K, Cohn J, Mniszewski S and Joslyn C (2006) A categorization approach to automated ontological function annotation. *Protein Sci* **15**, 1544–1549.

38 Clark WT and Radivojac P (2013) Information-theoretic evaluation of predicted ontological annotations. *Bioinformatics* **29**, i53–i61.

39 Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A *et al*. (2013) A large-scale evaluation of computational protein function prediction. *Nat Methods* **10**, 221–227.

40 Saerens M, Latinne P and Decaestecker C (2002) Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Comput* **14**, 21–41.

41 Denis F, Gilleron R and Letouzey F (2005) Learning from positive and unlabeled examples. *Theor Comput Sci* **348**, 70–83.

42 Blanchard G, Lee G and Scott C (2010) Semi-supervised novelty detection. *J Mach Learn Res* **11**, 2973–3009.

43 Jain S, White M, Trosset MW and Radivojac P (2016) Nonparametric semi-supervised learning of class proportions. arXiv preprint arXiv:1601.01944.

44 Elkan C and Noto K (2008) Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD 2008, Las Vegas, NV, pp. 213–220. Association for Computing Machinery, New York, NY.

45 Ward G, Hastie T, Barry S, Elith J and Leathwick JR (2009) Presence-only data and the EM algorithm. *Biometrics* **65**, 554–563.

46 du Plessis MC and Sugiyama M (2014) Class prior estimation from positive and unlabeled data. *IEICE Trans Inf Syst* **E97-D**, 1358–1362.

47 Scott C, Blanchard G and Handy G (2013) Classification with asymmetric label noise: consistency and maximal denoising. *J Mach Learn Res W&CP* **30**, 489–511.

48 Jain S, White M and Radivojac P (2016) Estimating the class prior and posterior from noisy positives and unlabeled data. arXiv preprint arXiv:1606.08561.

49 http://www.hopkinsmedicine.org/about/history/history5.html

50 Roses AD (2000) Pharmacogenetics and the practice of medicine. *Nature* **405**, 857–865.

51 Katsnelson A (2013) Momentum grows to make 'personalized' medicine more 'precise'. *Nat Med* **19**, 249.

52 Egelmeer AG, Velazquez ER, de Jong JM, Oberije C, Geussens Y, Nuyts S, Kremer B, Rietveld D, Leemans CR, de Jong MC *et al*. (2011) Development and validation of a nomogram for prediction of survival and local control in laryngeal carcinoma patients treated with radiotherapy alone: a cohort study based on 994 patients. *Radiother Oncol* **100**, 108–115.

53 Colombo J, Fachel AA, De Freitas Calmon M, Cury PM, Fukuyama EE, Tajara EH, Cordeiro JA, Verjovski-Almeida S, Reis EM and Rahal P (2009) Gene expression profiling reveals molecular marker candidates of laryngeal squamous cell carcinoma. *Oncol Rep* **21**, 649–663.

54 Pauling L, Itano HA, Singer SJ and Wells IC (1949) Sickle cell anemia, a molecular disease. *Science* **110**, 543–548.

55 Perutz M and Mitchison JM (1950) State of haemoglobin in sickle-cell anaemia. *Nature* **166**, 677–679.

56 Hunt JA and Ingram VM (1958) Allelomorphism and the chemical differences of the human haemoglobins A, S and C. *Nature* **181**, 1062–1063.

57 Stenson PD, Mort M, Ball EV, Shaw K, Phillips A and Cooper DN (2014) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* **133**, 1–9.

58 Amberger JS, Bocchini CA, Schiettecatte F, Scott AF and Hamosh A (2015) OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* **43**(Database issue), D789–D798.

59 Lenffer J, Nicholas FW, Castle K, Rao A, Gregory S, Poidinger M, Mailman MD and Ranganathan S (2006) OMIA (Online Mendelian Inheritance in Animals): an enhanced platform and integration into the Entrez search interface at NCBI. *Nucleic Acids Res* **34**(Database issue), D599–D601.

60 Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H, Ding M, Bamford S, Cole C, Ward S *et al*. (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* **43**(Database issue), D805–D811.

61 Fermi G, Perutz MF, Shaanan B and Fourme R (1984) The crystal structure of human deoxyhaemoglobin at 1.74 A resolution. *J Mol Biol* **175**, 159–174.

62 Harrington DJ, Adachi K and Royer WE Jr (1997) The high resolution crystal structure of deoxyhemoglobin S. *J Mol Biol* **272**, 398–407.

63 1000 Genomes Project Consortium; Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA and Abecasis GR (2015) A global reference for human genetic variation. *Nature* **526**, 68–74.

64 Reeb J, Hecht M, Mahlich Y, Bromberg Y and Rost B (2016) Predicted molecular effects of sequence variants link to system level of disease. *PLoS Computat Biol*, doi: 10.1371/journal.pcbi.1005047.

65 de Beer TA, Laskowski RA, Parks SL, Sipos B, Goldman N and Thornton JM (2013) Amino acid changes in disease-associated variants differ radically from variants observed in the 1000 Genomes Project Dataset. *PLoS Comput Biol* **9**, e1003382.

66 Bordoli L and Schwede T (2012) Automated protein structure modeling with SWISS-MODEL Workspace and the Protein Model Portal. *Methods Mol Biol* **857**, 107–136.

67 Thusberg J, Olatubosun A and Vihinen M (2011) Performance of mutation pathogenicity prediction methods on missense variants. *Hum Mutat* **32**, 358–368.

68 Cline MS and Karchin R (2011) Using bioinformatics to predict the functional impact of SNVs. *Bioinformatics* **27**, 441–448.

69 Capriotti E, Nehrt NL, Kann MG and Bromberg Y (2012) Bioinformatics for personal genome interpretation. *Brief Bioinform* **13**, 495–512.

70 Frousios K, Iliopoulos CS, Schlitt T and Simpson MA (2013) Predicting the functional consequences of non-synonymous DNA sequence variants – evaluation of bioinformatics tools and development of a consensus strategy. *Genomics* **102**, 223–228.

71 Peterson TA, Doughty E and Kann MG (2013) Towards precision medicine: advances in computational approaches for the analysis of human variants. *J Mol Biol* **425**, 4047–4063.

72 Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K and Liu X (2015) Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet* **24**, 2125–2137.

73 Bromberg Y and Capriotti E (2015) VarI-SIG 2014 – from SNPs to variants: interpreting different types of genetic variants. *BMC Genomics* **16**(Suppl 8), I1.

74 Rost B (2014) Personalized heatlh: harnessing the power of diversity. TedX TUM Munich, Aug 11, 2014.Available at: http://tedxtalks.ted.com/video/Personalized-health-harnessing.

75 Masica DL and Karchin R (2016) Towards increasing the clinical relevance of *in silico* methods to predict pathogenic missense variants. *PLoS Comput Biol* **12**, e1004725.

76 Schaefer C, Bromberg Y, Achten D and Rost B (2012) Disease-related mutations predicted to impact protein function. *BMC Genomics* **13**(Suppl 4), S11.

77 Ofran Y and Rost B (2007) Protein–protein interaction hot spots carved into sequences. *PLoS Comput Biol* **3**, e119.

78 Bromberg Y and Rost B (2007) SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res* **35**, 3823–3835.

79 Hecht M, Bromberg Y and Rost B (2015) Better prediction of functional effects for sequence variants. *BMC Genomics* **16**(Suppl 8), S1.

80 Bromberg Y, Kahn PC and Rost B (2013) Neutral and weakly nonneutral sequence variants may define individuality. *Proc Natl Acad Sci USA* **110**, 14255–14260.

81 Gibson G (2011) Rare and common variants: twenty arguments. *Nat Rev Genet* **13**, 135–145.

82 Lamparter D, Marbach D, Rueedi R, Kutalik Z and Bergmann S (2016) Fast and rigorous computation of gene and pathway scores from snp-based summary statistics. *PLoS Comput Biol* **12**, e1004714.

83 Evangelou M, Smyth DJ, Fortune MD, Burren OS, Walker NM, Guo H, Onengut-Gumuscu S, Chen WM, Concannon P, Rich SS *et al.* (2014) A method for gene-based pathway analysis using genomewide association study summary statistics reveals nine new type 1 diabetes associations. *Genet Epidemiol* **38**, 661–670.

84 Holmans P, Green EK, Pahwa JS, Ferreira MA, Purcell SM, Sklar P; Consortium Wellcome Trust Case-Control, Owen MJ, O'Donovan MC and Craddock N (2009) Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *Am J Hum Genet* **85**, 13–24.

85 Segre AV; Diagram Consortium; Magic investigators, Groop L, Mootha VK, Daly MJ and Altshuler D (2010) Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet* **6**. doi: 10.1371/journal.pgen.1001058.

86 Jensen AB, Moseley PL, Oprea TI, Ellesoe SG, Eriksson R, Schmock H, JensenPB, Jensen LJ and Brunak S (2014) Temporal disease trajectories condensed from population-wide registry data covering 6.2 million patients. *Nat Commun* **5**, 4022.

87 Rost B, Liu J, Nair R, Wrzeszczynski KO and Ofran Y (2003) Automatic prediction of protein function. *Cell Mol Life Sci* **60**, 2637–2650.

88 Gaudet P, Michel PA, Zahn-Zabal M, Cusin I, Duek PD, Evalet O, Gateau A, Gleizes A, Pereira M, Teixeira D *et al*. (2015) The neXtProt knowledgebase on human proteins: current status. *Nucleic Acids Res* **43**(Database issue), D764–D770.
89 Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, Funk CS, Kahanda I, Verspoor KM, Ben-Hur A *et al*. (2016) An expanded evaluation of protein function prediction methods shows an improvement in accuracy. arXiv preprint arXiv:1601.00891.
90 Schnoes AM, Brown SD, Dodevski I and Babbitt PC (2009) Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. *PLoS Comput Biol* **5**, e1000605.
91 Rentzsch R and Orengo CA (2009) Protein function prediction – the power of multiplicity. *Trends Biotechnol* **27**, 210–219.
92 Gobeill J, Gaudinat A, Pasche E, Vishnyakova D, Gaudet P, Bairoch A and Ruch P (2015) Deep question answering for protein annotation. *Database (Oxford)*, doi: 10.1093/database/bav081.
93 Babbitt PC, Bagos PG, Bairoch A, Bateman A, Chatonnet A, Chen MJ, Craik DJ, Finn RD, Gloriam D, Haft DH *et al*. (2015) Creating a specialist protein resource network: a meeting report for the protein bioinformatics and community resources retreat. *Database (Oxford)*, **2015**, bav063.
94 Xin F and Radivojac P (2011) Computational methods for identification of functional residues in protein structures. *Curr Protein Pept Sci* **12**, 456–469.
95 Radivojac P, Baenziger PH, Kann MG, Mort ME, Hahn MW and Mooney SD (2008) Gain and loss of phosphorylation sites in human cancer. *Bioinformatics* **24**, i241–i247.
96 Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD and Radivojac P (2009) Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* **25**, 2744–2750.
97 De Baets G, Van Durme J, Reumers J, Maurer-Stroh S, Vanhee P, Dopazo J, Schymkowitz J and Rousseau F (2012) SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Res* **40**(Database issue), D935–D939.
98 Yates CM and Sternberg MJ (2013) The effects of non-synonymous single nucleotide polymorphisms (nsSNPs) on protein-protein interactions. *J Mol Biol* **425**, 3949–3963.
99 Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, and Dunker AK (2007) Intrinsic disorder and functional proteomics. *Biophys J* **92**, 1439–1456.
100 Mohan A, Uversky VN and Radivojac P (2009) Influence of sequence changes and environment on intrinsically disordered proteins. *PLoS Comput Biol* **5**, e1000497.
101 Mort M, Evani US, Krishnan VG, Kamati KK, Baenziger PH, Bagchi A, Peters BJ, Sathyesh R, Li B, Sun Y *et al*. (2010) *In silico* functional profiling of human disease-associated and polymorphic amino acid substitutions. *Hum Mutat* **31**, 335–346.
102 Vacic V, Markwick PR, Oldfield CJ, Zhao X, Haynes C, Uversky VN, and Iakoucheva LM (2012) Disease-associated mutations disrupt functionally important regions of intrinsic protein disorder. *PLoS Comput Biol* **8**, e1002709.
103 Vacic V and Iakoucheva LM (2012) Disease mutations in disordered regions – exception to the rule? *Mol Biosyst* **8,** 27–32.
104 Uversky VN, Oldfield CJ and Dunker AK (2008) Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annu Rev Biophys* **37**, 215–246.
105 Kirchmair J, Goller AH, Lang D, Kunze J, Testa B, Wilson ID, Glen RC and Schneider G (2015) Predicting drug metabolism: experiment and/or computation? *Nat Rev Drug Discov* **14,** 387–404.
106 Dean L (2015) Irinotecan Therapy and UGT1A1 Genotype. National Center for Biotechnology Information, Bethesda, MD.
107 Swain SM (2001) Tamoxifen for patients with estrogen receptor-negative breast cancer. *J Clin Oncol* **19** (Suppl), 93S–97S.
108 Langer CJ (2011) Roles of egfr and kras mutations in the treatment of patients with non-small-cell lung cancer. *P T* **36**, 263–279.
109 Wilke RA, Mareedu RK and Moore JH (2008) The pathway less traveled: moving from candidate genes to candidate pathways in the analysis of genome-wide data from large scale pharmacogenetic association studies. *Curr Pharmacogenomics Person Med* **6**, 150–159.
110 Yahi A and Tatonetti NP (2015) A knowledge-based, automated method for phenotyping in the EHR using only clinical pathology reports. *AMIA Jt Summits Transl Sci Proc*, **2015**, 64–68.
111 Nissim N, Boland MR, Tatonetti NP, Elovici Y, Hripcsak G, Shahar Y, and Moskovitch R (2016) Improving condition severity classification with an efficient active learning based framework. *J Biomed Inform* **61**, 44–54.
112 van Belle TL, Coppieters KT and von Herrath MG (2011) Type 1 diabetes: etiology, immunology, and therapeutic strategies. *Physiol Rev* **91**, 79–118.
113 Gabaldon T and Huynen MA (2004) Prediction of protein function and pathways in the genome era. *Cell Mol Life Sci* **61**, 930–944.

114 Marbach D, Costello JC, Kuffner R, Vega NM, Prill RJ, Camacho DM, Allison KR; Dream Consortium, Kellis M, Collins JJ and Stolovitzky G (2012) Wisdom of crowds for robust gene network inference. *Nat Methods* **9**, 796–804.
115 Tilford CA and Siemers NO (2009) Gene set enrichment analysis. *Methods Mol Biol* **563**, 99–121.
116 Khatri P, Sirota M and Butte AJ (2012) Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol* **8**, e1002375.
117 Jin L, Zuo XY, Su WY, Zhao XL, Yuan MQ, Han LZ, Zhao X, Chen YD and Rao SQ (2014) Pathway-based analysis tools for complex diseases: a review. *Genomics Proteomics Bioinformatics* **12**, 210–220.
118 Kanehisa M, Goto S, Kawashima S, Okuno Y and Hattori M (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res* **32**(Database issue), D277–D280.
119 Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU and Eisenberg D (2004) The database of interacting proteins: 2004 update. *Nucleic Acids Res* **32** (Database issue), D449–D451.
120 Choi C, Krull M, Kel A, Kel-Margoulis O, Pistor S, Potapov A, Voss N, and Wingender E (2004) TRANSPATH – a high quality database focused on signal transduction. *Comp Funct Genomics* **5**, 163–168.
121 Sun J and Zhao Z (2010) A comparative study of cancer proteins in the human protein–protein interaction network. *BMC Genomics* **11**(Suppl 3), S5.
122 Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, Mohan SS, Sharma S, Pinkert S, Nagaraju S, Periaswamy B *et al.* (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat Genet* **38**, 285–293.
123 Cordell HJ (2009) Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* **10**, 392–404.
124 Moore JH and Williams SM (2009) Epistasis and its implications for personal genetics. *Am J Hum Genet* **85**, 309–320.
125 Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP *et al.* (2015) String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43**(Database issue), D447–D452.
126 Kamburov A, Stelzl U, Lehrach H and Herwig R (2013) The consensuspathdb interaction database: 2013 update. *Nucleic Acids Res* **41**(Database issue), D793–D800.
127 Brodie A, Tovia-Brodie O and Ofran Y (2014) Large scale analysis of phenotype-pathway relationships based on GWAS results. *PLoS One* **9**, e100887.
128 Marbach D, Lamparter D, Quon G, Kellis M, Kutalik Z and Bergmann S (2016) Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat Methods* **13**, 366–370.
129 Ramanan VK, Shen L, Moore JH and Saykin AJ (2012) Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends Genet* **28**, 323–332.
130 http://www.proteinlounge.com/pathwaybuilder.asp
131 D'Eustachio P (2011) Reactome knowledgebase of human biological pathways and processes. *Methods Mol Biol* **694**, 49–61.
132 Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM and Lee JJ (2015) Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7.
133 Zhang K, Chang S, Guo L and Wang J (2015) I-GSEA4GWAS v2: a web server for functional analysis of SNPs in trait-associated pathways identified from genome-wide association study. *Protein Cell* **6**, 221–224.
134 Willis JC and Lord GM (2015) Immune biomarkers: the promises and pitfalls of personalized medicine. *Nat Rev Immunol* **15**, 323–329.
135 Dalkilic MM, Costello JC, Clark WT and Radivojac P (2008) From protein-disease associations to disease informatics. *Front Biosci* **13**, 3391–3407.
136 Kann MG (2007) Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief Bioinform* **8**, 333–346.
137 Bromberg Y (2013) Disease gene prioritization. *PLoS Comput Biol* **9**, e1002902.
138 Moreau Y and Tranchevent LC (2012) Computational tools for prioritizing candidate genes: boosting disease gene discovery. *Nat Rev Genet* **13**, 523–536.