News from the Protein Mutability Landscape

Maximilian Hecht[1], **Yana Bromberg**[2] **and Burkhard Rost**[1,3,4]

*1 - **Department of Bioinformatics and Computational Biology I12,** Technische Universität München, Boltzmannstrasse 3, 85748 Garching, Germany*
*2 - **Department of Biochemistry and Microbiology,** Rutgers University, 76 Lipman Drive, New Brunswick, NJ 08901, USA*
*3 - **Institute of Advanced Study,** Technische Universität München, Boltzmannstrasse 3, 85748 Garching, Germany*
*4 - **Institute for Food and Plant Sciences,** Life Science Center Weihenstephan, Alte Akademie 8, 85354 Freising, Germany*

Correspondence to Maximilian Hecht: *hecht@rostlab.org*
http://dx.doi.org/10.1016/j.jmb.2013.07.028
Edited by E. Alexov

Abstract

Some mutations of protein residues matter more than others, and these are often conserved evolutionarily. The explosion of deep sequencing and genotyping increasingly requires the distinction between effect and neutral variants. The simplest approach predicts all mutations of conserved residues to have an effect; however, this works poorly, at best. Many computational tools that are optimized to predict the impact of point mutations provide more detail. Here, we expand the perspective from the view of single variants to the level of sketching the entire mutability landscape. This landscape is defined by the impact of substituting every residue at each position in a protein by each of the 19 non-native amino acids. We review some of the powerful conclusions about protein function, stability and their robustness to mutation that can be drawn from such an analysis. Large-scale experimental and computational mutagenesis experiments are increasingly furthering our understanding of protein function and of the genotype–phenotype associations. We also discuss how these can be used to improve predictions of protein function and pathogenicity of missense variants.

Introduction

> *Lewis Carroll—Through the Looking Glass: "Why, sometimes I've believed as many as six impossible things before breakfast."*

Understanding Genetic Diversity—A Central Challenge for Deep Sequencing

Elucidating which human genetic variations have which phenotypic effect and how the variation impacts disease is one of the major scientific challenges in the 21st century. While the vast majority of genetic variants are hypothesized to be neutral [1], that is, are assumed not to contribute to any phenotype, the relative percentage of neutral, near-neutral [2] and non-neutral variants remains unclear. Most likely, the precise ratios heavily depend on the particular protein under investigation (e.g., the human immunodeficiency virus gp120 is likely to be much more robust against mutation than p53 simply because many of the p53 residues are involved in binding and therefore "vulnerable" to mutation). A key aspect in the development of strategies for diagnosis and treatment of genetic diseases is to further our understanding of the underlying mechanisms that link genotypes and phenotypes.

Sequence variants such as single nucleotide polymorphisms (SNPs) are the most prevalent form of human genetic variation [3]. It has been estimated that more than 11 million SNPs will be observed among people; 7 million of these are frequent (common variants), that is, occur with a minor allele frequency above 5%, while the remaining (minor allele frequency, $<5\%$) are considered as rare [4]. Many of both rare and common variants may be instrumental in defining individual's differences [5–7]. Increasingly, however,

researchers begin to suspect that every possible point mutation might ultimately be observed.

For medical biology, non-synonymous SNPs (nsSNPs) or missense variants that change the amino acid sequence of the protein are particularly interesting. These variants are more likely to affect function than synonymous SNPs. Single-amino-acid variants can change the resulting phenotype, for example, by altering protein function directly or indirectly by impacting structure and/or binding. Such changes can lead to pathogenic phenotypes [8]. Recent studies suggest that every pair of individuals differs by almost one amino acid variant in each protein while individuals have about 1.2–1.7 variants (nsSNPs) that are novel with respect to both parents, that is, not observed in either parent [7].

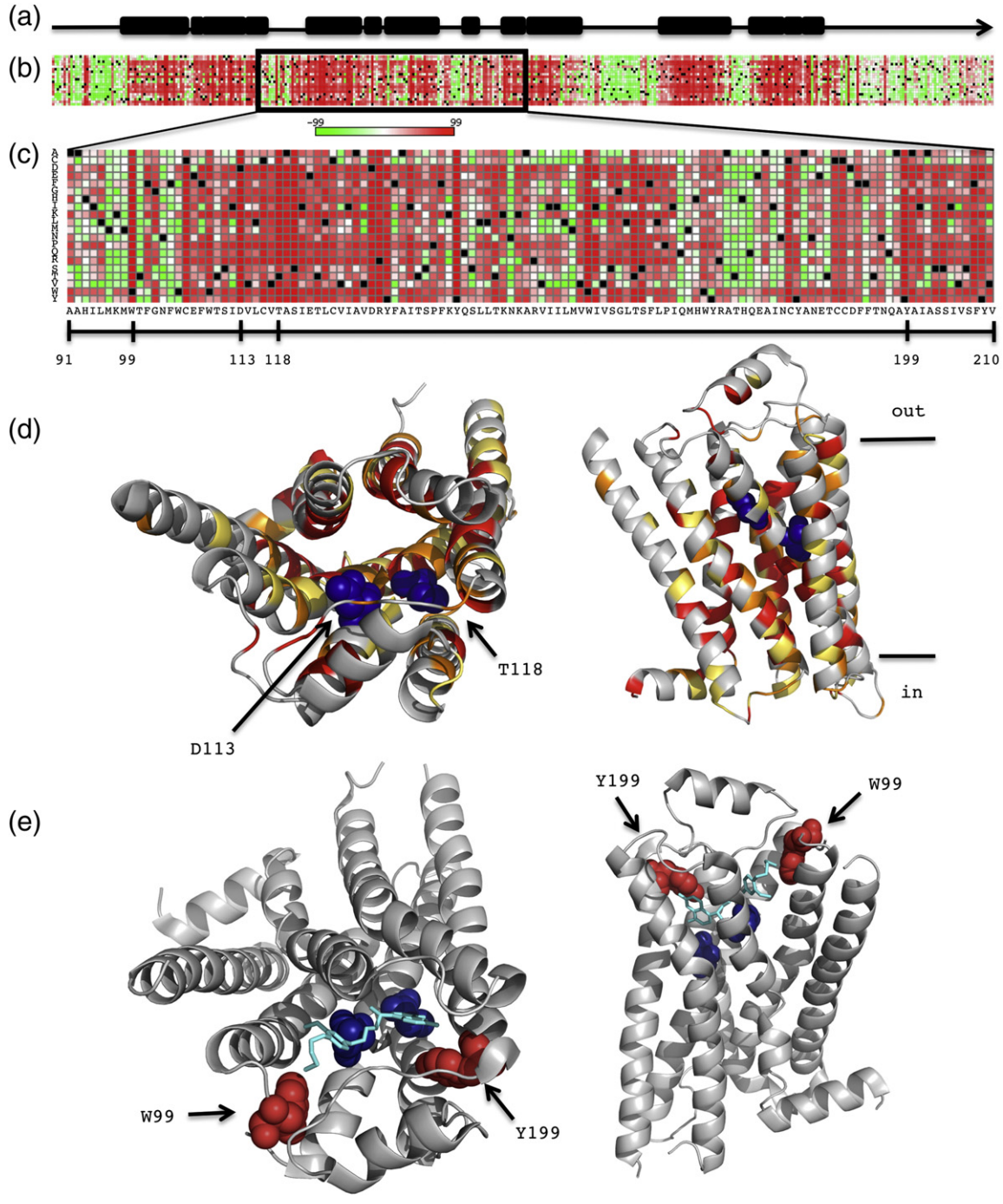

Fig. 1 (*legend on next page*)

Knowing how these changes affect function can give, for instance, insight into a child's disease predisposition.

GWAS (*g*enome-*w*ide *a*ssociation *s*tudies) has evolved as the most widely used approach relating human genetic variation to phenotypic diversity [9]. Results of these studies greatly increased our understanding of molecular pathways underlying specific human diseases. Most common SNPs have been assessed for statistical associations with many complex traits and common diseases. However, for the vast majority of complex trait associations, the underlying mechanisms remain unknown, and for many of the known common SNPs related to complex disease phenotypes, GWAS misses the known associations [10]. Rare variants are missed due to limited numbers [4,11]. The hope is that deep sequencing will address some of these issues by revealing variations between the full sequences.

Mutability Landscapes May Be THE Key to Understanding Diversity

Meanwhile, a different approach toward understanding the genotype–phenotype association is to study functionally important regions, robustness and evolvability of proteins by investigating the mutability landscape. The mutability landscape of, for example, protein function, can be defined as the effect of all possible point mutations/variants upon protein function, that is, of substituting the native amino acid at each residue position against all 19 non-native amino acids, one at a time (Fig. 1 [20,21]). Studying such a landscape may help us a lot in understanding protein function and evolution.

In this review, we exclusively focus on the effects of varying the protein sequence by single-amino-acid substitutions (SAASs). In order to avoid obscuring acronyms, we will simply use the term variant as synonym for SAAS. We review comprehensive mutagenesis in which each position in a protein is changed and complete mutagenesis in which each position is replaced by every non-native amino acid. However, we largely discard effects of varying multiple residues at the same time. Obviously, even such a reduced version of the *protein mutability* landscape already carries very important information about protein function. We attempt to sketch how this landscape brings about new challenges and new possibilities. In fact, we have to learn to understand what we see in this new looking glass.

Most Variant Effects Predicted Correctly *In Silico*

Several computational methods predict the effect of variants (SAAS or nsSNPs). Some predict the effect on protein function {e.g., sorting intolerant from tolerant (SIFT) [22,23] or screening for non-acceptable polymorphisms (SNAP) [24,25]}, others predict the effect with respect to their pathogenicity (e.g., MutPred [26], SNPs&GO [27], Mutation Assesor [28] or MutationTaster [29]) and others yet predict the effect on protein structure directly [30] or cannot be easily fit into these categories (e.g., PolyPhen-2 [31] or PON-P [32]). These methods use a diverse spectrum of input features, typically combining evolutionary information with biophysical features and experimental information about protein structure and function where available. There are several outstanding reviews on the prediction of functional effects [33–35], and the community puts great effort into assessing such predictions. For instance, CAGI (*C*ritical *A*ssessment of *G*enome *I*nterpretation) aspires to assessing method performance in predicting phenotypic impacts of genomic variation [36]. More formal studies assess predictors specifically with respect to their performance in identifying pathogenic variants [37]. The results of these studies suggest that each method has strengths and weaknesses, possibly resulting from the data used for development and the types of information

Fig. 1. Mutability landscape of a protein. The top line (a) sketches the sequence and secondary structure (transmembrane helices) of the adrenergic receptor (ADRB2_HUMAN, ID: P07550 [12]; assignment of secondary from the high-resolution structure PDB ID: 3PDS [13] using DSSP [14]). For each of the 413 residues (*x*-axis) of the receptor, (b) shows the predictions for the effects of all 19 non-native variants (*y*-axis; the stronger the predicted effect, the redder; the stronger the predicted neutrality, the greener). (c) Zoom of the fragment spanning from residue 91 to residue 210 and the relative positions of binding sites (D113 and T118) and proposed target residues (W99 and Y199). (d) The predicted functional effect of variants for the 3D structure (PDB ID: 2RH1 [15]); both known binding sites (positions 113 and 118) are shown as blue spheres. Shown are the average scores [SNAP score ranges from the most neutral (−100) to the strongest effect (+100)] over amino acids that would be considered as "neutral" given the biophysical amino acid features as captured in the PHAT substitution matrix [16] for transmembrane regions and in the BLOSUM62 matrix [17] for all other residues. Red depicts high average scores (score > 60), orange depicts intermediate scores (40 < score < 60), yellow depicts low scores (20 < score < 40) and gray marks sites with SNAP scores <20 (predicted as neutral or with little effect). (e) The 3D structure (PDB ID: 3PDS [13]) with a bound agonist and the two residues (W99 and Y199) that exhibit a high overall predicted effect and are under strong evolutionary constraint (predicted by EVfold [18,19]) with each other and the two binding sites.

included in prediction. Good *in silico* methods correctly predict the experimentally observed effects for most variants.

Typically, only 5–10% of all residues relate directly to function [38–40]. Some of these are revealed in the substitution profiles of protein families. A whole generation of methods targets the prediction of such functional sites through analyzing evolutionary information (e.g., ET [41], INTREPID [42] or DISCERN [43]). Since these methods predict functional sites, it is not surprising that they also capture some of the signal that variants impact function (V. Link and K. Sjölander, unpublished results). Thus, the ability to predict the functional effect of variants is clearly related to predicting protein function.

Predictions of variant effects have helped us prioritize mutations for large-scale reverse genetics projects, where mutations are randomly introduced into the genome. An example for such strategies is TILLING (*t*argeting *i*nduced *l*ocal *l*esions *in g*enomes), a method that combines chemical mutagenesis with a sensitive DNA screening technique in order to allow direct identification of mutations in a specific gene. TILLING uses the functional effect predictions from SIFT [22,23] to prioritize the post-processing of variants [44]. Another important application is to the assessment of disease-related human variants [20,45,46]. For instance, mutations that directly cause a disease, such as those found in OMIM [47], are clearly identified by methods that predict the functional effect of variants [48]. Existing *in silico* methods can even be good enough to reveal problems with experimental data: today's assessment of functional neutrality of variants seems particularly problematic [100]

Peeking Experimentally into the Protein Mutability Landscape

Alanine scans reveal function and interaction hot spots

The experimental study on how site-directed mutagenesis affects phenotypes may be THE most essential experimental tool for determining protein function. By substituting residues that are assumed to be important and measuring substitution effect, researchers identify the residues that are important for the hypothesized protein function. Over the last decade, the power of experimental and computational mutagenesis has grown considerably: a decade ago, many publications reported on single point mutations; today, 50 times more may no longer satisfy reviewers.

The ability of proteins to interact with substrates or other proteins is essential. The most important function of a protein can therefore be defined as its role within an interaction pathway [49]. Typically, only a few residues in a protein interaction interface contribute most of the binding affinity. These can be identified by the change in binding free energy upon mutation to alanine and are often referred to as binding *hot spots* [50,51]. One definition for a hot spot is that the binding free energy is altered by $\geq$1 kcal/mol upon mutation [52]. While the precise definition might be subject to debate, hot spots are "real" in the sense that they can be predicted accurately by methods that do not even assume that hot spots exist [53]. We moreover might anticipate the observation that the residues or positions contributing most to the energy of binding might also be the residues used more frequently when choosing sites that bind to many binding partners. Indeed, hot spots have been observed to have a high propensity for interaction with multiple partners [54].

Substituting the native amino acid by alanine is typically experimentally easiest and expected to be most revealing. Thus, alanine scans are most common, but increasingly, glycine, proline and cysteine scans are also carried out [55–57]. In these scans, all native residues of a protein are individually substituted by one of the above amino acids and the effect upon a given functional assay is measured. ASEdb, the Alanine Scanning Energetics database, provides a central repository for such data [58]. Residues that significantly change protein function are usually considered important. What constitutes a significant change depends on the type of function. *In silico* predictions suggest that when looking at the effect of all 19-non-native mutations, alanine substitutions are most representative (correlate most with the average over all mutations) [21]. Although this observation is based on one single protein (HXK4) and may therefore not be representative, the fact that *in silico* methods accurately reproduce such expert knowledge should be appreciated as an independent evidence of their success in predicting essential aspects of the mutability landscape.

Mutability landscape constrained by correlation networks?

Comprehensive experimental mutagenesis studies confirmed that the effect of point mutations (SAAS) upon function depends crucially on their positions in the protein sequence [59–61]. Even within a unit as familiar as the DNA binding domain of the *Escherichia coli* LacI [62,63] repressor, almost any variant can be tolerated at some positions while, at others, all variants affect function (Fig. 2a). Simple structural constraints might suffice to explain this variability: to accommodate the negatively charged DNA, binding regions of the repressor contain positively charged residues. Furthermore, binding

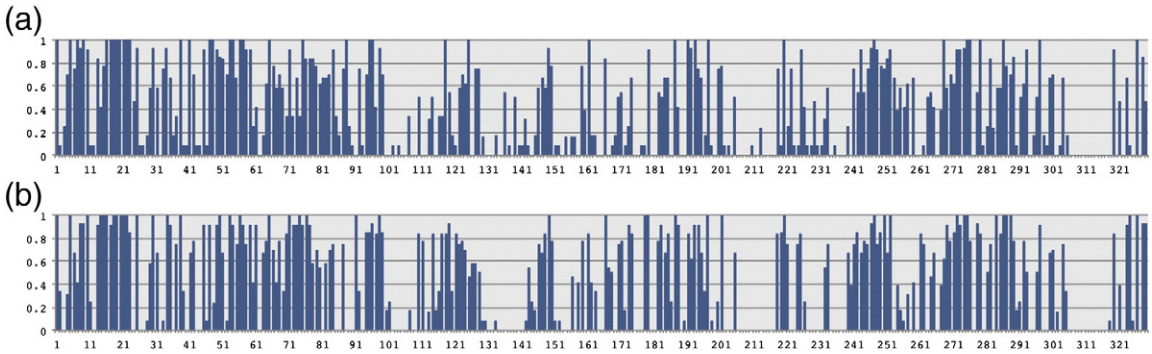

Fig. 2. Mutagenesis of *E. coli* LacI repressor. At each position between residue 2 and 329, 12–13 amino acid substitutions are displayed as a bar. The height of a bar depicts the relative percentage of substitutions that alter the repressor function as determined (a) experimentally [59] or (b) by computational prediction using SNAP2. With a correlation of 0.76 over all residues and an accuracy of 78.2% over all variants, this constitutes a below-average prediction of SNAP2 (~82% estimated overall accuracy).

requires helix formation. These two simple biophysical realities constrain the mutability landscape significantly in a specific, identity-revealing manner like a fingerprint. The differential sensitivity to mutation might just be a complex overlay of many such simple biophysical constraints.

The same constraints are written into the profile of evolutionary conservation of changes observed within families of related proteins [64–66]. These evolutionary imprints are strong enough to aid the prediction of protein structure [67–70] and function [39]. One particular idea that uses the constraints imposed by the mutability landscape is that of compensating/correlated mutations [71–73]. To simplify this, imagine a salt bridge, that is, the interaction between a positively charged residue and a negatively charged residue. If the negative one is mutated into a positively charged amino acid, the affected protein may malfunction. A compensatory mutation that also flips the charge of the positive position will again allow salt-bridge formation. If we could identify correlated mutations, we could use them to predict inter-residue contacts within [72] and between proteins [74,75]. After many years of development [18,76], this concept has finally brought about *de novo* predictions for three-dimensional (3D) structure of globular [77] and large membrane proteins [18,19], several of those are not similar to any protein structure we know today. Such predictions may even lead the way beyond structure [19]: many residues that are evolutionarily coupled and not close in space may be relevant for protein function. In fact, in a study of over 14,000 variants related to disease from over 1000 human proteins, correlated positions appeared significantly more likely to harbor disease mutations than average positions [78]. Compensatory mutations involve coupled variants and might rashly be considered to go beyond the focus of this review. We show that the correlated mutations perfectly highlight the importance of analyzing the mutability landscape.

Other recent studies carry the theme of coevolving positions even further. Patterns of correlated mutations in the WW domain nearly suffice to synthetize artificial WW domains with native-like folding and function [79,80]. Applying statistical coupling analysis to the S1A protein family, the Ranganathan laboratory introduced the "sector hypothesis" [81] that proteins are organized into distinct subunits or networks (sectors) of coevolving residues that are essential to structure and function. Such sectors involve only about every fifth residue; they are built around active sites, and they connect to other functional sites distant in sequence and structure through "networks" of contiguous residue interactions in the protein core [82]. These networks of coevolving residues may have resulted from the need for rapid adaptive variation arising from fluctuating selection pressure and that the organization into networks of cooperatively acting residues may provide such rapid adaptive potential through only a few mutations [82]. If so, structure and function may mostly be affected by mutations at sector positions while non-sector positions may tolerate variation. This hypothesis was tested through a complete single mutagenesis (individually substituting each residue by all 19 non-native amino acids) in one representative member of the PDZ family ($PSD95^{pdz3}$) [82]. The study showed that the statistical correlation between mutations with significant functional effect and sector positions was very strong; it was, in fact, stronger than that between mutations in the protein core (buried positions) and positions with ligand contacts. Moreover, a combination of two mutations at sector positions was sufficient to change the binding specificity of $PSD95^{pdz3}$ for a class-switching ligand. This adaption is exclusively initiated through mutations in the sector. While awaiting large-scale confirmation,

these findings already highlight the importance of annotating correlated mutational behavior for the prediction of pathogenicity/functional effects of missense variants.

Penetrating the Protein Mutability Landscape *In Silico*

Predicting the mutability landscape of the human exome

Ultimately, we want to study the entire protein mutability landscapes for at least some hundreds of representative proteins by assaying changes in protein function and their impact upon the organism. Despite tremendous breakthroughs in high-throughput experimentation, this analysis falls more into the world of Lewis Carroll than into that of a scientific grant proposal. However, such a landscape can be easily predicted, for example, for all human proteins. The downside is that we do not yet fully understand how to interpret the results. Nevertheless, in the context of understanding the deep sequencing data, such views are needed.

Finding the causal variants for a particular disease continues to be a challenging endeavor despite the continued decrease of the cost in sequencing entire genomes and entire exomes [83]. Accordingly, researches prioritize zooming in onto candidate variants in these studies by including computational effect predictions [84]. It has been suggested to combine several prediction methods (e.g., through majority vote) in order to overcome individual weaknesses and obtain most reliable predictions [85]. The dbNSFP [86] database is built to simplify this endeavor by providing effect predictions and scores from various methods for every potential variant in the human genome (approximately 76 million variants). Differences between methods become most apparent when comparing predictions on this large scale. The pairwise agreement between the four methods in dbNSFP ranges from 61% to 77%. The fraction of all potential substitutions predicted to be deleterious by individual methods ranges from 40% to 56%, suggesting that methods disagree strongly. Overall, methods accurately predict Mendelian disease-causing variants to strongly effect function. Unfortunately, this does not imply that the same methods can find a single disease-causing variant among the thousands of variants observed between any pair of individuals from the same population.

To visualize the predictive behavior of the two widely used methods SNAP [24,25] and SIFT [22,23], we compiled a pairwise amino acid substitution matrix over all theoretically possible variants in the human proteome based on the predictions of each method (Fig. 3; note that not all of those variants can be observed since not all amino acids can be transformed into all others through a SNP). Although SNAP predicts more effect substitutions than SIFT, trends appear to be largely similar. For instance, both methods predict substitutions from and to tryptophan as highly damaging on average. This is plausible due to its structural importance to proteins. Similarly, both methods predict substitutions of phenylalanine by any other residue, except for leucine and tyrosine, as rather damaging. Phenylalanine is preferentially exchanged with tyrosine, which differs only in that it contains a hydroxyl group in place of the *ortho* hydrogen on the benzene ring. The preference for leucine seems plausible due to its hydrophobic character. Examples of SNAP and SIFT differences are in the predictions for substitutions of arginine and by proline. SNAP might be closer to the truth for these two because they may be difficult to treat via a purely evolution based method. “To proline” mutations are likely to be rare due to their disruptions. For arginine, the explanation seems less clear. How can we cast such predictions into new methods that, for example, predict active sites? How can we use them to guide protein design?

Outcome of alanine scans predicted

Methods that predict functional effects have rarely been assessed in large-scale mutagenesis experiments. One reason is obviously the shortage of such experiments. Another might be the perception that computational methods typically predict neither the severity nor the direction of the effect (increase or decrease of function/affinity). It is true that today's prediction methods cannot directly distinguish between variants that increase and those that decrease binding. Instead, both tend to be predicted as effects. Nevertheless, prediction scores (i.e., the signal strength) on average correlate with the severity of the effect [24]. The concept of “importance for function” never entered the data set choice or development phase when creating SNAP. Still, when applied for residues in ASEdb that the method had never “seen” before, it correctly identified over 70% of the functionally important sites and correctly predicted many to-alanine variants (up to 84%, depending on cutoff) [21].

Comprehensive *in silico* mutagenesis helps studying disease-related proteins

A detailed study of the human melanocortin 4 receptor (hMC4R) demonstrated the value of studying the mutability landscape *in silico* [20]. hMC4R is related to diabetes and to weight regulation. Mutations in hMC4R have been shown to account for approximately 3% of all severe obesity cases (body

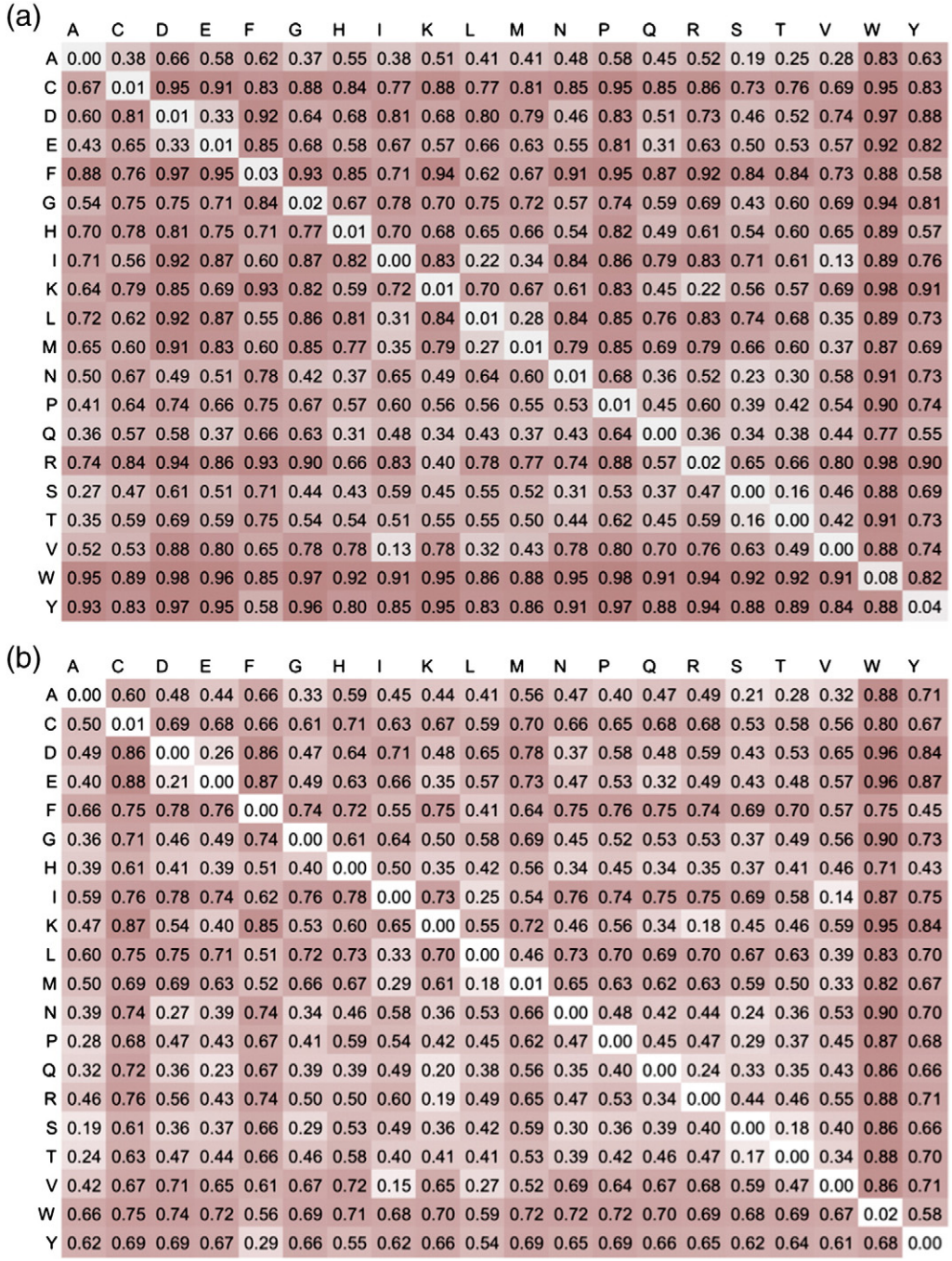

Fig. 3. Effect of pairwise amino acid substitutions in the human exome. Shown is the fraction of substitutions predicted to have an effect for every substitution of every amino acid (*y*-axis) by any other (*x*-axis) in the entire human exome. Results were obtained by locally calculating the predictions for (a) SNAP and (b) SIFT for every possible SAAS in every reviewed protein in the Swiss-Prot database [12] with human origin. Cells are colored according to the fraction of deleterious predictions with high values in red and low values in white. For every prediction for substitutions of amino acid "m" (*y*-axis) by "n" (*x*-axis), we applied the default threshold for each method (SNAP, 0; SIFT, 0.05).

mass index, >40), and consequently, they are the most frequent cause of monogenic obesity in humans [87,88]. MC4R, a member of the G-protein-coupled receptor (GPCR) family, is an integral membrane protein that crosses the lipid bilayer with seven transmembrane helices. SNAP assessed the functional essentiality for each of the 332 residues in hMC4R and the functional impact of all possible variants; predictions were compared to all available experimental data. The predictions of variants with

functional effect and predictions of important regions in hMC4R largely agreed with experimental evidence [20]. Toward this end, we down-weighted mutations expected to be neutral for structure (e.g., hydrophobic to hydrophobic in membrane regions).

Despite this scoring, the computational mutagenesis predicted as many as 118 residues to be functionally important. This seems a substantial over-prediction. Indeed, so far, we have experimental evidence for only 18 residues to be important for function; 15 of these 18 were in the set of 118 residues predicted to have strong impact [20], which is not an impressive performance but much higher than the random 6 in 18. The nsSNP database of effects (SNPdbe [89]) provides experimental links to obesity for 27 residues, 17 of those were in the 118. Only one single residue is found in both sets. Thus, *in silico* mutagenesis correctly predicted 31 of the known 44 positions reported to influence function if mutated.

What about the 74 residues with predictions but without observation (118 predicted, 44 so far experimentally known)? At this point, 74 mutations constitute a relatively large number of high-effect predictions, which cannot be verified due to lack of data. Re-evaluating the predictions, we might apply a more stringent threshold to consider an effect important. For instance, at a threshold with an expected accuracy >95%, 22 residues are predicted to impact function; 10 of those correspond to experimentally known sites, 1 corresponds to a site implicated with obesity and 11 remain without experimental annotations. These might constitute ideal starting points for designing new experiments [20].

Detailed analysis of mutability landscape for a GPCR, the beta-2-adrenergic receptor

To visualize the results of such an *in silico* mutagenesis, we applied SNAP2 (M.H., unpublished results) to another GPCR, the beta-2 adrenergic receptor for which experimental high-resolution 3D structures are available in the Protein Data Bank (PDB) [90] (PDB ID: 2RH1 [15], Fig. 1d; PDB ID: 3PDS [13], Fig. 1e). The predicted high-effect regions cluster around the binding sites and are significantly more abundant on the inside (facing the binding sites) than elsewhere. Strong effects (SNAP > 60; note: score ranges [−100,+100]) are predicted for 57 residues (Fig. 1d, red highlighting) including the two Swiss-Prot annotated [12,91] binding sites D113 and T118. Nine more sites with functional effect annotation in SNPdbe [89] were found. Among these, we find (1) D79, for which a mutation to N was shown to affect binding of catecholamines and to produce an uncoupling between the receptor and stimulatory G-proteins [92,93], and (2) D130, for which mutations to A or N were shown to increase pindolol-stimulated cAMP accumulation [94,95]. Although located in the cytoplasmic region, strong signals also highlight (1) Y141, for which a substitution by F is known to abolish insulin-induced tyrosine phosphorylation and insulin-induced receptor super sensitization [96], and (2) C341, for which a mutation to G was shown to alter binding (uncoupling of receptor) [97].

Thus, 46 sites (57 predicted, 11 experimentally observed) remain with strong effect predictions for which no variants have been tested experimentally. Again we observe a rather large discrepancy between observed and predicted “sensitive to mutation” positions. Some of these predictions will likely just be false positives. However, due to being located in the protein core, others may in fact affect function by structural alterations/misfolding. Application of an even more stringent threshold (score > 80 at >95% expected accuracy) weeds out 38 of these, leaving 12 residues with very strong effect predictions and without current variant annotations. We studied these also in light of EVfold [18,19] (prediction of inter-residue contacts through correlated mutations). Only 2 of the 12 had residue couplings in the realm of the top 5%, namely, W99 and Y199 (Fig. 1e). We could not find any experimental annotation about these two. However, a visual inspection (Fig. 1e) of a 3D structure with irreversibly bound agonist (PDB ID: 3PDS [13]) appears to suggest the two as reasonable targets for experimental verification.

This detailed view of the beta-2-adrenergic receptor provides another example for how useful it might be to analyze the mutability landscape through a complete *in silico* mutagenesis; it highlights functionally important regions and may help in experiment design to probe function locally or test entire regions for docking and drug development. The example also suggests that variant effect prediction might benefit from including inter-residue contact/evolutionary coupling predictions.

Perspective

Comprehensive mutagenesis experiments have furthered our understanding of protein function and continue to provide insight into the mechanisms of pathogenicity and adaption. Novel methodologies and technical advancements reduce the cost of experimental mutagenesis and enable research that was previously impossible. Still, studying the cooperative behavior of amino acids and the combined effect of mutations will remain a laborious and costly task. This is where computational methods are useful to predict the effects of variants upon protein function, structure and pathogenicity. These methods have grown in accuracy both in predicting functional effect (Fig. 2b) and disease-causing

mutations. Can they reach the next level? Can they be used to study the mutability landscape of a protein, that is, to unravel the effects of all possible variants? Here, we argue that the study of such a mutability landscape provides immensely important value and that currently neither experimental nor computational methods completely mine the potential of studying this landscape. Experimental methods remain constrained by the substantial amount of resources such studies would consume. Computational methods remain constrained by the degree to which we can interpret their results. At this point, lack of comprehensive experimental data seems a crucial problem for the development of better computational tools. However, *in silico* analyses of mutability landscape already help to design experiments and are crucial for the intelligent interpretation of deep sequencing/next generation sequencing data.

Acknowledgements

Thanks to Tim Karl, Guy Yachdav and Laszlo Kajan (Technische Universität München) for invaluable help with hardware and software; to Marlena Drabik (Technische Universität München) for administrative support; and to Thomas Hopf and Laszlo Kajan (both Technische Universität München) for helpful discussions and help with the manuscript. Thanks to the developers of PyMOL [98] (Fig. 1d and e) and Matrix2png [99] (Fig. 1b and c) for providing great tools. This work was supported by a grant from the Alexander von Humboldt Foundation through the German Ministry for Research and Education (Bundesministerium fuer Bildung und Forschung). Last, not the least, thanks to all those who deposit their experimental data in public databases and to those who maintain these databases.

Supplementary Data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jmb.2013.07.028

Received 30 April 2013;
Received in revised form 8 July 2013;
Accepted 19 July 2013
Available online 26 July 2013

Keywords:
complete single mutagenesis;
alanine scanning;
in silico mutagenesis;
exome-wide mutagenesis;
SNP effects

Abbreviations used:
3D, three-dimensional; hMC4R, human melanocortin 4 receptor; GPCR, G-protein-coupled receptor; nsSNP, non-synonymous SNP; PDB, Protein Data Bank; SAAS, single-amino-acid substitution; SIFT, sorting intolerant from tolerant; SNAP, screening for non-acceptable polymorphisms; SNP, single nucleotide polymorphism.

References

[1] Kimura M. Evolutionary rate at the molecular level. Nature 1968;217:624–6.
[2] Ohta T. Near-neutrality in evolution of genes and gene regulation. Proc Natl Acad Sci USA 2002;99:16134–7.
[3] Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74.
[4] Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. Nat Rev Genet 2009;10:241–51.
[5] Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 2012;337:64–9.
[6] O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, et al. Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature 2012;485:246–50.
[7] Rauch A, Wieczorek D, Graf E, Wieland T, Endele S, Schwarzmayr T, et al. Range of genetic mutations associated with severe non-syndromic sporadic intellectual disability: an exome sequencing study. Lancet 2012;380:1674–82.
[8] Thusberg J, Vihinen M. Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. Hum Mutat 2009;30:703–14.
[9] McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 2008;9:356–69.
[10] Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, et al. Exome sequencing and the genetic basis of complex traits. Nat Genet 2012;44:623–30.
[11] Samani NJ, Erdmann J, Hall AS, Hengstenberg C, Mangino M, Mayer B, et al. Genomewide association analysis of coronary artery disease. N Engl J Med 2007;357:443–53.
[12] Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res 2000;28:45–8.
[13] Rosenbaum DM, Zhang C, Lyons JA, Holl R, Aragao D, Arlow DH, et al. Structure and function of an irreversible agonist-beta(2) adrenoceptor complex. Nature 2011;469:236–40.
[14] Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. Biopolymers 1983;22:2577–637.

[15] Cherezov V, Rosenbaum DM, Hanson MA, Rasmussen SG, Thian FS, Kobilka TS, et al. High-resolution crystal structure of an engineered human beta2-adrenergic G protein-coupled receptor. Science 2007;318:1258–65.
[16] Ng PC, Henikoff JG, Henikoff S. PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. Bioinformatics 2000;16:760–6.
[17] Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89: 10915–9.
[18] Marks DS, Hopf TA, Sander C. Protein structure prediction from sequence variation. Nat Biotechnol 2012;30: 1072–80.
[19] Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS. Three-dimensional structures of membrane proteins from genomic sequencing. Cell 2012;149:1607–21.
[20] Bromberg Y, Overton J, Vaisse C, Leibel RL, Rost B. *In silico* mutagenesis: a case study of the melanocortin 4 receptor. FASEB J 2009;23:3059–69.
[21] Bromberg Y, Rost B. Comprehensive *in silico* mutagenesis highlights functionally important residues in proteins. Bioinformatics 2008;24:i207–12.
[22] Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009;4:1073–81.
[23] Ng PC, Henikoff S. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res 2003;31: 3812–4.
[24] Bromberg Y, Rost B. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 2007;35:3823–35.
[25] Bromberg Y, Yachdav G, Rost B. SNAP predicts effect of mutations on protein function. Bioinformatics 2008;24: 2397–8.
[26] Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, et al. Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 2009;25:2744–50.
[27] Calabrese R, Capriotti E, Fariselli P, Martelli PL, Casadio R. Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat 2009;30:1237–44.
[28] Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 2011;39:e118.
[29] Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 2010;7:575–6.
[30] Schaefer C, Rost B. Predict impact of single amino acid change upon protein structure. BMC Genomics 2012;13: S4.
[31] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods 2010;7:248–9.
[32] Olatubosun A, Valiaho J, Harkonen J, Thusberg J, Vihinen M. PON-P: integrated predictor for pathogenicity of missense variants. Hum Mutat 2012;33:1166–74.
[33] Stitziel NO, Kiezun A, Sunyaev S. Computational and statistical approaches to analyzing variants identified by exome sequencing. Genome Biol 2011;12:227.
[34] Cline MS, Karchin R. Using bioinformatics to predict the functional impact of SNVs. Bioinformatics 2011;27: 441–8.
[35] Mah JT, Low ES, Lee E. *In silico* SNP analysis and bioinformatics tools: a review of the state of the art to aid drug discovery. Drug Discovery Today 2011;16:800–9.
[36] Oetting WS. Exploring the functional consequences of genomic variation: the 2010 Human Genome Variation Society Scientific Meeting. Hum Mutat 2011;32:486–90.
[37] Thusberg J, Olatubosun A, Vihinen M. Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat 2011;32:358–68.
[38] Lesk AM, Levitt M, Chothia C. Alignment of the amino acid sequences of distantly related proteins using variable gap penalties. Protein Eng 1986;1:77–8.
[39] Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y. Automatic prediction of protein function. Cell Mol Life Sci 2003;60: 2637–50.
[40] Rost B, O'Donoghue S, Sander C (1998). Midnight zone of protein structure evolution. EMBL Heidelberg.
[41] Lichtarge O, Bourne HR, Cohen FE. Evolutionarily conserved Galphabetagamma binding surfaces support a model of the G protein-receptor complex. Proc Natl Acad Sci 1996;93: 7507–11.
[42] Sankararaman S, Sjolander K. INTREPID—INformation-theoretic TREe traversal for Protein functional site IDentification. Bioinformatics 2008;24:2445–52.
[43] Sankararaman S, Sha F, Kirsch JF, Jordan MI, Sjolander K. Active site prediction using evolutionary and structural information. Bioinformatics 2010;26:617–24.
[44] Henikoff S, Comai L. Single-nucleotide mutations for plant functional genomics. Annu Rev Plant Biol 2003;54: 375–401.
[45] Zimprich A. Genetics of Parkinson's disease and essential tremor. Curr Opin Neurol 2011;24:318–23.
[46] Zimprich A, Benet-Pages A, Struhal W, Graf E, Eck SH, Offman MN, et al. A mutation in VPS35, encoding a subunit of the retromer complex, causes late-onset Parkinson disease. Am J Hum Genet 2011;89:168–75.
[47] Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res 2009;37:D793–6.
[48] Schaefer C, Bromberg Y, Achten D, Rost B. Disease-related mutations predicted to impact protein function. BMC Genomics 2012;13:S11.
[49] Eisenberg D, Marcotte EM, Xenarios I, Yeates TO. Protein function in the post-genomic era. Nature 2000;405:823–6.
[50] Bogan AA, Thorn KS. Anatomy of hot spots in protein interfaces. J Mol Biol 1998;280:1–9.
[51] Clackson T, Wells JA. A hot spot of binding energy in a hormone-receptor interface. Science 1995;267:383–6.
[52] Kortemme T, Baker D. A simple physical model for binding energy hot spots in protein–protein complexes. Proc Natl Acad Sci USA 2002;99:14116–21.
[53] Ofran Y, Rost B. Protein–protein interaction hot spots carved into sequences. PLoS Comput Biol 2007;3:e119.
[54] DeLano WL, Ultsch MH, de Vos AM, Wells JA. Convergent solutions to binding at a protein–protein interface. Science 2000;287:1279–83.
[55] Konishi S, Iwaki S, Kimura-Someya T, Yamaguchi A. Cysteine-scanning mutagenesis around transmembrane segment VI of Tn10-encoded metal-tetracycline/H(+) antiporter. FEBS Lett 1999;461:315–8.
[56] Qin L, Cai S, Zhu Y, Inouye M. Cysteine-scanning analysis of the dimerization domain of EnvZ, an osmosensing histidine kinase. J Bacteriol 2003;185: 3429–35.

[57] Gardsvoll H, Gilquin B, Le Du MH, Menez A, Jorgensen TJ, Ploug M. Characterization of the functional epitope on the urokinase receptor: complete alanine scanning mutagenesis supplemented by chemical cross-linking. J Biol Chem 2006;281:19260–72.
[58] Thorn KS, Bogan AA. ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. Bioinformatics 2001;17:284–5.
[59] Markiewicz P, Kleina LG, Cruz C, Ehret S, Miller JH. Genetic studies of the *lac* repressor. XIV. Analysis of 4000 altered *Escherichia coli lac* repressors reveals essential and non-essential residues, as well as "spacers" which do not require a specific sequence. J Mol Biol 1994;240:421–33.
[60] Loeb DD, Swanstrom R, Everitt L, Manchester M, Stamper SE, Hutchison CA. Complete mutagenesis of the HIV-1 protease. Nature 1989;340:397–400.
[61] Rennell D, Bouvier SE, Hardy LW, Poteete AR. Systematic mutation of bacteriophage T4 lysozyme. J Mol Biol 1991;222:67–88.
[62] Gottesman ME, Yarmolinsky MB. Integration-negative mutants of bacteriophage lambda. J Mol Biol 1968;31:487–505.
[63] Gottesman S, Gottesman ME. Elements involved in site-specific recombination in bacteriophage lambda. J Mol Biol 1975;91:489–99.
[64] Epstein CJ. Role of the amino acid "code" and of selection for conformation in the evolution of proteins. Nature 1966;210:25–8.
[65] Woese CR, Dugre DH, Dugre SA, Kondo M, Saxinger W. On the fundamental nature and evolution of the genetic code. Cold Spring Harbor Symp Quant Biol 1966;31:723–36.
[66] Dayhoff MO, Barker WC, Hunt LT. Establishing homologies in protein sequences. Method Enzymol 1983;91:524–45.
[67] Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJE. Prediction of protein secondary structure and active sites using alignment of homologous sequences. J Mol Biol 1987;195:957–61.
[68] Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–99.
[69] Rost B. PHD: predicting one-dimensional protein structure by profile based neural networks. Methods Enzymol 1996;266:525–39.
[70] Rost B, Sander C. Bridging the protein sequence-structure gap by structure predictions. Annu Rev Biophys Biomol Struct 1996;25:113–36.
[71] Altschuh D, Lesk AM, Bloomer AC, Klug A. Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. J Mol Biol 1987;193:693–707.
[72] Goebel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins Struct Funct Genet 1994;18:309–17.
[73] Altschuh D, Vernet T, Berti P, Moras D, Nagai K. Coordinated amino acid changes in homologous protein families. Protein Eng 1988;2:193–9.
[74] Pazos F, Ranea JA, Juan D, Sternberg MJ. Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. J Mol Biol 2005;352:1002–15.
[75] Pazos F, Valencia A. *In silico* two-hybrid systemfor the selection of physically interacting protein pairs. Proteins Struct Funct Genet 2002;47:219–27.
[76] de Juan D, Pazos F, Valencia A. Emerging methods in protein co-evolution. Nat Rev Genet 2013;14:249–61.
[77] Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D structure computed from evolutionary sequence variation. PLoS One 2011;6:e28766.
[78] Kowarsch A, Fuchs A, Frishman D, Pagel P. Correlated mutations: a hallmark of phenotypic amino acid substitutions. PLoS Comput Biol 2010;6:e1000923.
[79] Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. Evolutionary information for specifying a protein fold. Nature 2005;437:512–8.
[80] Russ WP, Lowery DM, Mishra P, Yaffe MB, Ranganathan R. Natural-like function in artificial WW domains. Nature 2005;437:579–83.
[81] Halabi N, Rivoire O, Leibler S, Ranganathan R. Protein sectors: evolutionary units of three-dimensional structure. Cell 2009;138:774–86.
[82] McLaughlin RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. The spatial architecture of protein function and adaptation. Nature 2012;491:138–42.
[83] Maxmen A. Exome sequencing deciphers rare diseases. Cell 2011;144:635–7.
[84] Li MX, Gui HS, Kwan JS, Bao SY, Sham PC. A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Res 2012;40:e53.
[85] Chun S, Fay JC. Identification of deleterious mutations within three human genomes. Genome Res 2009;19:1553–61.
[86] Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 2011;32:894–9.
[87] Farooqi IS, Keogh JM, Yeo GS, Lank EJ, Cheetham T, O'Rahilly S. Clinical spectrum of obesity and mutations in the melanocortin 4 receptor gene. N Engl J Med 2003;348:1085–95.
[88] Lubrano-Berthelier C, Le Stunff C, Bougneres P, Vaisse C. A homozygous null mutation delineates the role of the melanocortin-4 receptor in humans. J Clin Endocrinol Metab 2004;89:2028–32.
[89] Schaefer C, Meier A, Rost B, Bromberg Y. SNPdbe: constructing an nsSNP functional impacts database. Bioinformatics 2012;28:601–2.
[90] Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR, et al. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–42.
[91] Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 2006;34:D187–91.
[92] Chung FZ, Wang CD, Potter PC, Venter JC, Fraser CM. Site-directed mutagenesis and continuous expression of human beta-adrenergic receptors. Identification of a conserved aspartate residue involved in agonist binding and receptor activation. J Biol Chem 1988;263:4052–5.
[93] Moffett S, Rousseau G, Lagace M, Bouvier M. The palmitoylation state of the beta(2)-adrenergic receptor regulates the synergistic action of cyclic AMP-dependent protein kinase and beta-adrenergic receptor kinase involved in its phosphorylation and desensitization. J Neurochem 2001;76:269–79.

[94] Ballesteros JA, Jensen AD, Liapakis G, Rasmussen SG, Shi L, Gether U, et al. Activation of the beta 2-adrenergic receptor involves disruption of an ionic lock between the cytoplasmic ends of transmembrane segments 3 and 6. J Biol Chem 2001;276:29171–7.

[95] Rasmussen SG, Jensen AD, Liapakis G, Ghanouni P, Javitch JA, Gether U. Mutation of a highly conserved aspartic acid in the beta2 adrenergic receptor: constitutive activation, structural instability, and conformational rearrangement of transmembrane segment 6. Mol Pharmacol 1999;56:175–84.

[96] Valiquette M, Parent S, Loisel TP, Bouvier M. Mutation of tyrosine-141 inhibits insulin-promoted tyrosine phosphorylation and increased responsiveness of the human beta 2-adrenergic receptor. EMBO J 1995;14:5542–9.

[97] O'Dowd BF, Hnatowich M, Caron MG, Lefkowitz RJ, Bouvier M. Palmitoylation of the human beta 2-adrenergic receptor. Mutation of Cys341 in the carboxyl tail leads to an uncoupled nonpalmitoylated form of the receptor. J Biol Chem 1989;264:7564–9.

[98] DeLano WL. The PyMOL molecular graphics system; 2002 .

[99] Pavlidis P, Noble WS. Matrix2png: a utility for visualizing matrix data. Bioinformatics 2003;19:295–6.

[100] Bromberg Y, Kahn PC, Rost B. Neutral and weakly nonneutral sequence variants may define individuality. Proceedings of the National Academy of Sciences; 2003.