Utility of gene-specific algorithms for predicting pathogenicity of uncertain gene variants

David K Crockett,1,2 Elaine Lyon,2 Marc S Williams,1,3 Scott P Narus,1 Julio C Facelli,1,4 Joyce A Mitchell1

ABSTRACT

The rapid advance of gene sequencing technologies has produced an unprecedented rate of discovery of genome variation in humans. A growing number of authoritative clinical repositories archive gene variants and disease phenotypes, yet there are currently many more gene variants that lack clear annotation or disease association. To date, there has been very limited coverage of generic-specific predictors in the literature. Here the evaluation is presented of “gene-specific” predictor models based on a naive Bayesian classifier for 20 gene—disease datasets, containing 3986 variants with clinically characterized patient conditions. The utility of gene-specific prediction is then compared with “all-gene” generalized prediction and also with existing popular predictors. Gene-specific computational prediction models derived from clinically curated gene variant disease datasets often outperform established generalized algorithms for novel and uncertain gene variants.

BACKGROUND AND SIGNIFICANCE

Personalized medicine implies that all relevant clinical information is available on demand for effective patient treatment. Proper interpretation of gene test results is a key component in customizing patient therapy. Efforts such as the Human Variome Project, 1000 Genomes, and NCBI Genetic Testing Registry highlight a growing interest in annotation and clinical interpretation of gene variants in human disease.1–3 As genetic information is incorporated into the electronic medical record, new decision support approaches are needed to provide clinicians with a preferred course of treatment.4 For decision support rules to add value, the clinical relevance of laboratory information must be well understood.5 6

Furthermore, with rapidly evolving technologies such as single-nucleotide polymorphism (SNP) chip genome-wide association studies and next-generation sequencing, genomic analysis is trending faster and cheaper and yielding much larger datasets. As such, gene variants are being discovered at an almost astronomical pace, with one recent report finding an average of 3 million variants per personal genome.7 More importantly, for genomic variation to be of real clinical utility, laboratory interpretation and disease association must be well understood for each new gene variant found.8 9

Unfortunately, an increasingly apparent gap exists between the rapidly growing collections of genetic variation and practical clinical implementation. Although collections of human genome variation have been underway for years, authoritative repositories of gene variants with clear association with disease phenotype are only now beginning to emerge.10–14 This is in contrast with existing collections of genome-wide mutations, such as dbSNP15 or OMIM,16 that are not curated using consistent, systematic or transparent methods. Focusing computer predictive algorithms on authoritative and specific gene—disease settings has the potential to bridge this knowledge gap.

Prediction algorithms for computing mutation severity have been used for many years.17–20 Despite their use in laboratories, they do not have sufficient accuracy to predict disease phenotype to the degree necessary to be clinically applicable. This prompts opportunities to explore the application of advanced informatics approaches to this problem.21–23 This study expands the recently reported primary sequence amino acid properties (PSAAP) algorithm,24 25 which uses a gene-specific classification approach utilizing amino acid physicochemical properties of the primary amino acid sequence to predict pathogenicity of novel and uncertain gene variants. To date, gene-specific approaches have been applied only to the RET proto-oncogene and hypertrophic cardiomyopathy.25 26

To evaluate the generalizability of our gene-specific PSAAP algorithm, we extended its use to a set of 20 genes with clinically curated disease variants (table 1). The analyses also compared the effectiveness of generic gene versus gene-specific approaches using a minimum (non-redundant) set of amino acid properties to describe exonic non-synonymous variants coupled with evaluation of overlap and/or trends of biochemical properties of mutation.

METHODS

Gene variant data relating well-characterized patient conditions to genotypes (genotype—phenotype) were assembled from multiple sources including: cystic fibrosis mutation database curated by Ruslan Dorfman (Hospital for Sick Children, Toronto);27 BioPKU database curated by Nenad Blau (University Children’s Hospital, Zurich);28 neurofibromatosis type 1 database curated by Ophelia Maertens (Center for Medical Genetics, University Hospital, Ghent); collagen, type IV, α5 (COL4A5) Mental Retardation Database curated by Judy Savage (Department of Medicine, University of Melbourne) as hosted by Leiden Open Source Variation Database (LOVD);29–31 biotinidase (BTD) curated by Barry Wolf (Medical Genetics, Henry Ford Hospital, Detroit);32 aryl hydrocarbon receptor

Brief communication

J Am Med Inform Assoc 2012;19:207–211. doi:10.1136/amiajnl-2011-000309 207
interacting protein curated by Rodrigo Toledo (Endocrine Genetics Unit, University of Sao Paulo Medical School) (personal communication); Disease Databases hosted by the Department of Pathology, University of Utah School of Medicine33 and genetic testing results archived at ARUP Laboratories (Salt Lake City). The clinically curated gene–disease datasets (n=20) containing some 3986 curated variants are summarized in table 1. This 20-gene collection contained 1659 exonic non-synonymous (ns) SNPs with known outcomes of benign (n=607) and pathogenic (n=1052). The gene variants were characterized using physicochemical properties of the substituted amino acid as recently reported.24, 25 Briefly, gene-specific clinically curated missense variants (nsSNPs) were characterized using a naïve Bayes classification scheme of primary amino acid sequence only and delta differences in physical, chemical, conformational, or energetic properties between the amino acid present in the wild-type and the variant. Descriptors were attributes derived from 544 amino acid properties archived in AIndex v9.4.34 AIndex is a database of numerical indices representing various physicochemical and biochemical properties of amino acids. For each gene variant, vectors of delta values for each biochemical property of the substituted amino acid were calculated and the resulting description was used to calculate the absolute value of the difference between wild-type and mutant—as trained in a gene-specific setting. Based on curated clinical outcomes of benign or pathogenic, the minimum (non-redundant) set of amino acid properties needed to describe pathogenicity of gene variants was investigated using various attribute selection methods such as correlation-based feature subset selection, SVM-RFE and Relief-F and various classifiers. Thresholds of 95% (or 0.95) for Greedy-Stepwise and Ranker were used during this analysis. The best performing correlation-based feature subset selection and naïve Bayes classification were implemented using the Weka software package.35 For each of the 20 genes, random selection was used to build a 2/3 training set and a 1/3 test set with known class labels (benign, pathogenic). Training and test sets were constructed to keep the original ratio of benign and pathogenic constant, but without regard to functional motif or protein location. Next, based on curated clinical classification of benign or pathogenic, algorithm training and pathogenicity prediction were performed gene by gene. Gene-specific models were also tested for prediction of other gene–disease outcomes, by using the training set of one gene and a test set from a second gene. In a similar fashion, an “all-gene” model was constructed using all the available training sets. This all-gene model was then tested by making gene-by-gene predictions. Owing to a low number of nsSNP exonic substitution variants, five genes (MECP2, MSH2, MSH6, PLOD1 and SPINK1) were only included in the all-gene training set, and not used for gene-specific training. Algorithm performance was evaluated using each gene test set, with sensitivity (true positive rate), specificity (true negative rate), and positive predictive value (PPV or precision) calculated for each classifer algorithm and gene-specific and all-gene permutations. Well-established prediction tools such as PolyPhen18 and SIFT17 are primarily based on multiple alignment, and amino acid substitution penalties have been available for many years. More recently, MutPred20 calculates probability of deleterious mutations by disrupted molecular mechanism. Additionally, PMut19 is neural net based and trained on human mutations. A more detailed description of each prediction algorithm is given in the online supplementary data. Lastly, gene-specific algorithm performance was compared with well-established prediction algorithms such as SIFT,17 PolyPhen,18 PMUT19, and MutPred20. Comparison of established prediction tools with gene-specific trained algorithms may increase our understanding of predicting mutation status. For all genes, the full-length protein isoform was used for this study. Splice variants were not considered. All gene variants were mapped to their reference amino acid sequence from UniProtKB (http://www.uniprot.org). Protein reference sequences are summarized in online supplementary table 1.

RESULTS AND DISCUSSION
Overall, the performance of the gene-specific trained algorithm was significantly better (8% to 15%) than the all-gene model,
with p values of 0.00001 (sensitivity), 0.00113 (specificity) and 0.00012 (PPV) as shown in figure 1. For the genes evaluated, the PPV of our gene-specific PSAAP algorithm averaged 89% (82–94%). This was on average 11% higher than the all-gene model, where PPV ranged from 62% to 86%. The one exception was SLC22A5, where PPV remained constant. Sensitivity averaged 13% higher than the all-gene model, except for SPRED1, which was decreased by 6%. Specificity was also generally improved (9% average) for all but PMS2 (no increase) and NF1 (which was decreased by 5%).

For the genes studied here, the PSAAP gene-specific prediction performs well. PPV values are displayed in online supplementary table 2. The self against self is plotted on the diagonal in blue with PPV >80 in bold. Other gene predictor performance with PPV above 80 is shaded in orange. Interestingly, gene-specific prediction models do not seem to generalize well—even across similar protein functional families. For instance, online supplementary table 2 shows that the RET kinase trained model (94% PPV) performed lower for the ACVRL1 kinase (84% PPV), while the ACVRL1 trained predictor (88% PPV) only predicted RET with 80% PPV. Additionally, the carboxylase enzyme BTD (91% PPV) only predicted the hydroxylase PAH gene variant outcome with 76% PPV, while the PAH trained predictor (89% PPV) only predicted BTD with 59% PPV. It is notable, however, that three out of 15 genes (SPRED1, NF1 and GALT) yielded comparable numbers for predicting disease association across other genes.

The improved performance of gene-specific algorithms may be explained in part by an important observation that biochemical and/or structural characteristics of mutation specific to one disease may be lost or diluted when combined with large genome-wide datasets for algorithm development. This can be illustrated by plotting non-synonymous variants specific to a gene—disease condition compared with random amino acid substitutions. When 1000 random amino acid changes were plotted (online supplementary figure 1A), a wide distribution evenly covers the entire range of possible substitutions. In contrast, when 1000 pathogenic mutations are plotted, characteristic trends of specific residues and frequency of substitution are readily seen (online supplementary figure 1B). More importantly, disease-specific examples of this concept are shown in figure 2. In the RET proto-oncogene (associated with medullary thyroid cancers), some 79% of all pathogenic changes were found to involve cysteine (C) to some other residue (X) as displayed in figure 2A. In the COL4A5 gene (associated with Alport syndrome), 84% of pathogenic changes involve glycine (G) to other residues (X) as shown in figure 2B. To confirm this trend, further experiments should be performed as additional curated gene–disease collections become available.
Although the majority of the PSAAP models did not perform as well for predicting pathogenicity in other genes—diseases, most still outperformed established algorithms. As shown in table 2, the majority of genes (13 out of 15) analyzed using the gene-specific PSAAP trained algorithm had improved PPV compared with other algorithms, with the overall PPV increasing by 8.8–22.0%. For example, the PSAAP model specific for SPRED1 (95% PPV as seen in table 2), when analyzed using established prediction algorithms yielded precision scores from 56% to 71%. As mentioned above, the PSAAP model specific for RET kinase (94% PPV) underperformed for the ACVRL1 kinase (84% PPV); however, both still outperformed established algorithms, where on-line predictions for ACVRL1 only ranged from 57% to 81% PPV. Two exceptions to this trend were GALT and SMAD, in which MutPred and/or PMut scored slightly higher as shown in bold in table 2.

It is important to note that the all-gene trained Bayes predictor also compares favorably with established algorithms, with the mean, minimum and maximum PPV for each predictor also summarized in table 2. For instance, although the gene-specific trained PSAAP model yielded the best PPV, the all-gene trained model outscores three of four established predictors, with MutPred being the exception. This observation may highlight the importance of authoritative variant data and amino acid physicochemical properties being used to develop/train algorithms. It also demonstrates that primary acid sequence only, when coupled with amino acid properties, can be successfully used to develop predictor algorithms.

Finally, a minimum attribute set of amino acid properties seems specific to each gene—disease, with overlap found among different genes using three feature selection methods ranging from 11% to 80%, as summarized in online supplementary table 3. Representative examples are shown in figure 3. Interestingly, the gene models with more shared amino acid attributes (GALT, 80%; NF1, 62%; SPRED1, 60%) also had the best generalizability.

**Table 2** Gene-specific and all-gene algorithm PPV compared with established algorithms

<table>
<thead>
<tr>
<th>Gene</th>
<th>PSAAP*</th>
<th>All-gene†</th>
<th>SIFT‡</th>
<th>PolyPhen§</th>
<th>PMut§</th>
<th>MutPred**</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACVRL1</td>
<td>88</td>
<td>77</td>
<td>57</td>
<td>67</td>
<td>69</td>
<td>81</td>
</tr>
<tr>
<td>AIP</td>
<td>91</td>
<td>71</td>
<td>71</td>
<td>73</td>
<td>80</td>
<td>79</td>
</tr>
<tr>
<td>BTD</td>
<td>91</td>
<td>79</td>
<td>77</td>
<td>72</td>
<td>71</td>
<td>87</td>
</tr>
<tr>
<td>CFTR</td>
<td>90</td>
<td>63</td>
<td>68</td>
<td>74</td>
<td>70</td>
<td>89</td>
</tr>
<tr>
<td>COL4A5</td>
<td>88</td>
<td>82</td>
<td>58</td>
<td>74</td>
<td>62</td>
<td>73</td>
</tr>
<tr>
<td>ENG</td>
<td>92</td>
<td>83</td>
<td>62</td>
<td>64</td>
<td>73</td>
<td>65</td>
</tr>
<tr>
<td>GALT</td>
<td>86</td>
<td>77</td>
<td>66</td>
<td>65</td>
<td>58</td>
<td>87</td>
</tr>
<tr>
<td>GJB2</td>
<td>87</td>
<td>77</td>
<td>69</td>
<td>74</td>
<td>67</td>
<td>83</td>
</tr>
<tr>
<td>NF1</td>
<td>89</td>
<td>70</td>
<td>64</td>
<td>70</td>
<td>70</td>
<td>84</td>
</tr>
<tr>
<td>PAH</td>
<td>89</td>
<td>80</td>
<td>59</td>
<td>76</td>
<td>77</td>
<td>84</td>
</tr>
<tr>
<td>PMS2</td>
<td>88</td>
<td>63</td>
<td>64</td>
<td>74</td>
<td>74</td>
<td>72</td>
</tr>
<tr>
<td>RET</td>
<td>94</td>
<td>84</td>
<td>78</td>
<td>54</td>
<td>72</td>
<td>84</td>
</tr>
<tr>
<td>SLC22A5</td>
<td>90</td>
<td>82</td>
<td>74</td>
<td>76</td>
<td>53</td>
<td>82</td>
</tr>
<tr>
<td>SMAD4</td>
<td>84</td>
<td>82</td>
<td>71</td>
<td>70</td>
<td>85</td>
<td>86</td>
</tr>
<tr>
<td>SPRED1</td>
<td>93</td>
<td>86</td>
<td>71</td>
<td>65</td>
<td>56</td>
<td>71</td>
</tr>
<tr>
<td>Mean</td>
<td>89.3</td>
<td>77.1</td>
<td>67.3</td>
<td>69.9</td>
<td>69.1</td>
<td>80.5</td>
</tr>
<tr>
<td>Min</td>
<td>84.0</td>
<td>63.0</td>
<td>57.0</td>
<td>54.0</td>
<td>53.0</td>
<td>65.0</td>
</tr>
<tr>
<td>Max</td>
<td>94.0</td>
<td>86.0</td>
<td>78.0</td>
<td>76.0</td>
<td>85.0</td>
<td>89.0</td>
</tr>
</tbody>
</table>

Bold values indicate the highest PPVs.

*PSAAP algorithm, gene-specific trained.
†PSAAP algorithm, all-gene (n = 20) trained.
‡Analyzed with default settings at http://sift.jcvi.org.
§Analyzed with default settings at http://genetics.bwh.harvard.edu/pph.
**Analyzed with default settings at http://mutdb.org/mutpred.

PPV, positive predictive value; PSAAP, primary sequence amino acid properties.

Of note, both SMAD4 and GALT did well using the established on-line prediction tools, where SMAD4 also had 58% overlap. Without considering the above mentioned four genes, the overlap ranged from only 11% to 57%. Overlap for the all-gene dataset follows the same trend, showing only 38% overlap between the feature selection methods.

**CONCLUSION**

The number of authoritative disease- and locus-specific gene variant collections in use for clinical diagnostics is rapidly growing. These clinically curated gene variant datasets, with reliable genotype–phenotype association, can readily be utilized...
for training and test set performance of machine classifiers. The generalizability of classification rules across multiple genes and diseases may be strengthened as the number of curated disease variants continues to increase, although our analysis suggests that gene-specific approaches will, with few exceptions, outperform generic approaches. Nonetheless, the recognition that the proposed classifier outperforms existing tools is important, given that it will take time for disease-specific curated genotype—phenotype databases to be developed, and, for some ultra-rare diseases, such databases may never be realistic.

For machine learning classifiers, amino acid attributes characterize substitution mutations for a given disease may be lost or diluted when combined with multiple genes and diseases. A key distinguishing feature of this gene-specific classifier methodology is that algorithms are trained explicitly to curated monogenic disease outcomes. While this methodology is complementary to established generalized prediction tools, algorithms should take advantage of authoritative (clinically curated) gene variant collections where they exist. This is especially important when pathologic variants exhibit characteristic trends or properties specific to a given disease.

This study included only gene variant collections with clearly documented disease association and known to the authors—36 and may lead to disease-specific mutation-guided management strategies, appropriate caution is justified when clinicians are asked to trust computational outcomes for determining patient care.56 Continued interaction between clinicians and laboratorians to refine mutation-specific clinical classification is imperative to optimal patient care.

Acknowledgments We gratefully acknowledge the extensive disease curation of gene variants by Drs Dorfman, Blau, Maertens, Savage, Wolf, Toledo and others.

Funding This work has been supported by ARUP Institute for Clinical and Experimental Pathology, National Library of Medicine Training grant (grant number LM007124) and NCRR Clinical and Translational Science Award (grant number 1KL2RR025763-01).

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES