Journal of Data Mining in Genomics & Proteomics - Open Access

www.omicsonline.org

Research Article
OPEN ACCESS Freely available online
doi:10.4172/2153-0602.1000103

Computational Feature Selection and Classification of *RET* Phenotypic Severity

David K. Crockett[1,2*], Stephen R. Piccolo[1], Scott P. Narus[1], Joyce A. Mitchell[1] and Julio C. Facelli[1,3]

1University of Utah School of Medicine, Biomedical Informatics, Salt Lake City, UT, USA
2University of Utah School of Medicine, Pathology, Salt Lake City, UT, USA
3University of Utah Center for High Performance Computing, Salt Lake City, UT, USA

Abstract

Although many reported mutations in the *RET* oncogene have been directly associated with hereditary thyroid carcinoma, other mutations are labelled 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, and regression, nearest neighbour, 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.

Keywords: Classification; Physicochemical properties of amino acids; Gene variant; Machine learning; Phenotype; Prediction; *RET*

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://

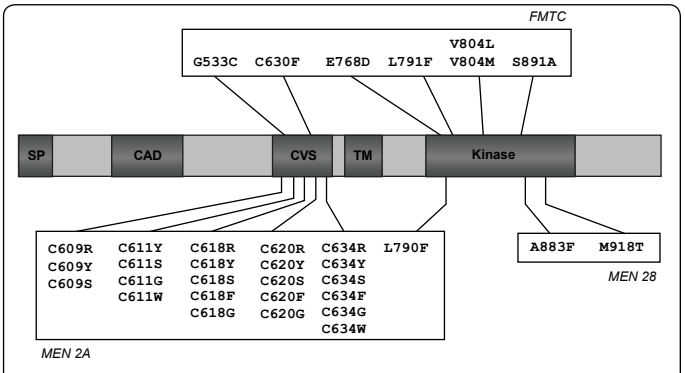

Figure 1: *RET* protein domains. Schematic view of the RET oncoprotein showing conserved domains of signal peptide (SP), cadherin repeat domains (CAD), cysteine rich region (CYS), transmembrane domain (TM), and protein tyrosine kinase (Kinase) where reported variants associate to three specific disease phenotypes; familial medullary thyroid cancer (FMTC), multiple endocrine neoplasia type 2A (MEN2A) and multiple endocrine neoplasia type 2B (MEN2B).

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. [20]

Table 1: *RET* mutation guided therapy for surgical removal of the thyroid[a].

***Corresponding author:** David K. Crockett, ARUP Laboratories, University of Utah School of Medicine, Biomedical Informatics, 500 Chipeta Way, Salt Lake City, Utah 84108, USA, Tel: 801-583-2787; Fax: 801-584-5109; Email: david.crockett@utah.edu

Received October 20, 2010; **Accepted** December 16, 2010; **Published** December 16, 2010

Citation: Crockett DK, Piccolo SR, Narus SP, Mitchell JA, Facelli JC (2010) Computational Feature Selection and Classification of *RET* Phenotypic Severity. J Data Mining in Genom Proteomics 1:103. doi:10.4172/2153-0602.1000103

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 1) displays reported disease causing variants as associated with different MEN2 phenotypes. (Table 1) summarizes mutation-guided therapy for thyroid cancer where surgical removal of thyroid is guided by codon position of the *RET* mutation.

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 [14,15]. The Correlation-based Feature Subset Selection algorithm [16], 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

Propertya	Original Source	PubMed ID
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/

Table 2: Feature selection (n=23) from 544 amino acid properties from AAindex.

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

Table 3: Summary of classification performance for machine learning algorithms.

RET Gene Variant[a]	PolyPhen	SIFT	MutPred
Curated Outcome	Prediction[b]	Prediction[c]	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)	probably 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

Table 4: Comparison of mutation prediction for selected *RET* mutations using PolyPhen, SIFT and MutPred.

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 [17]. 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 [18].

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). 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 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 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 [19]. 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 3). Analysis of the *RET* mutations using PolyPhen correctly identified 68 out 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 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.

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.

References

1. Ramensky V, Bork P, Sunyaev S (2002) Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30: 3894-3900.
2. Ng PC, Henikoff S (2003) SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31: 3812-3814.
3. Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, et al. (2009) Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics 25: 2744-2750.
4. Spencer DS, Stites WE (1996) The M32L substitution of staphylococcal nuclease: disagreement between theoretical prediction and experimental protein stability. J Mol Biol 257: 497-499.
5. Kang HH, Williams R, Leary J, kConFab Investigators, Ringland C, et al. (2006) Evaluation of models to predict BRCA germline mutations. Br J Cancer 95: 914-920.
6. Engelhardt BE, Jordan MI, Muratore KE, Brenner SE (2005) Protein molecular function prediction by Bayesian phylogenomics. PLoS Comput Biol 1: 45.
7. Woolley S, Johnson J, Smith MJ, Crandall KA, McClellan DA (2003) TreeSAAP: selection on amino acid properties using phylogenetic trees. Bioinformatics 19: 671-672.
8. Georgiev AG (2009) Interpretable numerical descriptors of amino acid space. J Comput Biol 16: 703-723.
9. UniProt consortium (2008) The universal protein resource (UniProt). Nucleic Acids Res 36:190-195.
10. Eng C, Clayton D, Schuffenecker I, Lenoir G, Cote G, et al. (1996) The relationship between specific RET proto-oncogene mutations and disease phenotype in multiple endocrine neoplasia type 2. International RET mutation consortium analysis. Jama 276: 1575-1579.
11. Kouvaraki MA, Shapiro SE, Perrier ND, Cote GJ, Gagel RF, et al. (2005) RET proto-oncogene: a review and update of genotype-phenotype correlations in hereditary medullary thyroid cancer and associated endocrine tumors. Thyroid 15: 531-544.
12. Kloos RT, Eng C, Evans DB, Francis GL, Gagel RF, et al. (2009) Medullary thyroid cancer: management guidelines of the American Thyroid Association. Thyroid 19: 565-612.
13. Margraf RL, Crockett DK, Krautscheid PM, Seamons R, Calderon FR, et al. (2009) Multiple endocrine neoplasia type 2 RET protooncogene database: repository of MEN2-associated RET sequence variation and reference for genotype/phenotype correlations. Hum Mutat 30: 548-556.
14. Kawashima S, Kanehisa M (2000) AAindex: amino acid index database. Nucleic Acids Res 28: 374.
15. Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, Katayama T, et al. (2008) AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 36: 202-205.
16. Hall MA (2000) Correlation-based feature selection of discrete and numeric class machine learning, in Computer Science Working Papers. University of Waikato, Department of Computer Science: Hamilton, New Zealand.
17. Frank E, Hall M, Trigg L, Holmes G, Witten IH (2004) Data mining in bioinformatics using Weka. Bioinformatics 20: 2479-2481.
18. Hand DJ (2009) Measuring classifier performance: a coherent alternative to the area under the ROC curve. Machine Learning 77: 103-123.
19. Lai AZ, Gujral TS, Mulligan LM (2007) RET signaling in endocrine tumors: delving deeper into molecular mechanisms. Endocr Pathol 18: 57-67.
20. Massoll N, Mazzaferri EL (2004) Diagnosis and management of medullary thyroid carcinoma. Clin Lab Med 24: 49-83.