IMPROVING DECISION SUPPORT FOR UNCERTAIN GENE VARIANTS

by

David K. Crockett

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Biomedical Informatics

University of Utah

August 2011

Copyright © David K. Crockett 2011

All Rights Reserved

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **David K. Crockett**

has been approved by the following supervisory committee members:

Joyce A. Mitchell	, Chair	**May 13, 2011** Date Approved
Julio C. Facelli	, Member	**May 13, 2011** Date Approved
Scott P. Narus	, Member	**May 13, 2011** Date Approved
Marc S. Williams	, Member	**May 13, 2011** Date Approved
Elaine Lyon	, Member	**May 13, 2011** Date Approved

and by **Joyce A. Mitchell** , Chair of

the Department of **Biomedical Informatics**

and by Charles A. Wight, Dean of The Graduate School.

ABSTRACT

Rapidly evolving technologies such as chip arrays and next-generation sequencing are uncovering human genetic variants at an unprecedented pace. Unfortunately, this ever growing collection of gene sequence variation has limited clinical utility without clear association to disease outcomes. As electronic medical records begin to incorporate genetic information, gene variant classification and accurate interpretation of gene test results plays a critical role in customizing patient therapy. To verify the functional impact of a given gene variant, laboratories rely on confirming evidence such as previous literature reports, patient history and disease segregation in a family. By definition variants of uncertain significance (VUS) lack this supporting evidence and in such cases, computational tools are often used to evaluate the predicted functional impact of a gene mutation.

This study evaluates leveraging high quality genotype-phenotype disease variant data from 20 genes and 3986 variants, to develop gene-specific predictors utilizing a combination of changes in primary amino acid sequence, amino acid properties as descriptors of mutation severity and Naïve Bayes classification. A Primary Sequence Amino Acid Properties (PSAAP) prediction algorithm was then combined with well established predictors in a weighted Consensus sum in context of gene-specific reference intervals for known phenotypes. PSAAP and Consensus were also used to evaluate known variants of uncertain significance in the *RET* proto-oncogene as a model gene.

The PSAAP algorithm was successfully extended to many genes and diseases. Gene-specific algorithms typically outperform generalized prediction tools. Characteristic mutation properties of a given gene and disease may be lost when diluted into genomewide data sets.

A reliable computational phenotype classification framework with quantitative metrics and disease specific reference ranges allows objective evaluation of novel or uncertain gene variants and augments decision making when confirming clinical information is limited.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

This work has been partially supported by ARUP Institute for Clinical and Experimental Pathology® (Salt Lake City, UT), National Center for Research Resources (NCRR) Clinical and Translational Science Award #1KL2RR025763-01 and National Library of Medicine (NLM) Training grant #LM007124. Open Access permission for previously published material from the *Journal of Data Mining in Genomics and Proteomics* and *PLoS ONE* is acknowledged. Lastly, I gratefully acknowledge the guidance, support, mentoring and expertise of my graduate committee. Their enthusiasm for excellence in research and writing has been graciously shared.

CHAPTER 1

INTRODUCTION

Medical genetics involves diagnosis, management, and determining risk of hereditary disorders.[1, 2] The genotype-phenotype correlation of gene variants in disease is a major component of medical genetics. In monogenic diseases, gene mutations are typically curated as either pathogenic or benign. However, many gene variants must be classified as “unknown” or “uncertain” significance because they have not been clearly associated with a clinical phenotype.[3] Accurate interpretation of gene testing is a key component in customizing patient therapy.

The investment of time and labor to validate disease association of a variant of uncertain significance (VUS) within the coding portion of a gene can be daunting and cost prohibitive.[4, 5] This is due in part to the nature of rare mutations, the lack of available functional assays or lack of power from a limited number of family studies, as well as the interactions between clinicians and laboratory geneticists needed to characterize novel gene variants.[6, 7] To help bridge this genotype-phenotype gap, the use of machine learning classification algorithms to narrow the uncertain “grey area” between pathogenic and benign sequence variants warrants careful evaluation.[8-11] Reliable machine-based classification may augment costly patient recruitment, family histories, and biochemical confirmation of a gene variant with no associated disease correlation.[12-14]

Machine learning is common to many industries with well known applications including speech recognition, internet search engines and detecting credit card fraud.[15, 16] In medicine and biology, computers often assist in tasks such as natural language processing, medical records and DNA sequence alignment.[17, 18]

Rapidly evolving technologies such as SNP chip genomewide association studies and next-generation sequencing has lowered the cost and increased the speed of genomic analysis yielding much larger data sets. Currently, gene variants are being discovered at an unprecedented pace. One recent report found an average of 3 million variants per personal genome.[19] Unfortunately, an ever-widening gap exists between this fast growing collection of genetic variation and practical clinical implementation due to a lack of understanding of the phenotypic consequences (if any) of any given variant.[20, 21]

Early efforts to develop clinical laboratory systems date back nearly 50 years, with electronic processing and reporting of laboratory test results.[22, 23] Additional efforts continued to improve and refine computer information systems used in laboratory settings and to enhance patient care.[24, 25] As genetic information began to influence patient care, laboratory information systems again required adjustments.[26, 27] Currently, the vast amount of genomic data on the horizon for medical records and patient treatment highlights increasing opportunities for genomic data decision support in the laboratory setting. Furthermore, where traditional decision support augments choosing the correct decision or task, moving toward a more complete environment of gene variant information that supports cognition in context of problem and workflow, may enhance accurate gene test interpretation and necessary laboratory recommendations - prior to a clinician acting in appropriate patient care.[28-30]

Recent endeavors such as the NCBI Genetic Testing Registry, 1000 Genomes and the Human Variome Project highlight this growing interest in gene variant annotation and clinical interpretation in human disease.[31-33] Over the past several decades, literature reports of genotype-to-phenotype (G2P) correlation have grown exponentially. Figure 1.1 shows nearly 17,000 such studies reported in PubMed in the last decade alone.

A given gene variant may be commonly referred to as a single nucleotide polymorphism (SNP) or single nucleotide variant (SNV). More specifically, a nonsynonymous SNP (nsSNP) refers to a point mutation or change in amino acid sequence as compared to a wild type or reference sequence.

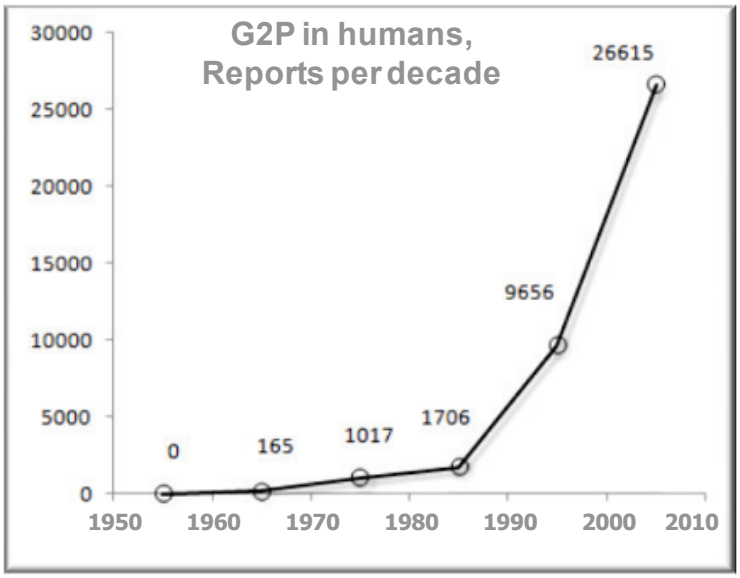


Figure 1.1 Growing wave of scientific literature dealing with gene variant information (genotype) and disease association (phenotype), from PubMed (March 2011) with search terms “genotype phenotype” and filtered to “human” and “English” only.

The fact that certain nsSNPs are causative of disease is well known.[34] Thus, investigating SNP functional effect has been ongoing for many years.[35-37] Due to the cost, labor and expertise required for wet-bench molecular evaluation, much of this effort has been done using computational tools. These prediction tools often focus on SNPs in protein coding regions that change one amino acid for another.[36] The severity of a given amino acid sequence change may range from mild to severe, and has been reported to impact various medical areas including genetic disease susceptibility such as sickle cell anemia, common disease risks like Alzheimer’s or drug sensitivities as seen in warfarin treatment.[38] It is not surprising then, that physical and chemical properties of amino acids have been used historically as a proxy to assess the functional impact of these substitution mutations.[35]

Early efforts in predicting amino acid substitution effects were based on metrics of estimating expected evolutionary distance between each possible amino acid pair. One of the first amino acid substitution matrices, Point Accepted Mutation (PAM), approximated the evolutionary

distance and frequency of amino acids for equivalent protein positions in closely related species.[39, 40] Later, Blocks of Amino Acid Substitution Matrix (BLOSUM) included more distantly related species, but only considered highly conserved protein regions.[41] Interestingly, both approaches used raw mutation rates to compute a score for each amino acid substitution and calculating likelihood that the mutation was caused by an evolutionary change (over time), rather than by sheer chance. The assumption was that substitutions more consistent with evolutionary trends conserved across many species are less likely to disrupt protein function. Conversely, substitutions not consistent with evolution (nonconserved substitutions) were more likely associated with disease.

Alternative approaches utilizing amino acid properties have considered how physiochemical properties differ with changes in volume, hydrophobicity, net charge, packing density and solvent accessibility all shown to correlate with predicted functional impact of SNPs.[37] A representative method, the Grantham distance, combines both biophysical properties and evolutionary distance where the significance of the amino acid substitution was quantified in a 3D space as weighted Euclidean distance of side chain composition, polarity and volume as coordinates.[42] Weights were modeled to estimated amino acid substitution mutation rates.[43] Importantly, while PAM and BLOSUM matrix likelihood scores or biophysical properties changes seen as large Grantham distances may be able to predict effects across large populations of SNPs, these computational metrics were not sufficiently accurate for predictions of specific and individual SNPs.[44, 45] One recent report used a combined conservation score of 16 amino acid properties as descriptors of mutation to improve the prediction of T4 lysozyme missense mutations.[46]

Subsequent efforts then focused on the fact that the importance of the evolutionary distance separating a pair of amino acids depends on the position where an amino acid substitution occurs.[44] More specifically, amino acid distribution at equivalent positions in a protein family is functionally or structurally important, where these positions may not tolerate a variety of amino acid changes. These equivalent positions were found by constructing an alignment from multiple related protein sequences. Thus, amino acid residues in highly

conserved alignment regions were assumed to be under some purifying evolutionary selection and important for normal protein function. Algorithms were designed to quantify this conserved evolutionary selection in protein activity, such as calculating the frequency of the most common amino acid in an alignment column. For example, Shannon entropy computes the distribution of all amino acids at a specific aligned position.[47] This idea was further improved by using relative entropy to augment comparing Shannon entropy of a conserved alignment against the Shannon entropy of the amino acid background distribution.

Several resulting algorithms and prediction scoring included both physicochemical properties of amino acid substitution and evolutionary conservation. For example, the Sorting Intolerant From Tolerant (SIFT) algorithm computes a weighted frequency average of which amino acid residue appears in a multiple alignment position, coupled with an estimate of unobserved variant frequencies.[48] The Position-Specific Independent Counts (PSIC) profile score considered the difference of likelihood between reference and variant amino acid at a given aligned position using a position-specific scoring matrix (PSSM).[49, 50] Another example is Align-GVGD (AGVGD), an extension of the original Grantham difference (GD) to multiple sequence alignments and true simultaneous multiple comparisons, where the Grantham variation (GV) is computed by replacing each value-pair of a given amino acid residue component for composition, polarity and charge with the maximum and minimum value in that alignment position.[51] A final example, the Multivariate Analysis of Protein Polymorphism (MAPP) score constructs a statistical summary of an alignment column by use of phylogenetic tree and tree topology weighting each sequence by branch length.[52, 53]

Lastly, it is important to mention algorithms for protein structure–function relationships. Where a nsSNP of interest can be mapped onto a known 3D protein structure, it is possible to determine several properties useful in predicting functional impact of amino acid substitution. Solvent accessibility of an amino acid is one of the strongest predictors of functional impact, where substituting various amino acid residues may disrupt the hydrophobic core of a soluble protein. Structural modeling of disease proteins can be used to determine whether a nsSNP results in backbone strain or leads to overpacking.[54] Importantly, a large number of X-ray

crystal structures have been determined which often include protein interacting partners, and/or small molecule, peptide ligands or inhibitors. The ability to locate a nsSNP on a computational protein structure also makes it possible to evaluate whether the amino acid substitution occurs in or near a binding or catalytic site or at a domain–domain interface of protein interaction. One popular example of an algorithm taking advantage of structural modeling is Polymorphism Phenotyping (PolyPhen). This is an automated tool for evaluating possible impact of amino acid substitution on the structure and function of a human protein that uses a Dictionary of Secondary Structure in Proteins (DSSP) to map a given substitution site to known protein 3D structures.[14]

Although clinicians rely on patient history, family segregation, literature review and trusted colleagues to stay informed of the phenotypic consequences of a given gene variant, when traditional evidence is lacking, well-established machine learning or computational tools are also employed – both for prediction and to assess likelihood.[3, 55-57] Established methods for predicting mutation severity based on amino acid substitution penalties, structural disruption, or sequence homology (ortholog conservation) include tools such as PolyPhen [14], SIFT [48], MutPred [10] and PMut [58]. Efforts such as dbNFSP have also been reported to archive multiple predictors into a single resource.[59] However, established algorithms will not always complete the prediction – due to lack of adequate homolog sequence alignment or availability of a solved protein structure. Furthermore, predictor results are not always in agreement with curated data or each other.[60-62] Thus, there are opportunities to explore the use of other informatics approaches to this problem.

An obvious key to improving prediction algorithms is finding the most authoritative source of gene variant data with clear association to disease outcomes. Optimal training sets can then be developed for use in machine learning efforts. Recent literature indicates that gene test reporting of uncertain gene variants range widely. One laboratory reported that between 30% to 50% of sequence variants reported for *BRCA1* and *BRCA2* were reported as variants of uncertain significance.[63] A second laboratory reported test orders for *BRCA1* and *BRCA2* had an equal chance (13%) of receiving an uncertain variant result as seeing a report for a known pathogenic gene variant.[64] More recent data indicate that reports of uncertain gene variants have

continued to decline to approximately 5% of *BRCA* tests performed, thus, highlighting the importance of maintaining and updating variant databases.[65]

Although collections of human genome variation have been underway for years, authoritative repositories of gene variants with clear association to disease phenotype are only now beginning to emerge.[66-70] This is in contrast to existing collections of genome-wide mutations such as dbSNP[71] or OMIM[72] that are not curated using consistent, systematic or transparent methods. As well-curated locus-specific databases of disease causing gene variants become more widely available and relied upon by clinical laboratories, there may be opportunities to improve prediction algorithms in a gene-specific manner - without dependence on multiple species conservation or solved protein crystal structure. Examples of this gene-disease specific focus using computational prediction have recently been shown for hypertrophic cardiomyopathy and in the *RET* proto-oncogene.[73-75]

When confirming evidence is obvious lacking, how does the laboratory decide what disease classification to report for a gene variant? The key process is gathering and evaluating the strength of evidence for related disease association for a gene variant in question – distancing the variant from uncertain significance to a confirmed phenotype. Figure 1.2 displays this concept. While a conclusive laboratory gene test result will prompt appropriate treatment for a patient, an inconclusive gene test interpretation may leave the clinician or patient with indecision and/or a frustrating lack of treatment options.

Guidelines and terminology for improved classification of gene variants have been recently proposed.[3, 76] Clinician frustration and obstacles to wide adoption of proposed guidelines may include the lack of a quantitative metric or standardized scale for evaluation of novel or uncertain gene variants.

A closely related challenge is an objective and standardized context or framework to make that metric meaningful. This “strength of evidence framework” becomes especially critical for interpretation of uncertain gene variants where there is an obvious lack of existing evidence. A reliable computational phenotype classification framework with a quantitative metric for evaluation of novel and uncertain gene variants can augments limited clinical information.

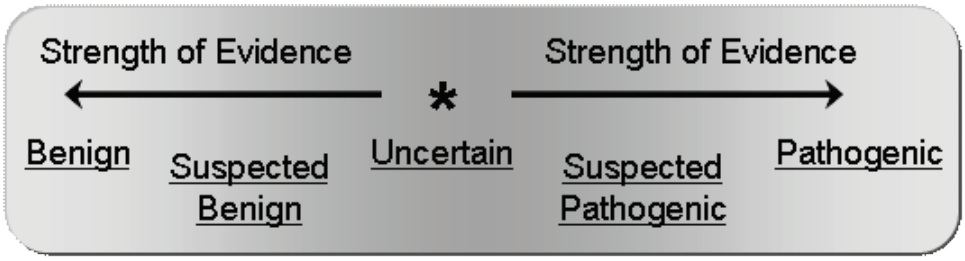


Figure 1.2 Gathering strength of evidence for gene variant classification through multiple literature reports, functional studies, genetic concordance studies, family history, clinical details, consulting colleagues, on-line databases and computational tools.

Hypothesis

Leveraging authoritative gene-disease collections and amino acid physicochemical properties as descriptors of wildtype and mutant, machine learning classification can be used to train gene-specific algorithms that outperform generic prediction tools. Integration into a standardized framework and quantitative metric for objective evaluation of uncertain gene variants may augment laboratory classification of gene test results.

Specific aims

Laboratory decision support for gene variant classification can be improved by:

Specific Aim 1. Primary protein sequence and biochemical properties of amino acid residues can be used as descriptors of differences between wild type and variant. Using the *RET* proto-oncogene as a model, machine-learning classification algorithms will be evaluated for their ability to distinguish benign and pathogenic gene variants as characterized by differences in values of physicochemical properties of the amino acid residue present in the wild type and mutant. Representative algorithms will be chosen from different categories of machine learning classification techniques, including rules, bayes, regression, nearest neighbor, support vector machines and trees. Machine-learning models will then be compared to well-established techniques used for mutation severity prediction.

Specific Aim 2. Machine learning classification can be used to predict pathogenicity of uncertain gene variants in authoritative gene-disease collections.

Reported gene variants in the *RET* proto-oncogene have been directly associated with multiple endocrine neoplasia type 2 and hereditary medullary thyroid carcinoma, yet some 46 non-synonymous variants of uncertain significance (VUS) exist in curated archives. In the absence of a reliable method for predicting phenotype outcomes, feature selected amino acid physical and chemical properties feeding a Bayes classifier will be used to predict disease association of uncertain gene variants into categories of benign and pathogenic. Algorithm performance and VUS predictions will be compared to established phylogenetic based mutation prediction algorithms. Gene-specific prediction will also be extended into 20 gene-disease data sets, containing 3,986 well characterized variants.

Specific Aim 3. A combined score of complementary predictors can be computed in a standard framework of gene-specific disease outcomes. Although proposed guidelines have recommended classification terminology and definitions for improving laboratory gene variant reporting, a standardized framework does not yet exist for quantitative evaluation of disease association for uncertain gene variants in an objective manner. Gene-specific prediction will be trained using clinically curated gene-disease data sets and implemented into a Consensus framework. This prediction model will include a weighted metric of existing and complementary prediction algorithms and calculated reference intervals from known disease outcomes specific to each gene.

Conclusions

As medical records increasingly incorporate genetic test information, improved decision support approaches are needed to provide clinicians with the preferred course of treatment.[77] Furthermore, for decision support rules to be of value, the clinical relevance of laboratory information must be well understood. Recent attention has focused on providing “on demand” gene variant information in medicine.[78] One-click interpretation or additional information available through such methods as “info buttons” and graphical summaries may augment clinical decision making.[78, 79]

Towards this goal, high quality genotype-phenotype disease variant data was leveraged to develop a gene-specific predictor utilizing a combination of primary amino acid sequence,

amino acid properties as descriptors of mutation severity and Naïve Bayes classification based on authoritative training sets. This Primary Sequence Amino Acid Properties (PSAAP) prediction algorithm was then used to evaluate known variants of uncertain significance in the *RET* proto-oncogene as a model. Where traditional confirming evidence was lacking, a weighted metric of PSAAP with other established and complementary predictors was computed for objective VUS evaluation. Finally, the Consensus interpretation framework was implemented using laboratory reference intervals of known disease outcomes for scoring uncertain variants to better communicate the gathered computational evidence to clinical decision makers. Data and methods used for this study were approved by the University of Utah Institutional Review Board (IRB #00035757).

References

1. Weinstein ND: **What does it mean to understand a risk? Evaluating risk comprehension.** *J Natl Cancer Inst Monogr* 1999**:**15-20.

2. Ensenauer RE, Michels VV, Reinke SS: **Genetic testing: practical, ethical, and counseling considerations.** *Mayo Clin Proc* 2005, **80:**63-73.

3. Richards CS, Bale S, Bellissimo DB, Das S, Grody WW, Hegde MR, Lyon E, Ward BE: **ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007.** *Genet Med* 2008, **10:**294-300.

4. Nowak R: **Genetic testing set for takeoff.** *Science* 1994, **265:**464-467.

5. Machens A, Gimm O, Hinze R, Hoppner W, Boehm BO, Dralle H: **Genotype-phenotype correlations in hereditary medullary thyroid carcinoma: oncological features and biochemical properties.** *J Clin Endocrinol Metab* 2001, **86:**1104-1109.

6. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, Dudley JT, Ormond KE, Pavlovic A, Morgan AA, et al: **Clinical assessment incorporating a personal genome.** *Lancet* 2010, **375:**1525-1535.

7. Tchernitchko D, Goossens M, Wajcman H: **In silico prediction of the deleterious effect of a mutation: proceed with caution in clinical genetics.** *Clin Chem* 2004, **50:**1974-1978.

8. Wei Q, Wang L, Wang Q, Kruger WD, Dunbrack RL, Jr.: **Testing computational prediction of missense mutation phenotypes: functional characterization of 204 mutations of human cystathionine beta synthase.** *Proteins* 2010, **78:**2058-2074.

9. Kumar P, Henikoff S, Ng PC: **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.** *Nat Protoc* 2009, **4:**1073-1081.

10. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P: **Automated inference of molecular mechanisms of disease from amino acid substitutions.** *Bioinformatics* 2009, **25:**2744-2750.

11. Dorfman R, Nalpathamkalam T, Taylor C, Gonska T, Keenan K, Yuan XW, Corey M, Tsui LC, Zielenski J, Durie P: **Do common in silico tools predict the clinical consequences of amino-acid substitutions in the CFTR gene?** *Clin Genet* 2010, **77:**464-473.

12. Ferreira-Gonzalez A, Teutsch S, Williams MS, Au SM, Fitzgerald KT, Miller PS, Fomous C: **US system of oversight for genetic testing: a report from the Secretary's Advisory Committee on Genetics, Health and Society.** *Per Med* 2008, **5:**521-528.

13. Williams MS: **Quality in clinical genetics.** *Am J Med Genet C Semin Med Genet* 2009, **151C:**175-178.

14. Ramensky V, Bork P, Sunyaev S: **Human non-synonymous SNPs: server and survey.** *Nucleic Acids Res* 2002, **30:**3894-3900.

15. Cristianini N: **Are we there yet?** *Neural Netw* 2010, **23:**466-470.

16. Biafore S: **Real-time fraud detection. Systems with predictive analytics can catch fraud and abuse before they happen.** *Healthc Inform* 2003, **20:**82.

17. Bhaskar H, Hoyle DC, Singh S: **Machine learning in bioinformatics: a brief survey and recommendations for practitioners.** *Comput Biol Med* 2006, **36:**1104-1125.

18. Rodriguez-Esteban R: **Biomedical text mining and its applications.** *PLoS Comput Biol* 2009, **5:**e1000597.

19. Moore B, Hu H, Singleton M, Reese MG, De La Vega FM, Yandell M: **Global analysis of disease-related DNA sequence variation in 10 healthy individuals: Implications for whole genome-based clinical diagnostics.** *Genet Med* 2011, **13:**210-217.

20. Li C: **Personalized medicine - the promised land: are we there yet?** *Clin Genet* 2011, **79:**403-412.

21. Angrist M: **Only connect: personal genomics and the future of American medicine.** *Mol Diagn Ther* 2010, **14:**67-72.

22. Lindberg DA: **Electronic processing and transmission of clinical laboratory data.** *Mo Med* 1965, **62:**296-302.

23. Lindberg DA: **Symposium on information science. VII. Electronic reporting, processing, and retrieval of clinical laboratory data.** *Bacteriol Rev* 1965, **29:**554-559.

24. Litzkow L, Ingram W, 3rd, Lezotte D: **The evolution of a functional real-time laboratory records retrieval and archival system.** *J Med Syst* 1977, **1:**177-186.

25. Pryor LR, Freeman VD: **An archival system for clinical laboratory data.** *Am J Clin Pathol* 1979, **72:**1013-1017.

26. Loughman WD, Mitchell JA, Mosher DC, Epstein CJ: **GENFILES: a computerized medical genetics information network. I. An overview.** *Am J Med Genet* 1980, **7:**243-250.

27. Mitchell JA, Loughman WD, Epstein CJ: **GENFILES: a computerized medical genetics information network. II. MEDGEN: the clinical genetics system.** *Am J Med Genet* 1980, **7:**251-266.

28. Weir CR, Nebeker JR: **Critical issues in an electronic documentation system.** *AMIA Annu Symp Proc* 2007**:**786-790.

29. Weir CR, Nebeker JJ, Hicken BL, Campo R, Drews F, Lebar B: **A cognitive task analysis of information management strategies in a computerized provider order entry environment.** *J Am Med Inform Assoc* 2007, **14:**65-75.

30. Kushniruk AW, Santos SL, Pourakis G, Nebeker JR, Boockvar KS: **Cognitive analysis of a medication reconciliation tool: applying laboratory and naturalistic approaches to system evaluation.** *Stud Health Technol Inform* 2011, **164:**203-207.

31. Javitt G, Katsanis S, Scott J, Hudson K: **Developing the blueprint for a genetic testing registry.** *Public Health Genomics* 2010, **13:**95-105.

32. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467:**1061-1073.

33. Cotton RG, Al Aqeel AI, Al-Mulla F, Carrera P, Claustres M, Ekong R, Hyland VJ, Macrae FA, Marafie MJ, Paalman MH, et al: **Capturing all disease-causing mutations for clinical and research use: toward an effortless system for the Human Variome Project.** *Genet Med* 2009, **11:**843-849.

34. Chen R, Davydov EV, Sirota M, Butte AJ: **Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association.** *PLoS One* 2010, **5:**e13574.

35. Cline M, Karchin R: **Using bioinformatics to predict the functional impact of SNVs.** *Bioinformatics* 2010.

36. Yue P, Moult J: **Identification and analysis of deleterious human SNPs.** *J Mol Biol* 2006, **356:**1263-1274.

37. Wang Z, Moult J: **SNPs, protein structure, and disease.** *Hum Mutat* 2001, **17:**263-270.

38. McClellan J, King MC: **Genetic heterogeneity in human disease.** *Cell* 2010, **141:**210-217.

39. McLaughlin PJ, Dayhoff MD: **Eukaryotes versus prokaryotes: an estimate of evolutionary distance.** *Science* 1970, **168:**1469-1471.

40. Schwartz RM, Dayhoff MO: **Origins of prokaryotes, eukaryotes, mitochondria, and chloroplasts.** *Science* 1978, **199:**395-403.

41. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci U S A* 1992, **89:**10915-10919.

42. Grantham R: **Amino acid difference formula to help explain protein evolution.** *Science* 1974, **185:**862-864.

43. McLachlan AD: **Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c 551.** *J Mol Biol* 1971, **61:**409-424.

44. Ng PC, Henikoff S: **Predicting deleterious amino acid substitutions.** *Genome Res* 2001, **11:**863-874.

45. Karchin R, Kelly L, Sali A: **Improving functional annotation of non-synonomous SNPs with information theory.** *Pac Symp Biocomput* 2005**:**397-408.

46. Lee TC, Lee AS, Li KB: **Incorporating the amino acid properties to predict the significance of missense mutations.** *Amino Acids* 2008, **35:**615-626.

47. Schneider TD: **Information content of individual genetic sequences.** *J Theor Biol* 1997, **189:**427-441.

48. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function.** *Nucleic Acids Res* 2003, **31:**3812-3814.

49. Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN: **PSIC: profile extraction from sequence alignments with position-specific counts of independent observations.** *Protein Eng* 1999, **12:**387-394.

50. Gribskov M, McLachlan AD, Eisenberg D: **Profile analysis: detection of distantly related proteins.** *Proc Natl Acad Sci U S A* 1987, **84:**4355-4358.

51. 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.

52. Stone EA, Cooper GM, Sidow A: **Trade-offs in detecting evolutionarily constrained sequence by comparative genomics.** *Annu Rev Genomics Hum Genet* 2005, **6:**143-164.

53. Stone EA, Sidow A: **Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity.** *Genome Res* 2005, **15:**978-986.

54. 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.

55. Chapman WW, Aronsky D, Fiszman M, Haug PJ: **Contribution of a speech recognition system to a computerized pneumonia guideline in the emergency department.** *Proc AMIA Symp* 2000**:**131-135.

56. Bayrak-Toydemir P, McDonald J, Mao R, Phansalkar A, Gedge F, Robles J, Goldgar D, Lyon E: **Likelihood ratios to assess genetic evidence for clinical significance of uncertain variants: hereditary hemorrhagic telangiectasia as a model.** *Exp Mol Pathol* 2008, **85:**45-49.

57. Malovini A, Nuzzo A, Ferrazzi F, Puca AA, Bellazzi R: **Phenotype forecasting with SNPs data through gene-based Bayesian networks.** *BMC Bioinformatics* 2009, **10 Suppl 2:**S7.

58. 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.

59. Liu X, Jian X, Boerwinkle E: **dbNSFP: a lightweight database of human non-synonymous SNPs and their functional predictions.** *Hum Mutat* 2011.

60. Spencer DS, Stites WE: **The M32L substitution of staphylococcal nuclease: disagreement between theoretical prediction and experimental protein stability.** *J Mol Biol* 1996, **257:**497-499.

61. Kang HH, Williams R, Leary J, Ringland C, Kirk J, Ward R: **Evaluation of models to predict BRCA germline mutations.** *Br J Cancer* 2006, **95:**914-920.

62. Engelhardt BE, Jordan MI, Muratore KE, Brenner SE: **Protein molecular function prediction by Bayesian phylogenomics.** *PLoS Comput Biol* 2005, **1:**e45.

63. Gomez-Garcia EB, Ambergen T, Blok MJ, van den Wijngaard A: **Patients with an unclassified genetic variant in the BRCA1 or BRCA2 genes show different clinical features from those with a mutation.** *J Clin Oncol* 2005, **23:**2185-2190.

64. Frank TS, Deffenbaugh AM, Reid JE, Hulick M, Ward BE, Lingenfelter B, Gumpper KL, Scholl T, Tavtigian SV, Pruss DR, Critchfield GC: **Clinical characteristics of individuals with germline mutations in BRCA1 and BRCA2: analysis of 10,000 individuals.** *J Clin Oncol* 2002, **20:**1480-1490.

65. Saam J, Burbidge L, Bowles K, Roa B, Pruss D, Schaller J, Reid J, Frye C, Hall MJ, Wenstrup RJ: **Decline in rate of BRCA1/2 variants of uncertain significance: 2002–2008. .** In *National Society of Genetic Counselors Annual Meeting; Los Angeles, CA*. 2008

66. Thony B, Blau N: **Mutations in the BH4-metabolizing genes GTP cyclohydrolase I, 6-pyruvoyl-tetrahydropterin synthase, sepiapterin reductase, carbinolamine-4a-dehydratase, and dihydropteridine reductase.** *Hum Mutat* 2006, **27:**870-878.

67. Calderon FR, Phansalkar AR, Crockett DK, Miller M, Mao R: **Mutation database for the galactose-1-phosphate uridyltransferase (GALT) gene.** *Hum Mutat* 2007, **28:**939-943.

68. Margraf RL, Crockett DK, Krautscheid PM, Seamons R, Calderon FR, Wittwer CT, Mao R: **Multiple endocrine neoplasia type 2 RET protooncogene database: repository of MEN2-associated RET sequence variation and reference for genotype/phenotype correlations.** *Hum Mutat* 2009, **30:**548-556.

69. Crockett DK, Pont-Kingdon G, Gedge F, Sumner K, Seamons R, Lyon E: **The Alport syndrome COL4A5 variant database.** *Hum Mutat* 2010, **31:**E1652-1657.

70. Li W, Sun L, Corey M, Zou F, Lee S, Cojocaru A, Taylor C, Blackman S, Stephenson A, Sandford A, et al: **Understanding the population structure of North American patients with cystic fibrosis.** *Clin Genet* 2011, **79:**136-146.

71. **Single Nucleotide Polymorphism Database.** [ncbi.nlm.nih.gov/projects/SNP/].

72. **Online Mendelian Inheritance in Man.** [ncbi.nlm.nih.gov/omim/].

73. Crockett DK, Piccolo SR, Narus SP, Mitchell JA, Facelli JC: **Computational Feature Selection and Classification of RET Phenotypic Severity.** *J Data Mining in Genom Proteomics* 2010, **1:**1-4.

74. Jordan DM, Kiezun A, Baxter SM, Agarwala V, Green RC, Murray MF, Pugh T, Lebo MS, Rehm HL, Funke BH, Sunyaev SR: **Development and validation of a computational**

method for assessment of missense variants in hypertrophic cardiomyopathy. *Am J Hum Genet* 2011, **88:**183-192.

75. Crockett DK, Piccolo SR, Ridge PG, Margraf RL, Lyon E, Williams MS, Mitchell JA: **Predicting phenotypic severity of uncertain gene variants in the RET proto-oncogene.** *PLoS One* 2011, **6:**e18380.

76. Goldgar DE, Easton DF, Byrnes GB, Spurdle AB, Iversen ES, Greenblatt MS: **Genetic evidence and integration of various data sources for classifying uncertain variants into a single model.** *Hum Mutat* 2008, **29:**1265-1272.

77. Hoffman MA: **The genome-enabled electronic medical record.** *J Biomed Inform* 2007, **40:**44-46.

78. Marshall E: **Human genome 10th anniversary. Human genetics in the clinic, one click away.** *Science* 2011, **331:**528-529.

79. Whiting PF, Sterne JA, Westwood ME, Bachmann LM, Harbord R, Egger M, Deeks JJ: **Graphical presentation of diagnostic information.** *BMC Med Res Methodol* 2008, **8:**20.

CHAPTER 2

FEATURE SELECTION AND CLASSIFICATION OF GENE VARIANT PHENOTYPE

*(*Reprinted with open access permission from Crockett DK, Piccolo SR, Narus SP, Mitchell JA, Facelli, JC. **Computational feature selection and classification of RET phenotypic severity.** *Journal of Data Mining in Genomics and Proteomics* 2010.1:103. doi:10.4172/2153-0602.1000103.)

Abstract

Although many reported mutations in the *RET* oncogene have been directly associated with hereditary thyroid carcinoma, other mutations are labeled as uncertain gene variants because they have not been clearly associated with a clinical phenotype. The process of determining the severity of a mutation is costly and time consuming. Informatics tools and methods may aid to bridge this genotype-phenotype gap. Towards this goal, machine-learning classification algorithms were evaluated for their ability to distinguish benign and pathogenic *RET* gene variants as characterized by differences in values of physicochemical properties of the residue present in the wild type and the one in the mutated sequence. Representative algorithms were chosen from different categories of machine learning classification techniques, including rules, bayes, regression, nearest neighbor, support vector machines and trees. Machine-learning models were then compared to well-established techniques used for mutation severity prediction. Machine-learning classification can be used to accurately predict *RET* mutation status using primary sequence information only. Existing algorithms that are based on sequence homology (ortholog conservation) or protein structural data are not necessarily superior.

Introduction

Accurate prediction of the functional severity for uncertain variants and novel mutations as relating to disease is of great importance to medicine and biology. Bridging the genotype-phenotype gap for uncertain gene variants and novel mutations provides a prime opportunity for application of informatics methods. The process of determining the severity of a mutation is costly and time consuming and informatics tools and methods may aid to bridge this genotype-phenotype gap. If proven sufficiently reliable, it may ultimately be possible to use these methods as diagnostic tools. At a minimum they can help to prioritize the studies of the mutations more likely associated with severe prognosis.

There are established methods for predicting mutation severity based on substitution penalties, structural disruption, or sequence homology (ortholog conservation), such as PolyPhen [1], SIFT [2] and MutPred [3]. However, prediction algorithms are not always in agreement with curated data or each other [4-6]. Thus, there are opportunities to explore the use of other

informatics approaches to this problem. Machine learning methods that can be trained on data available in well-curated gene variant collections are promising tools to improve the predictive capabilities available to the research community. While many existing models to predict severity of mutations are based on sequence similarities based on phylogenetic arguments, this approach attempts to use physicochemical properties of amino acids. Numerical values for amino acid properties have been previously reported as descriptors for classification[7, 8]. Our assumption is that because the physicochemical properties of amino acids define their binding properties, they may be better descriptors of the differences between wild type and mutant.

The *RET* oncogene is located on chromosome 10q11, with 21 exons coding a full length protein of 1,114 amino acids. Conserved functional domains found within the protein (RET_HUMAN, http://www.uniprot.org/uniprot/P07949) include a signal peptide, cadherin repeat domains, transmembrane domain, and protein tyrosine kinase [9]. Mutations in the *RET* oncogene (REarranged during Transfection; OMIM# 164761) have been directly associated with Multiple Endocrine Neoplasia type 2 (MEN2), a hereditary thyroid carcinoma syndrome [10, 11]. Although well known mutations often guide patient therapy and surgical options [12], other *RET* sequence mutations vary in functional severity. Some are pathogenic, some are benign, and some are of unknown significance. Curated *RET* oncogene mutations for MEN2 have been recently reported, many of which have documented phenotype outcomes [13]. Figure 2.1 displays reported disease causing variants as associated with different MEN2 phenotypes.

Table 2.1 summarizes mutation-guided therapy for thyroid cancer where surgical removal of thyroid is guided by codon position of the *RET* mutation.

Table 2.1 *RET* mutation guided therapy for surgical removal of the thyroid.[a]

RET Codon	Thyroidectomy	Phenotype
883, 918	within first 6 months	MEN 2B
609, 611, 618, 620, 630, or 634	within first 5 years	MEN 2A
768, 790, 804, or 891	within 5 - 10 years	FMTC

[a] Guidelines from 7th International Workshop on MEN2. [14]

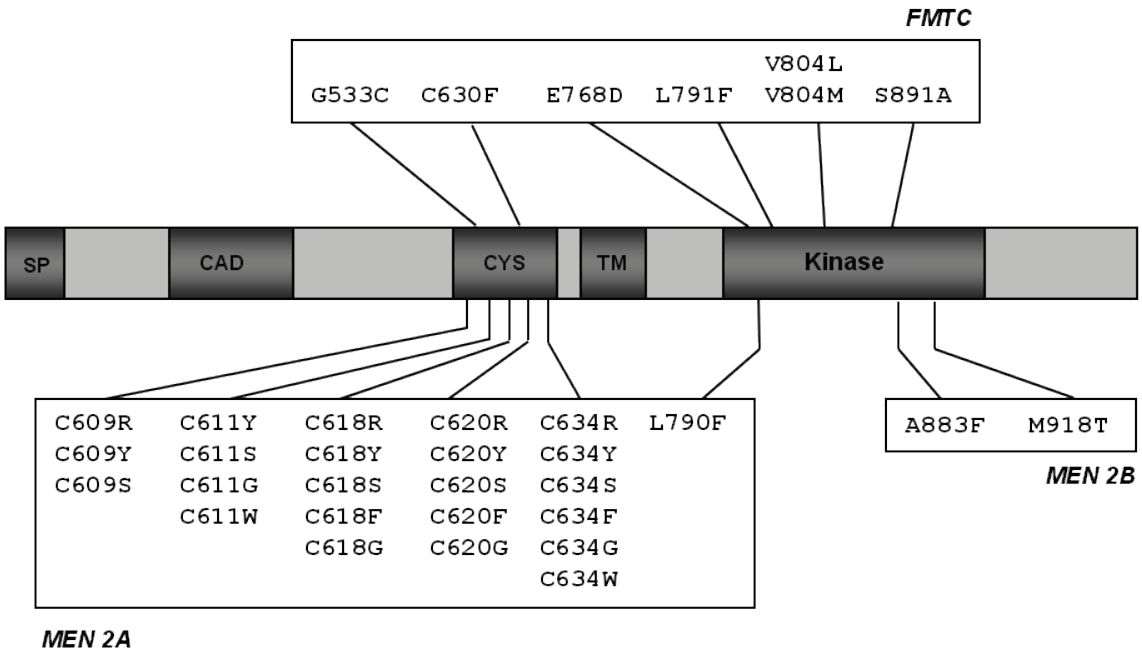


Figure 2.1 - RET protein domains. Schematic view of the RET oncoprotein showing conserved domains of signal peptide (SP), cadherin repeat domains (CAD), cysteine rich region (CYS), transmembrane domain (TM), and protein tyrosine kinase (Kinase) where reported variants associate to three specific disease phenotypes; familial medullary thyroid cancer (FMTC), multiple endocrine neoplasia type 2A (MEN2A) and multiple endocrine neoplasia type 2B (MEN2B).

Accurately predicting the mutation severity for gene variants in the *RET* oncogene could help clinicians identify patients less likely to respond to standard treatments, assist patients when making informed decisions about their care, and aid researchers in understanding mechanisms of disease severity.

Here we examine the hypothesis that novel informatics tools can take advantage of well-curated gene variant collections, utilizing physicochemical properties of the amino acids in the coded proteins to determine mutation severity. This study evaluates the performance of machine-learning classification algorithms for predicting mutational severity in *RET* oncogene variants with

known genotype-phenotype association when using representative chemical, physical, energetic, and conformational properties of amino acids as descriptors of the mutation.

Methods

A curated set of non-synonymous *RET* mutations with known phenotype severity ("pathogenic" or "benign"), publicly available at http://www.arup.utah.edu/database/, [13] was used to train and test representative machine learning classification algorithms. Archived *RET* gene variants were accessed from this database in January 2010. Sequence variants were verified for their position within the *RET* gene and named following standard Human Genome Organisation (HUGO) nomenclature.

RET mutations were characterized by the absolute differences between the values of 544 amino acid properties (AAIndex v9.4) of the residue present in the wild type and the one in the mutated sequence [15, 16]. The Correlation-based Feature Subset Selection algorithm [17], together with the Best First (greedy hillclimbing) search method, were used to identify the subset of properties that best differentiated benign mutations from pathogenic ones, based on the amino acid changes in *RET.* After feature selection was performed on training sets, selected properties specific to each training set (k=3) were carried forward as attributes for classification. Thus, each mutation was described by an array of variables, corresponding to the absolute value of the difference between the value of the property in the amino acid present in the wild type and the one in the mutant.

Due to the limited amount of clinically curated variants available publically, cross fold validation (k=3) was used to train and test classification of disease phenotype. The sample set (n=104) used 58 pathogenic variants specific to MEN2 phenotype and 46 benign variants. The data set only used nonsynonymous variants where one amino acid was substituted for another. Because of the limited sample size, we chose to perform cross validation rather than the ideal method of holding data separate for external validation.

For this study, five different machine-learning classification algorithms were evaluated including: ZeroR (zero rules), bayes (NaiveBayes), regression (SimpleLogistic), support vector machine (SMO), k nearest neighbor (IBk), and trees (RandomForest). Machine-learning

classification algorithms with their respective default settings as implemented in the Weka software package (v3.6) were used in this study [18]. Because "accuracy" is a term often plagued with misinterpretation, we choose to evaluate algorithm performance using previously reported and less ambiguous values of sensitivity, specificity, and positive predictive value [19].

Finally, the above classification models were also compared to existing mutation prediction algorithms based on sequence homology, amino acid substitution penalties or structural disruption using the full set of *RET* mutations with their curated outcomes. The SIFT algorithm is available on-line at http://sift.jcvi.org/ and gives outcomes of "tolerated" (meaning predicted benign) and "affects protein function" (meaning predicted pathogenic). PolyPhen was accessed at http://genetics.bwh.harvard.edu/pph and has outcomes of "benign" and "probably damaging" (meaning predicted pathogenic). MutPred is hosted at http://mutdb.org/mutpred and calculates the probability of a deleterious mutation with corresponding hypothesis of disrupted molecular mechanism when found. These algorithms were accessed during July/August 2010 and evaluated using their respective default settings.

Results

Utilizing a strategy of k-fold cross validation (k=3), the correlation-based feature selection chose 23 properties from the original 544 amino acid attributes in AAindex. These descriptors are summarized in Table 2.2. Overall, 8 properties were chosen using feature selection in 3 out of 3 folds, while some 15 properties were seen in 2 out of 3 folds. Amino acid properties relating to hydrophobicity or membrane buriedness, as well as positional or structural frequency seem to be representative of the features selected by this methodology.

To evaluate classifier performance, the weighted average from 3 fold cross validation of sensitivity (true positive rate), specificity (true negative rate), and positive predictive value (precision) were calculated for each classifier algorithm. Classifier performance is summarized in Table 2.3 as ranked by positive predictive value (PPV) or the percentage of variants classified as pathogenic that actually were pathogenic. For this data set, ZeroR (zero rules - which selects the majority class by default), yielded a baseline performance of 55.7%. The nearest neighbor, random forest, support vector machine, and regression models gave similar performance to each

Table 2.2 Feature selection (n=23) from 544 amino acid properties from AAindex.

Property[a]	Original Source	PubMed ID[b]
alpha NH chemical shifts	Bundi (1979)	7881270
Normalized frequency of C terminal helix	Chou (1978)	364941
Normalized frequency of chain reversal R	Tanaka (1977)	557155
Normalized positional frequency at helix termini N2	Aurora-Rose (1998)	9514257
Partition coefficient	Garel (1973)	4700470
Relative preference value at C2	Richardson (1988)	3381086
Relative preference value at N1	Richardson (1988)	3381086
Weights for beta sheet at the window position of 0	Qian (1988)	3172241
Amino acid distribution	Jukes (1975)	237322
Average relative fractional occurrence in A0(i)	Rackovsky (1982)	0903736
Average relative probability of inner beta sheet	Kanehisa (1980)	7426680
Composition	Grantham (1974)	4843792
Effective partition energy	Miyazawa (1985)	2004114
Free energy in alpha helical region	Munoz (1994)	7731949
Frequency of the 3rd residue in turn	Chou (1978)	364941
Helix formation parameters (delta delta G)	O Neil (1990)	2237415
Hydrophobicity	Prabhakaran (1990)	2390062
Membrane buried preference parameters	Argos (1982)	7151796
Normalized frequency of beta structure	Nagano 1973	4728695
Normalized frequency of coil	Nagano 1973	4728695
Normalized positional frequency at helix termini Cc	Aurora-Rose (1998)	9514257
STERIMOL maximum width of the side chain	Fauchere (1988)	3209351
Zimm Bragg parameter sigma x 1.0E4	Sueki (1984)	1004141

[a] Accessed August 2010 from http://www.genome.jp/aaindex/
[b] Accessed from http://www.ncbi.nlm.nih.gov/pubmed

other with 77.6%, 78.9%, 79.1%, and 81.4% respectively. Naïve Bayes was the best performing algorithm with a PPV of 82.7%, a gain in performance of 27% over the ZeroR classifier.

The machine learning algorithms constructed models that primarily used positional frequency and hydrophobicity related properties such as frequency of the 3rd residue in turn or membrane buried preference parameters as leading factors to classify the mutations. This may reinforce the importance of mutations in key residues responsible for proper transmembrane placement and strategic cysteine residues responsible for normal kinase dimerization function [20]. In other words, location of the change is not equal across the length of the protein sequence. Amino acid substitutions in key "hot spot" areas are thus more likely to result in pathogenic gain of function effects. Compared to the existing mutation prediction algorithms, we found that all the classifiers used here performed better than or similar to the well established algorithms (Table 2.3). Analysis of the *RET* mutations using PolyPhen correctly identified 68 out

Table 2.3 Summary of classification performance for machine learning algorithms.

Algorithm Name	Algorithm Sensitivity	Algorithm Specificity	Positive Predictive Value
ZeroR	1.00	0.00	0.557
IBk	0.896	0.674	0.776
RandomForest	0.776	0.739	0.789
SMO	0.914	0.696	0.791
SimpleLogistic	0.826	0.761	0.814
NaiveBayes	0.827	0.783	0.827
PolyPhen[a]	0.597	0.920	0.541
SIFT[b]	0.816	0.821	0.779
MutPred[c]	0.767	0.823	0.843

[a] http://genetics.bwh.harvard.edu/pph.
[b] http://sift.jcvi.org.
[c] http://mutdb.org/mutpred.

of 104 mutations as compared to the curated database entries (65% agreement). The MutPred algorithm performed similarly with 64% agreement (67 out of 104). It was unable, however, to complete predictions for 33 of the 104 mutations, although results for the remaining curated entries yielded 67 out of 71 (94% agreement). SIFT analysis correctly classified 75 of 104 cases when compared to the curated database for 72% agreement. To demonstrate disagreement when comparing existing algorithms to curated outcomes, results for selected *RET* mutations are summarized in Table 2.4. Discrepancies between the known phenotype and the existing prediction algorithms seemed to occur in cysteine related substitutions or where alignment to *RET* orthologs was not well conserved.

Table 2.4 Comparison of mutation prediction for selected *RET* mutations.

RET Gene Variant [a] Curated Outcome	PolyPhen Prediction[b]	SIFT Prediction[c]	MutPred Prediction[d]
G533C (pathogenic)	probably damaging	affects function	Not available
C609S (uncertain)	probably damaging	affects function	deleterious (0.90)
C611S (pathogenic)	probably damaging	affects function	deleterious (0.90)
C618G (pathogenic)	probably damaging	tolerated	deleterious (0.88)
C620R (pathogenic)	benign	tolerated	deleterious (0.75)
C630R (pathogenic)	probably damaging	tolerated	deleterious (0.70)
D631Y (pathogenic)	probably damaging	affects function	deleterious (0.69)
C634L (pathogenic)	probably damaging	tolerated	deleterious (0.69)
S649L (pathogenic)	possibly damaging	tolerated	deleterious (0.66)
G691S (benign)	benign	tolerated	benign (0.20)

[a] Curated *RET* variants from http://www.arup.utah.edu/database/MEN2/MEN2_welcome.php.
[b] Analyzed with default settings at http://genetics.bwh.harvard.edu/pph.
[c] Analyzed with default settings at http://sift.jcvi.org.
[d] Analyzed with default settings at http://mutdb.org/mutpred.

Discussion

One example that highlights the usefulness of predicting mutation severity was found in the *RET* codon 609. Although several changes in the codon 609 are known to be pathogenic, the variant C609S is currently listed as an uncertain variant in the curated database. The machine learning classifiers along with the mutation prediction tools labeled this variant as "predicted pathogenic," "probably damaging" (SIFT), "affects protein function" (PolyPhen) and mutation (0.90), with a gain of glycosylation site (MutPred). This example underscores the utility of computational prediction of mutations and suggests a need for careful evaluation of this C609S variant, including additional family outcome studies or further molecular confirmation of the resulting phenotype.

When mutations are characterized by the difference between the values in several amino acid properties in the wild type and the mutated sequence, machine-learning classification can be used to accurately predict *RET* mutation status using primary sequence information only. Existing algorithms that are based on sequence homology (ortholog conservation) or protein structural data are not necessarily superior - at least for this specific genotype-phenotype. These results indicate that using physiochemical properties of amino acids to characterize mutations is important and may be more relevant than evolutionary sequence conservation. Furthermore, the attributes found in AAIndex - in combination with feature selection - are a viable source of descriptors for use with machine learning tools and mutation prediction. Finally, several different types of algorithms worked similarly well, pointing to the robustness of this methodology.

Acknowledgements

Open Access permission for previously published material from the *Journal of Data Mining in Genomics and Proteomics* is acknowledged.

References

1. Ramensky V, Bork P, Sunyaev S: **Human non-synonymous SNPs: server and survey.** *Nucleic Acids Res* 2002, **30:**3894-3900.

2. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function.** *Nucleic Acids Res* 2003, **31:**3812-3814.

3. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P: **Automated inference of molecular mechanisms of disease from amino acid substitutions.** *Bioinformatics* 2009, **25:**2744-2750.

4. Spencer DS, Stites WE: **The M32L substitution of staphylococcal nuclease: disagreement between theoretical prediction and experimental protein stability.** *J Mol Biol* 1996, **257:**497-499.

5. Kang HH, Williams R, Leary J, Ringland C, Kirk J, Ward R: **Evaluation of models to predict BRCA germline mutations.** *Br J Cancer* 2006, **95:**914-920.

6. Engelhardt BE, Jordan MI, Muratore KE, Brenner SE: **Protein molecular function prediction by Bayesian phylogenomics.** *PLoS Comput Biol* 2005, **1:**e45.

7. Woolley S, Johnson J, Smith MJ, Crandall KA, McClellan DA: **TreeSAAP: selection on amino acid properties using phylogenetic trees.** *Bioinformatics* 2003, **19:**671-672.

8. Georgiev AG: **Interpretable numerical descriptors of amino acid space.** *J Comput Biol* 2009, **16:**703-723.

9. UniProt consortium: **The universal protein resource (UniProt).** *Nucleic Acids Res* 2008, **36:**D190-195.

10. Eng C, Clayton D, Schuffenecker I, Lenoir G, Cote G, Gagel RF, van Amstel HK, Lips CJ, Nishisho I, Takai SI, et al: **The relationship between specific RET proto-oncogene mutations and disease phenotype in multiple endocrine neoplasia type 2. International RET mutation consortium analysis.** *Jama* 1996, **276:**1575-1579.

11. Kouvaraki MA, Shapiro SE, Perrier ND, Cote GJ, Gagel RF, Hoff AO, Sherman SI, Lee JE, Evans DB: **RET proto-oncogene: a review and update of genotype-phenotype correlations in hereditary medullary thyroid cancer and associated endocrine tumors.** *Thyroid* 2005, **15:**531-544.

12. Kloos RT, Eng C, Evans DB, Francis GL, Gagel RF, Gharib H, Moley JF, Pacini F, Ringel MD, Schlumberger M, Wells SA, Jr.: **Medullary thyroid cancer: management guidelines of the American Thyroid Association.** *Thyroid* 2009, **19:**565-612.

13. Margraf RL, Crockett DK, Krautscheid PM, Seamons R, Calderon FR, Wittwer CT, Mao R: **Multiple endocrine neoplasia type 2 RET protooncogene database: repository of MEN2-associated RET sequence variation and reference for genotype/phenotype correlations.** *Hum Mutat* 2009, **30:**548-556.

14. Massoll N, Mazzaferri EL: **Diagnosis and management of medullary thyroid carcinoma.** *Clin Lab Med* 2004, **24:**49-83.

15. Kawashima S, Kanehisa M: **AAindex: amino acid index database.** *Nucleic Acids Res* 2000, **28:**374.

16. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, Kanehisa M: **AAindex: amino acid index database, progress report 2008.** *Nucleic Acids Res* 2008, **36:**D202-205.

17. Hall MA: **Correlation-based feature selection of discrete and numeric class machine learning.** In *Computer Science Working Papers*. Hamilton, New Zealand: University of Waikato, Department of Computer Science; 2000.

18. Frank E, Hall M, Trigg L, Holmes G, Witten IH: **Data mining in bioinformatics using Weka.** *Bioinformatics* 2004, **20:**2479-2481.

19. Hand DJ: **Measuring classifier performance: a coherent alternative to the area under the ROC curve.** *Machine Learning* 2009, **77:**103-123.

20. Lai AZ, Gujral TS, Mulligan LM: **RET signaling in endocrine tumors: delving deeper into molecular mechanisms.** *Endocr Pathol* 2007, **18:**57-67.

CHAPTER 3

PHENOTYPE PREDICTION OF NOVEL AND UNCERTAIN GENE VARIANTS

(Reprinted with open access permission from Crockett DK, Piccolo SR, Ridge PG, Margraf RL, Lyon E, Williams MS, Mitchell JA. **Predicting phenotypic severity of uncertain gene variants in the RET proto-oncogene.** *PLoS One* 2011. Mar 30;6(3):e18380.)

Abstract

Although reported gene variants in the *RET* oncogene have been directly associated with multiple endocrine neoplasia type 2 and hereditary medullary thyroid carcinoma, other mutations are classified as variants of uncertain significance (VUS) until the associated clinical phenotype is made clear. Currently, some 46 nonsynonymous VUS entries exist in curated archives. In the absence of a reliable method for predicting phenotype outcomes, this follow up study applies feature selected amino acid physical and chemical properties feeding a Bayes classifier to predict disease association of uncertain gene variants into categories of benign and pathogenic. Algorithm performance and VUS predictions were compared to established phylogenetic based mutation prediction algorithms. Curated outcomes and unpublished *RET* gene variants with known disease association were used to benchmark predictor performance. Reliable classification of *RET* uncertain gene variants will augment current clinical information of *RET* mutations and assist in improving prediction algorithms as knowledge increases.

Introduction

Medical genetics involves diagnosis, management, and determining risk of hereditary disorders [1, 2]. The genotype:phenotype correlation of gene variants in disease is a major component of medical genetics. In monogenic diseases, gene mutations are typically curated as either pathogenic or benign. However, many gene variants must be classified as “unknown” or “uncertain” significance because they have not been clearly associated with a clinical phenotype.

The outlay of time and labor to validate the disease association concerning a variant of uncertain significance (VUS) within the coding portion of a gene can be daunting and cost prohibitive [3, 4]. This is in large part, due to the communication between clinicians and laboratory geneticists needed to resolve these variants [5, 6]. To help bridge this genotype:phenotype gap, the use of machine learning classification algorithms to narrow the uncertain “grey area” between pathogenic and benign sequence variants warrants careful evaluation [7-10]. Reliable machine learning based classification may augment costly patient recruitment, family histories, and biochemical confirmation of a gene variant with no associated disease correlation [11-13].

There are established methods for predicting mutation severity based on amino acid substitution penalties, structural disruption, sequence homology (ortholog conservation) or neural nets, such as PolyPhen [13], SIFT [14], MutPred [9] and PMut [15]. However, prediction algorithms are not always in agreement with curated data or each other [16-18]. Thus, there are opportunities to explore the use of other informatics approaches to this problem. Machine learning methods that can be trained on data available in well-curated gene variant collections may be promising tools to improve the predictive capabilities available to the research community.

The human *RET* gene (REarranged during Transfection) is located on chromosome 10q.11 codes for 20 exons. The transcript length is 5,659 bps and translates to the 1,114 amino acid residue protein (UniProt RET_HUMAN, #P07949) as shown in Figure 3.1. The gene belongs to the cadherin superfamily and encodes a receptor tyrosine kinase which functions in signaling pathways for cell growth and differentiation. *RET* plays a critical role in neural crest development. It can also undergo oncogenic activation *in vivo* and *in vitro* by cytogenetic rearrangement. It can be further classified by Gene Ontology (GO) categories (www.geneontology.org) of biological process of homophilic cell adhesion, posterior midgut development, and protein amino acid phosphorylation. Its GO annotated cellular location is component integral to membrane and the GO category of molecular functions lists ATP binding, calcium ion binding and transmembrane receptor protein tyrosine kinase activity. Functional domains of the RET protein are also summarized in Figure 3.1.

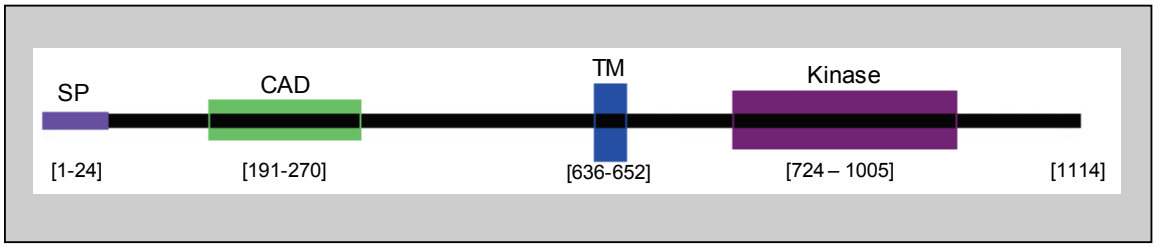


Figure 3.1 Schematic of the full length 1114 amino acid RET protein showing the signal peptide (SP, residues 1-24), cadherin domain (CAD, residues 191-270), transmembrane domain (TM, residues 636-652), and tyrosine kinase motif (Kinase, residues 724-1005).

RET is essential for the development of the sympathetic, parasympathetic and enteric nervous systems. Disruption of function by germline mutations in *RET* have been associated with several diseases in humans including three related inherited cancers: multiple endocrine neoplasia type IIA (MEN2A), multiple endocrine neoplasia type IIB (MEN2B), and familial medullary thyroid carcinoma (FMTC). [19, 20] RET has also been implicated in congenital aganglionosis (absence of enteric nerve cells) in the gastrointestinal tract (Hirschsprung's disease) lack of the neuroenteric plexi impairs smooth muscle activity of the intestines (particularly the colon) resulting in refractory constipation. [21]

Although well understood codon changes often guide patient therapy or surgical options [22], *RET* gene variants may vary in functional severity, where some are reported as benign, some pathogenic, and some of uncertain significance. Curated *RET* oncogene mutations have been recently reported by Margraf et al. [23] The disease classification of *RET* gene variants has been curated as benign (6%), pathogenic (52%) and VUS (42%), meaning unknown or uncertain association with disease or phenotype outcome. This archive currently hosts 146 *RET* variants, including 62 VUS entries that can be accessed at http://www.arup.utah.edu/database/.

Accurate prediction of disease association for novel mutations and uncertain gene variants is of great importance to medicine and biology. Informatics tools for predicting disease severity of uncertain gene variants will aid in the improvement of genetically-informed patient care. With a rapidly growing number of on-line resources for gene variants collections, the opportunity to apply machine learning algorithms to well curated disease causing gene sets becomes increasingly desirable.

The absence of any gold standard for predicting phenotype severity in uncertain gene variants prompts two questions. Are algorithms trained specific to a gene/disease setting more appropriate to use than generalized on-line prediction tools? Does agreement between several and varying algorithms influence clinician decision-making? This study expands a recently reported algorithm, we here term Primary Sequence Amino Acid Properties (PSAAP), which uses feature selected amino acid physicochemical properties of primary amino acid sequence. [24] This previous work detailed algorithm performance using only gene variants with known disease

association, while here we report applying the PSAAP algorithm classification for pathogenicity of novel and uncertain gene variants found in the *RET* proto-oncogene into categories of benign or pathogenic. The PSAAP algorithm performance has also been compared to four well-established prediction tools available on-line and agreement between algorithms summarized.

Methods

Nonsynonymous *RET* variants were characterized by physicochemical differences in primary amino acid sequence resulting from the mutation. Attributes of mutation status were characterized using values of 544 physical, chemical, conformational, or energetic properties (AAindex v9.4). [25] AAindex is a database of numerical indices representing various physicochemical and biochemical properties of amino acids and pairs of amino acids. For each *RET* variant, matrices of delta values for each biochemical property of the substituted amino acid were calculated by Python scripting and the resulting mutation described by an array of variables archived using SQL - where each matrix corresponds to the absolute value of the difference between the value of the property in the amino acid present in the wild type and the one in the mutant.

As previously described, representative algorithms from different categories of classification (such as nearest neighbor, bayes, regression, rule-based and support vector machine) were evaluated for their ability to correctly predict mutation status in the training set. [24] Briefly, a clinically curated set (n=84) of nonsynonymous *RET* mutations with known pathogenicity was used to train and test machine learning classification algorithms. Although training and test sets included different disease subtypes such as MEN2A (n=40), MEN2B (n=3), FMTC (n=5), MEN2A and FMTC (n=36) - class labels of “pathogenic” and “benign” were used to describe all curated disease association. Random selection was used to build a 2/3 training set (n=56) and 1/3 test set (n=28). Attribute selection (feature selection) was performed during classification training/testing. Machine classification algorithms were implemented using the Weka software package (v3.6). [26] When a given classification algorithm produced posterior probabilities of mutation status, we assigned each variant's mutation status according to the higher posterior probability (Weka's default behavior).

The PSAAP algorithm performance was evaluated using the test set, with sensitivity (true positive rate), specificity (true negative rate), and positive predictive value (precision) calculated. A data set of non-synonymous *RET* uncertain variants (n=46) was then analyzed using our PSAAP (Naïve Bayes, gene-specific trained) classification algorithm. The workflow of our PSAAP algorithm is summarized in Figure 3.2.

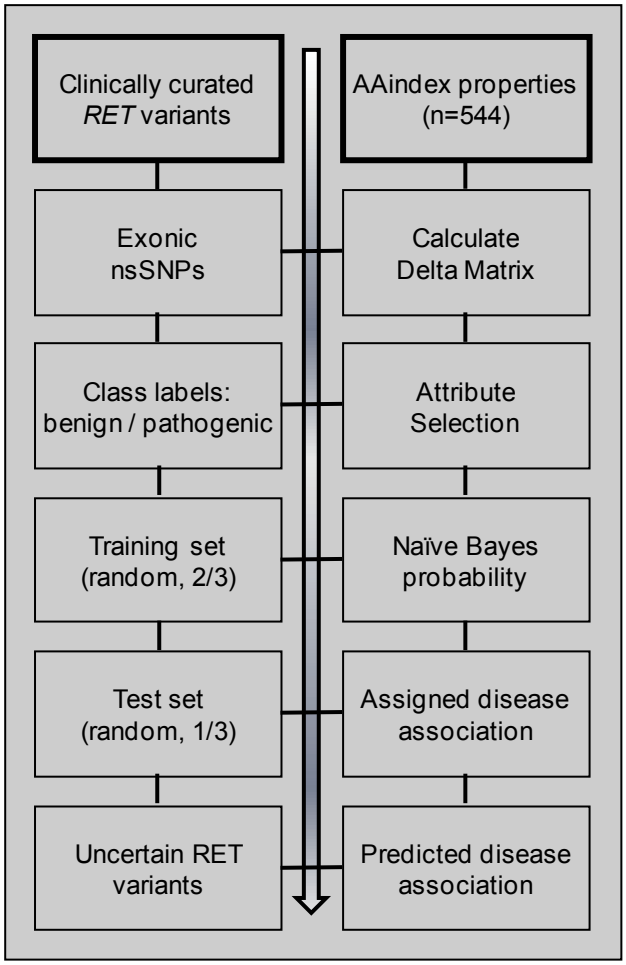


Figure 3.2 Overview of the PSAAP classifier workflow, highlighting the gene-specific algorithm training on clinically curated disease association.

Next, both curated *RET* mutations (known disease association) and *RET* uncertain variants (VUS data) were analyzed and compared using four existing mutation prediction

algorithms. These established prediction tools are mainly based on phylogenetic properties such as sequence homology, amino acid substitution penalties or structural disruption. MutPred (mutdb.org/mutpred) [9], PolyPhen (genetics.bwh.harvard.edu/pph) [13], SIFT (sift.jcvi.org) [14], and PMut (mmb2.pcb.ub.es:8080/PMut) [15] were accessed during July/August 2010. Both curated *RET* variants and *RET* VUS entries were evaluated using respective default settings.

Finally, several unpublished *RET* disease variants (n=5) with known pathogenic outcomes (by surgical pathology, molecular testing and family history) were identified during routine genetic testing at ARUP Laboratories. This nascent set of *RET* variants was also analyzed and compared by all prediction algorithms to further benchmark some standard of performance and precision. Data and methods used for this study were approved by the Institutional Review Board of the University of Utah (IRB #00035757).

Results

The independent test set of *RET* curated mutations was used to evaluate performance of different categories of classifier algorithms. The best performing algorithm (using Weka) was Naïve Bayes. Algorithm metrics for this novel Bayes classifier of *RET* disease outcome were calculated using the above test set data. Evaluation of the classifier yielded a sensitivity of 0.938, specificity of 0.867 and positive predictive value (precision) of 0.883. Performance for our Primary Sequence Amino Acid Properties (PSAAP) classifier is summarized in Table 3.1. A benchmark of prediction performance for the established algorithms (MutPred, PolyPhen, PMut and SIFT) was also performed using curated *RET* gene variants with known disease outcomes. Following the 88% of the PSAAP classifier, MutPred was next closest to predicting the correct disease outcomes for the known *RET* variants with 84% precision. PolyPhen yielded the highest specificity for *RET* variant disease association of 92%, yet had the lowest precision at 54%. PMut correctly predicted gene variant disease outcomes with 72% precision but had the lowest specificity at 59%. Table 3.1 also summarizes performance metrics (sensitivity, specificity, precision) for curated *RET* mutations using the four established prediction algorithms.

Table 3.1 PSAAP algorithm performance of predicted phenotypes using curated *RET* mutations.

	PSAAP Prediction[a]	MutPred Prediction[b]	PolyPhen Prediction[c]	PMut Prediction[d]	SIFT Prediction[e]
Sensitivity	0.938	0.767	0.597	0.783	0.816
Specificity	0.867	0.823	0.920	0.591	0.821
Precision	0.883	0.843	0.541	0.723	0.779

[a] Primary Sequence Amino Acid Properties (PSAAP) algorithm.
[b] Analyzed with default settings at http://mutdb.org/mutpred.
[c] Analyzed with default settings at http://genetics.bwh.harvard.edu/pph.
[d] Analyzed with default settings at http://mmb.pcb.ub.es/PMut.
[e] Analyzed with default settings at http://sift.jcvi.org.

Next, evaluation of *RET* non-synonymous VUS mutations (n=46) was performed using our recently reported algorithm .[24] The PSAAP algorithm classified 22 of the uncertain variants as pathogenic, while the remaining 24 fell within the benign grouping. For those variants classified as predicted pathogenic, the PSAAP algorithm estimated confidence remained above 90%. The classifier predicted disease outcome using our algorithm is listed in Table 3.2.

Results from analysis of the *RET* uncertain gene variants (VUS) using the established on-line prediction tools are also summarized in Table 3.2, with predicted pathogenic variants bolded and ranked by agreement. The MutPred tool calculates the probability of a deleterious mutation and corresponding hypothesis of disrupted molecular mechanism. We used MutPred's default probability cutoff of 0.75 for differentiating between benign and disrupted/pathogenic mutations. Our PSAAP algorithm agreed with MutPred in 16 benign and 8 pathogenic predictions for 52% agreement (24 out of 46). PolyPhen has outcomes of “benign,” “possibly damaging” and “probably damaging.” The PSAAP classifier agreed with PolyPhen in 13 benign and 22 pathogenic predictions for 76% agreement (35 out of 46). PMut yields outcomes of “pathological” or “neutral” and a corresponding reliability metric (lower is better). Our PSAAP trained algorithm was in concordance with PMut in 13 benign and 14 pathogenic predictions for 58% agreement (27 out of 46). The SIFT algorithm gives outcomes of “tolerated” and “affects protein function.” Our algorithm agreed with SIFT in 19 benign and 16 pathogenic predictions for 76% agreement (35 out of 46).

Table 3.2 Algorithm agreement for *RET* uncertain gene variants and predicted pathogenicity.

RET uncertain gene variant	PSAAP Prediction[a]	MutPred Prediction[b]	PolyPhen Prediction[c]	SIFT Prediction[d]	PMut Prediction[e]
5/5 agreement					
A510V	benign	benign	benign	tolerated	neutral
R600Q	benign	benign	benign	tolerated	neutral
K603Q	benign	benign	benign	tolerated	neutral
E632K	benign	benign	benign	tolerated	neutral
A640G	benign	benign	benign	tolerated	neutral
V648I	benign	benign	benign	tolerated	neutral
Y791N	pathogenic	disrupted	prob damaging	affects function	pathological
E843D	benign	benign	benign	tolerated	neutral
R844L	pathogenic	disrupted	prob damaging	affects function	pathological
R844W	pathogenic	disrupted	prob damaging	affects function	pathological
R886W	pathogenic	disrupted	prob damaging	affects function	pathological
R912Q	pathogenic	disrupted	prob damaging	affects function	pathological
4/5 agreement					
C611S	pathogenic	disrupted	prob damaging	affects function	neutral
D631G	pathogenic	benign	prob damaging	affects function	pathological
E805K	benign	disrupted	prob damaging	affects function	pathological
S819I	pathogenic	disrupted	prob damaging	affects function	neutral
R833C	pathogenic	benign	prob damaging	affects function	pathological
S904C	pathogenic	benign	prob damaging	affects function	pathological
S904F	pathogenic	benign	prob damaging	affects function	pathological
3/5 agreement					
Y606C	pathogenic	benign	prob damaging	tolerated	pathological
C531R	pathogenic	benign	prob damaging	tolerated	pathological
G533S	pathogenic	benign	prob damaging	affects function	neutral
D631A	pathogenic	benign	prob damaging	affects function	neutral
D631V	pathogenic	benign	prob damaging	affects function	neutral
R635G	pathogenic	benign	prob damaging	tolerated	pathological
P841L	pathogenic	benign	prob damaging	tolerated	pathological
L881V	benign	disrupted	prob damaging	affects function	neutral
K907M	pathogenic	benign	prob damaging	affects function	neutral
2/5 agreement					
C630S	pathogenic	benign	prob damaging	tolerated	neutral
D631E	benign	benign	prob damaging	tolerated	pathological
S649L	pathogenic	benign	prob damaging	tolerated	neutral
H665Q	benign	benign	prob damaging	tolerated	pathological
R844Q	benign	benign	prob damaging	tolerated	pathological
M848T	benign	benign	prob damaging	tolerated	pathological
I852M	benign	benign	prob damaging	affects function	neutral
K907E	benign	benign	prob damaging	affects function	neutral
1/5 agreement					
G321R	benign	benign	benign	tolerated	pathological
E511K	benign	benign	benign	tolerated	pathological
D631N	benign	benign	benign	tolerated	pathological
A641S	benign	benign	poss damaging	tolerated	neutral
K666N	benign	benign	prob damaging	tolerated	neutral
R770Q	benign	benign	prob damaging	tolerated	neutral
N777S	benign	benign	poss damaging	tolerated	neutral
V778I	benign	benign	benign	affects function	neutral
E818K	benign	benign	poss damaging	tolerated	neutral

[a] Primary Sequence Amino Acid Properties (PSAAP) algorithm.
[b] Analyzed with default settings at http://mutdb.org/mutpred.
[c] Analyzed with default settings at http://genetics.bwh.harvard.edu/pph.
[d] Analyzed with default settings at http://sift.jcvi.org.
[e] Analyzed with default settings at http://mmb.pcb.ub.es/PMut.

Of special interest, for predicted *RET* benign variants, 7 of 24 agreed across all algorithms, while only 6 of 22 predicted pathogenic *RET* variants showed agreement across the different methods. Although only 13 out of 46 (28%) were concordant, these variants may count as having a higher degree of confidence in prediction due to the varied methodologies and basis of classification. Importantly, the focus of molecular research and clinical efforts could therefore be directed to this prioritized listing of *RET* uncertain variants. Curated variants are shown mapped across the length of the protein in Figure 3.3A. This graphing visually highlights the cysteine rich region just prior to the transmembrane domain, and the transmembrane domain itself which contain the majority of pathogenic variants. Our predictions for the uncertain RET variants (VUS) are also mapped by location across the length of the protein as added into Figure 3.3B.

Finally, several unpublished *RET* gene variants with known pathological (MEN2) outcomes (n=5) were identified during routine genetic testing at ARUP Laboratories. To further benchmark *RET* mutation prediction, all five algorithms were used to classify this set of not yet seen variants. Our novel Bayes trained PSAAP classifier correctly identified all five variants as pathogenic. PMut called three disease causing variants correctly, but classified two others as "neutral" mutations, when in fact these changes were known to be associated with disease. PolyPhen also correctly identified three as probably damaging (pathogenic), but missed classified the same two variants as PMut. SIFT predicted four of these variants would affect function (pathogenic), but called one of the same variants "tolerated." MutPred correctly predicted all five as pathogenic.

Discussion

Mutations in the *RET* proto-oncogene have been directly associated with MEN2 and hereditary medullary thyroid carcinoma, and provide guidance for patient care. Accurate classification of phenotype severity for novel mutations and uncertain variants as relating to disease is of great importance to proper patient care. Although correlation of genotype-phenotype offers therapy options that would otherwise remain hidden and may lead to disease

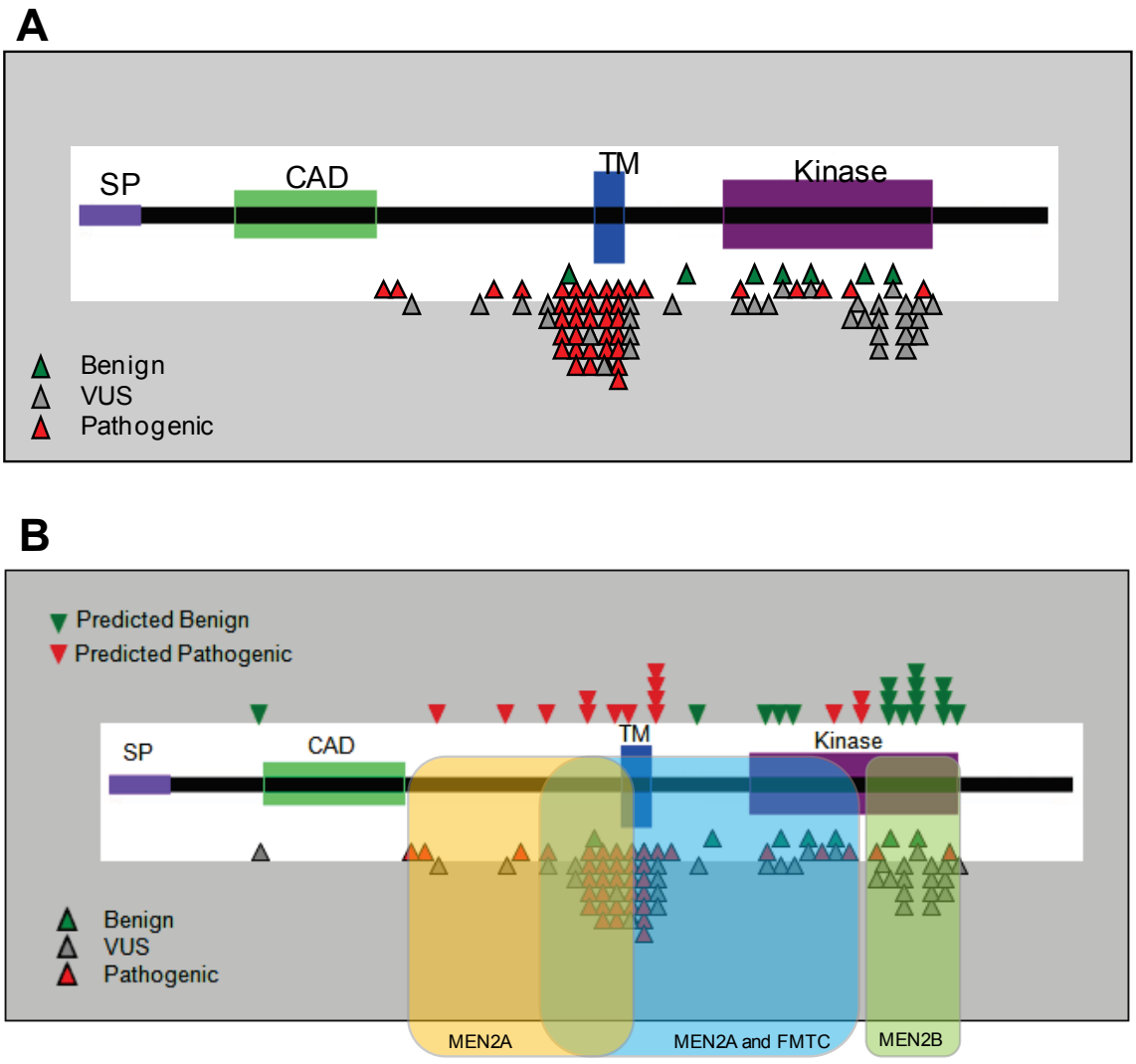


Figure 3.3 Schematic of the RET protein with A) clinically curated variants and B) predicted disease association for uncertain variants mapped across protein location. The phenotype overlay shows regions of reported MEN2A, MEN2B and FMTC disease.

specific mutation-guided management strategies, appropriate caution is justified when clinicians are asked to trust computational outcomes for determining patient care. [6]

On-line mutation prediction tools have been available for many years. Prediction tools such as PolyPhen [13] and SIFT [14] are primarily based on multiple alignment and amino acid substitution penalties. More recently, MutPred [9] which calculates probability of deleterious mutations by disrupted molecular mechanism. Additionally, PMut [15] is neural net based and

trained on human mutations. We recently reported classification of curated *RET* gene variants using primary amino acid sequence properties and Naïve Bayes.[24] A key feature to highlight is the fact that the PSAAP algorithm relies on Bayes probability trained on gene-specific and clinically curated disease outcomes. Comparison of this recent PSAAP algorithm with established on-line prediction tools may improve our understanding of predicting mutation status in the *RET* proto-oncogene.

Sorting Intolerant From Tolerant (SIFT) was first published in 2003 by Ng and Heinikoff from work done at the Fred Hutchinson Cancer Research Center in Seattle. [14] The algorithm predicts whether an amino acid substitution will affect the function of a protein based on both sequence homology to various orthologs and physical properties of amino acids. SIFT is a multistep procedure that (1) searches for and chooses similar sequences, (2) makes an alignment of these sequences, and (3) calculates scores based on the amino acids appearing at each position in the alignment. It was initially developed and trained on nsSNP data sets from LacI, Lysozyme, and HIV protease. [27] This algorithm works especially well when adequate numbers of sequence homologs are available for multiple alignment. Conversely, poor performance is seen when multiple alignment in not reliable or completely unavailable.

Polymorphism Phenotyping (PolyPhen) is an EMBL based tool from 2002 from Ramensky et al. [13] It was developed to predict the possible impact of an amino acid substitution on the structure and function of a human protein using physical and comparative considerations. It was originally developed from a set of disease-causing mutations in human proteins with known structures extracted from the SWISS-PROT database, and correlated to the Online Mendelian Inheritance in Man (OMIM) database. [28] Since the algorithm relies on predicted structural disruption, it works especially well where protein structure is known and less reliable when a solved protein structure is not available.

MutPred is a recently developed prediction algorithm by Li, Mooney and Radivojac. [9] It builds on the established SIFT method but offers improved classification accuracy based upon protein sequence, and models changes of structural features and functional sites between wild-type and mutant sequences with output of probabilities of gain or loss of structure and function. It

was trained on a set of disease SNPs from cancer and the OMIM disease archive. This predicted disruption of molecular function again work especially well for well studied proteins, where homolog and solved structure is available.

PMut was first published in 2005 by the Molecular Modeling Unit at the Institut de Recerca Biomédica, Parc Científic de Barcelona, Spain. [15] It is based on a two layer neural network and was trained using human mutational data. It allows for either prediction of single point amino acidic mutations or scanning of mutational hot spots. Results are obtained by alanine scanning, identifying massive mutations and genetically accessible mutations. A graphical interface for Protein Data Bank (PDB) structures, when available, and a database containing hot spot profiles for all non-redundant PDB structures are also accessible from the PMut server.

Benchmarking the established prediction algorithms with curated *RET* variants and associated MEN2 disease demonstrates our PSAAP classifier model compares very well to other established prediction tools. A distinguishing feature of the PSAAP model herein reported is the algorithm was trained specifically to curated *RET* disease outcomes, as summarized in Figure 3.2. This is in contrast to the less robust curated collections of mutations such as OMIM or dbSNP. Further, no homolog alignment or solved protein structure is necessary. Rather, it relies on primary sequence information only - with calculated delta matrices of substituted amino acid properties , and is therefore not limited to scenarios where SIFT or PolyPhen (and others) have traditional been used. These facts may explain the improved performance when classifying *RET* variants as compared to generalized prediction tools available on-line.

Ranking agreement of predicted phenotype severity across several complimentary algorithms may provide an additional level of clinical confidence in computational classifiers. At a minimum, these five all-in-agreement “predicted pathogenic” *RET* variants warrant closer investigation by traditional and molecular techniques. Furthermore, algorithm agreement in a clinical setting may be just as important for “benign” as it might be for “pathogenic.”

Personalized treatment in genomic medicine cannot advance until questions such as *what was found*, *what does it mean* and *what to do about it* can be answered for each individual

patient and genetic test result. Among the key features critical for a decision support framework in clinical genetic testing is a reliable phenotype classification tool and scoring metric to predict consequences of a variation that alters protein structure. For these uncertain gene variants, the in-house algorithm trained specifically on available *RET* curated outcomes seems to outperform well-established and generalized prediction tools available on-line. More importantly, agreement between several predictors may provide research priority for novel and uncertain gene variants.

The use of machine learning algorithms to classify uncertain gene variants in disease is a promising tool to strengthen our underlying knowledge of disease pathogenesis. Software algorithms to better classify gene variants of uncertain significance are necessary to move translational research forward. This follow up study used the PSAAP algorithm to "reclassify" 46 variants of uncertain significance within the *RET* proto-oncogene into categories of benign or pathogenic. This novel application of classification algorithms for computational prediction of phenotype severity in uncertain gene variants could be generally applied to any gene-disease setting where a corpus of curated gene variants are trusted and where reported mutations impact clinical care.

Acknowledgements

The authors gratefully acknowledge Dr. Becky Margraf for curation of the *RET* proto-oncogene disease variants. Open Access permission for previously published material from *PLoS ONE* is acknowledged.

References

1. Weinstein ND: **What does it mean to understand a risk? Evaluating risk comprehension.** *J Natl Cancer Inst Monogr* 1999**:**15-20.

2. Ensenauer RE, Michels VV, Reinke SS: **Genetic testing: practical, ethical, and counseling considerations.** *Mayo Clin Proc* 2005, **80:**63-73.

3. Nowak R: **Genetic testing set for takeoff.** *Science* 1994, **265:**464-467.

4. Machens A, Gimm O, Hinze R, Hoppner W, Boehm BO, Dralle H: **Genotype-phenotype correlations in hereditary medullary thyroid carcinoma: oncological features and biochemical properties.** *J Clin Endocrinol Metab* 2001, **86:**1104-1109.

5. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, Dewey FE, Dudley JT, Ormond KE, Pavlovic A, Morgan AA, et al: **Clinical assessment incorporating a personal genome.** *Lancet* 2010, **375:**1525-1535.

6. Tchernitchko D, Goossens M, Wajcman H: **In silico prediction of the deleterious effect of a mutation: proceed with caution in clinical genetics.** *Clin Chem* 2004, **50:**1974-1978.

7. Wei Q, Wang L, Wang Q, Kruger WD, Dunbrack RL, Jr.: **Testing computational prediction of missense mutation phenotypes: functional characterization of 204 mutations of human cystathionine beta synthase.** *Proteins* 2010, **78:**2058-2074.

8. Kumar P, Henikoff S, Ng PC: **Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm.** *Nat Protoc* 2009, **4:**1073-1081.

9. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P: **Automated inference of molecular mechanisms of disease from amino acid substitutions.** *Bioinformatics* 2009, **25:**2744-2750.

10. Dorfman R, Nalpathamkalam T, Taylor C, Gonska T, Keenan K, Yuan XW, Corey M, Tsui LC, Zielenski J, Durie P: **Do common in silico tools predict the clinical consequences of amino-acid substitutions in the CFTR gene?** *Clin Genet* 2010, **77:**464-473.

11. Ferreira-Gonzalez A, Teutsch S, Williams MS, Au SM, Fitzgerald KT, Miller PS, Fomous C: **US system of oversight for genetic testing: a report from the Secretary's Advisory Committee on Genetics, Health and Society.** *Per Med* 2008, **5:**521-528.

12. Williams MS: **Quality in clinical genetics.** *Am J Med Genet C Semin Med Genet* 2009, **151C:**175-178.

13. Ramensky V, Bork P, Sunyaev S: **Human non-synonymous SNPs: server and survey.** *Nucleic Acids Res* 2002, **30:**3894-3900.

14. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function.** *Nucleic Acids Res* 2003, **31:**3812-3814.

15. 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.

16. Spencer DS, Stites WE: **The M32L substitution of staphylococcal nuclease: disagreement between theoretical prediction and experimental protein stability.** *J Mol Biol* 1996, **257:**497-499.

17. Kang HH, Williams R, Leary J, Ringland C, Kirk J, Ward R: **Evaluation of models to predict BRCA germline mutations.** *Br J Cancer* 2006, **95:**914-920.

18. Engelhardt BE, Jordan MI, Muratore KE, Brenner SE: **Protein molecular function prediction by Bayesian phylogenomics.** *PLoS Comput Biol* 2005, **1:**e45.

19. Eng C, Clayton D, Schuffenecker I, Lenoir G, Cote G, Gagel RF, van Amstel HK, Lips CJ, Nishisho I, Takai SI, et al: **The relationship between specific RET proto-oncogene mutations and disease phenotype in multiple endocrine neoplasia type 2. International RET mutation consortium analysis.** *Jama* 1996, **276:**1575-1579.

20. Kouvaraki MA, Shapiro SE, Perrier ND, Cote GJ, Gagel RF, Hoff AO, Sherman SI, Lee JE, Evans DB: **RET proto-oncogene: a review and update of genotype-phenotype correlations in hereditary medullary thyroid cancer and associated endocrine tumors.** *Thyroid* 2005, **15:**531-544.

21. Attie T, Pelet A, Edery P, Eng C, Mulligan LM, Amiel J, Boutrand L, Beldjord C, Nihoul-Fekete C, Munnich A, et al.: **Diversity of RET proto-oncogene mutations in familial and sporadic Hirschsprung disease.** *Hum Mol Genet* 1995, **4:**1381-1386.

22. Kloos RT, Eng C, Evans DB, Francis GL, Gagel RF, Gharib H, Moley JF, Pacini F, Ringel MD, Schlumberger M, Wells SA, Jr.: **Medullary thyroid cancer: management guidelines of the American Thyroid Association.** *Thyroid* 2009, **19:**565-612.

23. Margraf RL, Crockett DK, Krautscheid PM, Seamons R, Calderon FR, Wittwer CT, Mao R: **Multiple endocrine neoplasia type 2 RET protooncogene database: repository of MEN2-associated RET sequence variation and reference for genotype/phenotype correlations.** *Hum Mutat* 2009, **30:**548-556.

24. Crockett DK, Piccolo SR, Narus SP, Mitchell JA, Facelli JC: **Computational feature selection and classification of RET phenotypic severity.** *J Data Mining in Genom Proteomics* 2010, **1:**1-4.

25. Kawashima S, Kanehisa M: **AAindex: amino acid index database.** *Nucleic Acids Res* 2000, **28:**374.

26. Frank E, Hall M, Trigg L, Holmes G, Witten IH: **Data mining in bioinformatics using Weka.** *Bioinformatics* 2004, **20:**2479-2481.

27. Ng PC, Henikoff S: **Accounting for human polymorphisms predicted to affect protein function.** *Genome Res* 2002, **12:**436-446.

28. Sunyaev S, Ramensky V, Koch I, Lathe W, 3rd, Kondrashov AS, Bork P: **Prediction of deleterious human alleles.** *Hum Mol Genet* 2001, **10:**591-597.

CHAPTER 4

PREDICTING PATHOGENICITY: UTILITY OF GENE SPECIFIC ALGORITHMS

(April 2011 manuscript submission to *Journal of American Medical Informatics Association*.)

Abstract

Accurate interpretation of gene test results is a key component in customizing patient therapy for personalized medicine. As electronic medical records begin to incorporate genetic information, gene variant annotation has far reaching implications when informing physicians on the most adequate course of treatment. While a growing numbered of authoritative clinical repositories with gene variants and clear association to disease phenotype are beginning to emerge, still there are many more gene variants that have not been annotated with disease association. Computer based mutation severity prediction models derived from clinically curated gene-disease data sets can help bridge this uncertainty gap for novel and uncertain gene variants. Here we present the evaluation of “gene-specific” and “all-gene” predictor models based on a Naïve Bayesian classifier for 20 gene-disease data sets, containing 3,986 variants with well characterized patient conditions.

Introduction

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.

Furthermore, with rapidly evolving technologies such as SNP chip genome wide association studies and next-generation sequencing, genomic analysis is trending faster and cheaper and yielding much larger data sets. 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.[5]

Unfortunately, an increasingly apparent gap exists between 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 to disease phenotype are only now beginning to emerge.[6-10] This is in contrast to existing collections of genome-wide mutations such as dbSNP[11] or OMIM[12] 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.[13-16] Despite their use in laboratories, they do not have sufficient accuracy to predict disease phenotype to the degree necessary to be clinically applicable. This allows opportunities to explore the application of advanced informatics approaches to this problem. This study expands the recently reported Primary Sequence Amino Acid Properties (PSAAP) algorithm [17], which uses a classification approach and amino acid physicochemical properties of the primary amino acid sequence to predict pathogenicity of novel and/or uncertain gene variants. To date, the approach has been applied only to the *RET* proto-oncogene.

To evaluate the generalizability of our gene-specific PSAAP algorithm, we extend its use to a set of 20 genes with clinically curated disease variants. The analyses also compare 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 condition to genotype (genotype-phenotype) were assembled from multiple sources including: cystic fibrosis mutation database curated by Ruslan Dorfman (Hospital for Sick Children, Toronto)[18]; BioPKU database curated by Nenad Blau (University Children's' Hospital, Zurich)[19]; neurofibromatosis type 1 database curated by Ophélia Maertens (Center for Medical Genetics, University Hospital, Ghent) and Collagen, type IV, alpha 5 (*COL4A5*) Mental Retardation Database curated by Judy Savige (Department of Medicine, University of Melbourne) as hosted by Leiden Open Source Variation Database (LOVD)[20-22]; biotinidase (*BTD*) curated by Barry Wolf (Medical Genetics, Henry Ford

Hospital, Detroit)[23]; aryl hydrocarbon receptor interacting protein (*AIP*) curated by Rodrigo Toledo (Endocrine Genetics Unit, University of Sao Paulo Medical School) (personal communication); Disease Databases hosted by Department of Pathology, University of Utah School of Medicine[24] and genetic testing results archived at ARUP Laboratories (Salt Lake City). The clinically curated gene-disease data sets (n=20) containing some 3986 curated variants are summarized in Table 4.1.

This 20 gene collection contained 1639 exonic non-synonymous SNP's (nsSNP) with known outcomes of benign (n=607) and pathogenic (n=1032). The gene variants were characterized using physicochemical properties of the substituted amino acid as previously reported. [25] Briefly, all nsSNP's were characterized by the differences in their 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 AAindex v9.4. [26] AAindex 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 mutation described by an array of variables, corresponding to the absolute value of the difference between wild type and mutant.

Next, 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 the attribute selection methods of correlation-based feature subset selection, SVM-RFE and Relief-F. Thresholds of 95% (or 0.95) for Greedy-Stepwise and Ranker were used during this analysis. All feature selection and Naïve Bayes classification was implemented using the Weka software package.[27]

For each of the 20 genes, random selection was used to build a 2/3 training and a 1/3 test sets with known class labels (benign, pathogenic) and constructed to keep the original ratio of benign and pathogenic constant. Next, based on curated clinical classification of benign or pathogenic, algorithm training and pathogenicity prediction was performed gene-by-gene. Gene-specific models were also tested for prediction of other gene-disease outcomes, by using the

Table 4.1. Clinically-curated gene variant data sets (n=20) with known disease association.

Gene Symbol *Biological Function*	Gene Name *Disease Association*	Curated Variants	Exonic nsSNPs
ACVRL1 activin receptor activity, type 1	activin A receptor type II-like 1 *hereditary hemorrhagic telangiectasia*	332	192
AIP transcription coactivator activity	aryl hydrocarbon receptor interacting protein *familial pituitary adenoma*	102	84
BTD biotin carboxylase activity	biotinidase *biotinidase deficiency*	155	105
CFTR chloride channel regulator activity	cystic fibrosis transmembrane conductance regulator *cystic fibrosis*	252	121
COL4A5 extracellular matrix structural	collagen, type IV, alpha 5 *X-linked Alport syndrome (hereditary nephritis)*	600	266
ENG TGF β receptor activity	endoglin *hereditary hemorrhagic telangiectasia*	397	124
GALT uridylyltransferase activity	galactose-1-phosphate uridylyltransferase *galactosemia*	247	168
GJB2 gap junction channel activity	gap junction protein, beta 2 (connexin 26) *hereditary sensorineural hearing loss*	61	43
MECP2 transcription co-repressor activity	methyl CpG binding protein 2 *Rett syndrome*	26	14
MSH2 guanine/thymine mispair binding	mutS homolog 2 *hereditary nonpolyposis colonrectal cancer*	89	8
MSH6 guanine/thymine mispair binding	mutS homolog 6 *hereditary nonpolyposis colonrectal cancer*	34	10
NF1 Ras GTPase activator activity	neurofibromin 1 *neurofibromatosis type 1*	125	121
PAH phenylalanine catabolism	phenylalanine hydroxylase *phenylketonuria (PKU)*	730	126
PLOD1 procollagen-dioxygenase activity	procollagen-lysine 1, 2-oxoglutarate 5-dioxygenase 1 *Ehlers-Danlos syndrome type VI*	34	12
PMS2 mismatched DNA binding	postmeiotic segregation increased 2 *hereditary nonpolyposis colorectal cancer*	348	45
RET receptor kinase activity	ret proto-oncogene *multiple endocrine neoplasia, (MTC)*	146	97
SLC22A5 carnitine transporter activity	solute carrier family 22, member 5 *primary carnitine deficiency*	95	57
SMAD4 transcription activator activity	SMAD family member 4 *juvenile polyposis syndrome, pancreatic cancer*	86	23
SPINK1 endopeptidase inhibitor activity	serine peptidase inhibitor, Kazal type 1 *hereditary pancreatitis*	73	5
SPRED1 inactivation of MAPK activity	sprouty-related, EVH1 domain containing 1 *Legius syndrome (neurofibromatosis type-like syndrome)*	54	18

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. Due 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) calculated for each classifier algorithm and gene-specific and all-gene permutations. Lastly, our PSAAP gene-specific algorithm performance was compared to well established prediction algorithms such as SIFT[13], PolyPhen[14], PMUT[15] and MutPred[16].

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 Table 4.2.

Table 4.2 Reference amino acid sequence from UniProtKB[a].

Gene symbol	UniProt #	Protein name	AA length	Date accessed
ACVRL1	P37023	ACVL1_HUMAN	503	December 6, 2010
AIP	O00170	AIP_HUMAN	330	January 5, 2011
BTD	P43251	BTD_HUMAN	543	December 6, 2010
CFTR	P13569	CFTR_HUMAN	1480	December 6, 2010
COL4A5	P29400	CO4A5_HUMAN	1685	December 7, 2010
ENG	P17813	EGLN_HUMAN	658	December 7, 2010
GALT	P07902	GALT_HUMAN	379	December 7, 2010
GJB2	P29033	CXB2_HUMAN	226	December 7, 2010
MECP2	P51608	MECP2_HUMAN	486	December 7, 2010
MSH2	P43246	MSH2_HUMAN	934	December 8, 2010
MSH6	P52701	MSH6_HUMAN	1360	December 8, 2010
NF1	P21359	NF1_HUMAN	2839	January 5, 2011
PAH	P00439	PH4H_HUMAN	452	January 6, 2011
PLOD1	Q02809	PLOD1_HUMAN	727	December 9, 2010
PMS2	P54278	PMS2_HUMAN	862	December 9, 2010
RET	P07949	RET_HUMAN	1114	December 9, 2010
SLC22A5	O76082	S22A5_HUMAN	557	December 9, 2010
SMAD4	Q13485	SMAD4_HUMAN	552	January 7, 2011
SPINK1	P00995	ISK1_HUMAN	79	December 9, 2010
SPRED1	Q7Z699	SPRE1_HUMAN	444	December 9, 2010

[a] http://www.uniprot.org.

Results

Overall, the performance of the PSAAP gene-specific trained algorithm was significantly better (8% to 13%) than the "all-gene" model, with p values of 0.00001 (sensitivity), 0.00113 (specificity) and 0.00012 (PPV) as shown in Figure 4.1. For the genes evaluated, the PPV of our gene-specific PSAAP algorithm averaged 89% (82% to 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 6% decreased. Specificity was also generally improved (9% average) for all but *PMS2* (no increase) and *NF1*, which was 5% decreased.

For the genes studied here, the PSAAP gene-specific prediction performs well. PPV values are displayed in Table 4.3. The self against self is plotted on the diagonal in blue with ppv>80 bolded. 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, Table 4.3 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 3 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 genomewide data sets for algorithm development. This can be illustrated by plotting nonsynonymous variants specific to a gene-disease condition as compared to random amino acid substitutions (Figure 4.2). When 1000 random amino acid changes were plotted (Figure 4.2A), a wide distribution evenly covers the entire range of possible substitutions. In contrast, when 1000 pathogenic mutations are graphed,

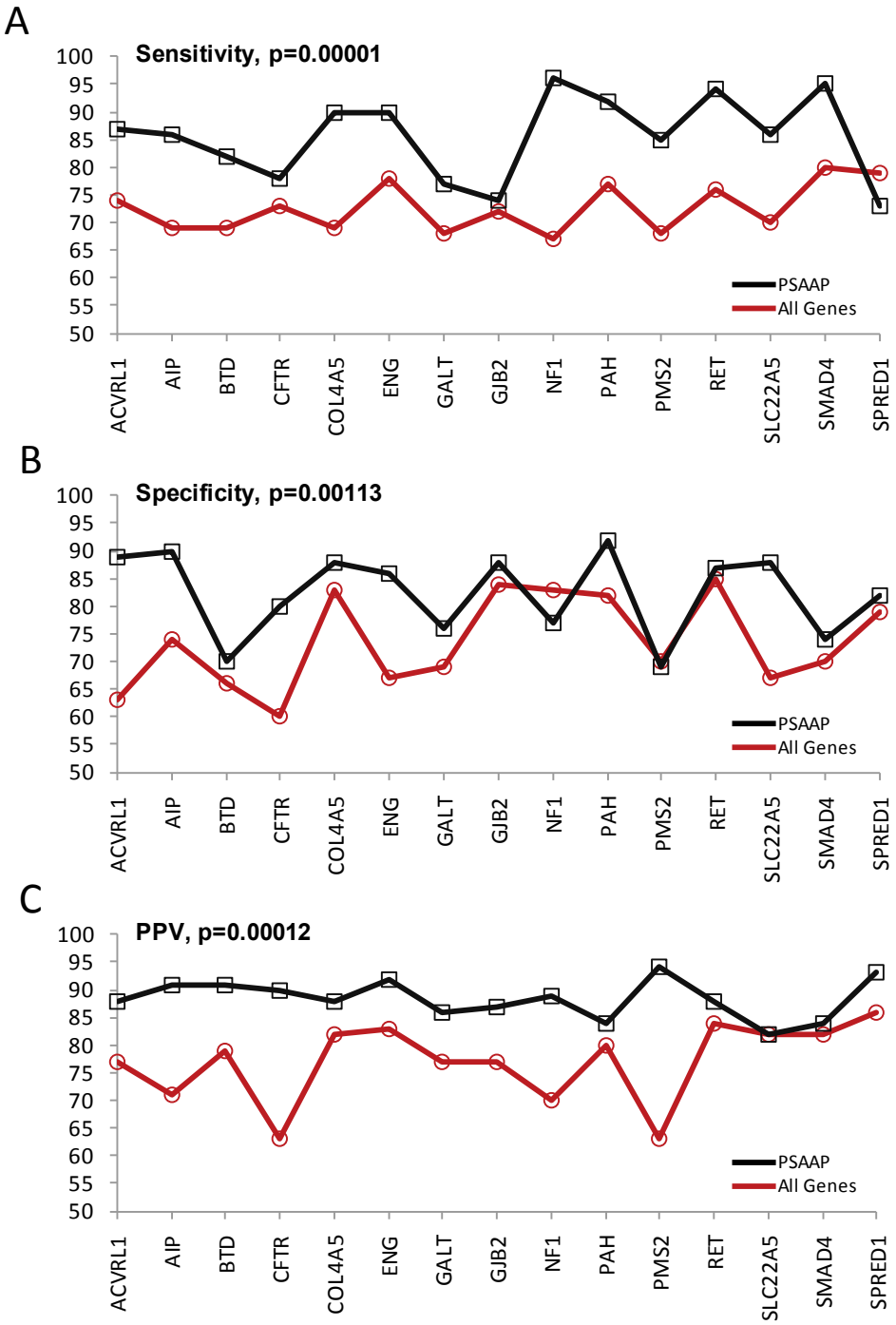


Figure 4.1 Performance of the gene-specific algorithm (PSAAP) as compared to an all-gene algorithm plotted to show A) sensitivity, B) specificity and C) positive predictive value (PPV). Significance of any improvement was calculated using a 2 tailed paired t-test.

Table 4.3 Positive prediction value (PPV) of gene-specific algorithms to predict pathogenicity in other genes, where the blue diagonal represents gene-specific prediction. Bolded results shown in tan squares denote PPV > 80%.

	ACVRL1	AIP	BTD	CFTR	COL4A5	ENG	GALT	GJB2	NF1	PAH	PMS2	RET	SLC22A5	SMAD4	SPRED1
ACVRL1	**88**	**83**	74	70	**84**	77	79	79	**85**	74	76	**80**	**81**	72	78
AIP	72	**91**	62	62	69	59	66	55	68	57	65	63	62	58	62
BTD	77	79	**91**	77	**85**	73	**82**	**81**	**85**	76	70	70	71	**81**	**85**
CFTR	53	62	56	**90**	56	54	59	55	51	54	47	60	53	57	61
COL4A5	47	58	62	51	**88**	**83**	55	61	52	57	46	56	57	56	50
ENG	48	47	62	57	**84**	**92**	49	55	51	56	50	60	54	60	61
GALT	**83**	**82**	**85**	**80**	77	74	**86**	77	**80**	**81**	**85**	**80**	**81**	77	**84**
GJB2	67	56	73	54	56	70	73	**87**	55	66	69	64	62	56	71
NF1	**90**	76	**84**	75	**90**	**89**	75	79	**89**	**83**	75	73	78	**81**	**84**
PAH	62	74	59	55	63	58	64	60	**82**	**89**	58	71	65	60	59
PMS2	66	62	63	61	61	70	55	69	62	71	**88**	66	70	63	56
RET	**84**	69	62	42	64	57	46	72	66	72	45	**94**	49	68	59
SLC22A5	74	66	63	73	72	71	69	68	73	70	68	72	**82**	71	**81**
SMAD4	49	53	65	61	49	64	47	53	67	67	56	52	64	**84**	67
SPRED1	**82**	**85**	**85**	**87**	**87**	**87**	**80**	**84**	**81**	**83**	77	**86**	**84**	**80**	**93**

characteristic trends of specific residues and frequency of substitution are readily seen (Figure 4.2B). Disease-specific examples of this concept are shown in Figure 4.3. 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 4.3A. In the *COL4A5* gene (associated with Alport syndrome), 84% of pathogenic changes involve glycine (G) to other residues (X) as shown in Figure 4.3B. To confirm this trend, further experiments should be performed as additional curated gene-disease collections become available.

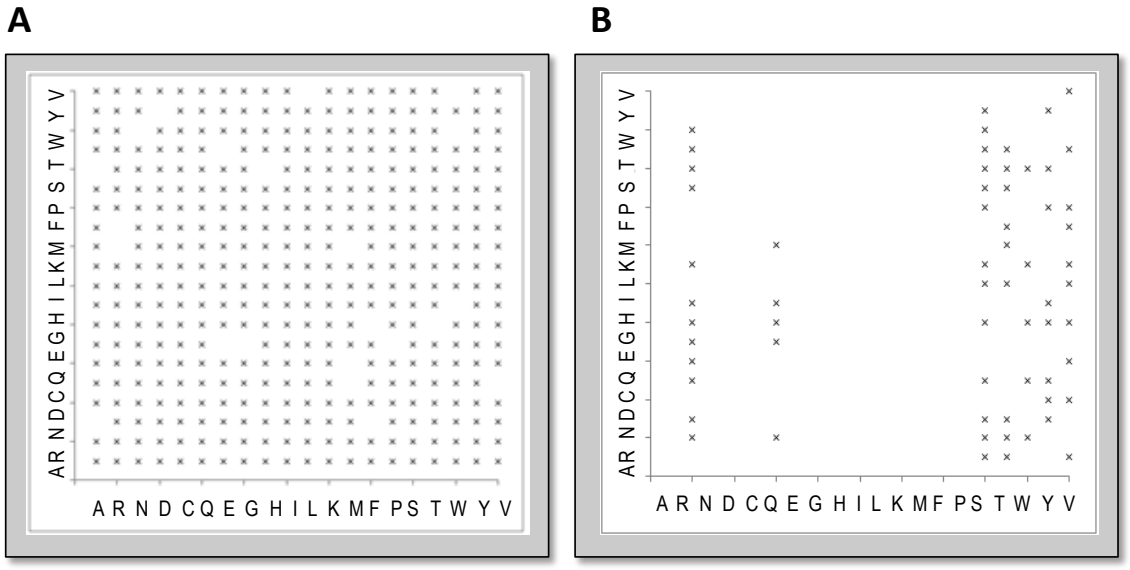


Figure 4.2 Specificity of pathogenic mutations demonstrated by plotting A) simulated random amino acid substitutions (n=1000) showing a wide distribution covering the entire range of possible substitutions and B) known pathogenic mutations (n=1000) showing characteristic trends of specific residues and frequency of substitution.

Although the majority of the gene-specific trained PSAAP models did not perform as well for predicting pathogenicity in other genes-diseases, most still outperformed established algorithms. As shown in Table 4.4, a majority of genes (13 out of 15) analyzed using the gene-specific PSAAP trained algorithm had improved PPV as compared to other algorithms, with the overall PPV increasing 8.8% to 22.0%. For example, the PSAAP model specific for *SPRED1* (93% PPV as seen in Table 4.3), 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 bolded/underlined in Table 4.4.

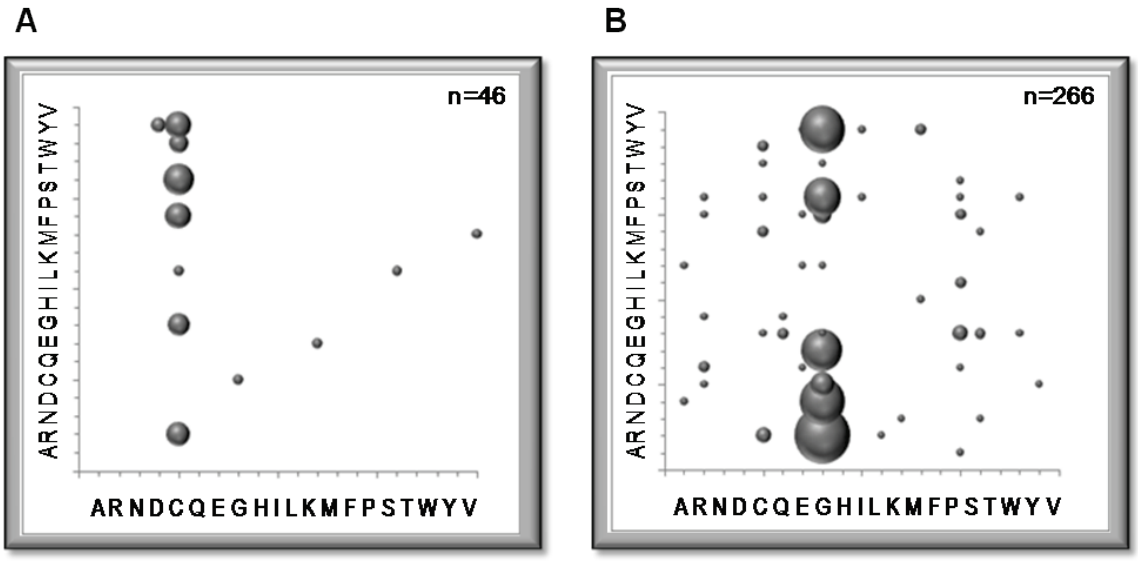


Figure 4.3 Disease specificity of pathogenic mutations demonstrated by plotting A) the *RET* proto-oncogene variants where 79% of pathogenic changes are cysteine [C] to another residue [X] and B) *COL4A5* where 84% pathogenic changes are glycine [G] to another residue [X] again showing characteristic trends of specific residues and frequency of substitution that may be lost when diluting gene-specific data into genome wide computational methods.

Table 4.4 Gene-specific and all-gene algorithm PPV as compared to established algorithms.

Gene	**PSAAP[a]**	**All-gene[b]**	**SIFT[c]**	**PolyPhen[d]**	**PMut[e]**	**MutPred[f]**
ACVRL1	**88**	77	57	67	69	81
AIP	**91**	71	71	73	80	79
BTD	**91**	79	77	72	71	87
CFTR	**90**	63	68	74	70	89
COL4A5	**88**	82	58	74	62	73
ENG	**92**	83	62	64	73	65
GALT	86	77	66	65	58	**87**
GJB2	**87**	77	69	74	67	83
NF1	**89**	70	64	70	70	84
PAH	**89**	80	59	76	77	84
PMS2	**88**	63	64	74	74	72
RET	**94**	84	78	54	72	84
SLC22A5	**90**	82	74	76	53	82
SMAD4	84	82	71	70	85	**86**
SPRED1	**93**	86	71	65	56	71
(avg	**89.3**	77.1	67.3	69.9	69.1	80.5)
(min	**84.0**	63.0	57.0	54.0	53.0	65.0)
(max	**94.0**	86.0	78.0	76.0	85.0	89.0)

[a] Primary Sequence Amino Acid Properties (PSAAP) algorithm, gene-specific trained.
[b] Primary Sequence Amino Acid Properties (PSAAP) algorithm, all-gene (n=20) trained.
[c] Analyzed with default settings at http://sift.jcvi.org.
[d] Analyzed with default settings at http://genetics.bwh.harvard.edu/pph.
[e] Analyzed with default settings at http://mmb.pcb.ub.es/PMut.
[f] Analyzed with default settings at http://mutdb.org/mutpred.

It is important to note that the all-gene trained Bayes predictor also compares favorably to established algorithms, with the average, minimum and maximum PPV for each predictor summarized in Table 4.4. 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% (Table 4.5). Representative examples are shown in Figure 4.4. Interestingly, the gene models with more shared amino acid attributes (*GALT*, 80%; *NF1*, 62%; *SPRED1*, 60%) also had the best generalizability. 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 4 genes, the overlap ranged from only 11% to 37%. Overlap for the all gene data set follows this same trend, showing only 38% overlap between the feature selection methods.

Discussion

The number of authoritative disease and locus specific gene variant collections in use for clinical diagnostics is rapidly growing. These clinically-curated gene variant data sets, 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, with few exceptions, outperforms 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.

Table 4.5 Overlap of minimum set of amino acid properties describing disease association.

	CfsSubset	**Relief**-F	**SVM-RFE**	**Overlap**
ACVRL1	7	39	49	7
AIP	90	29	117	25
BTD	41	20	39	8
CFTR	19	161	139	12
COL4A5	63	65	88	21
ENG	13	82	59	9
GALT	46	40	45	35
GJB2	11	37	145	11
NF1	28	20	39	18
PAH	29	73	129	24
PMS2	13	58	107	11
RET	87	56	47	9
SLC22A5	76	96	87	13
SMAD4	63	65	88	42
SPRED1	59	44	31	27
All GENE	**25**	**56**	**135**	**23**

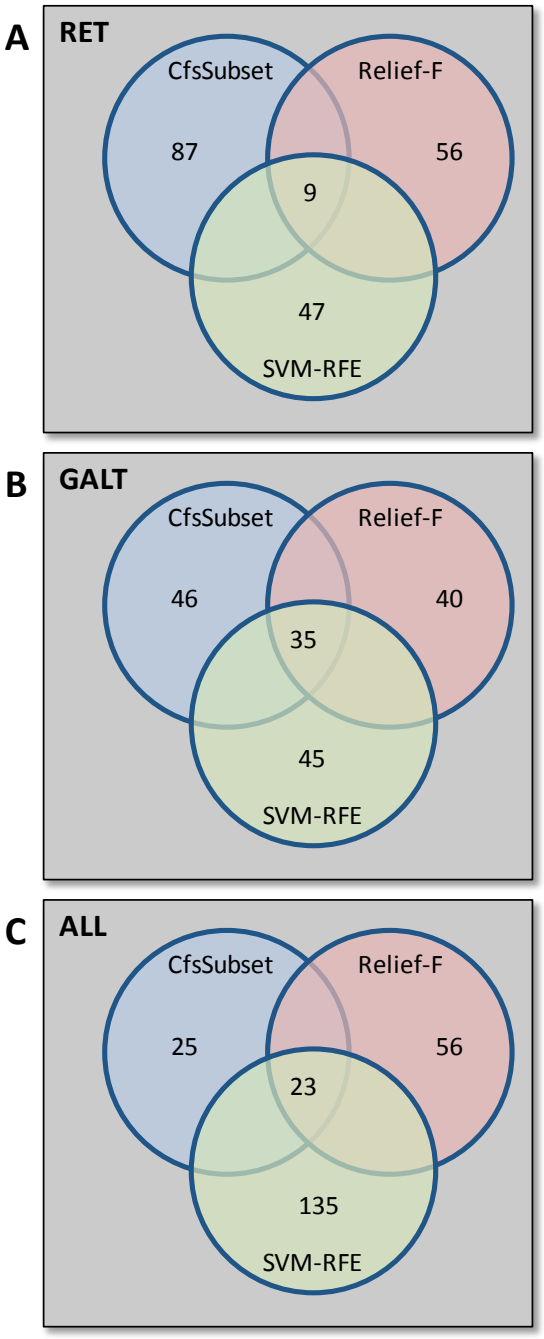


Figure 4.4 Venn diagram showing overlap of amino acid properties to characterize benign and pathogenic gene variants using three feature selection methods (CfsSubset, Relief-F, SVM-RFE). Overlap for A) *RET* with only 14% shared attributes, B) *GALT* with a much higher 80% overlap and C) the all-gene data set with only 38% shared attributes.

For machine learning classifiers, amino acid attributes characteristic of 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. Taking advantage of authoritative (clinically-curated) gene variant collections may avoid some inherent limitations of established prediction algorithms. For example, since this methodology relies on primary sequence information only and uses calculated delta matrices of substituted amino acid properties, no homolog alignment or solved protein structure is necessary.

This study included only gene variant collections with clearly documented disease association and known to the authors – and represents the largest collections to-date of clinically curated gene-disease results as used for diagnostic and gene test reporting purposes. Although correlation of genotype-phenotype offers therapeutic options that would otherwise remain hidden 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. [28] Continued interaction between clinicians and laboratorians to refine mutation-specific clinical classification is imperative to optimal patient care.

Acknowledgements

The authors gratefully acknowledge the extensive disease curation of gene variants by Drs. Dorfman, Blau, Maertens, Savige, Wolf, Toledo and others. This work has been partially supported by ARUP Institute for Clinical and Experimental Pathology, National Library of Medicine Training grant LM007124 and NCRR Clinical and Translational Science Award 1KL2RR025763-01.

References

1. Cotton RG, Al Aqeel AI, Al-Mulla F, Carrera P, Claustres M, Ekong R, Hyland VJ, Macrae FA, Marafie MJ, Paalman MH, et al: **Capturing all disease-causing mutations for clinical and research use: toward an effortless system for the Human Variome Project.** *Genet Med* 2009, **11:**843-849.

2. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467:**1061-1073.

3. Javitt G, Katsanis S, Scott J, Hudson K: **Developing the blueprint for a genetic testing registry.** *Public Health Genomics* 2010, **13:**95-105.

4. Hoffman MA: **The genome-enabled electronic medical record.** *J Biomed Inform* 2007, **40:**44-46.

5. Moore B, Hu H, Singleton M, Reese MG, Yandell M: **Global analysis of disease-related DNA sequence variation in 10 healthy individuals: Implications for whole-genome-based clinical diagnostics** *Genet Med* 2011**:**In Press.

6. Thony B, Blau N: **Mutations in the BH4-metabolizing genes GTP cyclohydrolase I, 6-pyruvoyl-tetrahydropterin synthase, sepiapterin reductase, carbinolamine-4a-dehydratase, and dihydropteridine reductase.** *Hum Mutat* 2006, **27:**870-878.

7. Calderon FR, Phansalkar AR, Crockett DK, Miller M, Mao R: **Mutation database for the galactose-1-phosphate uridyltransferase (GALT) gene.** *Hum Mutat* 2007, **28:**939-943.

8. Margraf RL, Crockett DK, Krautscheid PM, Seamons R, Calderon FR, Wittwer CT, Mao R: **Multiple endocrine neoplasia type 2 RET protooncogene database: repository of MEN2-associated RET sequence variation and reference for genotype/phenotype correlations.** *Hum Mutat* 2009, **30:**548-556.

9. Crockett DK, Pont-Kingdon G, Gedge F, Sumner K, Seamons R, Lyon E: **The Alport syndrome COL4A5 variant database.** *Hum Mutat* 2010, **31:**E1652-1657.

10. Li W, Sun L, Corey M, Zou F, Lee S, Cojocaru A, Taylor C, Blackman S, Stephenson A, Sandford A, et al: **Understanding the population structure of North American patients with cystic fibrosis.** *Clin Genet* 2011, **79:**136-146.

11. **Single Nucleotide Polymorphism Database.** [ncbi.nlm.nih.gov/projects/SNP/].

12. **Online Mendelian Inheritance in Man.** [ncbi.nlm.nih.gov/omim/].

13. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function.** *Nucleic Acids Res* 2003, **31:**3812-3814.

14. Ramensky V, Bork P, Sunyaev S: **Human non-synonymous SNPs: server and survey.** *Nucleic Acids Res* 2002, **30:**3894-3900.

15. 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.

16. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P: **Automated inference of molecular mechanisms of disease from amino acid substitutions.** *Bioinformatics* 2009, **25:**2744-2750.

17. Crockett DK, Piccolo SR, Ridge PG, Margraf RL, Lyon E, Williams MS, Mitchell JA: **Predicting phenotypic severity of uncertain gene variants in the *RET* proto-oncogene.** *PLoS One* 2011**:**In Press.

18. **Cystic Fibrosis Mutation Database.** [http://www.genet.sickkids.on.ca/cftr/app].

19. **BIOPKU:International Database of Patients and Mutations causing BH4-responsive HPA/PKU.** [http://www.biopku.org/biopku/].

20. **Neurofibromatosis Type 1 Database - germline.** [http://medgen.ugent.be/LOVD2/home.php?select_db=NF1_germline].

21. **Neurofibromatosis Type 1 Database - somatic.** [http://medgen.ugent.be/LOVD2/home.php?select_db=NF1_somatic].

22. **Collagen, type IV, alpha 5 (COL4A5) Mental Retardation Database** [http://www.LOVD.nl/COL4A5].

23. **Biotinidase Deficiency and BTD database.** [http://www.arup.utah.edu/database/BTD/BTD_welcome.php].

24. **ARUP online scientific resource, disease databases.** [http://www.arup.utah.edu/database/index.php].

25. Crockett DK, Piccolo SR, Narus SP, Mitchell JA, Facelli JC: **Computational Feature Selection and Classification of RET Phenotypic Severity.** *J Data Mining in Genom Proteomics* 2010, **1:**1-4.

26. Kawashima S, Kanehisa M: **AAindex: amino acid index database.** *Nucleic Acids Res* 2000, **28:**374.

27. Frank E, Hall M, Trigg L, Holmes G, Witten IH: **Data mining in bioinformatics using Weka.** *Bioinformatics* 2004, **20:**2479-2481.

28. Tchernitchko D, Goossens M, Wajcman H: **In silico prediction of the deleterious effect of a mutation: proceed with caution in clinical genetics.** *Clin Chem* 2004, **50:**1974-1978.

CHAPTER 5

CONSENSUS: A FRAMEWORK FOR REPORTING UNCERTAIN GENE VARIANTS

(May 2011 manuscript submission to *Genetics in Medicine*.)

Abstract

As electronic medical records incorporate genetic sequence information, gene variant classification is critical to inform clinicians on the most appropriate course of treatment. Proposed guidelines have recommended classification terminology and definitions for improving laboratory gene variant reporting. A standardized framework however, does not yet exist for quantitative evaluation of disease association for novel gene variants in an objective manner. Gene-specific prediction was trained using clinically curated gene-disease data from the *RET* proto-oncogene. This predictor output was implemented into a Consensus framework, including a weighted metric of existing and complementary prediction algorithms and calculated reference intervals from known disease outcomes. The Consensus model yields accurate evaluation of uncertain or novel gene variants against the backdrop of calculated reference intervals from known benign and pathogenic gene variants. Where existing evidence for a gene variant is scarce, visualization of the Consensus output is also proposed for augmenting diagnostic decisions. Accurate interpretation of gene testing is a key component in customizing patient therapy. A reliable phenotype classification framework with a quantitative metric for evaluation of novel or uncertain gene variants can augment limited clinical information and assist in improving prediction algorithms as gene variant knowledge increases.

Introduction

For appropriate and effective patient treatment, relevant clinical information should be available to the clinician on demand. Accurate interpretation of gene test results, including phenotype association of gene variants, is an important component in customizing patient therapy. Recent endeavors such as the NCBI Genetic Testing Registry, MutaDATABASE, 1000 Genomes and the Human Variome Project draw attention to this growing interest in gene variant annotation and clinical interpretation in human disease.[1-4] Ongoing efforts to catalog human genome variation for many years have led to authoritative repositories of gene variants with clear association to disease phenotype finally beginning to emerge.[5-8]

Rapidly evolving technologies such as SNP chip genome wide association studies and next-generation sequencing has lowered the cost and increased the speed of genomic analysis

yielding much larger data sets.[9] Currently, gene variants are being discovered at an unprecedented pace. One recent report found an average of 3 million variants per personal genome.[10] Unfortunately, an ever-widening gap exists between this fast growing collection of genetic variation and practical clinical implementation due to a lack of understanding of the phenotypic consequences (if any) of any given variant. Although the number of genetic testing laboratories has remained around 600 over the past several years, recent data shows that clinical testing is currently available for well over 2200 different genes or genetic conditions (www.genetests.org). As medical records increasingly incorporate genetic test information, improved decision support approaches are needed to provide clinicians with the preferred course of treatment.[11] Furthermore, for decision support rules to be of value, the clinical relevance of laboratory information must be well understood.[12]

Updated recommendations have been proposed from the American College of Medical Geneticists (ACMG) on reporting and classification of sequence variants, including approaches to help determine the clinical significance of variants of uncertain significance.[13] These guidelines delineate six interpretative categories of gene sequence variation, with defined classifications outlined and the hope of a unified standard terminology in gene test reporting. For improving interpretation of unclassified genetic variants, definitions and terminology has also been recommended by the International Agency for Research on Cancer (IARC), part of the World Health Organization).[14]

Despite these recommendations, however, for genetic laboratories to unify and standardize terminology and classification of gene variant test reporting, various terms such as *deleterious*, *mutation*, *pathogenic* or *causative* of *disease* are still being used.[15] In a similar vein, test results such as *indeterminate*, *unknown*, *uncertain, unclassified* and *undetermined* make it difficult to interpret the significance of a gene test result. Further compounding this issue, word modifiers such as *likely*, *suspected*, *predicted* and *mild*, *moderate* or *severe* sometimes also accompany variant classification. Of this environment, one recent study perceptively noted, “The outcome of this inconsistency for clinicians and patients in such cases is uninformative; unhelpful

at best and, at worst, open to misinterpretation."[16] In this light, the prevailing question becomes how to best help clinicians faced with decisions around gene variants of uncertain significance.

A brief review of the literature indicates that gene test reports of variants of uncertain significance range widely. One laboratory site reported that from 30% to 50% of sequence changes reported for *BRCA1* and *BRCA2*, respectively, were reported as variants of uncertain significance.[17] Similarly, analysis of a second laboratory revealed that a physician who orders *BRCA1* and *BRCA2* testing had an equal likelihood (13%) of receiving a uncertain variant result as seeing a test report containing a known pathogenic mutation.[18] More recent data indicates that identification of variants of uncertain significance has continued to decline to approximately 5% of *BRCA* tests performed – a testament to the importance of maintaining and updating variant databases.[19]

Another well known example is hereditary nonpolyposis colorectal cancer (HNPCC) syndrome, where according to the U.S. Preventive Services Task Force and others, a clinician may expect some 13% to 31% of tests reports to say mutation of unknown significance (uncertain variant).[20, 21] An uncertain variant indicates that the risk of cancer is not fully defined and patient treatment is then based on personal and family history of cancer. Clinicians may be further frustrated when the chance of receiving a test report containing an uncertain variant is even higher for individuals from under-represented ethnic groups due to insufficient data on common polymorphisms from that population.[22] Additionally, newly identified variants from known genes present a greater challenge for interpretation of sequence-based results because they lack traditional confirming evidence of disease association.[23]

Clinician frustration and obstacles to wide adoption of proposed guidelines may be two-fold. First, the lack of any quantitative metric or standardized scale for evaluation of novel or uncertain gene variants make each difficult test result interpretation subjective to location and expertise at hand. A second and closely related challenge is the lack of an objective and standardized framework or context to make that metric meaningful. This quantitative metric and framework for evaluation become especially critical for interpretation of novel and uncertain gene variants where by definition, traditional confirming evidence such as family history, pedigree trios

or sib pairs, confirming literature reports, bench assay biochemical evidence or colleague consensus of disease association is lacking.

With analogy to conventional laboratory testing, a given analysis will have well characterized instrument and assay performance, as well as context for appropriate interpretation such as age and gender specific reference intervals. This allows the laboratory to then standardize with relative ease what is "normal" or "abnormal."

Medical geneticists rely on patient history, family segregation, literature review and trusted colleagues to stay informed of the phenotypic consequences of a given gene variant. In addition, well established computer prediction tools are also employed.[13, 24] Is there a supporting computational method that can serve to replicate this same mental process of gathering evidence from complementary sources, assessing agreement of the evidence and summarizing this evidence into a clinical context for interpretation of the gene variant finding?

In an effort to increase the transparency of providing gene variant evidence in test reporting to the clinical setting, we here propose a practical implementation of our recently reported Primary Sequence Amino Acid Properties (PSAAP) gene-specific predictor into a standardized framework. This combined model of complementary predictors and calculated gene variant reference intervals for a given disease we here term Consensus. Examples of Consensus visualization are also explored for augmenting diagnostic decision making.

Methods

Several clinically curated disease sets of gene variants with known pathogenicity are publicly available at http://www.arup.utah.edu/database/. Each database relies on both medical and molecular expertise, and uniquely displays mutation and clinical information together. All sequence variants are verified for genomic position within a given reference gene. Variants are named following standard Human Genome Organization (HUGO) nomenclature. Archived non-synonymous substitution variants were accessed from the *RET* proto-oncogene database in January 2011.[25]

Various established prediction algorithms were chosen with orthogonal and complementary methodologies such as amino acid substitution penalties, structural disruption,

sequence homology (ortholog conservation) and neural nets. Mutation prediction was then performed for known benign (n=46), known pathogenic (n=51) and uncertain variants (n=45) using our gene-specific PSAAP algorithm, and established algorithms MutPred[26], PMUT[27], PolyPhen[28] and SIFT[29] as previously described.[30, 31] Prediction analysis was performed during January and February 2011 using the respective default settings for each algorithm.

Descriptive statistics (mean, median, standard deviation, minimum and maximum) were calculated for the numerical output from all five predictor algorithms. Spearman correlation coefficients were then determined to evaluate correlation between predictors. Principle components were calculated using factor analysis to determine independence of the five predictors. To compensate for lack of independence between variables, the resulting analysis of variance was used for linear transformation into a weighted sum of predictor values. The weighted average of the five predictor scores was then calculated as the "Consensus" score. All calculations were performed using SAS software, Version 9.1 of the SAS System Copyright © 2002-2003 SAS Institute Inc, Cary, NC, USA.

Next, with analogy to calculating analyte reference intervals for age or gender in traditional laboratory testing, a "reference range" of Consensus scores for *RET* gene variants with known disease outcome was calculated using EP Evaluator 8 (Data Innovations, South Burlington, VT). A nonparametric reference interval was used for benign (n=46) and pathogenic (n=51) with 95% confidence intervals (CI) for the lower and upper bounds. The confidence ratio of the reference interval was also calculated. Due to the reciprocal nature of the SIFT score (where a lower prediction value corresponds to more "pathogenic"), 1-SIFT was used. All predictor scores were normalized to a scale of 0 to 100.

Finally, in order to confirm performance of the Consensus framework, we retrospectively removed seven *RET* gene variants with known disease association (2 benign, 5 pathogenic) from training and test sets and repeated analysis using the proposed model framework. Disease outcome predictions and Consensus scoring was evaluated for each of these variants.

Appropriate graphical summary of diagnostic information, including predictive algorithms are key for visualization and interpretation of any results generated.[32] We have loosely based

the Consensus display on output from representative algorithms such as Scolioscore (http://www.axialbiotech.com) and FibroTest (http://www.biopredictive.com). Finally, the use of radar (radial) plots is well documented and serves to preserve contribution of each predictor in the weighted Consensus sum.[33, 34]

Results

Prediction results (numerical output) from the five algorithms were obtained for *RET* gene variants with known disease association of benign (Table 5.1) and pathogenic (Table 5.2). Predictor results for *RET* gene variants with no reported disease association (uncertain) are summarized in Table 5.3. Results of correlation between predictors and significance of correlation are summarized in Table 5.4. Substantial correlation was seen in at least three of the five predictors (MutPred, PSAAP, and PMUT). This significant correlation between variables indicates that a simple linear sum of predictors could not be used to combine the prediction scores. A weighted predictor sum (Consensus) required linear transformation of predictor outputs as determined by factors analysis.

Factor analysis was performed using principal components to determine weights of association between the five different predictors. More specifically, a set of eigenvectors was applied to weight each predictor accordingly by eigenvalues from principal components, with >80% cumulative variance explained reached using only the first three eigenvalues. Factors analysis and percent variance explained is detailed in Figure 5.1. PRINCOMP results and eigenvalues are summarized in Table 5.5.

A working example of the Consensus score for both a known benign and known pathogenic *RET* gene variant is detailed in Table 5.6, where each predictor sum is weighed and scaled to 100. Using this same method to sum each of the 5 predictors for each gene variant, we then computed reference range metrics for benign and pathogenic variants for the *RET* proto-oncogene. Benign variants ranged from 85 to 243, while pathogenic variants ranged from 305 to 462. Confidence ratios for the calculated reference intervals were 0.09 and 0.16 respectively. The *RET* gene variant Consensus reference intervals are summarized in Table 5.7. Scatter plot distribution of the *RET* Consensus scores for benign and pathogenic is displayed in Figure 5.2A.

Table 5.1 Five predictor results for benign *RET* gene variants.

Variant[a]	PSAPP[b]	MutPred[c]	PMUT[d]	PolyPhen[e]	SIFT[f]
A432E	0.12	0.00	0.63	0.79	1.00
C609W	0.24	0.91	0.89	0.99	0.50
D489N	0.97	0.00	0.24	1.00	0.20
E251K	0.16	0.00	0.63	1.00	0.64
E623K	0.14	0.51	0.83	1.00	0.80
E762Q	0.23	0.68	0.04	1.00	0.09
F1112Y	0.23	0.00	0.03	1.00	0.37
F174S	0.19	0.00	0.11	1.00	0.09
F393L	0.17	0.00	0.39	1.00	0.20
G691S	0.26	0.20	0.66	1.00	0.66
G894S	0.58	0.99	0.77	1.00	0.50
G93S	0.11	0.00	0.51	1.00	0.00
L1061P	0.09	0.75	0.65	1.00	0.10
L40P	0.19	0.00	0.08	1.00	0.00
M1064T	0.14	0.58	0.70	0.75	0.13
N359K	0.16	0.00	0.64	1.00	0.25
N394H	0.25	0.00	0.40	1.00	0.01
N394K	0.12	0.00	0.52	1.00	0.01
P1039Q	0.07	0.72	0.66	0.98	0.30
P1049L	0.09	0.57	0.84	1.00	0.20
P1067S	0.13	0.29	0.79	1.00	0.21
P198T	0.21	0.00	0.62	1.00	0.01
P20L	0.23	0.00	0.27	1.00	0.16
P399L	0.03	0.00	0.72	1.00	0.03
P64L	0.16	0.00	0.17	1.00	0.01
R114H	0.18	0.00	0.61	1.00	0.16
R163Q	0.03	0.00	0.53	0.00	0.25
R180P	0.25	0.00	0.68	1.00	0.24
R231H	0.10	0.00	0.48	0.98	0.56
R287Q	0.12	0.00	0.47	0.75	0.02
R313Q	0.25	0.00	0.57	0.37	0.03
R330Q	0.10	0.00	0.57	0.96	0.44
R360W	0.17	0.00	0.94	0.65	0.00
R475Q	0.08	0.00	0.71	0.79	0.29
R67H	0.15	0.00	0.25	1.00	0.55
R77C	0.11	0.00	0.26	1.00	0.06
R972G	0.52	0.91	0.71	1.00	0.10
R982C	0.16	0.89	0.65	1.00	0.70
S493E	0.07	0.00	0.66	1.00	0.97
S690P	0.62	0.53	0.46	1.00	0.14
S836Y	0.13	0.46	0.78	1.00	0.77
S922Y	0.61	0.92	0.77	0.82	0.80
T278N	0.12	0.00	0.22	1.00	0.29
V145G	0.12	0.00	0.31	1.00	0.00
V376A	0.07	0.13	0.19	0.04	0.60
Y806C	0.67	0.90	0.91	0.43	0.81

[a] RET_HUMAN (UniProt #P07949) used as reference amino acid sequence.
[b] Primary Sequence Amino Acid Properties (PSAAP) algorithm, gene-specific trained.
[c] Analyzed with default settings at http://mutdb.org/mutpred.
[d] Analyzed with default settings at http://mmb.pcb.ub.es/PMut.
[e] Analyzed with default settings at http://genetics.bwh.harvard.edu/pph.
[f] Analyzed with default settings at http://sift.jcvi.org.

Table 5.2 Five predictor results for pathogenic *RET* gene variants.

Variant[a]	PSAPP[b]	MutPred[c]	PMUT[d]	PolyPhen[e]	SIFT[f]
A640G	0.82	0.58	0.13	0.86	0.34
A883F	0.95	0.91	0.64	0.99	0.00
C515S	0.91	0.00	0.83	0.02	0.19
C609F	0.77	0.79	0.84	0.99	0.00
C609G	0.94	0.89	0.88	0.99	0.00
C609R	0.91	0.87	0.93	0.01	0.00
C609S	0.87	0.90	0.61	0.97	0.00
C609Y	0.85	0.90	0.98	0.97	0.00
C611F	0.92	0.93	0.73	1.00	0.00
C611G	0.86	0.92	0.80	1.00	0.00
C611R	0.77	0.92	0.86	1.00	0.00
C611S	0.77	0.90	0.42	0.00	0.00
C611W	0.78	0.94	0.80	0.31	0.00
C611Y	0.74	0.96	0.96	0.29	0.00
C618F	0.79	0.90	0.86	0.20	0.58
C618G	0.88	0.88	0.89	0.97	0.41
C618R	0.97	0.85	0.94	0.99	0.28
C618S	0.97	0.85	0.80	0.94	0.50
C618W	0.89	0.77	0.90	1.00	0.13
C618Y	0.75	0.89	0.98	0.98	0.84
C620F	0.81	0.86	0.81	1.00	0.23
C620G	0.87	0.83	0.70	1.00	0.46
C620R	0.79	0.75	0.88	1.00	0.69
C620S	0.91	0.81	0.70	1.00	0.74
C620W	0.85	0.84	0.81	0.00	0.06
C620Y	0.80	0.84	0.96	0.99	0.00
C630F	0.78	0.76	0.61	0.00	0.06
C630R	0.83	0.70	0.79	1.00	0.84
C630S	0.90	0.73	0.42	1.00	0.34
C630Y	0.87	0.75	0.94	1.00	1.00
C634F	0.85	0.83	0.75	1.00	0.04
C634G	0.82	0.78	0.84	1.00	0.02
C634L	0.77	0.69	0.39	1.00	0.80
C634R	0.85	0.79	0.88	1.00	0.02
C634S	0.87	0.82	0.43	1.00	0.04
C634W	0.89	0.83	0.81	1.00	0.01
C634Y	0.90	0.81	0.97	1.00	0.13
D631Y	0.87	0.69	0.68	1.00	0.02
E768D	0.47	0.56	0.16	1.00	0.13
G533C	0.78	0.00	0.79	1.00	0.00
K666E	0.85	0.51	0.24	1.00	0.31
L790F	0.97	0.80	0.12	1.00	0.01
M918T	0.41	0.54	0.79	1.00	0.00
R844L	0.39	0.84	0.83	1.00	0.10
S649L	0.87	0.66	0.46	0.99	0.65
S891A	0.78	0.67	0.08	1.00	0.31
S922F	0.42	0.89	0.68	1.00	0.00
T946M	0.89	0.76	0.78	0.89	0.60
V292M	0.95	0.50	0.96	1.00	0.14
V804L	0.84	0.73	0.17	1.00	0.03
V804M	0.90	0.77	0.11	1.00	0.09

[a] RET_HUMAN (UniProt #P07949) used as reference amino acid sequence.
[b] Primary Sequence Amino Acid Properties (PSAAP) algorithm, gene-specific trained.
[c] Analyzed with default settings at http://mutdb.org/mutpred.
[d] Analyzed with default settings at http://mmb.pcb.ub.es/PMut.
[e] Analyzed with default settings at http://genetics.bwh.harvard.edu/pph.
[f] Analyzed with default settings at http://sift.jcvi.org.

Table 5.3 Five predictor results for uncertain *RET* gene variants.

Variant[a]	PSAPP[b]	MutPred[c]	PMUT[d]	PolyPhen[e]	SIFT[f]
G321R	0.41	N/A	0.74	0.00	0.50
A510V	0.11	N/A	0.25	0.00	0.08
E511K	0.17	N/A	0.78	0.01	0.24
C531R	0.60	N/A	0.83	1.00	0.28
G533S	0.40	N/A	0.14	0.99	0.00
R600Q	1.00	N/A	0.41	0.00	0.27
K603Q	0.40	N/A	0.45	0.00	0.47
Y606C	0.38	0.54	0.91	0.90	0.23
C609S	0.92	0.90	0.61	1.00	0.00
C611S	0.77	0.90	0.42	1.00	0.00
C630S	0.79	0.73	0.42	1.00	0.34
D631N	0.89	0.61	0.18	0.05	0.19
D631A	0.53	0.65	0.75	0.98	0.01
D631G	0.31	0.62	0.51	0.93	0.02
D631V	0.77	0.60	0.64	0.98	0.00
D631E	0.14	0.39	0.24	0.85	0.07
E632K	0.40	0.63	0.16	0.13	0.19
R635G	0.41	0.71	0.76	1.00	0.10
A640G	0.22	0.58	0.13	0.00	0.50
A641S	0.26	0.59	0.02	0.29	0.09
V648I	0.30	0.58	0.10	0.00	0.75
S649L	0.46	0.66	0.46	1.00	0.32
H665Q	0.30	0.52	0.58	1.00	0.57
K666N	0.07	0.49	0.16	0.99	0.33
R770Q	1.00	0.44	0.49	1.00	0.08
N777S	0.27	0.65	0.35	0.28	0.64
V778I	0.30	0.52	0.02	0.00	0.02
Y791N	0.78	0.89	0.86	0.99	0.00
E805K	0.68	0.88	0.66	1.00	0.00
E818K	0.40	0.51	0.16	0.38	0.19
S819I	0.73	0.77	0.48	1.00	0.00
R833C	0.70	0.59	0.91	1.00	0.00
P841L	0.54	0.60	0.81	1.00	0.67
E843D	0.81	0.70	0.04	0.00	0.12
R844W	0.84	0.77	0.99	1.00	0.00
R844Q	1.00	0.64	0.65	1.00	0.06
R844L	0.83	0.84	0.83	1.00	0.02
M848T	0.62	0.67	0.80	1.00	0.08
I852M	0.84	0.74	0.07	0.99	0.04
L881V	0.60	0.81	0.11	0.96	0.00
R886W	0.74	0.78	0.98	1.00	0.00
S904C	0.82	0.72	0.61	0.98	0.01
S904F	0.75	0.65	0.83	1.00	0.00
K907E	0.60	0.71	0.27	0.99	0.00
K907M	0.58	0.64	0.30	0.99	0.00
R912Q	1.00	0.94	0.55	1.00	0.00

[a] RET_HUMAN (UniProt #P07949) used as reference amino acid sequence.
[b] Primary Sequence Amino Acid Properties (PSAAP) algorithm, gene-specific trained.
[c] Analyzed with default settings at http://mutdb.org/mutpred.
[d] Analyzed with default settings at http://mmb.pcb.ub.es/PMut.
[e] Analyzed with default settings at http://genetics.bwh.harvard.edu/pph.
[f] Analyzed with default settings at http://sift.jcvi.org.

Table 5.4 Descriptive statistics for *RET* gene variants with known disease association (including correlation of predictors and significance).

Simple Statistics

Variable	N	Mean	Std Dev	Median	Minimum	Maximum
MutPred	97	0.5072	0.3881	0.6900	0.0000	0.9900
PMut	97	0.6070	0.2729	0.6798	0.0342	0.9809
Poly	97	0.8537	0.3112	0.9990	0.0000	1.0000
PSAAP	97	0.5336	0.3446	0.6700	0.0300	0.9700
SIFT	97	0.2401	0.2927	0.1300	0.0000	1.0000

Spearman Correlation Coefficients

	MutPred	PMUT	PolyPhen	PSAAP
MutPred	**			
PMut	0.541	**		
Poly	-0.204	-0.289	**	
PSAAP	0.562	0.250	-0.134	**
SIFT	-0.296	-0.118	0.001	-0.118

P-value of correlation

MutPred	PMUT	PolyPhen	PSAAP
**			
<.0001	**		
0.0452	0.0041	**	
<.0001	0.0134	0.1898	**
0.0033	0.2496	0.9953	0.2497

Further demonstrating the utility of a reference interval metric for gene variants, the distribution of Consensus scores for prediction of *RET* uncertain gene variants shows approximate groupings into reference interval ranges as plotted in Figure 5.2B.

In combination, the overall Consensus score may augment the rare instance that a gene-specific prediction does not outperform the existing tools. This advantage of Consensus over a single predictor was seen by removing seven *RET* gene variants with known disease association where originally they were classified as variants of uncertain significance. After excluding these seven variants from the gene-specific training set, analysis using the Consensus framework was repeated. Due to the lack of a representative variant in the training data, PSAAP only called disease association correctly in five out of seven variants. However, in combination, the Consensus score correctly predicted the 6th variant. Closer inspection showed the remaining 7th variant was a nucleotide level "silent" polymorphism (no amino acid change), which could have been recognized by spice effect prediction software.

Finally, one common graphing display technique to preserve contribution of each variable (predictor) is the use of radial plots (also known as radar or spider plots). *RET* Consensus scoring results for the pathogenic variant C609Y and benign variant V376A are shown using radar plots in Figure 5.3. For augmenting clinical decision making, a more comprehensive display for Consensus scoring is shown in Figure 5.4 which incorporates algorithm output, predictor calls, weighted sum and colormetric scale.

Discussion

Currently, there is no widely accepted computational predictor in clinical use for evaluating uncertain gene variants. Furthermore, a lack of standardized framework and quantitative metric for evaluation of disease association of novel and uncertain variants remains an obstacle to widespread implementation of proposed guidelines and definitions of gene test reporting. The analogy of conventional laboratory analyte testing with established cutoffs and reference intervals may serve as a pattern for gene variant testing. In this regard, we have developed a standardized framework and metric for evaluation of uncertain gene variants, with the idea that rather than giving a clinician a 'black box' interpretation of uncertain gene variants,

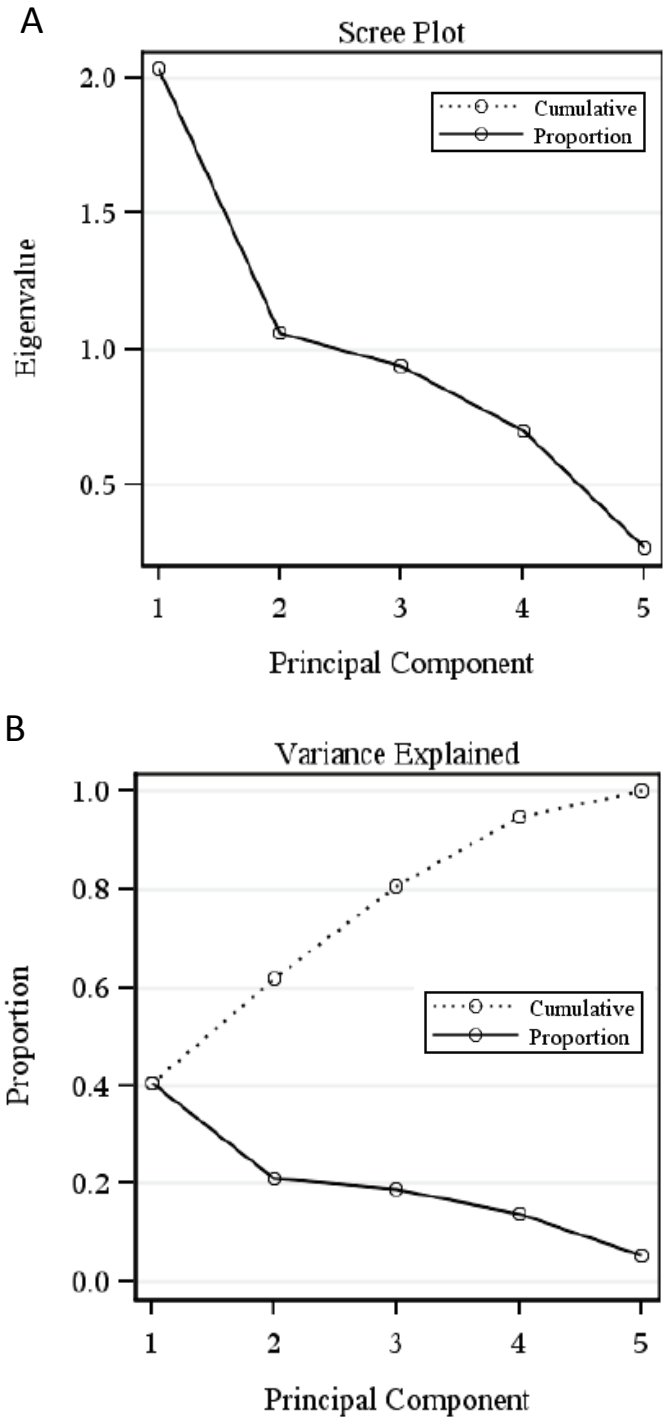


Figure 5.1 Analysis of variance explained as determined by using principal components. A) Scree plot of descending eigenvalues displaying the five principal components corresponding to the combined predictor algorithms. B) Percent variance explained corresponding to the proportion of cumulative input of five combine predictors.

Table 5.5 Principal components and eigenvalues of predictor scores from *RET* gene variants with known disease association.

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	2.03136792	0.97261793	0.4063	0.4063
2	1.05874999	0.11844368	0.2117	0.6180
3	0.94030630	0.23882133	0.1881	0.8061
4	0.70148497	0.43339416	0.1403	0.9464
5	0.26809082		0.0536	1.0000

Eigenvectors

	Prin1	Prin2	Prin3	Prin4	Prin5
MutPred	0.622645	0.141730	0.219197	-.017230	-.737483
PMut	0.464901	0.357804	-.064845	0.754559	0.286843
Poly	-.273897	0.401935	0.773679	0.400519	0.066595
PSAAP	0.563184	0.245172	0.247588	-.440005	0.606474
1 – SIFT	0.063118	0.793869	-.536535	0.276267	0.039931

the evidence and decision making is transparent to clinician so they can use this in consultation with the patient to make treatment decisions.

It is likely that providing this type of information will impact clinical decision making. While critics may argue that relying solely on a computational framework might 'mislead' clinicians in that we don't have the best evidence (i.e., a true known genotype-phenotype correlation), the reality is that clinicians still have to make treatment decisions based on any "uncertain significance" result. We propose that increasing the transparency of gene test evidence and interpretation would only help the clinician as compared to a situation where results that are on the border of benign and those on the border of pathogenic are treated the same. As Consensus is implemented into a laboratory setting, coordination with a clinical site to test how clinicians use the information would be an important and necessary follow-up study.

The lack of a widely accepted standard for computational predictors in a clinical setting remains a serious obstacle in the diagnostic utility of these algorithms. Gene-specific prediction algorithms been shown to be an improvement over existing generalized prediction tools, where a larger data set "n" for training algorithms may not compensate for lower quality of phenotype information. Examples of this gene-disease specific focus using computational prediction have recently been shown for hypertrophic cardiomyopathy and in the *RET* proto-oncogene.[31, 35] We have recently summarized similar efforts in gene-specific prediction for an authoritative 20 gene-disease data set showing similar improved prediction (submitted for publication). Focusing prediction algorithms on authoritative and specific gene-disease settings may aid to bridge this acceptance gap and shed additional light on clinical interpretation of uncertain gene variants. With ongoing efforts to amass gene variation in human disease, newly emerging "authoritative" or "diagnostic grade" clinically curated gene variant archives should be leveraged for training and testing machine learning classification tools.

Another key issue is that disease classification of gene variants evolves over time as new knowledge becomes available. We note that this is a problem whether one uses this proposed framework or the status quo system for dealing with gene test results of uncertain significance. At present, there is no way to communicate new variant knowledge effectively between gene test

Table 5.6 Example of Consensus weighted sum.

Predictor	Prin1	Prin2	Prin3
PSAAP	0.56	0.25	0.25
MutPred	0.62	0.14	0.22
PMUT	0.46	0.36	-0.06
PolyPhen	-0.27	0.40	0.77
one_minus_SIFT	0.06	0.79	-0.54

Variant[a]	PSAPP[b]	MutPred[c]	PMUT[d]	PolyPhen[e]	SIFT[f]	
Pathogenic						
C609Y	0.85	0.90	0.98	0.97	0.00	
Vector1 =	0.85*0.56 +	0.90*0.62 +	0.98*0.46 +	0.97*-0.27 +	(1-0.00)*0.06	= 1.283
Vector2 =	0.85*0.25 +	0.90*0.14 +	0.98*0.36 +	0.97*0.40 +	(1-0.00)*0.79	= 1.869
Vector3 =	0.85*0.25 +	0.90*0.22 +	0.98*-0.06 +	0.97*0.77 +	(1-0.00)*-0.54	= 0.559
						Weighted sum (x 100) = 371.1
Benign						
V376A	0.07	0.00	0.19	1.00	0.60	
Vector1 =	0.07*0.56 +	0.00*0.62 +	0.19*0.46 +	1.00*-0.27 +	(1-0.60)*0.06	= -0.119
Vector2 =	0.07*0.25 +	0.00*0.14 +	0.19*0.36 +	1.00*0.40 +	(1-0.60)*0.79	= 0.786
Vector3 =	0.07*0.25 +	0.00*0.22 +	0.19*-0.06 +	1.00*0.77 +	(1-0.60)*-0.54	= 0.560
						Weighted sum (x 100) = 146.5

[a] RET_HUMAN (UniProt #P07949) used as reference amino acid sequence.
[b] Primary Sequence Amino Acid Properties (PSAAP) algorithm, gene-specific trained.
[c] Analyzed with default settings at http://mutdb.org/mutpred.
[d] Analyzed with default settings at http://mmb.pcb.ub.es/PMut.
[e] Analyzed with default settings at http://genetics.bwh.harvard.edu/pph.
[f] Analyzed with default settings at http://sift.jcvi.org.

Table 5.7 Consensus score reference intervals for *RET* gene variants.

Disease Outcome	N	Lower Limit value	95% CI	Upper Limit value	95% CI	Confidence Ratio
Benign	46	85	<76 to 98	243	231 to >255	>0.09
Pathogenic	51	305	<287 to 319	462	458 to >470	>0.16

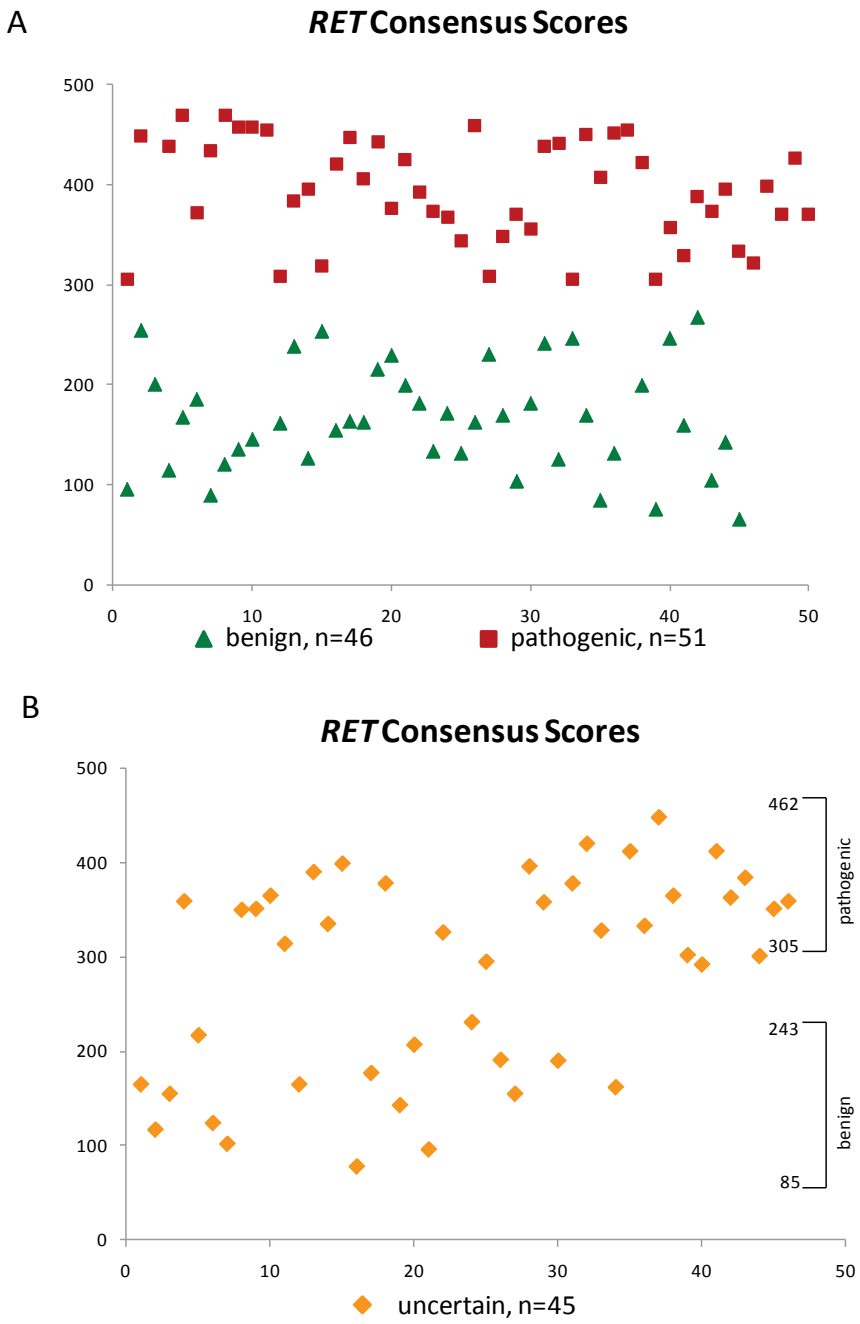
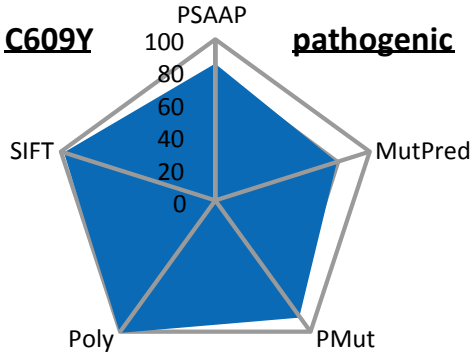


Figure 5.2 Scatter plot visualization of Consensus scores for *RET* gene variants including A) known benign, known pathogenic disease association and B) *RET* gene variants of uncertain significance, showing the utility of reference interval metrics for predicted benign and predicted pathogenic.

A



B

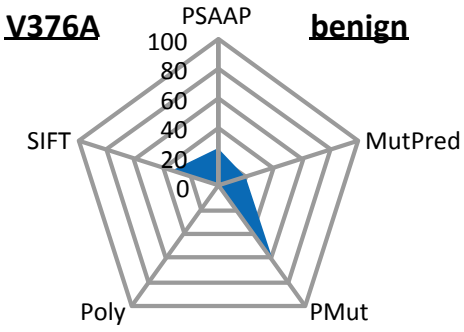
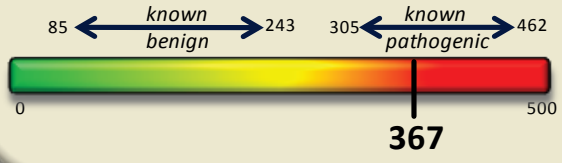


Figure 5.3 Using radar plots for Consensus scoring preserves the contribution of each predictor to the total sum. A) Consensus score plot of 470 (85, 90, 98, 97, 100) for the pathogenic gene variant C609Y. B) Consensus output of 83 (7, 13, 19, 4, 40) for a benign variant V376A.

A

RET – C634R

Predictor	Raw	Call	Consensus	Weights
PSAAP	0.85	pathogenic	89.7	PSAAP(0.563+0.245+0.247)*100
MutPred	0.79	pathogenic	77.7	MutPred(0.622+0.141+0.219)*100
PMUT	0.88	pathological	78.1	PMUT(0.464+0.357-0.064)*100
PolyPhen	1.00	prob damaging	90.1	PolyPhen(-0.237+0.401+0.77)*100
SIFT	0.02	affects function	31.4	1-SIFT(0.063+0.793-0.536)*100
			367	

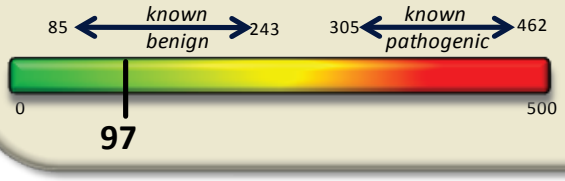


PATHOGENIC

B

RET – G691S

Predictor	Raw	Call	Consensus	Weights
PSAAP	0.26	benign	27.4	PSAAP(0.563+0.245+0.247)*100
MutPred	0.20	benign	19.1	MutPred(0.622+0.141+0.219)*100
PMUT	0.66	neutral	50.1	PMUT(0.464+0.357-0.064)*100
PolyPhen	0.00	benign	0.0	PolyPhen(-0.237+0.401+0.77)*100
SIFT	0.66	tolerated	10.9	1-SIFT(0.063+0.793-0.536)*100
			97	



BENIGN

Figure 5.4 Visualization of the 5 predictor Consensus model including algorithm output, predictor calls, weighting sum and colormetric scale for A) pathogenic gene variant C634R scoring 367 and B) benign variant G691S with a Consensus score of 97.

laboratories and clinicians. Thus, a standardized framework would allow for consistent and objective data provenance for longitudinal tracking of both variants and patient results, where notifying interested parties in updated variant classification and disease association would be more feasible. We also note that developing this framework now using monogenic diseases may allow increased understanding that could eventually be applied to multi-gene panel or whole genome approaches.

For scenarios lacking conventional gene variant evidence, the five specific predictors used for Consensus were carefully chosen due to the varied computational approach of each algorithm. Analysis of variance shows the majority of the weighted average stems from three of the five predictors (PSAAP, MutPred and PMUT). This may be indicative of the unique and varied approach of the three predictors. SIFT and PolyPhen were also included in the Consensus score for a “wisdom of the crowd” historical context due to the fact that many laboratories may already have these prediction algorithms in use.

One limitation of this methodology is the fact that although several popular gene variants collections are ongoing (i.e., dbSNP has recently passed the 12 million unique human gene variant milestone), a relatively small number of clinically-curated and authoritative gene-disease collections exist as used for diagnostic purposes. Fortunately, this number will continue to expand over time, not diminish, as gene-disease associations are better understood and personalized patient treatments advance. Another limitation is that mutation archives often have an unbalanced proportion of disease causing gene variants, and appropriate machine learning techniques must be used to compensate for uneven training and test data. Perhaps the most important limitation to acknowledge is how can we know whether a prediction for a gene variant of uncertain significance is truly correct? The honest response is likely “we can’t.” While only the passage of time may confirm the accuracy of a computational prediction, an important point not to dismiss is – would this approach (or similar) likely lead to better or worse decision making by providers? There may be analogous situations in other existing laboratory tests, where for example, anatomic pathology may yield some ideas that clinicians rely on for decision making.

The pathology report contains all information, not just the “interpretation.” This would imply that more information (not less) is appropriate for clinician decision making.[16, 36]

There may be some perceived liability for a laboratory that would report using this augmented methodology as compared to existing gene test reporting approaches. Although correlation of genotype-phenotype offers therapeutic options that would otherwise remain hidden 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.[37] On the other hand, when results are reported to clinicians and patients as variants of unknown significance, it may take years for sufficient molecular or family evidence to be confirmed for the laboratory to make a final determination. Interpretation of gene test results that are unclear or uncertain may be troubling for patients, and must have some effect (good or bad) on how clinicians manage these patients.[38] Transparent communication of summarized gene variant evidence and continued interaction between clinicians and laboratorians to refine mutation-specific clinical classification is imperative to optimal patient care. Recent examples of this importance have been detailed in newborn screening and case studies from cardiovascular genetics.[39, 40]

Gene variants are currently being identified at a tremendous pace. While many of these sequence changes may be considered as normal population allele variants, some percentage will certainly have disease association. Gene variants may be best leveraged for clinical utility by focusing on specific gene-disease areas. In concert, clinicians and diagnostic laboratories are the best source of authoritative gene variant annotation. Ranking agreement through the use of a weighted Consensus metric of predicted pathogenicity across several complementary algorithms may provide a level of clinical confidence in computational classifiers.

A proposed visual for augmenting the gene test report of an uncertain gene variant using known benign and pathogenic gene variants mapped onto a schematic of the RET protein is displayed in Figure 5.5. The protein diagram image is courtesy of the Human Protein Reference Database (HPRD.org). The variant being evaluated is denoted by “X” along the length of the protein and Consensus scoring of the variant is detailed using both the reference intervals with

colormetric scale and radial chart to show contribution of each predictor. Ongoing efforts include expanding the Consensus scoring framework and phenotype reference intervals to additional genes and diseases. Future efforts will be necessary to incorporate algorithm layers for nucleotide level prediction and functional protein motifs.

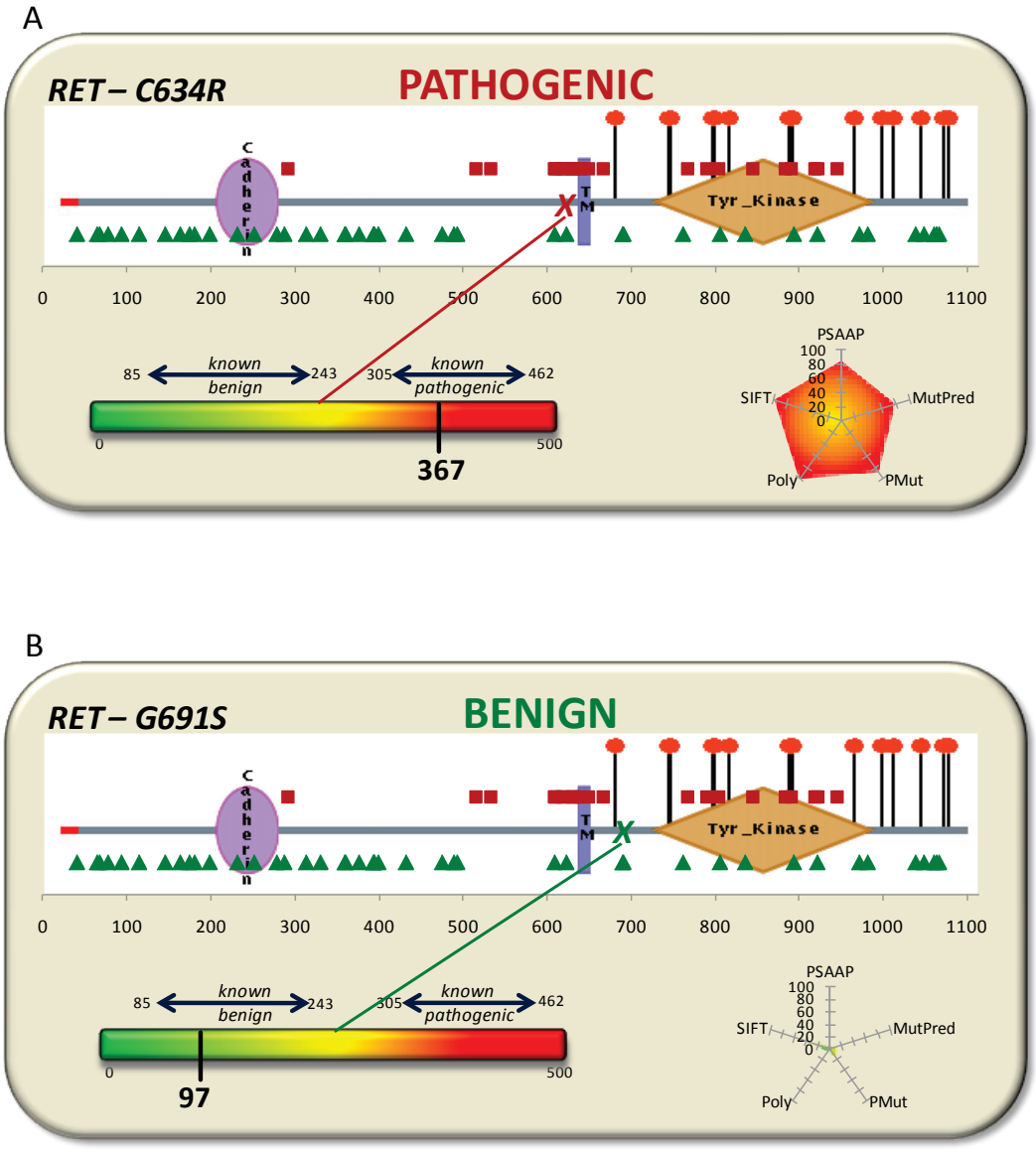


Figure 5.5 Proposed visualization of Consensus scoring using known gene variants plotted on the RET_HUMAN (UniProt #P07949) protein *(image courtesy of HPRD.org)* and algorithm output and spider plots to summarize predictor evidence for A) pathogenic variant C634R scoring 367 and B) benign variant G691S with a Consensus score of 97.

Acknowledgements

This work has been partially supported by ARUP Institute for Clinical and Experimental Pathology®, National Library of Medicine (NLM) training grant #LM007124 and National Center for Research Resources (NCRR) Clinical and Translational Science Award #1KL2RR025763-01.

References

1. Javitt G, Katsanis S, Scott J, Hudson K: **Developing the blueprint for a genetic testing registry.** *Public Health Genomics* 2010, **13:**95-105.

2. Bale S, Devisscher M, Van Criekinge W, Rehm HL, Decouttere F, Nussbaum R, Dunnen JT, Willems P: **MutaDATABASE: a centralized and standardized DNA variation database.** *Nat Biotechnol* 2011, **29:**117-118.

3. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA: **A map of human genome variation from population-scale sequencing.** *Nature* 2010, **467:**1061-1073.

4. Cotton RG, Al Aqeel AI, Al-Mulla F, Carrera P, Claustres M, Ekong R, Hyland VJ, Macrae FA, Marafie MJ, Paalman MH, et al: **Capturing all disease-causing mutations for clinical and research use: toward an effortless system for the Human Variome Project.** *Genet Med* 2009, **11:**843-849.

5. Thony B, Blau N: **Mutations in the BH4-metabolizing genes GTP cyclohydrolase I, 6-pyruvoyl-tetrahydropterin synthase, sepiapterin reductase, carbinolamine-4a-dehydratase, and dihydropteridine reductase.** *Hum Mutat* 2006, **27:**870-878.

6. Li W, Sun L, Corey M, Zou F, Lee S, Cojocaru A, Taylor C, Blackman S, Stephenson A, Sandford A, et al: **Understanding the population structure of North American patients with cystic fibrosis.** *Clin Genet* 2011, **79:**136-146.

7. Crockett DK, Pont-Kingdon G, Gedge F, Sumner K, Seamons R, Lyon E: **The Alport syndrome COL4A5 variant database.** *Hum Mutat* 2010, **31:**E1652-1657.

8. Calderon FR, Phansalkar AR, Crockett DK, Miller M, Mao R: **Mutation database for the galactose-1-phosphate uridyltransferase (GALT) gene.** *Hum Mutat* 2007, **28:**939-943.

9. Li C: **Personalized medicine - the promised land: are we there yet?** *Clin Genet* 2010.

10. Moore B, Hu H, Singleton M, Reese MG, De La Vega FM, Yandell M: **Global analysis of disease-related DNA sequence variation in 10 healthy individuals: Implications for whole genome-based clinical diagnostics.** *Genet Med* 2011, **13:**210-217.

11. Hoffman MA: **The genome-enabled electronic medical record.** *J Biomed Inform* 2007, **40:**44-46.

12. Marshall E: **Human genome 10th anniversary. Human genetics in the clinic, one click away.** *Science* 2011, **331:**528-529.

13. Richards CS, Bale S, Bellissimo DB, Das S, Grody WW, Hegde MR, Lyon E, Ward BE: **ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007.** *Genet Med* 2008, **10:**294-300.

14. Goldgar DE, Easton DF, Byrnes GB, Spurdle AB, Iversen ES, Greenblatt MS: **Genetic evidence and integration of various data sources for classifying uncertain variants into a single model.** *Hum Mutat* 2008, **29:**1265-1272.

15. Vos J, van Asperen CJ, Wijnen JT, Stiggelbout AM, Tibben A: **Disentangling the Babylonian speech confusion in genetic counseling: an analysis of the reliability and validity of the nomenclature for BRCA1/2 DNA-test results other than pathogenic.** *Genet Med* 2009, **11:**742-749.

16. Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, Hogervorst FB, Hoogerbrugge N, Spurdle AB, Tavtigian SV: **Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results.** *Hum Mutat* 2008, **29:**1282-1291.

17. Gomez-Garcia EB, Ambergen T, Blok MJ, van den Wijngaard A: **Patients with an unclassified genetic variant in the BRCA1 or BRCA2 genes show different clinical features from those with a mutation.** *J Clin Oncol* 2005, **23:**2185-2190.

18. Frank TS, Deffenbaugh AM, Reid JE, Hulick M, Ward BE, Lingenfelter B, Gumpper KL, Scholl T, Tavtigian SV, Pruss DR, Critchfield GC: **Clinical characteristics of individuals with germline mutations in BRCA1 and BRCA2: analysis of 10,000 individuals.** *J Clin Oncol* 2002, **20:**1480-1490.

19. Saam J, Burbidge L, Bowles K, Roa B, Pruss D, Schaller J, Reid J, Frye C, Hall MJ, Wenstrup RJ: **Decline in rate of BRCA1/2 variants of uncertain significance: 2002–2008. .** In *National Society of Genetic Counselors Annual Meeting; Los Angeles, CA*. 2008

20. [http://www.chomed.com/genetic-counseling/], Accessed March 29, 2011.

21. Chung DC, Rustgi AK: **The hereditary nonpolyposis colorectal cancer syndrome: genetics and clinical implications.** *Ann Intern Med* 2003, **138:**560-570.

22. John EM, Miron A, Gong G, Phipps AI, Felberg A, Li FP, West DW, Whittemore AS: **Prevalence of pathogenic BRCA1 mutation carriers in 5 US racial/ethnic groups.** *Jama* 2007, **298:**2869-2876.

23. Botkin JR, Teutsch SM, Kaye CI, Hayes M, Haddow JE, Bradley LA, Szegda K, Dotson WD: **Outcomes of interest in evidence-based evaluations of genetic tests.** *Genet Med* 2011, **12:**228-235.

24. Bayrak-Toydemir P, McDonald J, Mao R, Phansalkar A, Gedge F, Robles J, Goldgar D, Lyon E: **Likelihood ratios to assess genetic evidence for clinical significance of uncertain variants: hereditary hemorrhagic telangiectasia as a model.** *Exp Mol Pathol* 2008, **85:**45-49.

25. Margraf RL, Crockett DK, Krautscheid PM, Seamons R, Calderon FR, Wittwer CT, Mao R: **Multiple endocrine neoplasia type 2 RET protooncogene database: repository of MEN2-associated RET sequence variation and reference for genotype/phenotype correlations.** *Hum Mutat* 2009, **30:**548-556.

26. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD, Radivojac P: **Automated inference of molecular mechanisms of disease from amino acid substitutions.** *Bioinformatics* 2009, **25:**2744-2750.

27. 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.

28. Ramensky V, Bork P, Sunyaev S: **Human non-synonymous SNPs: server and survey.** *Nucleic Acids Res* 2002, **30:**3894-3900.

29. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function.** *Nucleic Acids Res* 2003, **31:**3812-3814.

30. Crockett DK, Piccolo SR, Narus SP, Mitchell JA, Facelli JC: **Computational Feature Selection and Classification of RET Phenotypic Severity.** *J Data Mining in Genom Proteomics* 2010, **1:**1-4.

31. Crockett DK, Lyon E, Williams MS, Narus SP, Facelli JC, Mitchell JA: **Predicting phenotypic severity of uncertain gene variants in the RET proto-oncogene.** *PLoS One* 2011, **6:**e18380.

32. Whiting PF, Sterne JA, Westwood ME, Bachmann LM, Harbord R, Egger M, Deeks JJ: **Graphical presentation of diagnostic information.** *BMC Med Res Methodol* 2008, **8:**20.

33. Cole WG: **Integrality and meaning: essential and orthogonal dimensions of graphical data display.** *Proc Annu Symp Comput Appl Med Care* 1993**:**404-408.

34. Saary MJ: **Radar plots: a useful way for presenting multivariate health care data.** *J Clin Epidemiol* 2008, **61:**311-317.

35. Jordan DM, Kiezun A, Baxter SM, Agarwala V, Green RC, Murray MF, Pugh T, Lebo MS, Rehm HL, Funke BH, Sunyaev SR: **Development and validation of a computational method for assessment of missense variants in hypertrophic cardiomyopathy.** *Am J Hum Genet* 2011, **88:**183-192.

36. Gulley ML, Braziel RM, Halling KC, Hsi ED, Kant JA, Nikiforova MN, Nowak JA, Ogino S, Oliveira A, Polesky HF, et al: **Clinical laboratory reports in molecular pathology.** *Arch Pathol Lab Med* 2007, **131:**852-863.

37. Tchernitchko D, Goossens M, Wajcman H: **In silico prediction of the deleterious effect of a mutation: proceed with caution in clinical genetics.** *Clin Chem* 2004, **50:**1974-1978.

38. van Dijk S, van Asperen CJ, Jacobi CE, Vink GR, Tibben A, Breuning MH, Otten W: **Variants of uncertain clinical significance as a result of BRCA1/2 testing: impact of an ambiguous breast cancer risk message.** *Genet Test* 2004, **8:**235-239.

39. Botkin JR, Clayton EW, Fost NC, Burke W, Murray TH, Baily MA, Wilfond B, Berg A, Ross LF: **Newborn screening technology: proceed with caution.** *Pediatrics* 2006, **117:**1793-1799.

40. Caleshu C, Day S, Rehm HL, Baxter S: **Use and interpretation of genetic tests in cardiovascular genetics.** *Heart*, **96:**1669-1675.

CHAPTER 6

SUMMARY AND PERSPECTIVES

Background

Rapidly evolving technologies such as DNA chip arrays and next-generation sequencing continue to decrease costs for genomic analysis yet yield much larger data sets.[1] As such, gene sequence variation in humans is being discovered at an unprecedented pace. A widening gap exists between this growing collection of genetic variation and meaningful clinical implementation due to a lack of clear phenotypic consequences (if any) of a given gene variant. As medical records increasingly incorporate genetic test information, improved decision support approaches are needed to augment gene variant interpretation to provide clinicians with the preferred course of treatment.[2] Furthermore, for laboratory decision support rules to be of greatest value, the clinical relevance of laboratory information must be well understood and prior to clinical decision making for patient treatment.[3-5]

Guidelines have been proposed from the American College of Medical Geneticists (ACMG) on reporting and classification of sequence variants have recently been published.[6] Similar recommendations by the International Agency for Research on Cancer (IARC) have also been published.[7] These recent efforts suggest definitions and approaches to help determine gene variant classification, including the clinical significance of variants of uncertain significance. Despite these recommendations however, genetic laboratories in large part, have failed to unify and standardize terminology and classification of gene variant test reporting. A related challenge is the lack of an objective disease context or decision support framework to make that metric useful. This quantitative metric and framework for evaluation are especially critical for interpretation of novel and uncertain gene variants where there is a lack of confirming evidence such as family history, literature reports or biochemical assays for the laboratory to assemble. Further, this absence of decision support framework and quantitative metric for evaluation of

disease association of novel and uncertain variants remains an obstacle to widespread implementation of proposed guidelines for gene test reporting. Currently, there is no widely accepted computational predictor in clinical use for evaluating uncertain gene variants.

As enormous amounts of genome variation become publically available, data sets for algorithm training typically number into the tens of 1000's or more. Generalized prediction algorithms are largely homolog or structural based, yet characteristic mutation properties of a given disease and gene may be lost when diluted into genomewide data sets. Importantly, there are a growing number of authoritative and clinically curated gene variants collections being assembled for diagnostic purposes. Gene variant phenotype information may be best leveraged for clinical utility by focusing on specific gene-disease areas.

Algorithm agreement in a clinical setting may be just as important for "benign" as it might be for "pathogenic." Ranking agreement of predicted phenotype severity across several complementary algorithms may provide an additional level of clinical confidence in computational classifiers. At a minimum, all-in-agreement "predicted pathogenic" gene variants warrant closer investigation by traditional and molecular techniques. More importantly, agreement between several predictors may provide research priority for novel and uncertain gene variants.

Contributions

An accurate gene-specific Primary Sequence Amino Acid Properties (PSAAP) prediction algorithm was developed using a combination primary amino acid sequence, amino acid properties as descriptors and Naïve Bayes classification based on authoritative training sets. This algorithm was implemented with a weighted average of complementary prediction tools into a Consensus framework of calculated reference intervals of specific disease outcomes for improving laboratory decision making in regards to uncertain gene variants. These efforts have led to the development of several key contributions to not only bioinformatics but also medical genetics and laboratory medicine. Specific contributions include:

1. Many established prediction tools for evaluating pathogenicity of gene variants rely on multiple alignment of sequence conservation across many species and/or solved protein structure to assess structural disruption. Leveraging authoritative gene-variant

collections with high quality genotype-phenotype disease outcomes allowed the development of gene-specific prediction algorithms with improved accuracy as compared to previous prediction tools. Importantly, limitations of sequence homology and known protein structure were not required for the successful use of PSAAP predictions.

2. A related contribution was the important observation that characteristic biochemical and/or structural changes specific to one disease may be lost or diluted when combined with large genome-wide data sets for algorithm development. This suggests that machine learning training sets using a very large "n", as common in many established prediction algorithms, may not always compensate for lower quality data. Stated conversely, characteristic trends of specific amino acid residues and frequency of substitution as found in smaller clinically-curated gene variant disease collections may add power, specificity and accuracy to the prediction tool.
3. Assembly of authoritative and clinically-curated collections of gene variant and disease outcomes including 20 genes and 3986 variants allowed the successful extension of PSAAP prediction in other gene-disease settings. This is the largest such collection of "diagnostic grade" gene variant phenotype information to date. As such, this corpus has proved valuable when building training and test data sets for machine learning classifiers. The generalizability of classification rules across multiple genes and diseases is another key contribution of this study.
4. Combining the PSAAP prediction and other established algorithms with complementary approaches allows a standard quantitative metric of gene variant pathogenicity to be computed. This Consensus score was made from scaling each predictor to 100 and summing a weighted average based on principle components contribution of each predictor.
5. Where traditional confirming evidence is lacking, a laboratory gene test result may be listed as "uncertain significance." To avoid this scenario and with analogy to conventional laboratory testing, disease reference intervals were calculated using Consensus scores of know benign and known pathogenic gene variants. This reference range of

standardized scores from known disease outcomes creates a novel framework for evaluation of gene variants of uncertain significance.

6. Appropriate graphic representation of medical data is vital to communicate disease evidence to clinical decision makers. A proposed visual summary of gathered computational evidence was developed for use in laboratory gene test decision making and to communicate gene test results to the clinician in a more transparent manner.

Significance

Personalized medicine implies all relevant information is available on demand to clinicians. Proper interpretation of gene test results is crucial when customizing patient therapy [8, 9]. Efforts such as the Human Gene Mutation Database, the Human Variome Project and NCBI's Genetic Testing Registry highlight a growing interest in this important area. Further, as electronic medical records begin to incorporate genome sequencing information, this topic has far reaching implications in today's world of whole genome sequencing and health care reforms. In concert, clinicians and diagnostic laboratories are the best source of authoritative gene variant annotation.

Initial efforts to develop clinical laboratory systems included electronic processing and reporting of laboratory test results.[10, 11] Subsequent efforts continued to improve and refine computer information systems used in laboratory settings and to enhance patient care.[12, 13] As genetic information began to influence patient care, laboratory information systems again required adjustments.[14, 15] Currently, the vast amount of genomic data on the horizon for electronic patient records and treatment highlights timely opportunities for genomic data decision support in the laboratory setting. Similar to developing the first clinical laboratory information systems, early efforts are beginning to move forward for laboratory decision support in a gene testing environment, including combining multiple predictors to improve reliability.[16, 17]

Of note, where traditional decision support assists in choosing a correct decision or task, presenting a rich environment of gene variant information may better support accurate gene test interpretation and necessary laboratory recommendations. Importantly, no pre-conceived filtering or formatting of gene variant meta-information should be performed as the various uses of this

data cannot be completely anticipated. Furthermore, this information should not only reduce the overall cognitive load of the test reporting process, but also be made available prior to a clinician acting in appropriate patient care.[4, 5, 18] In this light, laboratory decision support tools for genomics and sequence variant classification may be the next grand challenge for laboratory medicine.

The initial hypothesis of using machine learning classification to train gene-specific algorithms that outperform generalized prediction tools is accepted. This is due to leveraging authoritative gene variant collections, utilizing amino acid physicochemical properties as descriptors of wildtype/mutant and Naïve Bayes classification of benign and pathogenic. Integrating gene-specific prediction into a decision support framework with quantitative metrics for objective evaluation of uncertain gene variants augments laboratory classification of gene test results.

More specifically, laboratory decision support for gene variant classification was improved by using primary protein sequence and biochemical properties of amino acid residues as descriptors of differences between wild type and variant. Further, machine learning classification was used to predict pathogenicity of uncertain gene variants in authoritative gene-disease collections. Finally, a combined score of complementary predictors was computed in a decision support framework of gene-specific disease outcomes.

Limitations

An important limitation is the fact that despite several ongoing gene variants collections such as the 1000 Genomes and the Human Variome Project, a relatively small number of clinically-curated and authoritative gene-disease collections exist as used for diagnostic purposes. Fortunately, this number will likely grow over time, as gene-disease associations are better understood. Another limitation is that gene-specific mutation archives often have an unbalanced proportion of disease causing gene variants, and appropriate machine learning techniques must be used to compensate for uneven training and test data.[19, 20]

Another key issue is the “moving target” of disease classification of gene variants that evolves over time as new knowledge becomes available. We note that this is a problem whether

one uses this proposed framework or the status quo system for dealing with gene test results of uncertain significance. At present, there is no way to communicate new variant knowledge effectively between gene test laboratories and clinicians. Thus, a decision support framework should allow for consistent and objective data provenance for longitudinal tracking of both variants and patient results, where notifying interested parties in updated variant classification and disease association would be more feasible. We also note that developing this framework now, using well-understood monogenic diseases, may allow increased understanding that could eventually be applied to multigene panel or whole genome approaches. Thus, for algorithm training and testing to be as straightforward as possible, scenarios of single gene - single disease (monogenic) were chosen as a starting point for development of these computational tools.

Perhaps the most important limitation to acknowledge is how can one know whether a prediction for a gene variant of uncertain significance is truly correct? The honest response is likely "you can't." While only the passage of time may confirm the accuracy of a computational prediction, an important point not to dismiss is – would this approach (or similar) likely lead to better or worse decision making by providers? There may be analogous situations in other existing laboratory tests, where for example, anatomic pathology may yield some ideas that clinicians rely on for decision making. The pathology report contains all information, not just the "interpretation." This would imply that more information (not less) is appropriate for clinician decision making.[21, 22]

Future efforts

The challenge of clinical interpretation of genetic sequence remains foremost in personalized medicine, regardless if it involves only one gene or an entire genome. With successful efforts to generalize PSAAP and Consensus to other monogenic diseases, the extension of these tools into the pending diagnostic setting of multigene panels and next generation sequencing becomes even more important. Future efforts will also be needed to augment the PSAAP methodology with an algorithm layer for nucleotide level prediction, accounting for "silent" mutational effects not seen in the final translated amino acid product.

The use of machine learning algorithms to classify uncertain gene variants in disease is a promising decision support tool to enhance gene test interpretation and strengthen our underlying knowledge of disease pathogenesis. Furthermore, improved software algorithms to classify gene variants of uncertain significance are necessary to move translational research forward. This novel application of classification algorithms for computational prediction of phenotype severity in uncertain gene variants could be generally applied to any gene-disease setting where an authoritative corpus of curated gene variants exist and where reported mutations impact clinical care.

Personalized medicine cannot advance until the significance of every laboratory result is determined for each patient. Among the key features critical for a decision support framework in clinical genetic testing are first, a reliable scoring metric to predict consequences of a variation that alters protein structure and secondly, ease of access to pertinent gene variant information to aid the laboratory or clinician when performing clinical diagnoses. Thus, one crucial future effort will be to test the potential of Consensus in regards to improved workflow (laboratory audience) as compared to improved interpretation (clinician audience).

References

1. Li C: **Personalized medicine - the promised land: are we there yet?** *Clin Genet* 2010.

2. Hoffman MA: **The genome-enabled electronic medical record.** *J Biomed Inform* 2007, **40:**44-46.

3. Marshall E: **Human genome 10th anniversary. Human genetics in the clinic, one click away.** *Science* 2011, **331:**528-529.

4. Weir CR, Nebeker JJ, Hicken BL, Campo R, Drews F, Lebar B: **A cognitive task analysis of information management strategies in a computerized provider order entry environment.** *J Am Med Inform Assoc* 2007, **14:**65-75.

5. Kushniruk AW, Santos SL, Pourakis G, Nebeker JR, Boockvar KS: **Cognitive analysis of a medication reconciliation tool: applying laboratory and naturalistic approaches to system evaluation.** *Stud Health Technol Inform* 2011, **164:**203-207.

6. Richards CS, Bale S, Bellissimo DB, Das S, Grody WW, Hegde MR, Lyon E, Ward BE: **ACMG recommendations for standards for interpretation and reporting of sequence variations: Revisions 2007.** *Genet Med* 2008, **10:**294-300.

7. Goldgar DE, Easton DF, Byrnes GB, Spurdle AB, Iversen ES, Greenblatt MS: **Genetic evidence and integration of various data sources for classifying uncertain variants into a single model.** *Hum Mutat* 2008, **29:**1265-1272.

8. Del Fiol G, Williams MS, Maram N, Rocha RA, Wood GM, Mitchell JA: **Integrating genetic information resources with an EHR.** *AMIA Annu Symp Proc* 2006**:**904.

9. Lubin IM, McGovern MM, Gibson Z, Gross SJ, Lyon E, Pagon RA, Pratt VM, Rashid J, Shaw C, Stoddard L, et al: **Clinician perspectives about molecular genetic testing for heritable conditions and development of a clinician-friendly laboratory report.** *J Mol Diagn* 2009, **11:**162-171.

10. Lindberg DA: **Electronic processing and transmission of clinical laboratory data.** *Mo Med* 1965, **62:**296-302.

11. Lindberg DA: **Symposium on information science. VII. Electronic reporting, processing, and retrieval of clinical laboratory data.** *Bacteriol Rev* 1965, **29:**554-559.

12. Litzkow L, Ingram W, 3rd, Lezotte D: **The evolution of a functional real-time laboratory records retrieval and archival system.** *J Med Syst* 1977, **1:**177-186.

13. Pryor LR, Freeman VD: **An archival system for clinical laboratory data.** *Am J Clin Pathol* 1979, **72:**1013-1017.

14. Loughman WD, Mitchell JA, Mosher DC, Epstein CJ: **GENFILES: a computerized medical genetics information network. I. An overview.** *Am J Med Genet* 1980, **7:**243-250.

15. Mitchell JA, Loughman WD, Epstein CJ: **GENFILES: a computerized medical genetics information network. II. MEDGEN: the clinical genetics system.** *Am J Med Genet* 1980, **7:**251-266.

16. Gonzalez-Perez A, Lopez-Bigas N: **Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel.** *Am J Hum Genet* 2011, **88:**440-449.

17. Karchin R, Agarwal M, Sali A, Couch F, Beattie MS: **Classifying Variants of Undetermined Significance in BRCA2 with protein likelihood ratios.** *Cancer Inform* 2008, **6:**203-216.

18. Weir CR, Nebeker JR: **Critical issues in an electronic documentation system.** *AMIA Annu Symp Proc* 2007**:**786-790.

19. Saeys Y, Inza I, Larranaga P: **A review of feature selection techniques in bioinformatics.** *Bioinformatics* 2007, **23:**2507-2517.

20. Frank E, Hall M, Trigg L, Holmes G, Witten IH: **Data mining in bioinformatics using Weka.** *Bioinformatics* 2004, **20:**2479-2481.

21. Gulley ML, Braziel RM, Halling KC, Hsi ED, Kant JA, Nikiforova MN, Nowak JA, Ogino S, Oliveira A, Polesky HF, et al: **Clinical laboratory reports in molecular pathology.** *Arch Pathol Lab Med* 2007, **131:**852-863.

22. Plon SE, Eccles DM, Easton D, Foulkes WD, Genuardi M, Greenblatt MS, Hogervorst FB, Hoogerbrugge N, Spurdle AB, Tavtigian SV: **Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results.** *Hum Mutat* 2008, **29:**1282-1291.

APPENDIX

Summary of Fall 2009, BMI 6950 - Special Topics, Machine Learning with Dr. Scott Narus.

Machine Learning

9.8.2009 **Google** search "machine learning" Results 1 - 10 of about 26,600,000.
9.8.2009 **Google Scholar** search "machine learning" Results 1 - 10 of about 2,060,000.
9.8.2009 **PubMed** search "machine learning" Items 1 - 20 of 35174. (Including 2898 Review articles)

First few **original articles** were:
- Singleton WT. An experimental investigation of sewing-machine skill. *Br J Psychol.* 1957 May;48(2):127-32.
- Hively W. Programming stimuli in matching to sample. *J Exp Anal Behav.* 1962 Jul;5:279-98.
- Sterling JA, Lowenburg H. Increased Longevity In Congenital Biliary Atresia. *Ann N Y Acad Sci*. 1963 Dec 30;111:483-508.

Some **recent articles** were:
- Re M, Pesole G, Horner DS. Accurate discrimination of conserved coding and non-coding regions through multiple indicators of evolutionary dynamics. *BMC Bioinformatics*. 2009 Sep 8;10(1):282.
- Potamitis I, Ganchev T, Kontodimas D. On automatic bioacoustic detection of pests: the cases of Rhynchophorus ferrugineus and Sitophilus oryzae. *J Econ Entomol*. 2009 Aug;102(4):1681-90.
- Bostan B, Greiner R, Szafron D, Lu P. Predicting Homologous Signaling Pathways Using Machine Learning. *Bioinformatics.* 2009 Sep 7.

Definitions

Machine learning refers to a system capable of the autonomous acquisition and integration of knowledge. This capacity to learn from experience, analytical observation, and other means, results in a system that can continuously self-improve and thereby offer increased efficiency and effectiveness.

The ability of a machine to recognize patterns that have occurred repeatedly and improve its performance based on past experience.

Ability of a machine to improve its own performance through the use of a software that employs artificial intelligence techniques to mimic the ways by which humans seem to learn, such as repetition and experience.

A scientific discipline that is concerned with the design and development of algorithms that allow computers to learn based on data, such as from sensor data or databases. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data. It is closely related to fields such as statistics, probability theory, data mining, pattern recognition, artificial intelligence, adaptive control, and theoretical computer science.

Extracting useful information from large and complex machine-readable data sets is a problem faced by people in nearly every area of commerce, manufacturing, government, and in every academic discipline and science. Creating an environment conducive to solving such problems requires assembling a community with broad collective expertise in mathematics, statistics, computing, and a wide variety of different application areas.

Quotes I liked

If an expert system--brilliantly designed, engineered and implemented--cannot learn not to repeat its mistakes, it is not as intelligent as a worm or a sea anemone or a kitten.
-Oliver G. Selfridge, from The Gardens of Learning.

Find a bug in a program, and fix it, and the program will work today. Show the program how to find and fix a bug, and the program will work forever.
- Oliver G. Selfridge, in AI's Greatest Trends and Controversies

"Exactly what the computer provides is the ability not to be rigid and unthinking but, rather, to behave conditionally. That is what it means to apply knowledge to action: It means to let the action taken reflect knowledge of the situation, to be sometimes this way, sometimes that, as appropriate. . . . In sum, technology can be controlled especially if it is saturated with intelligence to watch over how it goes, to keep accounts, to prevent errors, and to provide wisdom to each decision."

- Allen Newell, from Fairy Tales

Applications

Applications for machine learning include machine perception, computer vision, natural language processing, syntactic pattern recognition, search engines, medical diagnosis, bioinformatics, brain-machine interfaces and cheminformatics, detecting credit card fraud, stock market analysis, classifying DNA sequences, speech and handwriting recognition, object recognition in computer vision, game playing, software engineering, adaptive websites and robot locomotion.

Algorithms

Machine learning algorithms are organized into a taxonomy, based on the desired outcome of the algorithm.

Common algorithm types include:

Supervised learning - Generates a function that maps inputs to desired outputs. For example, in a classification problem, the learner approximates a function mapping a vector into classes by looking at input-output examples of the function.

Unsupervised learning - Models a set of inputs: labeled examples are not available.

Semi-supervised learning - Combines both labeled and unlabeled examples to generate an appropriate function or classifier.

Reinforcement learning - Learns how to act given an observation of the world. Every action has some impact in the environment, and the environment provides feedback in the form of rewards that guides the learning algorithm.

Transduction - Tries to predict new outputs based on training inputs, training outputs, and test inputs.

Learning to learn - Learns its own inductive bias based on previous experience.

Supervised learning – Machine learning with some "supervision" in the form of pre-determined classifications. Often involves "training" and "test" sets of data with known labels or outcomes or classification which gives feedback about how learning is progressing. The goal is usually to get the computer to learn a classification system that we have created.

Algorithm examples

Artificial neural network, Backpropagation , Boosting, Bayesian statistics, Case-based reasoning, Decision tree learning, Inductive logic programming, Gaussian process regression, Minimum message length (decision trees, decision graphs, etc.), Naive bayes classifier, Nearest Neighbor Algorithm, Symbolic machine learning algorithms, Subsymbolic machine learning algorithms, Support vector machines, Random Forests, Ensembles of Classifiers, Ordinal Classification, Data Pre-processing, Principal Component Analysis (PCA).

Unsupervised learning - Models a set of inputs where labeled examples are not available. The goal is to have the computer learn how to do something that we don't tell it how to do.

Algorithm examples
Multivariate analysis, Artificial neural network, Data clustering, Expectation-maximization algorithm, Self-organizing map, Adaptive resonance theory (ART), Radial basis function network, Generative topographic map, Blind Source Separation or Independent Component Analysis (ICA).

Both techniques can be valuable and which one you choose should depend on the circumstances--what kind of problem is being solved, how much time is allotted to solving it (supervised learning or clustering is often faster than reinforcement learning techniques), and whether supervised learning is even possible**.**

From: http://www.aihorizon.com/essays/generalai/supervised_unsupervised_machine_learning.htm

Variations
In the field of machine learning, semi-supervised learning (SSL) occupies the middle ground, between supervised learning (in which all training examples are labeled) and unsupervised learning (in which no label data are given). Olivier Chapelle, *Semi-Supervised Learning* (MIT Press, 2006). Weakly supervised learning - bootstrapping models from small sets of annotated data. Indirectly supervised learning - end-to-end task evaluation drives learning in an embedded language interpretation module. Semi-supervised Learning - uses both labeled and unlabeled data to perform an otherwise supervised learning or unsupervised learning task. http://pages.cs.wisc.edu/~jerryzhu/pub/SSL_EoML.pdf

Academic
"The Journal of Machine Learning Research (JMLR) provides an international forum for the electronic and paper publication of high-quality scholarly articles in all areas of machine learning."
http://jmlr.csail.mit.edu/
MIT Computer Science and Artificial Intelligence Laboratory http://www.csail.mit.edu/
Alberta Ingenuity Centre for Machine Learning http://www.machinelearningcentre.ca/
Carnegie Mellon University's School of Computer Science, Machine Learning Department
http://www.ml.cmu.edu/
Cornell University's Department of Computer Science http://www.cs.cornell.edu/
University of Cambridge, Computational and Biological Learning Lab http://mlg.eng.cam.ac.uk/
UC Irvine Center for Machine Learning and Intelligent Systems http://cml.ics.uci.edu/
UC Irvine Machine Learning Repository http://archive.ics.uci.edu/ml/
Classification Society of North America (CSNA) http://www.classification-society.org/csna/
International Machine Learning Society http://www.machinelearning.org/
International Conference on Machine Learning http://www.machinelearning.org/icml.html
Machine Learning Summer Schools http://mlss.cc/
Vertical News http://www.verticalnews.com/search_results.php?search_term=Machine+Learning

Commercial/Industry
IBM, Microsoft, Yahoo, and Google have substantial efforts in the area of machine learning algorithms.
IBM - Machine Learning Grouphttp://www.haifa.ibm.com/dept/vst/simulation_vsml_ml.html
IBM - Mathematical & Computational Sciences http://www.zurich.ibm.com/mcs/

HAIFA, Israel and ARMONK, N.Y. - 30 Jun 2009: IBM (NYSE: IBM) today announced the public availability of Milepost GCC, the world's first open source machine learning compiler. The compiler intelligently optimizes applications, translating directly into shorter software development times and bigger performance gains.

Machine Learning for Embedded Programs Optimisation http://www.milepost.eu/

Microsoft Research Cambridge http://research.microsoft.com/en-us/labs/Cambridge/
Machine Learning and Applied Statistics (MLAS) http://research.microsoft.com/en-us/groups/mlas/
The Machine Learning and Perception group http://research.microsoft.com/en-us/groups/mlp/

Yahoo! Research http://research.yahoo.com/Machine_Learning

The Machine Learning group is a team of experts in computer science, statistics, mathematical optimization, and automatic control. We focus on making computers learn abstractions, patterns, conditional probability distributions, and policies from web scale data with the goal to improve the online experience for Yahoo users, partner publishers, and advertisers.

Military

Air Force Office of Scientific Research http://www.afisr.af.mil/units/nasic.asp
Navy Center for Applied Research in Artificial Intelligence http://www.nrl.navy.mil/aic/as/
DARPA http://www.darpa.mil/

Main Funding Sources

Alberta Ingenuity fund http://www.albertaingenuity.ca/programs/funding/centres/machine/learning
Alexa http://www.alexa.com/
ATT/Bell Labs http://www.research.att.com/
DARPA http://www.darpa.mil/
DOD http://www.defense.gov/
Google http://www.google.com/
IBM/Watson http://researchweb.watson.ibm.com/
NIH/NLM http://www.nlm.nih.gov/ep/Grants.html#research,
http://www.nlm.nih.gov/ep/AwardsResearch.html
Netflix http://www.netflixprize.com/
NSF http://www.nsf.gov/

"Machine learning is really good at partially solving just about any problem" (August 20th, 2009 by Partner from cdixon.org)
There's a saying in artificial intelligence circles that techniques like machine learning (and NLP) can very quickly get you, say, 80% of the way to solving just about any (real world) problem, but going beyond 80% is extremely hard, maybe even impossible. The Netflix Challenge is a case in point: hundreds of the best researchers in the world worked on the problem for 2 years and the (apparent) winning team got a 10% improvement over Netflix's in-house algorithm. This is consistent with my own experience, having spent many years and dollars on machine learning projects.
This doesn't mean machine learning isn't useful - it just means you need to apply it to contexts that are fault tolerant: for example, online ad targeting, ranking search results, recommendations, and spam filtering. Areas where people aren't so fault tolerant and machine learning usually disappoints include machine translation, speech recognition, and image recognition.
That's not to say you can't use machine learning to attack these non-fault tolorant problems, but just that you need to realize the limits of automation and build mechanisms to compensate for those limits. One great thing about most machine learning algorithms is you can infer confidence levels and then, say, ship low confidence results to a manual process.
A corollary of all of the above is that it is very rare for startup companies to ever have a competitive advantage because of their machine learning algorithms. If a worldwide concerted effort can only improve Netflix's algorithm by 10%, how likely are 4 people in an R+D department in a startup going to have a significant breakthrough. Modern ML algorithms are the product of thousands of academics and billions of dollars of R+D and are generally only improved upon at the margins by individual companies.
The NetFlix $1million dollar prize was recently rewarded – with round 2 to begin soon…
http://www.wired.com/epicenter/2009/09/how-the-netflix-prize-was-won/

Here at the University of Utah

Machine Learning Group in CS http://www.cs.utah.edu/~cindi/machine-learning.html
Cindi Thompson http://www.cs.utah.edu/~cindi/index.html
(currently with PriceWaterhouse in San Jose)
Suresh Venkatasubramanian http://www.cs.utah.edu/~suresh/research.html
Kiri L. Wagstaff AI/ML Scholarship http://www.wkiri.com/projects/klw-scholarship.html
Hal Daume http://www.cs.utah.edu/~hal/
CS course listing for Business students. http://www.cs.utah.edu/~hal/courses/2008S_ML/,

Other software available for machine learning:

Torch3: a generic machine learning library, particularly good for neural networks, but also a lot more!
MegaM: Optimization software for maximum entropy models, uses conjugate gradient for binary/binomial problems and LM-BFGS for multiclass problems
FastDT: Very fast decision tree learner that implements bagging and boosting
libSVM: a very efficient library for SVMs
SVM-Light: another efficient library for SVMs
Weka: the "defacto" machine learning/datamining library
Mallet: a library for structured prediction with CRFs (plus other stuff)

WEKA Summary

Algorithm Details
NAME
weka.classifiers.bayes.NaiveBayes

SYNOPSIS
Class for a Naive Bayes classifier using estimator classes. Numeric estimator precision values are chosen based on analysis of the training data. For this reason, the classifier is not an UpdateableClassifier (which in typical usage are initialized with zero training instances) -- if you need the UpdateableClassifier functionality, use the NaiveBayesUpdateable classifier. The NaiveBayesUpdateable classifier will use a default precision of 0.1 for numeric attributes when buildClassifier is called with zero training instances.

For more information on Naive Bayes classifiers, see

George H. John, Pat Langley: Estimating Continuous Distributions in Bayesian Classifiers. In: Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 338-345, 1995.

OPTIONS
debug -- If set to true, classifier may output additional info to the console.
displayModelInOldFormat -- Use old format for model output. The old format is better when there are many class values. The new format is better when there are fewer classes and many attributes.
useKernelEstimator -- Use a kernel estimator for numeric attributes rather than a normal distribution.
useSupervisedDiscretization -- Use supervised discretization to convert numeric attributes to nominal ones.

CAPABILITIES
Class -- Binary class, Missing class values, Nominal class
Attributes -- Numeric attributes, Missing values, Empty nominal attributes, Unary attributes, Binary attributes, Nominal attributes
Additional
min # of instances: 0

Naïve Bayes (http://weka.sourceforge.net/doc/weka/classifiers/bayes/NaiveBayes.html)

Package weka.classifiers.bayes

Class Summary	
AODE	AODE achieves highly accurate classification by averaging over all of a small space of alternative naive-Bayes-like models that have weaker (and hence less detrimental) independence assumptions than naive Bayes.
BayesNet	Base class for a Bayes Network classifier.
ComplementNaiveBayes	Class for building and using a Complement class Naive Bayes classifier.

NaiveBayes	Class for a Naive Bayes classifier using estimator classes.
NaiveBayesMultinomial	The core equation for this classifier: P[Ci|D] = (P[D|Ci] x P[Ci]) / P[D] (Bayes rule) where Ci is class i and D is a document
NaiveBayesSimple	Class for building and using a simple Naive Bayes classifier.
NaiveBayesUpdateable	Class for a Naive Bayes classifier using estimator classes.

NAME
weka.classifiers.functions.SimpleLogistic

SYNOPSIS
Classifier for building linear logistic regression models. LogitBoost with simple regression functions as base learners is used for fitting the logistic models. The optimal number of LogitBoost iterations to perform is cross-validated, which leads to automatic attribute selection. For more information see:
Niels Landwehr, Mark Hall, Eibe Frank (2005). Logistic Model Trees.

Marc Sumner, Eibe Frank, Mark Hall: Speeding up Logistic Model Tree Induction. In: 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, 675-683, 2005.

OPTIONS
debug -- If set to true, classifier may output additional info to the console.
errorOnProbabilities -- Use error on the probabilties as error measure when determining the best number of LogitBoost iterations. If set, the number of LogitBoost iterations is chosen that minimizes the root mean squared error (either on the training set or in the cross-validation, depending on useCrossValidation).
heuristicStop -- If heuristicStop > 0, the heuristic for greedy stopping while cross-validating the number of LogitBoost iterations is enabled. This means LogitBoost is stopped if no new error minimum has been reached in the last heuristicStop iterations. It is recommended to use this heuristic, it gives a large speed-up especially on small datasets. The default value is 50.
maxBoostingIterations -- Sets the maximum number of iterations for LogitBoost. Default value is 500, for very small/large datasets a lower/higher value might be preferable.
numBoostingIterations -- Set fixed number of iterations for LogitBoost. If >= 0, this sets the number of LogitBoost iterations to perform. If < 0, the number is cross-validated or a stopping criterion on the training set is used (depending on the value of useCrossValidation).
useAIC -- The AIC is used to determine when to stop LogitBoost iterations (instead of cross-validation or training error).
useCrossValidation -- Sets whether the number of LogitBoost iterations is to be cross-validated or the stopping criterion on the training set should be used. If not set (and no fixed number of iterations was given), the number of LogitBoost iterations is used that minimizes the error on the training set (misclassification error or error on probabilities depending on errorOnProbabilities).
weightTrimBeta -- Set the beta value used for weight trimming in LogitBoost. Only instances carrying (1 - beta)% of the weight from previous iteration are used in the next iteration. Set to 0 for no weight trimming. The default value is 0.

CAPABILITIES
Class -- Binary class, Missing class values, Nominal class
Attributes -- Numeric attributes, Missing values, Date attributes, Empty nominal attributes, Unary attributes, Binary attributes, Nominal attributes
Additional
min # of instances: 1

SimpleLogistic (http://weka.sourceforge.net/doc/weka/classifiers/functions/SimpleLogistic.html)
Class for building a logistic regression model using LogitBoost. Incorporates attribute selection by fitting simple regression functions in LogitBoost. For more information, see master thesis "Logistic Model Trees" (Niels Landwehr, 2003).

Package weka.classifiers.functions

Class Summary	
LeastMedSq	Implements a least median sqaured linear regression utilising the existing weka LinearRegression class to form predictions.
LinearRegression	Class for using linear regression for prediction.
Logistic	Second implementation for building and using a multinomial logistic regression model with a ridge estimator.
MultilayerPerceptron	A Classifier that uses backpropagation to classify instances.
PaceRegression	Class for building pace regression linear models and using them for prediction.
RBFNetwork	Class that implements a normalized Gaussian radial basis function network.
SimpleLinearRegression	Class for learning a simple linear regression model.
SimpleLogistic	Class for building a logistic regression model using LogitBoost.
SMO	Implements John C.
SMOreg	Implements Alex J.Smola and Bernhard Scholkopf sequential minimal optimization algorithm for training a support vector regression using polynomial or RBF kernels.
VotedPerceptron	Implements the voted perceptron algorithm by Freund and Schapire.
Winnow	Implements Winnow and Balanced Winnow algorithms by N.

NAME
weka.classifiers.rules.JRip

SYNOPSIS
This class implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which was proposed by William W. Cohen as an optimized version of IREP.

The algorithm is briefly described as follows:

Initialize RS = {}, and for each class from the less prevalent one to the more frequent one, DO:

1. Building stage:
Repeat 1.1 and 1.2 until the descrition length (DL) of the ruleset and examples is 64 bits greater than the smallest DL met so far, or there are no positive examples, or the error rate >= 50%.

1.1. Grow phase:
Grow one rule by greedily adding antecedents (or conditions) to the rule until the rule is perfect (i.e. 100% accurate). The procedure tries every possible value of each attribute and selects the condition with highest information gain: p(log(p/t)-log(P/T)).

1.2. Prune phase:
Incrementally prune each rule and allow the pruning of any final sequences of the antecedents;The pruning metric is (p-n)/(p+n) -- but it's actually 2p/(p+n) -1, so in this implementation we simply use p/(p+n) (actually (p+1)/(p+n+2), thus if p+n is 0, it's 0.5).

2. Optimization stage:
after generating the initial ruleset {Ri}, generate and prune two variants of each rule Ri from randomized data using procedure 1.1 and 1.2. But one variant is generated from an empty rule while the other is generated by greedily adding antecedents to the original rule. Moreover, the pruning metric used here is (TP+TN)/(P+N).Then the smallest possible DL for each variant and the original rule is computed. The

variant with the minimal DL is selected as the final representative of Ri in the ruleset.After all the rules in {Ri} have been examined and if there are still residual positives, more rules are generated based on the residual positives using Building Stage again.
3. Delete the rules from the ruleset that would increase the DL of the whole ruleset if it were in it. and add resultant ruleset to RS.
ENDDO
Note that there seem to be 2 bugs in the original ripper program that would affect the ruleset size and accuracy slightly. This implementation avoids these bugs and thus is a little bit different from Cohen's original implementation. Even after fixing the bugs, since the order of classes with the same frequency is not defined in ripper, there still seems to be some trivial difference between this implementation and the original ripper, especially for audiology data in UCI repository, where there are lots of classes of few instances.

Details please see:
William W. Cohen: Fast Effective Rule Induction. In: Twelfth International Conference on Machine Learning, 115-123, 1995.

PS. We have compared this implementation with the original ripper implementation in aspects of accuracy, ruleset size and running time on both artificial data "ab+bcd+defg" and UCI datasets. In all these aspects it seems to be quite comparable to the original ripper implementation. However, we didn't consider memory consumption optimization in this implementation.

OPTIONS
checkErrorRate -- Whether check for error rate >= 1/2 is included in stopping criterion.
debug -- Whether debug information is output to the console.
folds -- Determines the amount of data used for pruning. One fold is used for pruning, the rest for growing the rules.
minNo -- The minimum total weight of the instances in a rule.
optimizations -- The number of optimization runs.
seed -- The seed used for randomizing the data.
usePruning -- Whether pruning is performed.

CAPABILITIES
Class -- Binary class, Missing class values, Nominal class

Attributes -- Numeric attributes, Missing values, Date attributes, Empty nominal attributes, Unary attributes, Binary attributes, Nominal attributes
Additional
min # of instances: 3

JRip (http://weka.sourceforge.net/doc/weka/classifiers/rules/JRip.html)
Xin Xu (xx5@cs.waikato.ac.nz), Eibe Frank (eibe@cs.waikato.ac.nz)

Package weka.classifiers.rules

Class Summary	
ConjunctiveRule	This class implements a single conjunctive rule learner that can predict for numeric and nominal class labels.
DecisionTable	Class for building and using a simple decision table majority classifier.
DecisionTable.hashKey	Class providing keys to the hash table
JRip	This class implements a propositional rule learner, Repeated Incremental Pruning to Produce Error Reduction (RIPPER), which is proposed by William W.

M5Rules	Generates a decision list for regression problems using separate-and-conquer.
NNge	NNge classifier.
OneR	Class for building and using a 1R classifier.
PART	Class for generating a PART decision list.
Prism	Class for building and using a PRISM rule set for classifcation.
Ridor	The implementation of a RIpple-DOwn Rule learner.
Rule	Abstract class of generic rule
RuleStats	This class implements the statistics functions used in the propositional rule learner, from the simpler ones like count of true/false positive/negatives, filter data based on the ruleset, etc.
ZeroR	Class for building and using a 0-R classifier.

NAME
weka.classifiers.trees.DecisionStump

SYNOPSIS
Class for building and using a decision stump. Usually used in conjunction with a boosting algorithm. Does regression (based on mean-squared error) or classification (based on entropy). Missing is treated as a separate value.

DecisionStump (http://weka.sourceforge.net/doc/weka/classifiers/trees/DecisionStump.html)
Eibe Frank (eibe@cs.waikato.ac.nz),

OPTIONS
debug -- If set to true, classifier may output additional info to the console.

CAPABILITIES
Class -- Date class, Binary class, Numeric class, Missing class values, Nominal class

Attributes -- Numeric attributes, Missing values, Date attributes, Empty nominal attributes, Unary attributes, Binary attributes, Nominal attributes

Additional
min # of instances: 1

Package weka.classifiers.trees

Class Summary	
ADTree	Class for generating an alternating decision tree.
DecisionStump	Class for building and using a decision stump.
Id3	Class implementing an Id3 decision tree classifier.
J48	Class for generating an unpruned or a pruned C4.5 decision tree.
LMT	Class for "logistic model tree" classifier.
M5P	M5P.
NBTree	Class for generating a Naive Bayes tree (decision tree with Naive Bayes classifiers at the leaves).

RandomForest	Class for constructing random forests.
RandomTree	Class for constructing a tree that considers K random features at each node.
REPTree	Fast decision tree learner.
UserClassifier	Class for generating an user defined decision tree.

Implements John C. Platt's sequential minimal optimization algorithm for training a support vector classifier using polynomial or RBF kernels. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. (Note that the coefficients in the output are based on the normalized/standardized data, not the original data.) Multi-class problems are solved using pairwise classification. To obtain proper probability estimates, use the option that fits logistic regression models to the outputs of the support vector machine. In the multi-class case the predicted probabilities will be coupled using Hastie and Tibshirani's pairwise coupling method. Note: for improved speed standardization should be turned off when operating on SparseInstances.

http://weka.sourceforge.net/doc/weka/classifiers/functions/SMO.html

NAME
weka.classifiers.lazy.IBk

SYNOPSIS
K-nearest neighbours classifier. Can select appropriate value of K based on cross-validation. Can also do distance weighting.

For more information, see

D. Aha, D. Kibler (1991). Instance-based learning algorithms. Machine Learning. 6:37-66.

OPTIONS
KNN -- The number of neighbours to use.

crossValidate -- Whether hold-one-out cross-validation will be used to select the best k value.

debug -- If set to true, classifier may output additional info to the console.

distanceWeighting -- Gets the distance weighting method used.

meanSquared -- Whether the mean squared error is used rather than mean absolute error when doing cross-validation for regression problems.

nearestNeighbourSearchAlgorithm -- The nearest neighbour search algorithm to use (Default: weka.core.neighboursearch.LinearNNSearch).

windowSize -- Gets the maximum number of instances allowed in the training pool. The addition of new instances above this value will result in old instances being removed. A value of 0 signifies no limit to the number of training instances.

http://weka.sourceforge.net/doc/weka/classifiers/lazy/IBk.html

Class Summary	
IB1	IB1-type classifier.
IBk	*K*-nearest neighbours classifier.

KStar	K* is an instance-based classifier, that is the class of a test instance is based upon the class of those training instances similar to it, as determined by some similarity function.
LBR	Lazy Bayesian Rules implement a lazy learning approach to lessening the attribute-independence assumption of naive Bayes.
LWL	Locally-weighted learning.

NAME
weka.classifiers.rules.ZeroR

SYNOPSIS
Class for building and using a 0-R classifier. Predicts the mean (for a numeric class) or the mode (for a nominal class).

OPTIONS
debug -- If set to true, classifier may output additional info to the console.

CAPABILITIES
Class -- Date class, Binary class, Numeric class, Missing class values, Nominal class

Attributes -- Numeric attributes, Missing values, Relational attributes, String attributes, Date attributes, Empty nominal attributes, Unary attributes, Binary attributes, Nominal attributes
Additional
min # of instances: 0

Strengths and limitations of common machine learning classifier algorithms.

Weka Classifier Category	Representative Algorithm	Description
Bayes	**Naïve Bayes**	Assumes that the conditional probabilities of the independent variables are statistically independent (class conditional independence). However, bias in estimating probabilities often may not make a difference in practice. It is the order of the probabilities, not their exact values that determine the classifications. Kernel density estimation is used for numeric attributes.

Algorithm Strengths
Naive Bayes classifiers can handle an arbitrary number of independent variables whether continuous or categorical. They have also exhibited high accuracy and speed when applied to large databases. Although relatively simple (mathematically) they are surprisingly effective for real-world problems.

Algorithm Limitations
Often not capable of solving more complex classification problems. Performance does not improve significantly with increasing sample size. Unrealistic assumption of attribute independence. If there are strong attribute dependencies, this algorithm may not perform as well as other algorithms designed to handle dependencies.

Weka Classifier Category	Representative Algorithm	Description
Regression	**SimpleLogistic**	Builds linear logistic regression models.

		LogitBoost with simple regression functions as base learners is used for fitting the logistic models. The optimal number of iterations to perform is cross-validated, which leads to automatic attribute selection.

Algorithm Strengths
Supports "automatic attribute selection" which places emphasis on variables that appear to influence the outcome more than others.

Algorithm Limitations
Uses a fairly complex approach for arriving at the weights and making the final predictions, so the overall model may not be intuitive to the user.

Weka Classifier Category	Representative Algorithm	Description
Rules	**JRip**	Induces rules from the data based on observed predictive patterns.

Algorithm Strengths
Rules for classifying instances are easily interpretable, i.e. "transparent" models that enable the user to understand exactly how they work. Rules are typically based on multiple attributes, so they explicitly account for dependencies between attributes.

Algorithm Limitations
Rules may become very complicated and thus be less interpretable and possibly less generalizable (called "overfitting"). Pruning can be used to simplify the rules, but often difficult to find a balance between pruning enough and pruning too much. Does not create a condition for a keyword which appears in more than two categories – which might lead to contradicting rules.

Weka Classifier Category	Representative Algorithm	Description
Functions	**SMO**	A sequential minimal optimization algorithm for training a support vector classifier using polynomial or RBF kernels. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default, so the coefficients in the output are based on the normalized/standardized data, not the original data. It uses quadratic programming optimization to break complex learning tasks down into smaller ones that are easier to solve.

Algorithm Strengths
Extremely powerful non-linear classifier (excellent performance in binary classification).

Algorithm Limitations
Computationally demanding to train and to run. Sensitive to noisy data. Prone to overfitting and thus may not generalization. Only predicts a class label but does not provide any estimate of the underlying probability.

Weka Classifier Category	Representative Algorithm	Description

Trees	**DecisionStump**	Builds a classification tree through a process known as binary recursive partitioning – basing its decision on only one attribute. Missing data is treated as a separate value.

Algorithm Strengths
Fast; simple to implement; can convert result to a set of easily interpretable rules; handles noisy data. A good algorithm choice when the data mining task is classification or prediction of outcomes and the goal is to generate rules that can be easily understood,. In cases where a single attribute is an excellent predictor, this algorithm performs well because it creates a simple model rather than overfitting.

Algorithm Limitations
"Univariate" splits/partitioning using only one attribute at a time so limits types of possible trees. When dependencies exist between attributes, this algorithm likely will not perform as well as others.

Weka Classifier Category	Representative Algorithm	Description
Lazy	**IBk**	K nearest neighbor is a very simple algorithm. It is based on the minimum distance from a query instance to the training samples and determine the K-nearest neighbors. Can also do distance weighting.

Algorithm Strengths
Can be robust to noisy training data, especially when using inverse square of weighted distance as the “distance”. Remains effective even with large training data sets. Nonparametric architecture, simple and powerful, requires no training time.

Algorithm Limitations
Strong dependence on a distance metric. Memory intensive, so classification and estimation are slow.

References

Dyer CR. http://pages.cs.wisc.edu/~dyer/cs540/notes/learning.html

Langley P, Iba W, and Thompson K. *An analysis of Bayesian classifiers*. In Proceedings of the Tenth National Conference on Artificial Intelligence. AAAI Press, San Jose, 1992 pages 223-228.

Statnikov A, Wang L, Aliferis CF. *A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification.* 2008. BMC bioinformatics 9: 319

Weka (sourceforge). http://weka.sourceforge.net/doc/weka/classifiers/

Witten and Frank: *Data Mining: Practical machine learning tools and techniques.* 2nd edition edn. San Francisco: Morgan Kaufmann; 2005.

Teknomo, Kardi. K-Nearest Neighbors Tutorial. http:\\people.revoledu.com\kardi\ tutorial\KNN\