Published in final edited form as:
Hum Mutat. 2008 November ; 29(11): 1327–1336. doi:10.1002/humu.20892.

In silico analysis of missense substitutions using sequence-alignment based methods

Sean V Tavtigian[1,*], **Marc S Greenblatt**[2], **Fabienne Lesueur**[1], and **Graham B Byrnes**[1] **for the IARC Unclassified Genetic Variants Working Group**[†]

[1]International Agency for Research on Cancer; Lyon 69372, France

[2]Vermont Cancer Center, University of Vermont; Burlington, Vermont 05401, USA

Abstract

Genetic testing for mutations in high-risk cancer susceptibility genes often reveals missense substitutions that are not easily classified as pathogenic or neutral. Among the methods that can help in their classification are computational analyses. Predictions of pathogenic vs neutral, or the probability that a variant is pathogenic, can be made based on 1) inferences from evolutionary conservation using protein multiple sequence alignments (PMSAs) of the gene of interest for almost any missense sequence variant, and 2) for many variants, structural features of wild type and variant proteins. These in silico methods have improved considerably in recent years. In this paper, we review and/or make suggestions with respect to 1) the rationale for using in silico methods to help predict the consequences of missense variants, 2) important aspects of creating PMSAs that are informative for classification, 3) specific features of algorithms that have been used for classification of clinically observed variants, 4) validation studies demonstrating that computational analyses can have predictive values of ~75–95%, 5) current limitations of data sets and algorithms that need to be addressed in order to improve the computational classifiers, and 6) how in silico algorithms can be a part of the "integrated analysis" of multiple lines of evidence to help classify variants. We conclude that carefully validated computational algorithms, in the context of other evidence, can be an important tool for classification of missense variants.

Keywords

unclassified variant; missense substitution; protein multiple sequence alignments; PMSAs

INTRODUCTION

Mutation screening of high-risk cancer susceptibility genes often reveals sequence variants that are labeled pathogenic essentially a priori because they either disrupt the gene's structure or truncate the open reading frame. However, many other sequence variants that turn up during clinical mutation screening are not readily classified due to the interplay between gene structure and the genetic code. Most of the initially unclassified sequence variants are missense substitutions arising from a single nucleotide substitution to the open reading frame.

On the basis of basic biochemistry, molecular biology, and general protein structural principles, we often "feel" that we can evaluate a missense substitution and its context in the

[*]Corresponding author: Sean V. Tavtigian, International Agency for Research on Cancer; Lyon 69372, France, Phone: +33 (0)4 72 73 85 12, FAX: +33 (0)4 72 73 83 88, tavtigian@iarc.fr.
[†]The Working Group members are listed in the Appendix.

protein of interest and judge whether or not it will affect function based on notions of “radical”, “non-conservative”, and “conservative” missense substitution. However, conclusions that are based solely on differences between wild type and variant amino acid for features such as substitution matrix comparisons, physico-chemical “scores”, and simple protein structural change have not proven sufficiently robust to be clinically useful [Goldgar et al., 2004; Tavtigian et al., 2006; Chan et al., 2007].

Over the last 10 years, computational approaches to in silico analysis of substitutions have improved considerably over older comparisons based on inferences from matrix scores or protein structure. Most of the currently available algorithms rely on protein multiple sequence alignments (PMSAs) of the gene of interest across multiple species. When carefully validated, they become a valuable tool for classification, as reviewed in this paper. A limitation of PMSA-based analyses is that the parameter being measured, evolutionary fitness - like functional assays, which measure damage to protein function -is only a surrogate for the parameter of interest, pathogenicity [Kryukov et al., 2007]. Thus, PMSA-based methods are indirect measures of pathogenicity. Their validation involves both careful application of computer algorithms and careful curation of sequence data, such as model organism genome sequences that contribute to the PMSAs.

A second important contribution has come from the idea of “integrated analysis” of multiple parameters to classify missense substitutions [Goldgar et al., 2004]. The integrated analysis approach, discussed briefly below and in more detail in a companion paper in this issue [Goldgar et al., 2008] can cope with some uncertainty within each parameter and does not require each method to output a perfect binary classification (“pathogenic” versus “benign”). The ability of the integrated analyses to cope with uncertainty in the individual analysis methods allows us to quantify the computational approaches that depend on PMSAs and to use them to classify variants alongside more direct measures of pathogenicity (statistical genetic, epidemiologic, tumor pathologic, etc.) [Goldgar et al., 2008; Hofstra et al., 2008].

This paper describes how combining improved in silico missense analysis algorithms with higher quality multiple sequence alignments should lead to better in silico assessment. Incorporated within an integrated analysis, in silico assessment of missense substitutions can indeed stand as a clinically useful first line of analysis for newly observed substitutions.

OVERVIEW OF IN SILICO APPROACHES TO MISSENSE SUBSTITUTION ANALYSIS

Fundamentally, there are four classes of amino acid, sequence, or structural attribute that have been used to try to distinguish between neutral and pathogenic missense substitutions in silico: (1) pairwise comparison of the physico-chemical characteristics or evolutionary substitution frequencies between the wild-type and variant amino acid, (2) evolutionary conservation at the position at which a missense substitution is observed, (3) comparison between the variant amino acid and the evolutionarily tolerated amino acid range of variation at its position in the protein, and (4) protein structural considerations. In silico missense analysis algorithms may use data from just one of these classes, or combine data from two, three, or all four of the classes.

Pairwise amino acid comparisons may be based on data from amino acid substitution scoring matrices (e.g., PAM250, BLOSUM62). These matrices were derived from the frequencies with which the 20 amino acids are observed to substitute for each other in multiple sequence alignments of related proteins [Dayhoff et al., 1978; Henikoff and Henikoff, 1992]. BLOSUM62 scores range from 4–11 for identities, from 0–3 for commonly observed substitutions, and from (−1) to (−4) for substitutions rarely observed in related proteins. The average BLOSUM62 score is lower for pathogenic substitutions than for neutral substitutions [Ferrer-Costa et al., 2002; Balasubramanian et al., 2005]. Alternatively, comparisons may be based on amino acid physical or chemical properties. A score called the Grantham Difference describes the difference in side chain atomic composition, polarity, and volume between two amino acids [Grantham, 1974]. Substitutions with Grantham Differences of 5–60 are generally considered “conservative”, 60–100 “non-conservative”, and >100 “radical”. The average Grantham Difference for pathogenic substitutions is higher than for neutral substitutions [Miller et al., 2001; Abkevich et al., 2004; Balasubramanian et al., 2005]. However, these methods of pairwise amino acid comparisons alone have not led to popular missense substitution classification algorithms.

The observation that disease associated missense variants are over abundant at the positions in human proteins that are evolutionarily conserved has led to the use of PMSAs to help analyze missense substitutions [Walker et al., 1999; Miller et al., 2001; Ferrer-Costa et al., 2002; Abkevich et al., 2004; Balasubramanian et al., 2005]. The logical basis for using PMSAs to help assess whether missense substitutions cause pathogenic loss of function in disease susceptibility genes traces back to work done between the mid-1960s and the early 1970s [Zuckerkandl and Pauling, 1965; Jukes and King, 1971] and can be summarized in two related statements: (1) missense substitutions falling at positions in the gene that are evolutionarily constrained are often pathogenic, whereas those falling at positions that are not constrained are often neutral or have minimal impact, and (2) missense substitutions falling outside of the cross-species range of variation observed at their position in the PMSA are often pathogenic, whereas substitutions falling within the cross-species range of variation are often neutral or have minimal impact. Alignment based prediction tools have been tested against a number of sets of variants thought to be associated with genetic disease (Table 1). Current versions of several algorithms appear to have an accuracy of about 80%, and in some cases over 90% in classifying variants as pathogenic or not [Chan et al., 2007; Balasubramanian et al., 2005; Chao et al., 2008].

In order to use this logic in practice, one must be able to answer three questions. (1) Is a particular PMSA is reasonably informative, i.e., has it sampled enough sequences at sufficient evolutionary remove from each other (“alignment depth”) to contribute to missense substitution analysis with reasonable sensitivity and specificity? (2) How does one use a PMSA to distinguish between positions that are functionally constrained or not? (3) Do different substitutions have different effects, and can we distinguish them based on variation observed in a PMSA?

Is the Protein Multiple Sequence Alignment (PMSA) informative?

For alignment based tools to have classification value, the PMSAs used with them must be of sufficient size and carefully constructed and curated. In almost all proteins, some amino acid positions are highly conserved and others show great variation. If an amino acid position is invariant, it could either be due to selection against variation or by chance due to the limited number of sequences sampled.

The statistical likelihood that an invariant amino acid position is truly functionally constrained, rather than a false positive, depends on the size of the evolutionary database that is used. The minimum number of substitutions that are represented in a PMSA can be calculated by a variety of methods and expressed as the number of substitutions per amino acid position, averaged across all positions in the gene. The expected number of invariant positions in the alignment can be calculated based on the null hypothesis that variants should be uniformly distributed. Comparing the observed and expected numbers of invariant amino acids tells us whether the alignment is deep enough to be informative. In general, if a PMSA contains >3.0 substitutions/position, then the probability that any amino acid will be invariant is <5%. Under ideal conditions with ~50% of the positions in the protein actually functionally constrained, an alignment with three times as many substitutions as the gene has codons will achieve probability >95% that a given invariant position is invariant because it is actually functionally constrained [Greenblatt et al., 2003, Cooper et al., 2003]. However, in real world alignments, the proportion of truly conserved positions may be smaller, requiring more substitutions per position. For example, a BRCA1 PMSA containing 7 vertebrate sequences (mammals through fish) had 137 observed invariant positions (about 14% of BRCA1's 1863 amino acids), compared with the expected number of 36.6 (p<10–8). Thus, up to 1/4 of the invariant positions may still have been invariant by chance [Abkevich et al., 2004]. Consequently, even though this alignment had about 4 substitutions per position, it did not meet the criterion that >95% of invariant positions were invariant because of functional constraint. Reaching that criterion required adding two more sequences (from opossum and sea urchin), and pushed the total number of substitutions per position past 5. The fraction of constrained amino acids varies by protein. In the case of BRCA1, about 15% of the positions in the protein appear to be under strong functional constraint and 85% not. On the other hand, about half of CDKN2A positions appear to be constrained [Chan et al., 2007]. In general, better conserved proteins will require fewer substitutions per position to reach a PPV of 95%. The absolute minimum is approximately 3 substitutions per position, and some genes will require substantially more.

An informative PMSA that meets these depth criteria must also be curated so that it uses sequence data properly and is biologically logical. Investigators who use PMSAs as a tool to analyze missense substitutions often treat alignments, and the individual sequences in them, as observational data. However, several recent papers serve as reminders that sequence alignment engines often create alignment errors, which can lead to incorrect phylogenies [Martin et al., 2007; Wong et al., 2008; Loytynoja and Goldman, 2008]. Likewise, alignment errors are sometimes present in the PMSAs used for analyses of missense substitutions and have the potential to adversely affect the results. In addition, even the best alignment engines will produce PMSAs that contain gross faults if fed incorrect sequences. Thus we should bear in mind that genuine cDNA sequences may contain sequencing errors or may represent unusual splice forms that are missing conserved exons. Moreover, some of the protein sequences included in PMSAs created explicitly for analysis of missense substitutions are actually gene models inferred by computer analysis of large scale genomic data and may contain gene assembly errors (see Figure 1 and, for an error correction strategy, Supplementary Figure S1). We believe that corruption of PMSAs by errors in GenBank cDNA sequences may be a pervasive problem that must be addressed by interpreters of in silico algorithms.

How to use a PMSA: Individual positions, range of variation, and evidence of constraint

Several different approaches have been taken to measuring either the range of variation or the evidence of constraint for a specific amino acid in a protein of interest. They range from 1) simply listing the different amino acids present at the position of interest, through 2) phylogenetic tree-based methods that count the minimum number of substitutions required to account for the observed amino acid diversity and estimate the likelihood that the position is functionally constrained [Goldgar et al., 2004], to calculating 3) the average BLOSUM62 score for all of the pairs of amino acids present at the position of interest in the PMSA [Greenblatt et al., 2003; Walker et al., 1999]. The average BLOSUM62 score serves as a measure of the relative amino acid interchangeability that has been evolutionarily tolerated; low scores are indicative of a broad range of tolerance whereas high scores are indicative of little or no tolerance for amino acid substitution. 4) The Grantham difference can be modified to measure the physico-chemical variation that has been evolutionarily tolerated at a particular position in a PMSA [Abkevich et al., 2004; Tavtigian et al, 2006].

Comparing missense substitutions to the evolutionary variation observed in a PMSA

Different substitutions at the same position can have different effects (see examples in [Greenblatt et al., 2003; Raevaara et al., 2005; Chan et al., 2007]). Methods that assess the relevant features of a missense substitution should improve predictions over methods that merely describe the range of variation observed at their position in a PMSA. The simplest method is to group physico-chemically similar amino acids together into sets. A substitution within the group would be considered probably neutral whereas an out-of-group substitution would be considered probably deleterious. This approach by itself has not been validated as a classifier, but it has been incorporated as one property among many in at least one complex, multivariate classifier algorithm [Ferrer-Costa et al., 2004].

More quantitative methods have been used successfully to classify variants. Grantham scores can measure the fit between a missense substitution, the human wild type sequence, and the range of variation at its position in a PMSA [Vitkup et al., 2003; Tavtigian et al., 2006]. Similar in spirit to the Grantham scores but mathematically more complex, the MAPP (Multivariate Analysis of Protein Polymorphism) impact score is a measure of amino acid fit that combines multiple sequence alignment with multiple amino acid physical properties [Stone and Sidow, 2005].

Missense analysis algorithms—The assessments of range of amino acid variation described above have led to four missense substitution analysis algorithms that depend very heavily on PMSAs: the BLOSUM62 method [Greenblatt et al., 2003], SIFT [Ng and Henikoff, 2001], Align-GVGD [Tavtigian et al., 2006], and MAPP [Stone and Sidow, 2005].

The BLOSUM62 method scores sequence conservation at positions in a PMSA, as described above, without consideration of the fit between missense substitutions and the observed range of variation. Yet, using a set of clinically observed, already classified, missense substitutions from five human genes, the predictive value (PV) of cross-species sequence conservation at the site of a missense substitution using an optimized cutoff value of BLOSUM62 (above or below 3.5) was around 75–80%. When methods that considered features of the substituted amino acid were used, there was consensus for ~63% of variants, but classification of “deleterious” versus “neutral” differed for over 35% of variants. However, no other method by itself was statistically superior to the BLOSUM62 cutoff score [Chan et al., 2007].

SIFT (Sorting Intolerant from Tolerant) uses Dirichlet mixtures extracted from PMSAs to create position specific scoring matrices (PSSM) and score missense substitutions. The algorithm is accessible through an easily used web server (http://blocks.fhcrc.org/sift/SIFT.html). The output score is a normalized probability for each of the 19 possible missense substitutions at each position in the aligned target gene. Important considerations with SIFT are: (1) the program has the capacity to query sequence databases and build a PMSA from a user supplied target sequence; however, better results are usually obtained if a user supplies his/her own curated alignment [Chan et al., 2007]. (2) SIFT scores of ≤0.05 are usually taken as indicative of deleterious substitutions. However, the authors specifically point out that in some situations higher or lower cutoffs might give a more accurate result for binary deleterious/neutral classifications [Ng and Henikoff, 2001; Ng and Henikoff, 2003].

A second missense analysis algorithm that uses scoring matrices (similar to the SIFT approach) is embedded in the PANTHER database [Thomas et al., 2003; Thomas and Kejariwal, 2004] (http://www.pantherdb.org/). The primary mission of the PANTHER database is to organize genes into families and subfamilies and to classify them according to inferred function. Much of the organization achieved by this database relies on making PMSAs across a large number of gene subfamilies and families. One important limitation, however, is that PANTHER's PMSAs generally cover only the most conserved portions of genes, limiting the fraction of missense substitutions to which it can be applied.

Align-GVGD calculates the Grantham Variation (GV) for positions in a PMSA and the Grantham Deviation (GD) for missense substitutions at those positions. Like SIFT, it is accessible through an easy to use web server (http://agvgd.iarc.fr/). The output is in the form of two variables, GV and GD; the scores from the two variables are combined to provide a classifier. The program has had two generations of classifiers. The original classifier generated five categories (Enriched Deleterious 1 and 2, "unclassified", and Enriched Neutral 1 and 2, [Tavtigian et al., 2006]). The newer classifier does not attempt a binary division into deleterious and neutral categories but rather provides a series of ordered grades ranging from the most likely deleterious "C65" to the least likely deleterious "C0". One important feature is that the Align-GVGD website houses curated protein multiple sequence alignments for several important cancer susceptibility genes. Users can score their missense substitutions against the alignments provided at the website.

MAPP is available as JAVA code at its creator's website [Stone and Sidow, 2005] (http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html). The program requires the user to define a PMSA and to specify the relationship among the sequences in the PMSA. The output is a single MAPP impact score for each substitution analyzed. The MAPP impact score is a continuous variable; while users interested in binary classification can conduct a sensitivity/specificity test to find appropriate binary cutoffs, the algorithm's creators have provided evidence that the impact score can also be used to stratify substitutions into a spectrum of graded risk. Recently, MAPP was applied to deep alignments of MLH1 and MSH2. Results from the analysis are posted on a website that scores all possible missense substitutions in those two genes (http://mappmmr.blueankh.com) [Chao et al., 2008].

Analysis via protein structure—A somewhat independent approach to analysis of missense substitutions depends more heavily on protein structural considerations, sometimes paired with machine learning algorithms. Although many permutations have been described, there seem to be three main strategies: (1) rule or decision tree based classifiers, (2) data vectors analyzed by machine learning algorithms to generate classifiers, and (3) molecular dynamics simulations.

Rule or decision tree based classifiers focus on a series of features annotated onto the individual amino acid positions of a human protein. Many features are best extracted from an annotated crystal structure, e.g., residues located in enzymatic active sites; in binding sites, or residues with very low or very high solvent accessibility, residues located in an alpha helix, etc. The features may also include a PSSM derived from a sequence alignment, though the sequence alignment used is likely to be less extensive than that employed by pure alignment-based analysis algorithms. An amino acid substitution is then predicted to affect protein function if the substitution violates an empirically determined condition. PolyPhen (Polymorphism Phenotyping) is the best known classifier from this family, and it is available through a user-friendly web site [Sunyaev et al., 2001; Ramensky et al., 2002] (http://coot.embl.de/PolyPhen/). Other decision-tree models [Mirkovic et al., 2004; Karchin et al., 2007] have apparently not been developed into web servers to date.

One weakness of the decision tree approach is that it is generally not very good at combining marginal results from two or more of the inputs to reach a stronger result. Machine learning algorithms such as Support Vector Machine (SVM) provide a route for analyzing the multiple data types and considering the joint effects of multiple inputs. Given a data vector of n components and a training set of missense substitutions all of which are known pathogenic or known neutral, SVM creates an n-dimensional space and then finds the hyperplane in that space that best separates the known neutral substitutions from the known pathogenic substitutions. Other machine learning algorithms that have been applied to the problem of missense substitution classification include Neural Nets, Random Forest, and Naive Bayes.

Yue et al. have created algorithms and a web site, SNPs3D (http://www.snps3d.org/), that carry out two independent analyses of missense substitutions: a structural analysis and a sequence alignment/PSSM analysis [Yue et al., 2006]. The structural analysis begins with either a crystal structure of the protein of interest or of a homolog having at least 40% sequence identity. A wild-type amino acid is then replaced with a missense substitution, and the difference between the two extracted in a series of 15 characters (hydrophobic burial, sidechain overpacking, etc., [Wang and Moult 2001; Yue et al., 2005]). Training sets were compiled of 3768 presumed pathogenic substitutions from 243 genes and 16,682 presumed neutral substitutions from an overlapping set of 346 genes [Yue et al., 2005], later expanded to 10,263 presumed pathogenic substitutions from 731 genes [Yue et al., 2006], using data from in the Human Gene Mutation Database (HGMD), annotated with appropriate structural information. The data vectors were used to train a SVM to distinguish between pathogenic and neutral substitutions. Probabilities were calculated for each possible substitution to every position in the alignment, and four manipulations of the Shannon entropy were also calculated for each position in the alignment.

The classifiers within SNPs3D consider either structural factors or sequence alignment/PSSM, but not both simultaneously. In contrast, Ferrer-Costa et al. have constructed a missense analysis classifier and online server, PMUT (http://mmb2.pcb.ub.es:8080/PMut/), that combines sequence alignment/PSSM with structural factors to characterize missense substitutions [Ferrer-Costa et al., 2004; Ferrer-Costa et al., 2005]. The PMUT classifier uses a feed-forward neural network to combine data across 19 parameters (analysis from an alignment but no structural information) or 23 parameters (analysis from an alignment plus crystal structure information, including protein structure; substitution matrix values; changes in amino acid physico-chemical properties between wild-type and mutant amino acid; and measures of sequence variation at the position of the mutation).

Other methods of protein structural analysis exist, such as molecular dynamics folding simulation as implemented in the program FoldX (http://foldx.crg.es/). This method has been used to assess protein missense variants [Pey et al., 2007; Tokuriki et al., 2007], but not in cancer susceptibility genes. However, assessments limited to protein folding would be blind to substitutions whose major effect is to alter an important protein:protein interaction.

BENCHMARKING – CAUTION IN USE OF DATA SETS

Development of missense classification algorithms has usually depended upon access to a dataset of missense variants that have been classified as functional/neutral vs nonfunctional/pathogenic. Validated sets of variants may be used to optimize the performance of score-based classifiers, train machine-learning based classifiers, and/or compare the performance of established classifiers. Of key importance, datasets that are used either to estimate the error rates or to compare the performance of classifier algorithms need to be drawn from a distribution of missense substitutions that closely resembles the distribution of substitutions in the real world data to which the classifiers will eventually be applied.

Most germline missense substitutions observed in high-risk cancer susceptibility genes in human subjects result from point mutations (single base substitutions and indels). Germline substitutions do not occur randomly with respect to the underlying DNA sequence. There are up to 100-fold differences between sequence substitution probabilities, based mostly on local sequence context. The most obvious example is that transitions at the dinucleotide CpG (specifically, CG to either CA or TG) happen at a much higher rate than most other substitutions. In general, transitions happen at a higher rate than do transversions. The interplay between germline nucleotide substitution frequencies and the genetic code results in mutational bias towards silent substitutions and towards relatively conservative amino acid substitutions, an effect easily observed when dinucleotide substitution rate constants (such as those determined by Lunter and Hein [Lunter and Hein, 2004]) are applied to the genetic code. For example, the average Grantham Difference between all possible amino acid pairs is 100 (by definition), whereas the average Grantham Difference between human canonical amino acids and the missense substitutions seen in large mutation screening series, such as Myriad Genetics' BRCA1 and BRCA2 mutation screening data, is close to 70 [Grantham, 1974; Abkevich et al., 2004].

Three types of missense substitution data sets have been used repeatedly during the creation, optimization, and comparison of missense classifiers: (1) datasets from systematic mutation of a phage or viral proteins, (2) sets of expected neutral pseudo-substitutions taken as the missense differences between human proteins and aligned mammalian ortholog sequences, and (3) datasets derived from the annotations of human missense variants present in the SWISS-PROT knowledgebase. Systematic mutation datasets contain many missense substitutions that require two nucleotide substitutions, and even pseudo-substitution datasets taken from aligned mammalian orthologs contain a substantial number of such missense substitutions. Consequently, human variants annotated in SWISS-PROT would appear to be better for measuring classifier performance or comparing different classifiers to each other [Care et al., 2007].

Capriotti et al. extracted a large set of annotated human missense substitutions from the SWISS-PROT database (8,987 substitutions in 1,434 genes) to train their SVM based classifier (SeqProfCod), and then used an independent set of human missense substitutions from the SWISS-PROT database (2,008 substitutions in 720 genes) to measure classifier performance [Capriotti et al., 2008]. Because we have been intimately involved in clinical classification of missense substitutions in high-risk cancer susceptibility genes, we examined the set of 177 classified substitutions in BRCA1 (n=43), BRCA2 (n=68), MLH1 (n=49), and MSH2 (n=17) present in the subset of SWISS-PROT that was used in Capriotti et al. study.

Of the 43 BRCA1 missense substitutions, all from exon 11, the SWISS-PROT database had annotated 17 as neutral polymorphisms and 26 as disease-associated. Over the last several years, 17 of these 43 substitutions have been securely classified, all as neutral variants (Class 1, Plon et al., 2008), either by Myriad Genetics or by members of the Breast Cancer Information Consortium (BIC) [Deffenbuagh et al., 2002; Goldgar et al., 2004; Judkins et al., 2005; Tavtigian et al., 2006; Easton et al., 2007; Spurdle et al., 2008]. However, the SWISS-PROT database had annotated 5 of these 17 Class 1 variants as pathogenic. Moreover, we now know that the prior probability that missense substitutions falling outside of the BRCA1 RING or BRCT domains are pathogenic is less than 0.01 [Easton et al., 2007; Tavtigian et al., 2008]. As all 43 of these substitutions fall between the BRCA1 RING and BRCT domains, it is likely that they are all neutral, so the other 21 variants classified as pathogenic by SWISS-PROT are also likely errors. Of the 68 BRCA2 missense substitutions, 17 have been securely classified: 2 as clearly pathogenic (Class 5) and the remaining 15 as neutral (Class 1) [Deffenbuagh et al., 2002; Goldgar et al., 2004; Chenevix-Trench et al., 2006; Easton et al., 2007; Spurdle et al., 2008]. The two Class 5 variants were correctly annotated in the SWISS-PROT database, but 8 of the 15 Class 1 variants were incorrectly annotated as pathogenic.

To estimate the SWISS-PROT database annotation accuracy for missense substitutions in MLH1 and MSH2, we merged classifications reported in three recent manuscripts [Chan et al., 2007; Barnetson et al., 2008; Chao et al., 2008] and the online MMR Gene Missense Mutation Database maintained by the University Medical Center Groningen (http://www.mmrmissense.org). One variant (MLH1 A681T) was excluded because the classification in one manuscript contradicted the classification in the other two, and three variants reported as neutral in the papers were excluded because the Groningen database recorded some evidence of functional deficit (MLH1 K618A, MLH1 Y646C, MSH2 G322D). The remaining set of 94 variants that had been validated by these four groups (61 likely pathogenic and 33 likely neutral substitutions) were cross-referenced against the annotated MLH1 and MSH2 substitutions used by Capriotti et al [Capriotti et al., 2008]. The validated data contained 18 likely pathogenic and 10 likely neutral substitutions that were also in the SWISS-PROT data set. All 18 of the likely deleterious substitutions were annotated as deleterious in SWISS-PROT, but 5 of the 10 likely neutral substitutions were labeled pathogenic in SWISS-PROT.

This validated sample of the SWISS-PROT dataset represents 0.7% of their training set and none of their benchmarking set; however, it is the subset that is most relevant to clinical cancer genetics. There is a pattern of significant Type 1 error. Based on validation by multiple expert groups, the PPV of a “pathogenic” classification in SWISS-PROT is at best 57% (if one uses only the 17 BRCA1 variants validated by Myriad or the BIC), or as low as 43% (if one uses all 43 BRCA1 variants and assumes that all are neutral based on Easton et al. [Easton et al., 2007]). Consequently, the apparent error rates in the database appear to significantly exceed the error rates in current computational classifiers, whose predictive values range from 75% to over 90%.

It would appear that the best data sets for testing classifiers are those from locus-specific databases that are curated by individuals or groups specialized in the analysis of one or a few genes [Chao et al., 2008; Chan et al., 2007; Barnetsen et al., 2008; Goldgar et al., 2004; Easton et al., 2007]. The sequence variants in such databases will be dominated by single nucleotide substitutions and the distribution of substitutions resembles what will be present in clinical mutation screening datasets, meeting the criteria set down by Care et al [Care et al., 2007]. Nonetheless, even these sets contain some uncertainty. For example, many variants are classified as pathogenic based on as few as two carriers plus some supporting clinicopathological data [Chao et al., 2008; Chan et al., 2007]. We hope that application of the suggestions present in the other manuscripts of this special issue will result in datasets from which one could extract sets of substitutions classified with better than 90% overall accuracy.

BENCHMARKING – RESULTS FROM CURRENT In silico METHODS

During the creation of PMUT, Ferrer-Costa et al compared the predictive performance of the individual components of their data vector to each other. The “pathogenic” substitutions that they used came from Swiss-Prot and the “neutral” substitutions were evolutionarily tolerated substitutions observed in proteins with >95% sequence identity to the human proteins from which the pathogenic substitutions were gathered. Thus, the underlying data set has all of the weaknesses discussed above and the results need to be treated with caution. Nonetheless, an interesting result emerged: classification based on the PSSMs alone out performed classifications based on any other single component. Classification based on BLOSUM62 alone came second, and the differences between classification based on the PSSMs alone and any other single component (save BLOSUM62) were as great or greater difference between the full multi-component model and the PSSM-only model [Ferrer-Costa et al., 2004].

More recently, curated locus specific databases have been used as data sources for comparing multiple classifiers. Chan et al. used a total of 254 substitutions from 5 different genes (CDKN2A, MLH1, MSH2, MECP2, and TYR) to compare two pairwise missense substitution scores (BLOSUM62 change and Grantham Difference) and four classifier algorithms (BLOSUM62 pairwise, SIFT, PolyPhen, and Align-GVGD). Five interesting results emerged from this analysis. (1) The four classifier algorithms all outperformed the two simple substitution scores. (2) BLOSUM62 pairwise and Align-GVGD showed better specificity than sensitivity and thus relatively low false-positive prediction of pathogenic. On the other hand, SIFT and PolyPhen showed better sensitivity than specificity and thus relatively low false negative prediction of neutral. (3) The sensitivity and specificity differences between the algorithms more or less balanced out, so that their overall predictive values, which ranged from 73%–82%, were not significantly different from each other. (4) Using informative sequence alignments that met a 3 substitutions per position criterion, invariance at a position in an alignment became a very good predictor of pathogenic substitutions (PPV= 96.8%, better than any single method). (5) Concordance among methods was a strong predictor. When all 4 methods agreed that a variant was deleterious, the positive predictive value was 94.6%, and when all 4 methods agreed that a variant was neutral, the negative predictive value was 73.5%. The concordant NPV and PPV were better than the corresponding values for any single prediction algorithm [Chan et al., 2007].

Using a set of 55 classified MLH1 missense substitutions and 21 classified MSH2 missense substitutions, Chao et al. optimized the MAPP algorithm and compared its performance to PolyPhen and SIFT. In optimizing MAPP, the authors allowed for variable depth of PMSA along the length of each protein based on the frequency of gaps in local segments of the alignment, and optimized the MAPP impact score that best partitioned neutral from pathogenic substitutions, which are more conventional optimizations to carry out with a score-based classifier. A cutoff score of 4.5 determined from a receiver-operator characteristic curve was used to distinguish predictions of pathogenic versus neutral. Perhaps because of these optimizations, MAPP significantly outperformed both SIFT and PolyPhen, yielding both sensitivity and specificity of >90%. The specificity of SIFT and PolyPhen were similar (81%), but the sensitivity of SIFT was much better than that of PolyPhen (82% vs. 58%). Notably, alignments were created by SIFT and results may have been better had it been run with optimized, curated PMSAs. MAPP-MMR results are now available online (http://mappmmr.blueankh.com/Impact.php). However, as of June 2008, variants with scores between 3 and 5 are designated as “borderline”.

Machine learning algorithms (e.g., Artificial Neural Network [ANN], SVM) are promising tools to improve predictions, but they have not yet been adequately validated. Chan et al found ANN and SVM that combined sequence alignment data with structural parameters improved prediction accuracy when the PMSAs were shallow, but did not improve prediction accuracy when the PMSAs were deep enough to meet their 3 substitutions per position criterion [Chan et al, 2007]. In a study of BRCA1 variants, supervised learners outperformed Align-GVGD and SIFT. However, the sequence variants used were only classified by functional assay, so these results should be repeated with more robust data sets before they are widely accepted [Karchin et al., 2007].

The Chan et al. and Chao et al. studies provide a solid starting point for comparing and systematically improving missense classification algorithms under conditions similar to those under which they would have to operate if used for clinical classification. In particular, it is encouraging that the 23 MLH1 missense substitutions and 5 MSH2 missense substitutions considered in common between the two studies carried the same classification in both studies and only two (MLH1 Y646C and MLH1 A681T) yielded contradictory results in our 4-study comparison (v.s.), consistent with the aspiration that the overall accuracy of classification of substitutions used for comparison studies should be better than 90%.

APPLICATION TO CLINICAL CLASSIFICATION

Few groups have explicitly used results from in silico analysis of missense substitutions for clinically relevant classification. There are two basic reasons for this: (1) the in silico analyses by themselves do not deliver strong enough likelihood ratios or predictive values to serve as stand-alone classifiers, and (2) integrated missense classification methods that can combine across several disparate data types to achieve final classification have not become widely available [Goldgar et al., 2008].

The clearest progress towards use of in silico analysis of missense substitutions for clinical classification is evident across a series of BRCA1 and BRCA2 missense substitution analysis papers coordinated by members of the BIC [Goldgar et al., 2004; Tavtigian et al., 2006; Chenevix-Trench et al., 2006; Lovelock et al., 2006; Spurdle et al., 2008]. These papers describe an integrated classification method that could incorporate the results from in silico classifiers if those results were formatted as a likelihood ratio. These methods should be validated using other data sets.

Recently, we developed a new classifier based on the Align-GVGD algorithm [Tavtigian et al., 2008]. Rather than attempting a binary classification, the output from Align-GVGD is now an ordered series of grades ranging from C65 (most likely deleterious) to C0 (most likely neutral). The grades were calibrated against family histories data from a dataset derived from 70,000 subjects tested fro BRCA mutations at Myriad Genetics [Easton et al., 2007]. The results of the calibration were formatted as posterior probabilities that can now serve as empirically determined prior probabilities in future analyses of individual substitutions. The probabilities range from 0.81 in favor of pathogenic for the grade C65 to 0.01 for the grade C0 [Tavtigian et al., 2008]. Since the prior probability of a missense variant in the dataset as a whole being pathogenic was about 0.13, we can calculate that the likelihood ratio for C65 was 28.5:1. The criterion for the grade C65 is very strict; almost all variants here occur at evolutionarily invariant positions. It is therefore reassuring that the LR determined for C65 substitutions from informative BRCA1 and BRCA2 PMSAs is very close to the LR for substitutions falling at invariant positions determined by Chan et al (30:1) using similarly informative PMSAs [Chan et al., 2007; Tavtigian et al., 2008].

Working with missense substitutions in mismatch repair genes, Barnetson et al. presented a qualitative, point-based, integrative analysis of missense substitutions [Barnetson et al., 2008]. Their analysis included assessment of sequence variation in PMSAs, prepared with ClustalW, that contained 4–5 orthologous sequences from human to yeast. The analysis also included assessment of the missense substitutions by PolyPhen and SIFT (with SIFT creating the alignment). Of 23 initially unclassified MLH1 or MSH2 missense substitutions considered, Barnetson et al classified 11 as “benign” and 2 as “pathogenic”. We note, however, that the classifications of 3 of these substitutions (MLH1 K618A, MLH1 A681T, MSH2 G322D) conflict with either functional assay results or classifications given in contemporaneous publications or the Groningen MMR database [Chan et al., 2007; Chao et al., 2008, <http://www.mmrmissense.org/default.aspx> as of 15 May 2008]. Validation of such results is critical. The point based method will be difficult to apply to other genes, and assessment of sequence conservation and the SIFT analysis would be easier if the alignments used were available in the publication record.

CALIBRATION

The likelihood ratio-based integrated assessment of variants in BRCA1 and BRCA2 provided a comfortable context in which to introduce in silico assessment of missense substitutions because the format accepts the input from any individual method with a weighting that is experimentally determined. Very effective individual methods will have very high or very low LRs and therefore will have strong effects on the prediction, and modestly effective methods will have LRs closer to 1.0 and exert modest influence. At this time, the LRs from in silico classifiers appear to range from ~2.5 (i.e., PV of 72%, seen in a number of methods) to ~25–30 (PV of 96–97%, substitution at an evolutionarily invariant amino acid).

The method by which the newly-defined Align-GVGD grades were calibrated against BRCA1 and BRCA2 mutation screening data is sufficiently general that other classifier algorithms could be calibrated against the same data set in the same way so long as the output of the classifier is either binary or expressed in a limited number (3–5) of ordered categories [Tavtigian et al., 2008]. However, whether expressed as likelihood ratios or probabilities, the quantitative results generated from specific BRCA1 and BRCA2 PMSAs cannot be assumed to transfer directly to other genes and other PMSAs. Accordingly, one way forward might involve the following steps:

1. Selection of a defined set of cancer susceptibility genes (BRCA1, BRCA2, MLH1, MSH2, CDKN2A, and perhaps others) to serve as a calibration step.
2. Creation of a set of reference sequence alignments that will be available online. Creation of the alignments should include both investigators who are specialized in analysis of each specific gene and investigators who have experience in gene model assembly so that model organism genomic sequences can be used appropriately. Different classifier algorithms that incorporate PMSAs may optimize at different depths of alignment. Therefore, an effort should be made use each algorithm near its optimum depth of alignment.
3. Use of the existing BRCA1 and BRCA2 mutation screening dataset, and in particularly the summary family history data contained therein, to calibrate an expanded set of classifier algorithms.
4. Creation of a sizeable set of variants in these genes that have been classified Class 1–2 and Class 4–5 [Plon et al., 2008] by specialists in each gene.
5. Use of this set of classified substitutions to calculate sensitivity, specificity, and predictive values of an expanded set of classifier algorithms and also obtain some measurements of sensitivity and specificity for concordant and discordant classification from a defined set of classifiers. Once large numbers of substitutions have been analyzed under well curated conditions, patterns of classification errors might emerge that provide clues to methodological shortcomings.

CONCLUSIONS

Since a number of promising in silico missense substitution analysis methods are dependent on PMSAs, creation of high quality PMSAs should be a priority. In particular, individual available sequences should be regarded as hypotheses, and conflicts visible in initial PMSAs may need to be resolved through scrutiny of experimentally determined primary sequence data (genomic, cDNA, other). A curation process that compares more types of data will resolve discrepancies and result in more valid alignments. Different classifier algorithms that incorporate PMSAs may optimize at different depths of alignment, and different genes clearly optimize at different depths of alignment. Therefore, careful study of each algorithm and gene will be needed to clarify optimum depth of alignment.

Training of missense classifiers that are based on machine learning algorithms and performance comparisons between existing algorithms both require large data sets of accurately classified missense substitutions. Existing large data sets, such as the annotations contained in SWISS-PROT, have high enough error rates to compromise either of these activities. Thus an important activity for the near future should be to expand the sets of securely classified substitutions in the high-risk susceptibility genes and then use the resulting data sets in new rounds of algorithm training and/or benchmarking.

Moving from the development and testing of missense substitution analysis algorithms to actually using them in a clinical genetics context involves crossing a psychological barrier as much as crossing a methodological barrier. The predictive values of 75–90+% obtained by some in silico classification algorithms compare favorably with many medical tests that are in current use [Chan et al., 2007; Plon et al., 2008]. Although important clinical decisions regarding variant interpretation should not be made based on a single test of 80% accuracy, in the context of other evidence (e.g., association of a genetic variant with disease in a family) these methods can be an important tool in medical decision-making. At this time, it appears that in some cases the accuracy is limited by the available evolutionary, mutation, or structural databases and in some cases by the limitations of each individual method.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by grants from the National Institutes of Health (CA 96536, MSG) and the Lake Champlain Cancer Research Organization (MSG).

References

Abkevich V, Zharkikh A, Deffenbaugh A, Frank D, Chen Y, Shattuck D, Skolnick MH, Gutin A, Tavtigian SV. Analysis of missense variation in human BRCA1 in the context of interspecific sequence variation. J Med Gen. 2004; 41:492–507.

Balasubramanian S, Xia Y, Freinkman E, Gerstein M. Sequence variation in G-protein-coupled receptors: analysis of single nucleotide polymorphisms. Nucleic Acids Res. 2005; 33:1710–1721. [PubMed: 15784611]

Barnetson RA, Cartwright N, van Vliet A, Haq N, Drew K, Farrington S, Williams N, Warner J, Campbell H, Porteous ME, Dunlop MG. Classification of ambiguous mutations in DNA mismatch repair genes identified in a population-based study of colorectal cancer. Hum Mutat. 2008; 29:367–374. [PubMed: 18033691]

Capriotti E, Arbiza L, Casadio R, Dopazo J, Dopazo H, Marti-Renom MA. Use of estimated evolutionary strength at the codon level improves the prediction of disease-related protein mutations in humans. Hum Mutat. 2008; 29:198–204. [PubMed: 17935148]

Care MA, Needham CJ, Bulpitt AJ, Westhead DR. Deleterious SNP prediction: be mindful of your training data! Bioinformatics. 2007; 23:664–672. [PubMed: 17234639]

Chan PA, Duraisamy S, Miller PJ, Newell JA, McBride C, Bond JP, Raevaara T, Ollila S, Nystrom M, Grimm AJ, Christodoulou J, Oetting WS, Greenblatt MS. Interpreting missense variants: comparing computational methods in human disease genes CDKN2A, MLH1, MSH2, MECP2, and tyrosinase (TYR). Hum Mutat. 2007; 28:683–693. [PubMed: 17370310]

Chao EC, Velasquez JL, Witherspoon MSL, Rozek LS, Peel D, Ng P, Gruber SB, Watson P, Rennert G, Anton-Culver H, Lynch H, Lipkin SM. Accurate classification of MLH1/MSH2 missense variants with multivariate analysis of protein polymorphisms-mismatch repair (MAPP-MMR). Human Mutation. 2008; 29:852–860. [PubMed: 18383312]

Chenevix-Trench G, Healey S, Lakhani S, Waring P, Cummings M, Brinkworth R, Deffenbaugh AM, Burbidge LA, Pruss D, Judkins T, Scholl T, Bekessy A, Marsh A, Lovelock P, Wong M, Tesoriero A, Renard H, Southey M, Hopper JL, Yannoukakos K, Brown M, Easton D, Tavtigian SV, Goldgar D, Spurdle AB. Genetic and histopathologic evaluation of BRCA1 and BRCA2 DNA sequence variants of unknown clinical significance. Cancer Res. 2006; 66:2019–2027. [PubMed: 16489001]

Cooper GM, Brudno M, NISC, Green ED, Batzoglou S, Sidow A. Quantitative estimates of sequence divergence for comparative analyses of mammalian genomes. Genome Res. 2003; 13:813–820. [PubMed: 12727901]

Dayhoff, MO.; Schwartz, RM.; Orcutt, BC. A model of evolutionary change in proteins. In: Dayhoff, MO., editor. Atlas of Protein Sequence and Structure. Vol. 5. Washington DC: National Biochemical Research Foundation; 1978. p. 345-352.

Deffenbaugh AM, Frank TS, Hoffman M, Cannon-Albright L, Neuhausen SL. Characterization of common BRCA1 and BRCA2 variants. Genet Test. 2002; 6:119–121. [PubMed: 12215251]

Easton DF, Deffenbaugh AM, Pruss D, Frye C, Wenstrup RJ, Allen-Brady K, Tavtigian SV, Monteiro AN, Iversen ES, Couch FJ, Goldgar DE. A Systematic Genetic Assessment of 1,433 Sequence Variants of Unknown Clinical Significance in the BRCA1 and BRCA2 Breast Cancer-Predisposition Genes. Am J Hum Genet. 2007; 81:873–883. [PubMed: 17924331]

Ferrer-Costa C, Orozco M, de la Cruz X. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. J Mol Biol. 2002; 315:771–786. [PubMed: 11812146]

Ferrer-Costa C, Orozco M, de la Cruz X. Sequence-based prediction of pathological mutations. Proteins. 2004; 57:811–819. [PubMed: 15390262]

Ferrer-Costa C, Gelpi JL, Zamakola L, Parraga I, de la Cruz X, Orozco M. PMUT: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics. 2005; 21:3176–3178. [PubMed: 15879453]

Goldgar DE, Easton DF, Byrnes GB, Spurdle AB, Iversen ES, Greenblatt MS. IARC Unclassified Genetic Variants Working Group. Integration of various data sources for classifying uncertain variants into a single model. Hum Mutat. 2008:29.

Goldgar DE, Easton DF, Deffenbaugh AM, Monteiro AN, Tavtigian SV, Couch FJ. Integrated evaluation of DNA sequence variants of unknown clinical significance: application to BRCA1 and BRCA2. Am J Hum Genet. 2004; 75:535–544. [PubMed: 15290653]

Grantham R. Amino acid difference formula to help explain protein evolution. Science. 1974; 185:862–864. [PubMed: 4843792]

Greenblatt MS, Beaudet JG, Gump JR, Godin KS, Trombley L, Koh J, Bond JP. Detailed computational study of p53 and p16: using evolutionary sequence analysis and disease-associated mutations to predict the functional consequences of allelic variants. Oncogene. 2003; 22:1150–1163. [PubMed: 12606942]

Henikoff S, Henikoff J. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA. 1992; 89:10915–10919. [PubMed: 1438297]

Hofstra RW, Spurdle AB, Eccles D, Foulkes WD, de Wind N, Hoogerbrugge N, Hogervorst FBL. IARC Unclassified Genetic Variants Working Group. Tumor characteristics as an analytic tool for classifying genetic variants of uncertain clinical significance. Hum Mutat. 2008:29.

Judkins T, Hendrickson BC, Deffenbaugh AM, Eliason K, Leclair B, Norton MJ, Ward BE, Pruss D, Scholl T. Application of embryonic lethal or other obvious phenotypes to characterize the clinical significance of genetic variants found in trans with known deleterious mutations. Cancer Res. 2005; 65:10096–10103. [PubMed: 16267036]

Jukes TH, King JL. Deleterious mutations and neutral substitutions. Nature. 1971; 231:114–115. [PubMed: 4930087]

Karchin R, Monteiro AN, Tavtigian SV, Carvalho MA, Sali A. Functional Impact of Missense Variants in BRCA1 Predicted by Supervised Learning. PLoS Comput Biol. 2007; 3:e26. [PubMed: 17305420]

Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet. 2007; 80:727–739. [PubMed: 17357078]

Lovelock PK, Healey S, Au W, Sum EY, Tesoriero A, Wong EM, Hinson S, Brinkworth R, Bekessy A, Diez O, Izatt L, Solomon E, Jenkins M, Renard H, Hopper J, Waring P, Tavtigian SV, Goldgar D, Lindeman GJ, Visvader JE, Couch FJ, Henderson BR, Southey M, Chenevix-Trench G, Spurdle AB, Brown MA. Genetic, functional, and histopathological evaluation of two C-terminal BRCA1 missense variants. J Med Genet. 2006; 43:74–83. [PubMed: 15923272]

Loytynoja A, Goldman N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science. 2008; 320:1632–1635. [PubMed: 18566285]

Lunter G, Hein J. A nucleotide substitution model with nearest-neighbour interactions. Bioinformatics. 2004; 20 (Suppl 1):I216–I223. [PubMed: 15262802]

Martin W, Roettger M, Lockhart PJ. A reality check for alignments and trees. Trends Genet. 2007; 23:478–480. [PubMed: 17825944]

Miller MP, Kumar S. Understanding human disease mutations through the use of interspecific genetic variation. Hum Mol Genet. 2001; 10:2319–2328. [PubMed: 11689479]

Mirkovic N, Marti-Renom MA, Weber BL, Sali A, Monteiro AN. Structure-based assessment of missense mutations in human BRCA1: implications for breast and ovarian cancer predisposition. Cancer Res. 2004; 64:3790–3797. [PubMed: 15172985]

Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. Genome Res. 2001; 11:863–874. [PubMed: 11337480]

Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003; 31:3812–3814. [PubMed: 12824425]

Pey AL, Stricher F, Serrano L, Martinez A. Predicted effects of missense mutations on native-state stability account for phenotypic outcome in phenylketonuria, a paradigm of misfolding diseases. Am J Hum Genet. 2007; 81:1006–1024. [PubMed: 17924342]

Plon SE, Eccles DM, Easton DF, Foulkes W, Genuardi M, Greenblatt MS, Hogervorst FBL, Hoogerbrugge N, Spurdle AB, Tavtigian S. IARC Unclassified Genetic Variants Working Group. Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results. Hum Mutat. 2008:29.

Raevaara TE, Korhonen MK, Lohi H, Hampel H, Lynch E, Lonnqvist KE, Holinski-Feder E, Sutter C, McKinnon W, Duraisamy S, Gerdes AM, Peltomaki P, Kohonen-Ccorish M, Mangold E, Macrae F, Greenblatt M, de la Chapelle A, Nystrom M. Functional significance and clinical phenotype of nontruncating mismatch repair variants of MLH1. Gastroenterology. 2005; 129:537–549. [PubMed: 16083711]

Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. Nucleic Acids Res. 2002; 30:3894–3900. [PubMed: 12202775]

Spurdle AB, Lakhani SR, Healey S, Parry S, Da Silva LM, Brinkworth R, Hopper JL, Brown MA, Babikyan D, Chenevix-Trench G, Tavtigian SV, Goldgar DE. Clinical classification of BRCA1 and BRCA2 DNA sequence variants: the value of cytokeratin profiles and evolutionary analysis--a report from the kConFab Investigators. J Clin Oncol. 2008; 26:1657–1663. [PubMed: 18375895]

Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. Genome Res. 2005; 15:978–986. [PubMed: 15965030]

Sunyaev S, Ramensky V, Koch I, Lathe Wr, Kondrashov AS, Bork P. Prediction of deleterious human alleles. Hum Mol Genet. 2001; 10:591–597. [PubMed: 11230178]

Tavtigian S, Byrnes GB, Goldgar DE, Thomas A. Classification of rare missense substitutions, using risk surfaces, with genetic- and molecular-epidemiology applications. Hum Mutat. 2008:29.

Tavtigian SV, Deffenbaugh AM, Yin L, Judkins T, Scholl T, Samollow PB, de Silva D, Zharkikh A, Thomas A. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. J Med Genet. 2006; 43:295–305. [PubMed: 16014699]

Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, Diemer K, Muruganujan A, Narechania A. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003; 13:2129–2141. [PubMed: 12952881]

Thomas PD, Kejariwal A. Coding single-nucleotide polymorphisms associated with complex vs. Mendelian disease: evolutionary evidence for differences in molecular effects. Proc Natl Acad Sci U S A. 2004; 101:15398–15403. [PubMed: 15492219]

Tokuriki N, Stricher F, Schymkowitz J, Serrano L, Tawfik DS. The stability effects of protein mutations appear to be universally distributed. J Mol Biol. 2007; 369:1318–1332. [PubMed: 17482644]

Vitkup D, Sander C, Church GM. The amino-acid mutational spectrum of human genetic disease. Genome Biol. 2003; 4:R72. [PubMed: 14611658]

Walker DR, Bond JP, Tarone RE, Harris CC, Makalowski W, Boguski MS, Greenblatt MS. Evolutionary conservation and somatic mutation hotspot maps of p53: correlation with p53 protein structural and functional features. Oncogene. 1999; 18:211–218. [PubMed: 9926936]

Wang Z, Moult J. SNPs, protein structure, and disease. Hum Mutat. 2001; 17:263–270. [PubMed: 11295823]

Wong KM, Suchard MA, Huelsenbeck JP. Alignment uncertainty and genomic analysis. Science. 2008; 319:473–476. [PubMed: 18218900]

Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol. 2005; 353:459–473. [PubMed: 16169011]

Yue P, Melamud E, Moult J. SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics. 2006; 7:166. [PubMed: 16551372]

Zuckerkandl E, Pauling L. Molecules as documents of evolutionary history. J Theor Biol. 1965; 8:357–366. [PubMed: 5876245]

APPENDIX. Members of the IARC Working Group on Unclassified Genetic Variants

Paolo Boffetta, IARC, France; Fergus Couch, Mayo Clinic, USA; Niels de Wind, Leiden University, the Netherlands; Diana Eccles, University of Southampton, UK; Douglas Easton, Cambridge University, UK; William Foulkes, McGill University, Canada; Maurizio Genuardi, University of Florence, Italy; David Goldgar, University of Utah, USA; Marc Greenblatt, University of Vermont, USA; Robert Hofstra, University Medical Center Groningen, the Netherlands; Frans Hogervorst, Netherlands Cancer Institute, the Netherlands; Nicoline Hoogerbrugge, University Medical Center Neimejen, the Netherlands; Sharon Plon, Baylor University, USA; Paolo Radice, Istituto Nazionale Tumori, Italy; Lene Rasmussen, Roskilde University, Denmark; Olga Sinilnikova, Hospices Civils de Lyon, France; Amanda Spurdle, Queensland Institute of Medical Research, Australia; Sean Tavtigian, IARC, France.

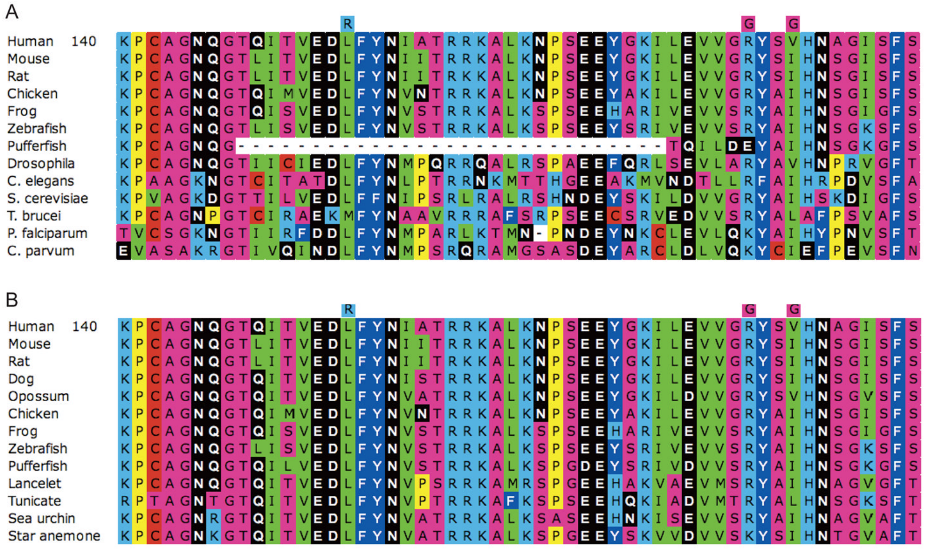
A
R
G
G
Human 140 KPCAGNQGTQITVEDLFYNIATRRKALKNPSEEYGKILEVVGRYSVHNAGISFS
Mouse KPCAGNQGTLITVEDLFYNIITRRKALKNPSEEYGKILEVVGRYSIHNSGISFS
Rat KPCAGNQGTLITVEDLFYNIITRKKALKNPSEEYGKILEVVGRYSIHNSGISFS
Chicken KPCAGNQGTQIMVEDLFYNVNTRRKALKNPSEEYAKILEVVGRYAIHNSGISFS
Frog KPCAGNQGTQISVEDLFYNVSTRRKALKSPSEEHARIVEVVSRYAIHNSGIGFS
Zebrafish KPCAGNQGTLISVEDLFYNVSTRRKALKSPSEEYSRIVEVVSRYAIHNSGKSFS
Pufferfish KPCAGNQG-----------------------------TQILDEYAIHNSGKGFS
Drosophila KPCAGNQGTIICIEDLFYNMPQRRQALRSPAEEFQRLSEVLARYAVHNPRVGFT
C. elegans KPAAGKNGTCITATDLFYNLPTRRNKMTTHGEEAKMVNDTLLRFAIHRPDVSFA
S. cerevisiae KPVAGKDGTTILVEDLFFNIPSRLRALRSHNDEYSKILDVVGRYAIHSKDIGFS
T. brucei KPCAGNPGTCIRAEKMFYNAAVRRRAFSRPSEECSRVEDVVSRYALAFPSVAFS
P. falciparum TVCSGKNGTIIRFDDLFYNMPARLKTMN-PNDEYNKCLEVLQKYAIHYPNVSFT
C. parvum EVASAKRGTIVQINDLFYNMPSRQRAMGSASDEYARCLDLVQKYCIEFPEVSFN
B
R
G
G
Human 140 KPCAGNQGTQITVEDLFYNIATRRKALKNPSEEYGKILEVVGRYSVHNAGISFS
Mouse KPCAGNQGTLITVEDLFYNIITRRKALKNPSEEYGKILEVVGRYSIHNSGISFS
Rat KPCAGNQGTLITVEDLFYNIITRKKALKNPSEEYGKILEVVGRYSIHNSGISFS
Dog KPCAGNQGTQITVEDLFYNISTRRKALKNPSEEYGKILEVVGRYSIHNSGISFS
Opossum KPCAGNQGTQITVEDLFYNVATRRKALKNPSEEYGKILDVVGRYSVHNSGVSFS
Chicken KPCAGNQGTQIMVEDLFYNVNTRRKALKNPSEEYAKILEVVGRYAIHNSGISFS
Frog KPCAGNQGTQISVEDLFYNVSTRRKALKSPSEEHARIVEVVSRYAIHNSGIGFS
Zebrafish KPCAGNQGTLISVEDLFYNVSTRRKALKSPSEEYSRIVEVVSRYAIHNSGKSFS
Pufferfish KPCAGNQGTQILVEDLFYNVSTRRKALKSPGDEYSRIVDVVSRYAIHNSGKGFS
Lancelet KPCAGNQGTQITVEDLFYNVPSRRKAMRSPGEEHAKVAEVMSRYAIHNAGVGFT
Tunicate RPTAGNTGTQITVEDLFYNVPTRRKAFKSPSEEHQKIADVMTRYALHNSGKSFT
Sea urchin KPCAGNRGTQITVEDLFYNVATRRKALKSASEEHNKISEVVSRYAIHNAGVAFT
Star anemone KPCAGNKGTQITVEDLFYNVATRRKALKSPGEEYSKVVDVVSKYSIHNTGVAFT

Figure 1.

A. A section of MLH1 alignment, from human residue K140, excerpted from Chan et al. 2007. Note that the alignment contains an apparent gap in the pufferfish sequence. The mutant residue from 3 human missense substitutions, p. L155R, p. R182G,. and p.V185G, are positioned above the alignment. B. The corresponding section of a newly prepared MLH1 alignment in which the apparent gap in the pufferfish sequence has been repaired via an analysis of the underlying nucleotide sequences from the T. nigroviridis and F. rubripes genomic sequences. Note that: (1) at the right boundary of the gap in panel A, the first 4 residues, TQIL, were misaligned and actually belonged at the left edge of the gap; (2) the next two residues, DE, were probably a sequence assembly artifact; and (3) in silico assessments of the 2 human missense substitutions p.L155R and p.R182G may differ depending on which alignment is used because of the changes introduced by repairing the artifactual gap.

Note also a difference in character between the two alignments. Alignment A is phylogenetically rather deep, containing sequences from long branch-length animals such as C. elegans as well as fungal and other non-animal sequences. Alignment B is phylogenetically less deep (though still deeper than those that have been used for BRCA1 and BRCA2) and tries to exploit sequences from non-vertebrate deuterostomates plus starlet anemone, organisms whose overall gene repertoire is more similar to that of vertebrates than are the genomes of protostomate animals or non-animal eukaryotes. Either way, recognizing the presence of a structural error in the T. nigroviridis sequence illustrates that an initial alignment can be used as a hypothesis test on the individual sequences it contains, providing clues to areas where sequence corrections may improve the alignment.

Table 1

Tools for in silico analysis of missense substitutions

Program name	URL and key reference	Operating principle	Output	Use notes
Align-GVGD	Web server: http://agvgd.iarc.fr/ [Tavtigian et al., 2006]	Combines an alignment with amino acid physico-chemical characteristics to calculate the range of variation present at each position in the alignment (GV) and the distance of missense substitutions from that range of variation (GD).	Two distance measurements and a grade: GV – range of variation GD – distance of a missense substitution from the edge of the range of variation. Grades – C0 to C65 – provide an empirical mapping from GV-GD to genetic risk.	Website has a small library of curated sequence alignments. Users must supply FASTA format alignments for genes not included in the library. True gaps are coded "–" Missing residues are coded "X"
MAPP	Program to download: http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html [Stone and Sidow, 2005]	Combines an alignment with amino acid physico-chemical characteristics to calculate the physico-chemical centroid of each position and the variance between each of the 20 amino acids and that centroid. This is the MAPP impact score.	A many-column table that gives: the physico-chemical characteristics of each position; the MAPP impact score, which is a continuous variable, for all 20 amino acids at each position; and a listing of which amino acids should be deleterious and which should be neutral.	Users must supply FASTA format alignments and a phylogenetic tree describing the relationships between the sequences sampled. True alignment gaps are coded "–" Can be optimized for individual genes [Chao et al., 2008], and users can optimize MAPP impact score thresholds for classification.
PANTHER	Web server: http://www.pantherdb.org/tools/csnpScoreForm.jsp [Thomas and Kejariwal, 2004]	The overall server uses sequence alignments to classify genes by deduced function. One service provided uses pre- built alignments to calculate a Hidden Markov Model based position specific scoring matrix, the subPSEC score.	A 7 column table that includes: subPSEC score, a position specific scoring matrix score; The estimated probability that a substitution is deleterious; and NIC, the balance of prior knowledge versus target gene alignment data used to calculate the subPSEC score.	Users provide their target sequence and a list of missense substitutions. Will generally only output results for well-conserved segments of target proteins, and the user notes point out that most substitutions falling at poorly conserved segments of the target protein will be neutral.
Pmut	Web server: http://mmb2.pcb.ub.es:8080/PMut/ [Ferrer-Costa et al., 2005]	Uses a feed-forward neural network and data from 19 parameters (alignment only) or 23 parameters (alignment + crystal structure) to analyze substitutions.	Analyzes 19 possible missense substitutions against the wild-type. Gives the neural net output score, ranging from 0 to 1; a reliability score which is small for variants with NN scores near 0.5; and a binary prediction of "neutral" vs "pathological".	Need a PDB structure ID to run the version that uses crystal structure data. For the version that takes a user- supplied alignment, true gaps are coded ".“; however, we have never succeeded to run this version of the program.
PolyPhen	Web server: http://coot.embl.de/PolyPhen/ [Sunyaev et al., 2001]	Based on a decision tree that combines a number of protein structural attributes with a pre- built sequence alignment, generally including only Mammalian sequences.	Calculates a PSIC score, which is the difference in fitness between wild type and mutant amino acid, and then converts to a 3 category classification: benign, possibly damaging, probably damaging. Supplementary output shows which data components were available.	User inputs a protein identifier or target sequence. The web portal accepts only one substitution at a time. However, software for mass submission of substitutions is available online.
SIFT	Web server: http://blocks.fhcrc.org/sift/SIFT.html [Ng and Henikoff, 2003]	Uses sequence alignments to create a Dirichlet mixtures-based score matrix for each position in the alignment. The score for each possible amino acid substitution is converted to a normalized probability that the substitution would be evolution- arily tolerated, the SIFT score.	Will output the SIFT score for each individual substitution submitted or all possible substitutions. Also provides the binary classification tolerated/predicted to affect protein function, plus the median sequence conservation from the alignment.	We recommend that users create their own alignments rather than using the auto-build feature. True gaps are coded "–" Missing residues are coded "X". The sequence conservation score provides a useful estimate of whether the alignment contains sufficient variation to support classification.
SNPs3D	Web server: http://www.snps3d.org/ [Yue et al., 2006]	Uses the Support Vector Machine and data from 15 parameters (structure based) or 5 parameters (alignment based) to analyze	Negative svm profiles are indicative of deleterious and positive profiles are indicative of neutral. Scores between −0.5 and +0.5 have reduced confidence. A summary of the	Fails to analyze an appreciable fraction of substitutions entered. If a substitution is on the surface of the protein, then the structural analysis will almost always point towards

Program name	URL and key reference	Operating principle	Output	Use notes
		substitutions. The output score is called the svm profile.	underlying data is available from both structural and alignment-based analyses.	neutral; in this case, the alignment-based analysis should be more reliable.