Predicting the Effects of Amino Acid Substitutions on Protein Function

Pauline C. Ng and Steven Henikoff

Fred Hutchinson Cancer Research Center, Seattle, Washington 98109;
email: sift@fhcrc.org, steveh@fhcrc.org

Annu. Rev. Genomics Hum. Genet. 2006. 7:61–80

The *Annual Review of Genomics and Human Genetics* is online at genom.annualreviews.org

This article's doi: 10.1146/annurev.genom.7.080505.115630

1527-8204/06/0922-0061$20.00

Key Words

nsSNP, nonsynonymous, single nucleotide polymorphism, coding SNP, missense

Abstract

Nonsynonymous single nucleotide polymorphisms (nsSNPs) are coding variants that introduce amino acid changes in their corresponding proteins. Because nsSNPs can affect protein function, they are believed to have the largest impact on human health compared with SNPs in other regions of the genome. Therefore, it is important to distinguish those nsSNPs that affect protein function from those that are functionally neutral. Here we provide an overview of amino acid substitution (AAS) prediction methods, which use sequence and/or structure to predict the effect of an AAS on protein function. Most methods predict approximately 25–30% of human nsSNPs to negatively affect protein function, and such nsSNPs tend to be rare in the population. We discuss the utility of AAS prediction methods for Mendelian and complex diseases as well as their broader applications for understanding protein function.

Online Mendelian Inheritance in Man (OMIM): database containing mutations discovered in patients and found in genes known to be involved in disease

Human Gene Mutation Database (HGMD): database containing mutations in genes known to be involved in disease

Nonsynonymous SNP (nsSNP): a single nucleotide polymorphism located in a coding region that causes an amino acid substitution in the corresponding protein

Amino acid substitution (AAS) prediction method: bioinformatics tool that evaluates whether an AAS affects protein function

INTRODUCTION

Most genetic variation is considered neutral but single base changes in and around a gene can affect its expression or the function of its protein products (11, 56). A nonsynonymous or missense variant is a single base change in a coding region that causes an amino acid change in the corresponding protein. If a nonsynonymous variant alters protein function, the change can have drastic phenotypic consequences. Most alterations are deleterious and so are eventually eliminated through purifying selection. However, beneficial mutations can sweep through the population and become fixed, thus contributing to species differentiation.

The importance of nonsynonymous substitutions in humans is illustrated by two databases containing disease-causing variants, Online Mendelian Inheritance in Man (OMIM) and Human Gene Mutation Database (HGMD) (24, 62). In both databases, nonsynonymous changes account for approximately half of the genetic changes known to cause disease. Although these databases contain information primarily concerning disorders caused by single Mendelian lesions, it is likely that nonsynonymous changes will play a similarly important role in complex diseases because of their potentially large impact.

The human population is estimated to have 67,000–200,000 common nonsynonymous SNPs (nsSNPs) (8, 23, 35) and each person is thought to be heterozygous for 24,000–40,000 nsSNPs (8). It would be time-consuming, difficult, and expensive to experimentally characterize the impact of each nsSNP on protein function. But because an amino acid change can have a large impact on fitness, a computational method that could predict whether an amino acid substitution (AAS) affects protein function would help researchers prioritize AASs for additional study. The observation that disease-causing mutations are more likely to occur at positions that are conserved throughout evolution, as compared with positions that are not conserved, suggested that prediction could be based on sequence homology (39). It was also observed that disease-causing AASs had common structural features that distinguished them from neutral substitutions, suggesting that structure could also be used for prediction (68, 77). Since these studies were performed, a plethora of AAS prediction methods based on sequence and/or structure have become available (7, 9, 14, 16–19, 27, 31, 33, 40–42, 45–47, 53, 58, 59, 64, 65, 68, 69, 72, 73).

In this review we first survey existing AAS prediction methods and summarize the history of the field. We also offer practical advice for researchers who would like to use AAS prediction methods. Next, we look at the usefulness of AAS prediction methods in identifying candidate mutations responsible for both Mendelian and complex diseases. Third, we discuss other applications of AAS prediction methods. Finally, we discuss likely future improvements of such methods.

METHODOLOGY OF AMINO ACID SUBSTITUTION PREDICTION

Basic Methodology

AAS prediction methods use sequence and/or structural information for prediction. Prediction is feasible because mutations that affect protein function tend to occur at evolutionarily conserved sites (**Figure 1*a***) and/or are buried in protein structure (**Figure 1*b***). These observations came from several early studies that used AASs found in disease genes in affected individuals (39, 68, 77). These studies assumed that these substitutions affected protein function, thereby causing disease. These studies also assumed that a majority of nsSNPs in humans or the substitutions observed between humans and closely related species are functionally neutral. When Wang & Moult (77) modeled disease-causing mutations

onto their corresponding wild-type protein structures, they found that 83% of disease-causing mutations affected protein stability. By applying the stability criterion and several other structural criteria, they could detect 90% of disease-causing mutations. In contrast, only 30% of neutral nsSNPs were detected with the same set of rules, which suggests that their rules could be used to distinguish disease-causing mutations from neutral nsSNPs. Using both structure and sequence, Sunyaev et al. (68) could detect 70% of disease-causing mutations and only 17% of neutral substitutions. Based on analysis of protein sequences, Miller & Kumar (39) showed that disease-causing AASs are overabundant at conserved sites.

Based on these observations, AAS prediction methods using either sequence and/or structural information were introduced, some of which were also implemented as Web servers [**Table 1**; **Supplemental Table 1** (follow the Supplemental Material link from the Annual Reviews home page at **http://www.annualreviews.org.**)] (7, 9, 14, 16–19, 27, 31, 33, 40–42, 45–47, 53, 58, 59, 64, 65, 68, 69, 72, 73). The typical procedure used by AAS prediction methods is shown in **Figure 2**. To make a prediction, AAS prediction methods can use sequence, structure, and/or annotation. Sequence-based AAS prediction methods (14, 18, 19, 42, 45, 58, 69, 73) accept an input sequence and search it against a sequence database to find homologous sequences. A multiple sequence alignment of the homologous sequences reveals what positions have been conserved throughout evolutionary time, and these positions are inferred to be important for function. The AAS prediction method then scores the AAS based on the amino acids appearing in the multiple alignment and the severity of the amino acid change. An amino acid that is not present at the substitution site in the multiple alignment can still be predicted to be tolerated if there are amino acids with similar physiochemical properties present in the alignment. For example, if a protein sequence alignment

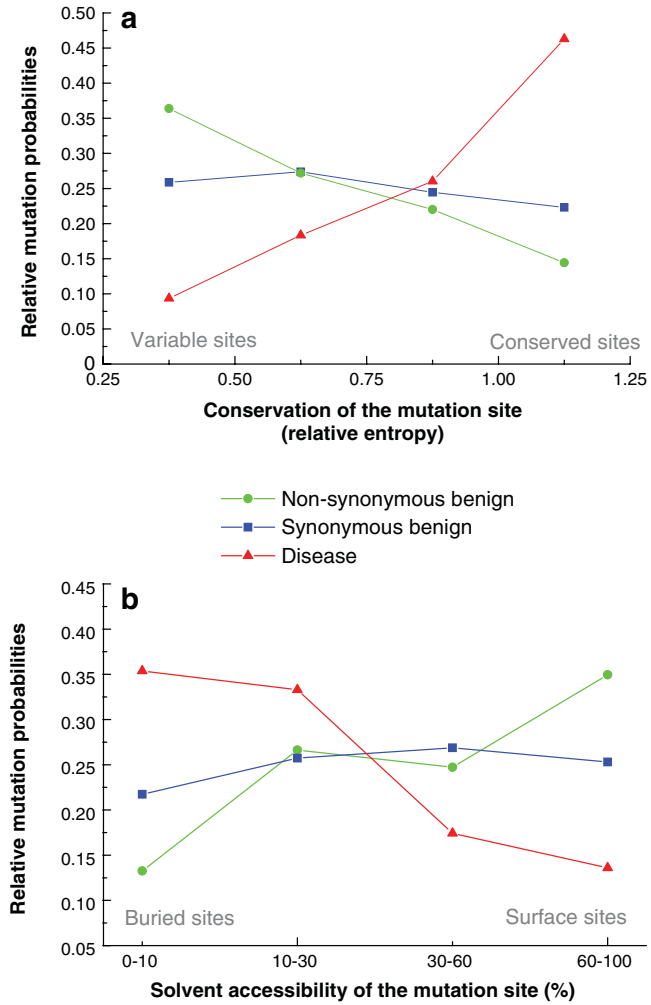

Figure 1

(*a*) The probability that a mutation will cause a disease increases monotonically with an increase in the degree of site conservation. Nonsynonymous SNPs found in normal individuals are assumed to be benign and their probabilities show the opposite trend where their occurrence decreases with the degree of site conservation. The benign synonymous SNPs do not change amino acids and should be predominantly neutral. As a result, their probability is uniform across sites, regardless of whether or not the site is conserved. Figure from Reference 75; licensee BioMed Central Ltd. (This is an Open Access Article: Verbatim copying and redistribution of this article are permitted in all media for any purpose.) (*b*) The solvent accessibility of an amino acid residue in a protein reflects the degree of the residue's exposure to the surrounding solvent in the protein structure. The relative probability of disease-causing mutations is highest in the protein interior and lowest on the protein surface. The benign SNPs show the reverse trend, as their relative probability is highest on the surface and lowest in the protein interior. Figure from Reference 75; licensee BioMed Central Ltd. (This is an Open Access Article: Verbatim copying and redistribution of this article are permitted in all media for any purpose.)

Table 1 Amino acid substitution (AAS) prediction methods available on the Internet*

Method and Web site	Interface	Performance	Algorithm
SIFT **http://blocks.fhcrc.org/sift/SIFT.html** (45–47)	**Input:** Protein sequence and AAS, protein sequence alignment and AAS, dbSNP id, or protein id **Output:** Score ranges from 0 to 1, where 0 is damaging and 1 is neutral	FN error: 31% FP error: 20% dbSNP: 25% predicted to be damaging Coverage: 60%	Using sequence homology, scores are calculated using position-specific scoring matrices with Dirichlet priors
PolyPhen **http://www.bork.embl-heidelberg.de/PolyPhen** (64, 65)	**Input:** Protein sequence and AAS, dbSNP id, HGVbASE id, or protein id **Output:** Score ranges from 0 to a positive number, where 0 is neutral, and a high positive number is damaging	FN error: 31% FP error: 9% dbSNP: 32% predicted damaging Coverage: 81%	Uses sequence conservation, structure to model position of amino acid substitution, and SWISS-PROT annotation
SNPs3D **http://www.snps3d.org/** (82, 83)	**Input:** dbSNP id, protein id, literature search, or gene ontology **Output:** Scores from structure-based SVM and sequence-based SVM reported separately. Score <0 is damaging. Mutation on protein structure can be visualized	Structure-based SVM FN error: 26% FP error: 15% Coverage: 14% Sequence-based SVM FN error: 20% FP error: 10% Coverage: 71% Predicted damaging in dbSNP: 25%	Structure-based support vector machine uses 15 structural factors Sequence-conservation support vector machine uses five features that capture sequence conservation
PANTHER PSEC **https://panther.appliedbiosystems.com/methods/csnpScoreForm.jsp** (73)	**Input:** Protein sequence and AAS **Output:** A negative score is damaging, zero is neutral, and positive is gain-of-function	FN error: 59% FP error: N/A Coverage: 40% dbSNP: 9% predicted damaging	Uses sequence homology; scores are calculated using PANTHER Hidden Markov Model families
PMUT **http://mmb2.pcb.ub.es:8080/PMut/** (16–18)	**Input:** Protein id, protein sequence, or multiple sequence alignment **Output:** Score ranges from 0 to 1, where 0 is neutral and high scores are predicted to be damaging. Mutation on protein structure is shown	FN error: 21% FP error: 17% When structure is included: FN error: 12% FP error: 10%	Prediction provided by one of two neural networks. Neural network uses internal databases, secondary structure prediction, and sequence conservation
TopoSNP **http://gila.bioengr.uic.edu/snp/toposnp** (64, 65)	**Input:** Protein id or protein sequence **Output:** Can view position of mutation. Location of substitution on protein (surface, internal, or pocket) and conservation reported separately Results are stored so an input protein sequence not in the database will not be processed	FN error: 12% FP error: N/A dbSNP: 68% predicted to be damaging	Classifies substitution as buried, on the surface, or in a pocket of the protein's structure. Also provides conservation score based on Pfam protein alignments

*False negative (FN) error rate is the percentage of substitutions predicted to be functionally neutral on a set of AASs that are known to affect protein function. These substitutions come from a mutagenesis set or those suspected to be involved in disease. False positive (FP) error rate is the percentage of substitutions predicted to be damaging on substitutions known to be functionally neutral. For some methods, error rates were not reported and are marked as N/A in the table. The coverage shown in this table is for dbSNP (60) and not for mutations in disease genes. Disease genes tend to have higher coverage because they are well studied. Coverage depends on databases, so methods published later tend to have more coverage than those published earlier. For a more complete list of AAS prediction methods, see **Supplemental Table 1**. (Follow the Supplemental Material link from the Annual Reviews home page at **http://www.annualreviews.org**.)

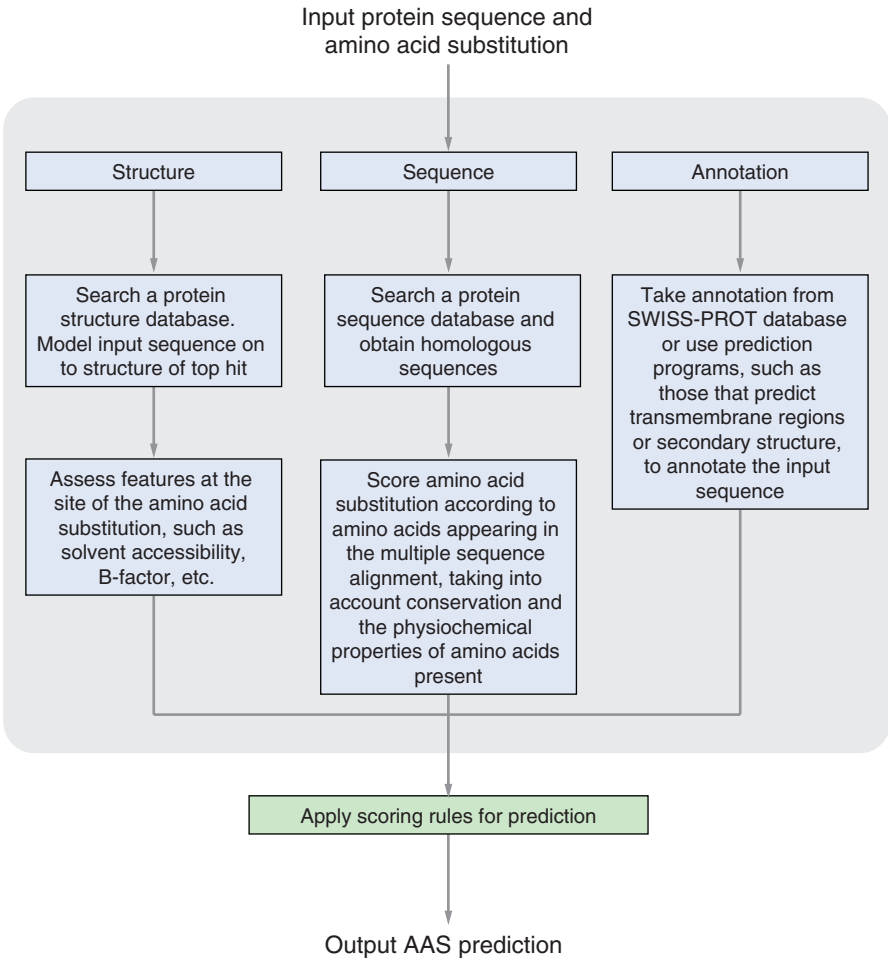

Figure 2

Flowchart for amino acid substitution (AAS) prediction. Input typically consists of the protein sequence and AASs. The method can use sequence and/or structural features for prediction. Some methods also use annotation to aid in prediction.

shows tyrosines and tryptophans at a particular site, one would expect that the other aromatic amino acid, phenylalanine, would also be tolerated at that site. The probability of observing a particular AAS at a site can be estimated from an appropriate model of protein conservation, and substitutions that are unlikely to be observed at a site are expected to reduce protein stability or function.

Structure-based AAS prediction methods (9, 27, 65, 69, 72) take an input sequence and find the best match against a protein structure database. Because most structure-based AAS prediction methods use general structural features surrounding the site of substitution and do not require detailed information at the atomic level, they can model the substitution onto the structure of a homologous protein rather than require the exact structure of the input sequence. AAS prediction methods then examine the position of the AAS and can take into account several structure factors such as solvent accessibility, carbon-beta density, crystallographic B-factor, and the difference in free energy between the new and the old amino acid. Based on these structural features, structure-based AAS prediction methods follow rules to arrive at a prediction.

AAS prediction methods can also incorporate annotations to refine prediction. The Swiss-Prot database annotates the positions of a protein that are located in the active site,

are involved in ligand binding, are part of a disulfide bridge, or are involved in other protein-protein interactions (1). AAS prediction methods can use this information to guide prediction (17, 18, 68, 69, 77). For example, if the position of the AAS is annotated as involved in ligand binding, then the AAS is predicted to affect the protein. Also, one can use sequence-based predictions of secondary structure and solvent accessibility and incorporate this annotation into the scoring scheme (18, 31).

Direct comparisons between AAS prediction methods are difficult because they were trained and tested on different data sets using different versions of sequence and structural databases as resources. The performance of an AAS prediction method depends on the data sets the method is tested on. AAS prediction methods are typically tested on two types of data sets: a nonneutral set, which contains substitutions assumed to affect protein function, and a neutral set, which contains substitutions assumed to have no effect. An AAS prediction method should predict the substitutions in the nonneutral set to be damaging to protein function. The percentage of non-neutral substitutions incorrectly predicted to be tolerated is an approximation of the false negative error rate. The AAS prediction method should also predict the majority of the substitutions in a neutral set as having no effect on protein function. The percentage of neutral substitutions incorrectly predicted to affect protein function approximates the false positive error rate. The best AAS prediction methods minimize both false negative and false positive error rates.

Popular nonneutral sets include data from laboratory mutagenesis experiments (36, 49, 54, 80), human disease proteins where many mutations have been characterized (3, 21, 22, 25, 32, 50, 55, 70), and human disease databases such as OMIM, HGMD, or Swiss-Prot (1, 24, 62), which contain AASs that have been found in patients. Currently, the mutagenesis data sets represent only a few proteins, so caution should be used in extrapolating results. For the data sets that contain mutations found in disease genes, it is assumed that the AAS found in the patient is the causative variant. However, the database entry is not necessarily the etiological variant. Instead, it could be in linkage disequilibrium with the causative variant. Alternatively, the causative mutation could be in an unscreened gene.

Popular neutral sets include substitutions that cause no phenotypic effect in mutagenesis experiments (36, 49, 54, 80) and nonsynonymous single-nucleotide mutations that have been fixed during divergence between human and a closely related species. The false positive error rate based on the first set tends to be higher than the second set because mutations deemed neutral in laboratory experiments are those that do not give a detectable phenotype. However, such mutations could have too small an impact on protein function or require alternative environmental conditions to reveal phenotypic effects (46). In the second set, substitutions between human and another species have undergone millions of years of evolutionary selective pressure, whereas those with negligible selection coefficients and very small effects on protein function have been eliminated.

There are many AAS prediction methods available and it is beyond the scope of this review to critique and compare every method. We have tried to summarize the major points of each method (**Supplemental Table 1**). It is encouraging that progress is being made in that overall prediction performance has improved over early methods, and many methods are now available to the research community as Web servers (**Table 1**).

Caveats for Using Structure in Prediction

AAS methods that use only protein structure provide fewer predictions than methods that use sequence because there are far fewer protein structures than sequences for

which homology can be found. The percentage of amino acid substitutions that can be predicted by an AAS method is defined as the method's coverage. Coverage for AAS methods that rely on protein structure only is approximately 14% (83), whereas coverage for AAS methods that use sequence can be as high as 81% (53). Notably, most methods now have a sequence-based score and analysis of structure has become an option offered by the methods. Sequence-based coverage should continue to improve, because large-scale sequencing projects are depositing predicted protein sequences into public databases at an increasing rate, whereas determination of three-dimensional (3D) structures remains a challenging endeavor.

Even when a structure is available, prediction based solely on protein structure can be misleading because the protein's structure is often determined in the isolated context of a crystal and cannot take into consideration supramolecular interactions. Structure-based AAS prediction methods tend to predict positions on the surface of the protein as neutral. Although substitutions at sites buried in protein structure are more likely to be damaging than surface residues, substitutions at sites that appear to be solvent-accessible in the crystal structure may also be important for function. These sites might be involved in intermolecular interactions with proteins that are absent from the 3D structure of the single protein. For example, the β-globin E6V substitution causes sickle-cell anemia. The substitution occurs on the surface of the protein and leads to formation of hemoglobin aggregates that underlie the sickling phenotype. E6V is incorrectly predicted to be benign by a structure-/sequence-based AAS prediction method, Polymorphism Phenotyping (PolyPhen) (71), whereas it is correctly predicted by a sequence homology-based AAS prediction method, Sorting Intolerant From Tolerant (SIFT). One possible way to compensate for such misprediction is to identify surface pockets or depressed regions in the protein's structure, which can be inferred to be potential functional binding regions, as done in the AAS prediction method topoSNP (64, 65).

Some AAS methods use structural and functional annotation from the Swiss-Prot database in addition to structure and sequence modeling (17, 18, 68, 69, 77). The functional annotation is used to identify the residues that are part of a binding site, active site, or disulfide bond. It is presumed that changes at these types of sites would have major effects on protein function. Hence, substitutions occurring at these positions are predicted to affect function. One would think that use of such annotation would improve prediction, but a recent study shows that use of Swiss-Prot functional annotation decreases the overall prediction accuracy (83). Although use of Swiss-Prot annotation reduces the false negative rate by 1.6%, it increases the false positive rate by 2.1%. Therefore, caution is warranted when interpreting results, based solely on functional annotation, in which an AAS is predicted to be damaging.

When the structure of the query protein is not available, AAS prediction methods model the query protein's structure based on a homolog. If the homolog is too distantly related, prediction accuracy can suffer. Chasman & Adams (9) obtained their best prediction results when the query protein was at least 60% identical to its homolog. Yue & Moult (82) found that prediction is best when the sequence identity between the query protein and the homologous protein is greater than 40%. The false positive rate for their AAS prediction method increased by 11% when structures with less than 40% sequence identity were used (82).

Caveats for Using Sequence in Prediction

The first step for all sequence-based AAS prediction methods is to choose homologous sequences, whether manually or automatically. Because the amino acids appearing in the aligned sequences form the basis of the

Polymorphism Phenotyping (PolyPhen): a popular structure-/sequence-based amino acid substitution prediction method available on the Internet (**http://www.bork.embl-heidelberg.de/PolyPhen/**)

Sorting Intolerant From Tolerant (SIFT): a popular sequence-based amino acid substitution prediction method available on the Internet (**http://blocks.fhcrc.org/sift/SIFT.html**)

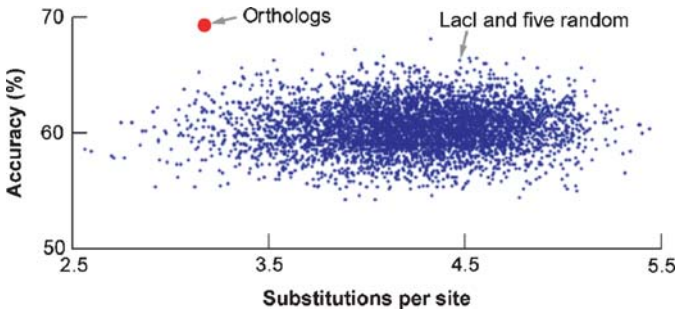

Figure 3

For some amino acid substitution (AAS) prediction methods, researchers can submit a multiple sequence alignment for their protein of interest; using orthologs in the multiple sequence alignment instead of paralogs gives better performance. Blue points represent prediction accuracy based on an alignment of the input protein and five sequences randomly chosen from a group of orthologs and paralogs. The red point is the prediction accuracy based on a multiple sequence alignment of the input protein sequence and five orthologs. Figure from Reference 67. Published with permission from *Genome Research*, Volume 15, copyright by Cold Spring Harbor Laboratory Press.

scoring and the prediction, the sequences and their alignment are extremely important, and users can take an active role in this prediction step, which determines the quality of their predictions. The optimal set of sequences is distantly related orthologs. Using orthologs instead of paralogs can improve performance by 8%, even when there are few sequences (67) (**Figure 3**).

Users should be cautious even with proteins that are judged to be orthologous based on phylogeny. Orthologous genes in different species are derived from a common ancestor, but they may not necessarily have the same function. If function has changed, then amino acids that are important for the function of one protein may not necessarily be important for the function of the ortholog, and hence may have changed without any selection pressure. For example, 2% of disease-causing mutations in human genes are identical to the sequences of their respective mouse orthologs, suggesting that even though these positions have huge phenotypic effects on human health, they have different roles or are no longer important in mice. If the orthologs in alignment have slightly different functions, then the positions that differentiate function among orthologs may be incorrectly predicted.

Although functional differences between orthologs can result in misprediction by AAS methods, prior knowledge that orthologs have different functions can be used to identify which amino acid(s) caused the functional change. For example, AlcR is a transcriptional regulator whose activation depends on the presence of an inducer in one *Bordetella* species but not in another species (5). Although the *AlcR* genes are clearly orthologs in the genus, they behave differently in different species. To find the amino acid change underlying this behavioral difference, the authors of this study used the AAS prediction method, SIFT, which predicted that a seemingly conservative change (S103T) should not be tolerated. The authors deduced that this residue was likely responsible for inducer dependence, which they subsequently proved by showing that S103T eliminated inducer dependence.

Most AAS prediction methods do not take DNA sequence into account. As a result, they can miss changes that alter splice sites or changes in regions under positive selection. By ignoring DNA sequence differences, these methods might incorrectly predict substitutions at sites under positive selection to be neutral because of the many amino acids present at that position. However, the AAS prediction method of Fleming et al. (19) uses the DNA sequences of homologous genes in addition to protein sequences to find sites under positive selection. The nonsynonymous-to-synonymous substitution rate ratio is allowed to vary at different amino acid sites and positions with high ratios are posited to be under positive selection. It is assumed that amino acid changes at these positions affect protein function. Thus, knowledge of positively selected sites can lower the false negative error.

It is often assumed that the protein sequence derived from the reference sequence genome is functional and the "disease"

mutation reduces protein function. But in some instances, the amino acid corresponding to the more common allele reduces protein function and a mutation causes gain-of-function. Thomas et al. (73) noticed this phenomenon when they applied their AAS prediction method PANTHER PSEC to mutations likely to cause disease. They found that the reference amino acid was predicted to damage protein function whereas the disease mutation was predicted to be functionally neutral for a small percentage of the cases (0.1%). This type of prediction reveals two things. First, it suggests that the human protein sequence has reduced function compared with orthologous proteins in other organisms. Second, it predicts that the "disease" mutation returns protein function to a level similar to that in orthologous proteins. Thus the mutation can be thought to have a gain-of-function effect in humans. Characterizing this mutation in model organisms may be inappropriate because the common human allele has reduced protein activity compared with that in model organisms.

Both structure and sequence are useful for prediction. Having a structure appears to provide prediction performance that is equivalent to having four homologous sequences, with information from structure and sequence complementing each other (59). Finally, it is important to understand that these are predictions only. They are meant to guide future experiments and not to be used directly in a clinical setting (71).

USEFULNESS OF AMINO ACID SUBSTITUTION PREDICTION METHODS IN HUMAN VARIATION AND DISEASE

Nonsynonymous SNPs

There are an estimated 67,000–200,000 common nsSNPs in the human population (8, 23, 35), with each person expected to be heterozygous for 24,000–40,000 nsSNPs (8). Given this large number of nsSNPs and the observation that single AASs can have a large effect on an organism or species, it is of great interest to identify which nsSNPs affect protein function and, consequently, may affect human health.

nsSNPs in the human population are observed less frequently than expected from the overall mutation rate, which is evidence that they are under strong purifying selection (8, 23, 63). Specifically, if a random mutation were to occur in a coding region, it should lead to an amino acid change two thirds of the time, but nsSNPs comprise only half of the observed coding SNPs in the human genome (8). Furthermore, nsSNPs that cause a nonconservative amino acid change in the corresponding protein (for example, hydrophobic amino acid to a charged amino acid) survive at approximately half the rate of conservative nsSNPs (for example, a hydrophobic amino acid changed to another hydrophobic amino acid). These data strongly support the notion that AASs play an important role in human health. By providing information about which substitutions are selected against, AAS prediction methods can help identify which nsSNPs may be involved in disease.

Putative nsSNPs are catalogued in the dbSNP database maintained by NCBI, which currently contains >50,000 nsSNPs (dbSNP build 124) (60). Of these nsSNPs, 25–30% are predicted to reduce protein function by most AAS prediction methods (46, 69, 83). Such nsSNPs have a lower minor allele frequency distribution than those that are predicted to be functionally neutral (33–35, 69, 78). This suggests that damaging nsSNPs are being actively selected against and confirms that AAS prediction methods can successfully identify putatively damaging nsSNPs that play an important role in health. In one study, the AAS prediction method SIFT was applied to nsSNPs found in genes implicated in DNA repair, cell cycle arrest, apoptosis, and detoxification (78). nsSNPs with low minor allele frequency (<6%) were predicted by SIFT to be damaging to the protein twice as often as

common nsSNPs (>10%). In a second study, less than 2% of common nsSNPs in environmental response genes were predicted to be damaging by AAS prediction methods PolyPhen and SIFT (34, 35). In a third study, SIFT was applied to nsSNPs found in membrane transporter genes (34). Of the nsSNPs with minor allele frequency <0.01, between 0.01 and 0.10, and >0.1, 40%, 13%, and 5% of nsSNPs were predicted to be damaging, respectively. The authors found that general-purpose AAS scoring matrices, such as BLOSUM62, could not distinguish nsSNPs by minor allele frequency. Thus, application of AAS prediction methods to large nsSNP data sets has confirmed that putatively damaging nsSNPs are selected against and are likely to have an impact on their respective proteins.

Experimental studies of individual proteins have also confirmed the accuracy of AAS prediction methods. In one study, Brooks-Wilson et al. (6) studied mutations in E-cadherin that cause hereditary diffuse gastric cancer. They used SIFT to predict that three missense mutations found in families with diffuse gastric cancer would be damaging to E-cadherin function, and all three were confirmed to be damaging using in vitro assays. In a second study, Zhang et al. (84) examined the variants in PEPT1, a protein involved in transporting drugs across the cell membrane. nsSNPs found in the gene for PEPT1 were tested in vitro. The single SNP that reduced transport capacity was predicted by SIFT to affect protein function. This polymorphism may be important in drug delivery for pharmacogenetics. In a third study, SIFT was used to predict categories of cancer risk. Mutations in the gene encoding melanocyte stimulating hormone receptor (MSHR) increase the risk of skin cancer. Kanetsky et al. (30) identified risk variants in the gene for MSHR from either published literature or using SIFT predictions. The ability to assign an individual to a risk category was found to be similar when using either published literature or SIFT prediction.

In summary, AAS prediction methods have proven useful for identifying damaging nsSNPs involved in human disease. Experimentally characterizing an AAS can be expensive and time-consuming, and AAS prediction methods provide a valuable resource to substantially reduce the effort.

Application to Mendelian Disease

AAS prediction methods have succeeded in distinguishing nonsynonymous changes that cause simple Mendelian diseases from neutral nsSNPs that do not. AAS prediction methods are trained and tested on AASs that were identified in disease genes in afflicted individuals from databases such as OMIM (24), HGMD (62), and Swiss-Prot (1). It is assumed that these substitutions found in patients affect protein function and cause disease as a result. Currently, most of the diseases represented by the genes in these databases segregate in a Mendelian manner, which suggests that they are caused by single deleterious lesions. Most AAS prediction methods predict that 70–90% of the AASs catalogued in these disease databases are damaging, whereas only 10–20% of variants in neutral data sets are predicted to be damaging (**Table 1**). This demonstrates that AAS prediction methods can distinguish between AASs that cause Mendelian disease and neutral AASs. In this way, AAS prediction methods can help narrow down candidate nsSNPs to identify the causative lesion within a large genomic region implicated in disease by linkage studies.

Although most simple Mendelian diseases remain rare because of purifying selection, some become relatively common in populations because of overdominant selection. Overdominant selection occurs when the heteroyzgote carrier has higher fitness than both the mutant and normal homozygotes. The E6V substitution in β-globin mentioned above is common in certain populations because heterozygous carriers are more resistant to malaria than normal homozygotes,

whereas individuals homozygous for the rare allele have sickle-cell anemia. In this example, the AAS affects protein function but is maintained in the population because reduced activity correlates with an advantageous effect. Another well-known example is overdominance associated with methylenetetrahydrofolate reductase (MTHFR) alleles. Variants that reduce MTHFR activity can cause mental retardation and cardiovascular disease in carriers. Reduced MTHFR activity is thought to have been beneficial to an individual's overall fitness during recent human evolution, so MTHFR variants damaging to protein function have became common in human populations. Because overdominant nsSNPs can severely affect protein function, they may be detected by AAS prediction methods, and both E6V in β-globin and overdominant MTHFR missense alleles are predicted to be damaging by SIFT (Reference 46 and data not shown).

Application to Complex Diseases

Although AAS prediction methods can identify nsSNPs involved in Mendelian diseases, their usefulness in studying complex diseases, such as hypertension, diabetes, and heart disease, is still being explored. Complex genetic diseases are those that cannot be mapped to single loci, which might indicate an interaction between multiple loci or an interaction with the environment, or both. Two general models have been proposed to explain the nature of genetic variation underlying complex disease (4, 51, 52, 85). The common disease-common variant model predicts that the variants of a particular locus that contribute to disease are few but common in the population and that complex disease results from interactions between variants of many different genes. The common disease-rare variant model predicts that there are many etiological variants at a locus, and each variant is present at a low allele frequency in the human population. In this section, we discuss how useful AAS prediction methods might be for detecting variants involved in complex disease according to the two models.

Common disease-common variant. According to the common disease-common variant hypothesis, variants that are common in the population, with minor allele frequency greater than 5% or 1%, contribute to disease. If nsSNPs involved in complex disease can be categorized according to the rules of structure and sequence conservation that AAS prediction methods have been trained on, then current AAS prediction methods will be useful in identifying these common etiological variants. A list of common nsSNPs that have been associated with complex disease has been compiled (28, 37). When the AAS prediction method PANTHER PSEC was applied to this set of disease-causing nsSNPs, the score distribution was different from the distribution for Mendelian disease mutations (74). More importantly, the score distribution of these disease-associated nsSNPs is indistinguishable from the distribution for normal human variation (**Figure 4**). The authors conclude that nsSNPs involved in complex diseases caused by common variants do not occur at highly conserved positions, and thus cannot be detected by current AAS prediction methods. The data set of nsSNPs known to be involved in common disease is relatively small compared with the set of mutations known to be involved in Mendelian disease. As more causative common variants implicated in complex disease are discovered, it will be interesting to see if future studies yield similar results. If so, new rules will need to be formulated to predict the common polymorphisms involved in complex disease.

AAS prediction methods can still be useful in identifying the etiological variants in haplotypes. Haplotypes are particular combinations of alleles that are observed in a population. When a haplotype is associated with affected individuals, the variants belonging to the haplotype are all candidates for being the causative mutation. If a haplotype contains a set of missense alleles, one can use AAS

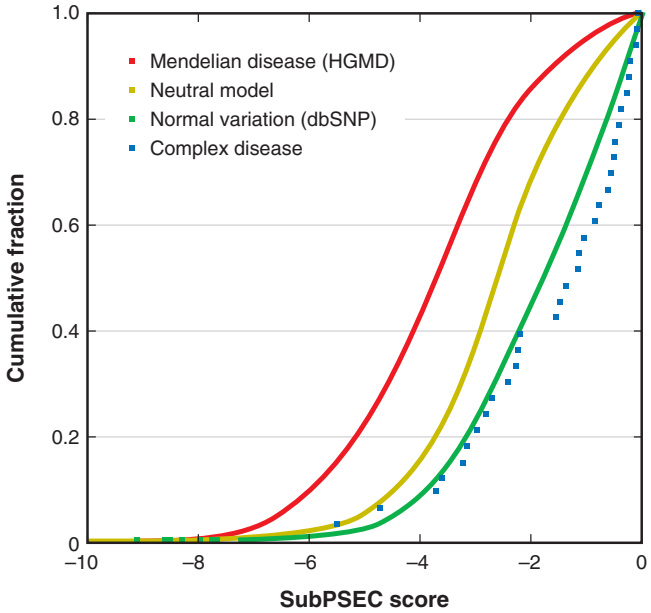

Figure 4

Amino acid substitution (AAS) prediction methods may not be able to identify disease variants in the common disease/common variant model. Cumulative distributions of scores from the AAS prediction method PANTHER subPSEC are shown. A subPSEC score of 0 is predicted to be functionally neutral and very negative scores are predicted to be damaging to protein function. Distributions are shown for mutations involved in Mendelian disease (*red*), common variants associated with complex disease (*blue squares*), neutral and "normal" human variation (*yellow* and *green*, respectively). Figure from Reference 74, Volumes 90–102. Copyright © 1993–2005. Natl. Acad. Sci. USA. All rights reserved.

prediction methods to prioritize which nsSNP may be the etiological variant (84).

Common disease-rare variant. According to the common disease/rare variant hypothesis, low-frequency variants with strong effects on a locus can contribute to disease (52). That is, some "complex" diseases could actually be simple Mendelian diseases, but are caused by different allelic variants in different individuals. Identifying these rare causative variants requires sequencing genes in many individuals. This process would uncover a large number of missense variants but only a few may contribute to disease. Below we discuss two studies that used AAS prediction methods to distinguish the causative missense variants from neutral variants.

A study on the plasma levels of HDL cholesterol (HDL-C) demonstrates the use of AAS prediction methods to detect rare deleterious alleles that contribute to common disease (10). Low levels of HDL-C are a major risk factor for coronary atherosclerosis. Three candidate genes involved in Mendelian forms of low HDL-C levels were sequenced in apparently normal individuals and a larger number of rare nsSNPs were found in people with low levels of HDL-C compared with those with high levels of HDL-C. The authors then applied the AAS prediction method PolyPhen on the nsSNPs. The fraction of nsSNPs predicted to be damaging in individuals with low levels of HDL-C is highly significant for two of the genes: $p = 9 \times 10^{-13}$ for *ABCA1* and $p = 0.0003$ for *LCAT* (**Table 2**). In the third gene, *APOA1*, only one nsSNP was detected, and although it was predicted to be damaging, this is not statistically significant ($p = 0.09$). If technological advances permit us to inexpensively sequence the coding regions of many individuals, one could possibly identify the genes and variants involved in disease by the increased number of nsSNPs and the significantly high proportion of nsSNPs predicted to be damaging in the affected population. Even with a Bonferroni correction taking into account the approximately 30,000 genes in the human genome, the result of *ABCA1* is still significant at $p = 10^{-8}$. The result of LCAT is no longer significant at the genome-wide level. This could be compensated for by detecting more variants, which requires sequencing, or improved performance of AAS prediction methods to reduce the false positive error.

Another study on the role of mitochondrial mutations in Parkinson's disease again shows that applying AAS prediction methods to rare variants can identify genes of interest (61). Seven genes in the mitochondrial genomes from normal individuals and patients with Parkinson's disease were sequenced. The authors' AAS prediction method could

Table 2 A significantly high fraction of rare variants in individuals at risk for coronary atherosclerosis are predicted to be damaging by the amino acid substitution (AAS) prediction method PolyPhen (10)*

Gene	Number of nsSNPs detected	Number predicted to be possibly or probably damaging by PolyPhen	Cumlative p-value, assuming 0.09 false positive rate
Low levels of HDL-C, at risk for coronary atherosclerosis			
ABCA1	25	17	$9*10^{-13}$
APOA1	1	1	0.09
LCAT	5	4	0.0003
High levels of HDL-C			
ABCA1	4	2	0.04
APOA1	0	0	NA
LCAT	1	0	0.91

*The list of nsSNPs and their predictions from Cohen et al. (10) was compiled. For clarity, the two populations studied (256 Dallas County residents and 263 Canadians) were combined and duplicated SNPs were counted only once. The p-value was calculated from a binomial distribution, with the probability that a nonsynonymous (nsSNP) would be predicted to be damaging set to 0.09, Polyphen's false positive rate on neutral substitutions (69).

distinguish patients from controls with 100% accuracy and two genes were identified to have more putative damaging missense variants compared with controls. In this case, all genes had a similar proportion of missense variants. Only by applying their AAS prediction method were the authors able to identify two genes that may be involved in Parkinson's disease. Because mutations in mitochondrial DNA become more abundant with increasing age, the role of rare missense variants may play an important role in late-onset diseases such as Parkinson's disease and Alzheimer's disease. Because there already exists technology for resequencing mitochondrial genomes (38), one may be able to study the common disease/rare variant model in diseases suspected to involve the mitochondria.

If common diseases cannot be explained by common variants, then there will be a strong incentive to sequence coding regions to test the common disease/rare variant hypothesis. Preliminary studies show that AAS prediction methods play an important role in predicting which rare missense variants are deleterious. Moreover, as technologies for discovering rare variants advance, AAS prediction methods will become increasingly important for identifying disease genes in genome-wide studies.

OTHER APPLICATIONS OF AAS PREDICTION METHODS

Interspecies Comparisons

AAS prediction methods are often applied to polymorphisms within a species, but they can also be used on the fixed substitutions across species. Because AAS prediction methods predict which amino acids are damaging to protein function, they can identify which genes are under relaxed or neutral selection in a particular species.

Interspecies comparisons show that domesticated species have a higher number of putative damaging SNPs than the wild species. The Chicken Polymorphism Consortium sequenced three domestic breeds of chicken and compared these data to those obtained from a wild line (79). For the nonsynonymous substitutions fixed between the domestic and wild lines, the alleles found in domestic breeds were more than twice as likely to be predicted as damaging by the AAS prediction method SIFT compared with those in the wild line. This result suggests that domestic breeds are under relaxed selection, perhaps because they live in a less harsh environment, which has allowed for damaging substitutions to become fixed in the domesticated species. In another interspecies

comparison, the missense substitutions between two species of domesticated rice were studied and a higher density of predicted damaging substitutions occurred in regions under positive selection (G.S.K. Wong, personal communication). Because gene loss may sometimes be adaptive (48), the result suggested reduced function of some genes in rice could be an adaptive response to domestication.

These studies suggest that AAS prediction methods can provide insights into phenotypic differences observed between species. Because of false positive error, it would be difficult to study this on an individual gene basis. However, by grouping genes within protein families or pathways, it may be possible to identify pathways that have undergone relaxed selection in certain species.

Large-Scale Mutagenesis

AAS prediction methods can be applied to large-scale, reverse genetics projects, in which mutations are introduced randomly in the genome of an experimental organism, and to random mutagenesis projects, when a gene of interest is saturated with mutations (4, 22, 26, 36, 49, 51, 52, 54, 80, 85). When there are many missense mutations in the gene(s) of interest, assaying all missense mutations can be expensive and time-consuming. AAS prediction methods can be used to prioritize missense mutations that are most likely to affect protein function and alter phenotype. TILLING is one example of a large-scale reverse genetic strategy that uses an AAS prediction method (SIFT) to prioritize which missense mutations are likely to reveal a phenotype (26). TILLING has been applied to a wide range of organisms: *Arabidopsis*, zebrafish, maize, *Drosophila*, and Lotus. Because AAS prediction methods are automated and general, they can be widely applied to help researchers prioritize which AAS to characterize in genes of interest.

CONCLUSIONS

The presence of many AAS prediction methods and their broad use underscores their importance. Prediction accuracy has gradually improved, but few head-to-head comparisons exist (29, 35, 71, 81). Moreover, as the number of servers providing AAS prediction increases, it will become increasingly difficult for investigators to interpret the predictions.

These problems are similar to those faced by the protein structure prediction community 10 years ago. Critical Assessment of Techniques for Protein Structure Prediction (CASP) (43) was motivated by the need to fairly assess structure prediction programs in order to advance structure prediction methods; we propose a similar solution for AAS prediction methods.

Every two years, CASP releases the sequences of proteins for which structure is known, but not yet available to the public. Researchers return predictions which are then compared with known structure, and the efficacy of each prediction method is assessed. Because investigators in the CASP program work on the same proteins during a specified period of time, the CASP program offers a valuable way to summarize which methods improve structure prediction and this ultimately advances the field. As automated structure prediction methods have become increasingly successful, there is now a server competition that runs continuously as structures are released, with results assessed automatically (57).

A similar server competition could be implemented for AAS prediction. Once a disease-associated variant is mapped, the responsible nsSNP and benign control nsSNPs could be sent to participating servers and the results immediately evaluated. Obtaining suitable data sets is difficult because each data set has its own advantages and disadvantages. For example, some of the AASs obtained from patients in the OMIM database may not necessarily be the causative variant, which will artifactually inflate the false

negative error. But if one method has a higher accuracy on both neutral and damaging data sets, then, despite the inaccuracies of the test sets, the prediction method can be assessed to be better. Moreover, as genome-wide mutagenesis projects generate phenotypic data (26), more accurate data can be obtained and tested.

In addition to accuracy, AAS prediction methods can be evaluated based on coverage. Coverage depends on the source and version of structure and sequence databases used. The number of sequences and structures increase every year so coverage is expected to increase simply because more information is available.

Most methods today offer confidence scores or an estimate of the degree to which a substitution is damaging. One could also test how well an AAS prediction method's score correlates with phenotypic severity; the MAPP method and the method by Mooney et al. (42, 67) have already demonstrated promising results as their scores correlate with phenotypic severity. AAS prediction methods could also be assessed according to their ability to predict how the substitution affects the protein: whether it increases or decreases function, affects structure, or which domain or function is affected for multifunctional enzymes.

Finally, the main benefit of such a competition is that investigators would have access to unified prediction. Currently, CASP manages a metaserver that collects predictions from other servers to obtain a consensus structure. A metaserver for AAS prediction could do the same, so that researchers studying protein variation could easily obtain a consensus prediction, with the best accuracy available.

The progress that has been made over the past few years with AAS prediction methods is promising: methodology has improved and applications have proliferated. AAS prediction methods have proven successful for Mendelian traits and may eventually play an important role in identifying complex disease variants. Because AASs are a source of fundamental changes between and within species, AAS prediction methods will continue to be of major importance in the future. These methods, in conjunction with those that predict gene regulatory and splicing variants (12, 13, 15, 20, 66, 76), will guide us to a better understanding of functional diversity in genomes.

SUMMARY POINTS

1. Approximately half of the known disease-causing mutations result from amino acid substitutions (AASs).
2. AAS prediction methods can successfully distinguish between AASs that cause Mendelian disease and functionally neutral AASs.
3. Automated AAS prediction methods have been applied on a genome-wide scale.
4. Nonsynonymous SNPs (nsSNPs) predicted to be damaging tend to have low minor allele frequencies, which indicates purifying selection and validates AAS prediction methods.
5. AAS prediction methods have been applied to individual genes and predictions have been experimentally confirmed.
6. Currently, AAS prediction methods cannot distinguish between common variants involved in common disease and normal variation.
7. AAS prediction methods can identify rare variants involved in common disease. When putative damaging rare variants are classified by gene, disease genes can be identified.

8. AAS prediction methods can be used on a wide range of problems, such as analysis of interspecies differences and large-scale mutagenesis projects.

FUTURE DIRECTIONS

1. Because there are many new and improved amino acid substitution (AAS) prediction methods with complementary strengths, better accuracy should be possible by combining prediction methods. A competition similar to the Critical Assessment of Techniques for Protein Structure Prediction (CASP) would help to evaluate progress and identify strengths and weaknesses of prediction algorithms.
2. While the genetics community is actively engaged in discovering genetic variants involved in complex disease, AAS prediction methods will need to be continually reassessed and possibly redesigned for optimal prediction of complex disease-causing variants.
3. Although AASs currently account for a large proportion of the genetic variation contributing to human disease, new bioinformatics methods that evaluate gene regulatory and splicing variants will broaden our understanding of functional variation.

ACKNOWLEDGMENTS

We thank Jorja Henikoff and Sarah Shaw Murray for manuscript comments.

LITERATURE CITED

1. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, et al. 2005. The Universal Protein Resource (UniProt). *Nucleic Acids Res.* 33:D154–59
2. Deleted in proof
3. Beutler E, Vulliamy TJ. 2002. Hematologically important mutations: glucose-6-phosphate dehydrogenase. *Blood Cells Mol. Dis.* 28:93–103
4. Botstein D, Risch N. 2003. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nat. Genet.* 33(Suppl.):228–37
5. Brickman TJ, Armstrong SK. 2002. Bordetella interspecies allelic variation in AlcR inducer requirements: identification of a critical determinant of AlcR inducer responsiveness and construction of an alcR(Con) mutant allele. *J. Bacteriol.* 184:1530–39
6. Brooks-Wilson AR, Kaurah P, Suriano G, Leach S, Senz J, et al. 2004. Germline E-cadherin mutations in hereditary diffuse gastric cancer: assessment of 42 new families and review of genetic screening criteria. *J. Med. Genet.* 41:508–17
7. Cai Z, Tsung EF, Marinescu VD, Ramoni MF, Riva A, Kohane IS. 2004. Bayesian approach to discovering pathogenic SNPs in conserved protein domains. *Hum. Mutat.* 24:178–84
8. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, et al. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* 22:231–38
9. **Chasman D, Adams RM. 2001. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.* 307:683–706**

9. One of the first papers to establish structure-/sequence-based rules to predict amino substitutions that affect protein function.

10. Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. 2004. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* 305:869–72
11. Collins FS, Guyer MS, Charkravarti A. 1997. Variations on a theme: cataloging human DNA sequence variation. *Science* 278:1580–81
12. Conde L, Vaquerizas JM, Ferrer-Costa C, de la Cruz X, Orozco M, Dopazo J. 2005. PupasView: a visual tool for selecting suitable SNPs, with putative pathological effect in genes, for genotyping purposes. *Nucleic Acids Res.* 33:W501–5
13. Conde L, Vaquerizas JM, Santoyo J, Al-Shahrour F, Ruiz-Llorente S, et al. 2004. PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level. *Nucleic Acids Res.* 32:W242–48
14. del Sol Mesa A, Pazos F, Valencia A. 2003. Automatic methods for predicting functionally important residues. *J. Mol. Biol.* 326:1289–302
15. Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004. Single nucleotide polymorphism-based validation of exonic splicing enhancers. *PLoS Biol.* 2:E268
16. Ferrer-Costa C, Gelpi JL, Zamakola L, Parraga I, de la Cruz X, Orozco M. 2005. PMUT: a web-based tool for the annotation of pathological mutations on proteins. *Bioinformatics* 21:3176–78
17. Ferrer-Costa C, Orozco M, de la Cruz X. 2002. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *J. Mol. Biol.* 315:771–86
18. Ferrer-Costa C, Orozco M, de la Cruz X. 2004. Sequence-based prediction of pathological mutations. *Proteins* 57:811–19
19. Fleming MA, Potter JD, Ramirez CJ, Ostrander GK, Ostrander EA. 2003. Understanding missense mutations in the *BRCA1* gene: an evolutionary approach. *Proc. Natl. Acad. Sci. USA* 100:1151–56
20. Freimuth RR, Stormo GD, McLeod HL. 2005. PolyMAPr: programs for polymorphism database mining, annotation, and functional analysis. *Hum. Mutat.* 25:110–17
21. Gottlieb B, Lehvaslaiho H, Beitel LK, Lumbroso R, Pinsky L, Trifiro M. 1998. The Androgen Receptor Gene Mutations Database. *Nucleic Acids Res.* 26:234–38
22. Hainaut P, Hernandez T, Robinson A, Rodriguez-Tome P, Flores T, et al. 1998. IARC Database of p53 gene mutations in human tumors and cell lines: updated compilation, revised formats and new visualisation tools. *Nucleic Acids Res.* 26:205–13
23. Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, et al. 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* 22:239–47
24. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 33:D514–17
25. Hardison RC, Chui DH, Giardine B, Riemer C, Patrinos GP, et al. 2002. HbVar: a relational database of human hemoglobin variants and thalassemia mutations at the globin gene server. *Hum. Mutat.* 19:225–33
26. Henikoff S, Comai L. 2003. Single-nucleotide mutations for plant functional genomics. *Annu. Rev. Plant Biol.* 54:375–401
27. Herrgard S, Cammer SA, Hoffman BT, Knutson S, Gallina M, et al. 2003. Prediction of deleterious functional effects of amino acid mutations using a library of structure-based function descriptors. *Proteins* 53:806–16
28. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. 2002. A comprehensive review of genetic association studies. *Genet. Med.* 4:45–61

10. An AAS prediction method predicts a significant number of the rare nonsynonymous variants found in individuals at risk for coronary atherosclerosis to be damaging. This paper shows that AAS prediction methods can distinguish the rare variants likely to cause common disease from neutral variants.

29. Johnson MM, Houck J, Chen C. 2005. Screening for deleterious nonsynonymous single-nucleotide polymorphisms in genes involved in steroid hormone metabolism and response. *Cancer Epidemiol. Biomarkers Prev*. 14:1326–29
30. Kanetsky PA, Ge F, Najarian D, Swoyer J, Panossian S, et al. 2004. Assessment of polymorphic variants in the melanocortin-1 receptor gene with cutaneous pigmentation using an evolutionary approach. *Cancer Epidemiol. Biomarkers Prev.* 13:808–19
31. Krishnan VG, Westhead DR. 2003. A comparative study of machine-learning methods to predict the effects of single nucleotide polymorphisms on protein function. *Bioinformatics* 19:2199–209
32. Kwok CJ, Martin AC, Au SW, Lam VM. 2002. G6PDdb, an integrated database of glucose-6-phosphate dehydrogenase (G6PD) mutations. *Hum. Mutat*. 19:217–24
33. Lau AY, Chasman DI. 2004. Functional classification of proteins and protein variants. *Proc. Natl. Acad. Sci. USA* 101:6576–81
34. Leabman MK, Huang CC, DeYoung J, Carlson EJ, Taylor TR, et al. 2003. Natural variation in human membrane transporter genes reveals evolutionary and functional constraints. *Proc. Natl. Acad. Sci.* 100:5896–901
35. Livingston RJ, von Niederhausern A, Jegga AG, Crawford DC, Carlson CS, et al. 2004. Pattern of sequence variation across 213 environmental response genes. *Genome Res.* 14:1821–31
36. Loeb DD, Swanstrom R, Everitt L, Manchester M, Stamper SE, Hutchison CA 3rd. 1989. Complete mutagenesis of the HIV-1 protease. *Nature* 340:397–400
37. Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. 2003. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet*. 33:177–82
38. Maitra A, Cohen Y, Gillespie SE, Mambo E, Fukushima N, et al. 2004. The Human MitoChip: a high-throughput sequencing microarray for mitochondrial mutation detection. *Genome Res.* 14:812–19
39. Miller MP, Kumar S. 2001. Understanding human disease mutations through the use of interspecific genetic variation. *Hum. Mol. Genet.* 10:2319–28
40. Mooney S. 2005. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. *Brief. Bioinform*. 6:44–56
41. Mooney SD, Altman RB. 2003. MutDB: annotating human variation with functionally relevant data. *Bioinformatics* 19:1858–60
42. Mooney SD, Klein TE, Altman RB, Trifiro MA, Gottlieb B. 2003. A functional analysis of disease-associated mutations in the androgen receptor gene. *Nucl. Acids Res.* 31:e42
43. Moult J. 2005. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol*. 15:285–89
44. Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–62
45. Ng PC, Henikoff S. 2001. Predicting deleterious amino acid substitutions. *Genome Res.* 11:863–74
46. Ng PC, Henikoff S. 2002. Accounting for human polymorphisms predicted to affect protein function. *Genome Res.* 12:436–46
47. Ng PC, Henikoff S. 2003. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res*. 31:3812–14
48. Olson MV. 1999. When less is more: gene loss as an engine of evolutionary change. *Am. J. Hum. Genet*. 64:18–23
49. Pace HC, Kercher MA, Lu P, Markiewicz P, Miller JH, et al. 1997. Lac repressor genetic map in real space. *Trends Biochem. Sci.* 22:334–39

39. An early paper that analyzed seven disease genes and recognized that disease mutations appear at positions conserved throughout evolution, indicating that sequence-based prediction was possible.

45. The first AAS prediction method implemented on a Web server.

46. A sequence-based AAS prediction method that successfully predicted disease-causing mutations. Because this method does not require protein structure, predictions were obtained for a majority of the nsSNPs in dbSNP.

50. Patrinos GP, Giardine B, Riemer C, Miller W, Chui DH, et al. 2004. Improvements in the HbVar database of human hemoglobin variants and thalassemia mutations for population and sequence variation studies. *Nucleic Acids Res.* 32:D537–41
51. Ponting CP, Goodstadt L. 2005. Statistical genetics: usual suspects in complex disease. *Eur. J. Hum. Genet*. 13:269–70
52. Pritchard JK. 2001. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet*. 69:124–37
53. Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res*. 30:3894–900
54. Rennell D, Bouvier SE, Hardy LW, Poteete AR. 1991. Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol*. 222:67–88
55. Rhee SY, Gonzales MJ, Kantor R, Betts BJ, Ravela J, Shafer RW. 2003. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res*. 31:298–303
56. Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. *Science* 273:1516–17
57. Rychlewski L, Fischer D. 2005. LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction. *Protein Sci*. 14:240–45
58. Santibanez Koref MF, Gangeswaran R, Santibanez Koref IP, Shanahan N, Hancock JM. 2003. A phylogenetic approach to assessing the significance of missense mutations in disease genes. *Hum. Mutat*. 22:51–58
59. Saunders CT, Baker D. 2002. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J. Mol. Biol*. 322:891–901
60. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, et al. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 29:308–11
61. Smigrodzki R, Parks J, Parker WD. 2004. High frequency of mitochondrial complex I mutations in Parkinson's disease and aging. *Neurobiol. Aging* 25:1273–81
62. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, et al. 2003. Human Gene Mutation Database (HGMD): 2003 update. *Hum. Mutat.* 21:577–81
63. Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, et al. 2001. Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 293:489–93
64. Stitziel NO, Binkowski TA, Tseng YY, Kasif S, Liang J. 2004. topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. *Nucleic Acids Res.* 32:D520–22
65. Stitziel NO, Tseng YY, Pervouchine D, Goddeau D, Kasif S, Liang J. 2003. Structural location of disease-associated single-nucleotide polymorphisms. *J. Mol. Biol*. 327:1021–30
66. Stone EA, Cooper GM, Sidow A. 2005. Trade-offs in detecting evolutionarily constrained sequence by comparative genomics. *Annu. Rev. Genomics Hum. Genet*. 6:143–64
67. Stone EA, Sidow A. 2005. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Res*. 15:978–86
68. Sunyaev S, Ramensky V, Bork P. 2000. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet*. 16:198–200
69. Sunyaev S, Ramensky V, Koch I, Lathe W 3rd, Kondrashov AS, Bork P. 2001. Prediction of deleterious human alleles. *Hum. Mol. Genet*. 10:591–97
70. Szabo C, Masiello A, Ryan JF, Brody LC. 2000. The breast cancer information core: database design, structure, and scope. *Hum. Mutat*. 16:123–31
71. Tchernitchko D, Goossens M, Wajcman H. 2004. In silico prediction of the deleterious effect of a mutation: proceed with caution in clinical genetics. *Clin. Chem*. 50:1974–78

59. Carefully examines the contribution of structure and sequence to prediction.

61. Shows that AAS prediction methods can predict the rare variants involved in common disease and the aggregation of these results can be used to identify the genes involved in disease.

68. One of the first papers to establish structure-/sequence-based rules to distinguish disease mutations from neutral amino acid substitutions.

72. Terp BN, Cooper DN, Christensen IT, Jorgensen FS, Bross P, et al. 2002. Assessing the relative importance of the biophysical properties of amino acid substitutions associated with human genetic disease. *Hum. Mutat.* 20:98–109
73. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, et al. 2003. PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* 13:2129–41
74. Thomas PD, Kejariwal A. 2004. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. *Proc. Natl. Acad. Sci. USA* 101:15398–403
75. Vitkup D, Sander C, Church GM. 2003. The amino-acid mutational spectrum of human genetic disease. *Genome Biol.* 4:R72
76. Wang X, Tomso DJ, Liu X, Bell DA. 2005. Single nucleotide polymorphism in transcriptional regulatory regions and expression of environmentally responsive genes. *Toxicol. Appl. Pharmacol.* 207:84–90
77. Wang Z, Moult J. 2001. SNPs, protein structure, and disease. *Hum. Mutat.* 17:263–70
78. Wong GK-S, Yang Z, Passey DA, Kibukawa M, Paddock M, et al. 2003. A population threshold for functional polymorphisms. *Genome Res.* 13:1873–79
79. Wong GK, Liu B, Wang J, Zhang Y, Yang X, et al. 2004. A genetic variation map for chicken with 2.8 million single-nucleotide polymorphisms. *Nature* 432:717–22
80. Wrobel JA, Chao SF, Conrad MJ, Merker JD, Swanstrom R, et al. 1998. A genetic approach for identifying critical residues in the fingers and palm subdomains of HIV-1 reverse transcriptase. *Proc. Natl. Acad. Sci. USA* 95:638–45
81. Xi T, Jones IM, Mohrenweiser HW. 2004. Many amino acid substitution variants identified in DNA repair genes during human population screenings are predicted to impact protein function. *Genomics* 83:970–79
82. Yue P, Li Z, Moult J. 2005. Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* 353:459–73
83. Yue P, Moult J. 2005. Identification and analysis of deleterious human SNPs. *J. Mol. Biol.* 356:1263–74
84. Zhang EY, Fu D-J, Pak YA, Stewart T, Mukhopadhyay N, et al. 2004. Genetic polymorphisms in human proton-dependent dipeptide transporter PEPT1: implications for the functional role of Pro586. *J. Pharmacol. Exp. Ther.* 310:437–45
85. Zwick ME, Cutler DJ, Chakravarti A. 2000. Patterns of genetic variation in Mendelian and complex traits. *Annu. Rev. Genomics Hum. Genet.* 1:387–407

74. An AAS prediction method is applied to common variants implicated in common disease and the prediction scores are no different from that of neutral variation. The authors conclude that current AAS prediction methods may not be useful for finding common variants involved in common disease.

77. One of the first papers to observe that a majority of disease mutations affect protein stability and establish structure-based rules that distinguished disease mutations from nonsynonymous SNPs.

Annual Review of Genomics and Human Genetics

Volume 7, 2006

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Genomics and Human Genetics* chapters may be found at http://genom.annualreviews.org/