Predicting Disease-Related Mutations Based on Protein Domain Framework

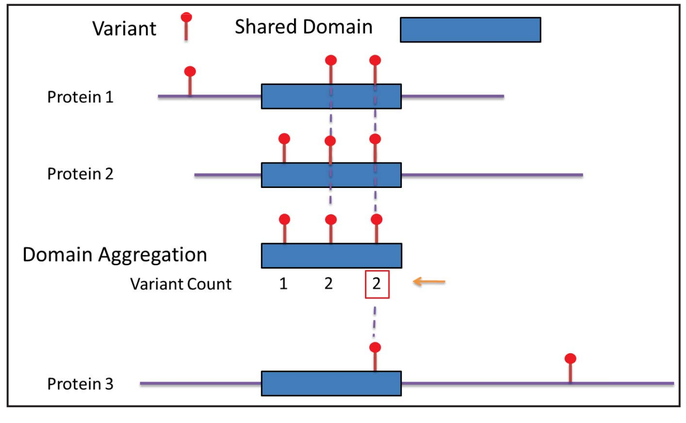
Thomas G. Coard, Thomas Peterson, and Maricel G. Kann
Department of Biological Sciences, University of Maryland, Baltimore County Baltimore, MD 21250

Abstract

Genomic information about individual patients can be used to significantly improve diagnosis, prognosis, and treatment of diseases. However, given the lack of known genomic associations with disease, this task remains a challenge. Due to the complexity and interconnectedness of genes and biochemical pathways it is difficult to predict previously unknown genotype-phenotype relationships when only one gene mutation is considered at a time. To overcome these problems, we analyze mutations in conserved protein domain regions, allowing us to compare to the mutation patterns in other human genes and also to genes in evolutionarily distant species such as yeast and mouse. We have mapped the location of disease-causing mutations to homologous protein domain regions in order to compare to variants of unknown significance using a statistical method called Domain Significance Score (DS-Score). We have tested three machine learning techniques to classify putative disease variants and have concluded that random forest is optimal for our purpose. Next, we compare our domain-based methodology against other traditional methods that use sequence conservation, structural properties, and other molecular properties to classify the variants. Our results will provide us with new insights into the molecular underpinnings of disease and will identify new biomarkers and drug targets, enabling therapeutic research.

Background

- Due to the complex nature of the genotype/phenotype relationship it is difficult to predict what genetic mutations cause genetic disorders.
- Much research exists that utilizes machine learning techniques to distinguish between variants that cause human disease and variants that have no known phenotypic affect
- These predictions can then be used to help diagnose potential candidates for genetic disorders and enable therapeutic research.
- In our research, comparing to homologous protein domain regions allows us to infer the effects of genetic variants via inference to the mutation patterns in other human genes and also to genes in evolutionarily distant species such as yeast and mouse.



Human Variant Databases

OMIM [1]: Online Mendelian Inheritance in Man. A comprehensive, authoritative compendium of human genes and genetic phenotypes

HGMD [2]: Human Gene Mutation Database. Represents an attempt to collate known (published) gene lesions responsible for human inherited disease.

UniProtKB/Swiss-Prot [3] A comprehensive, high-quality database of protein sequence, functional information, and variants.

ClinVar [4]: Aggregates information about genomic variation and its relationship to human health.

dbSNP[5]: Tracks variants that are common in the population

1,000 Genomes Project[6]: Sequenced genomes of over 1,000 individuals

Wellderly Project[7]: Sequenced genomes of over 1,000 individuals

Model Organism Variant Databases

Mouse Genome Database [8]: variants known to be phenotpyically altering in mouse

Saccharomyces Genome Database [9]: variants known to be phenotpyically altering in yeast

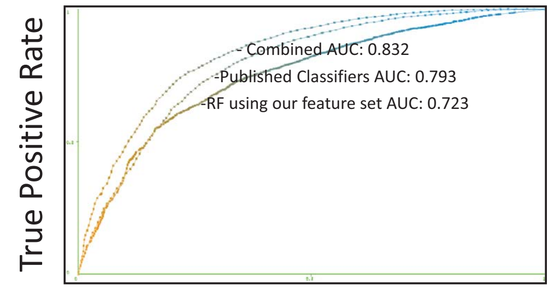
Published Classifiers Compared

- CADD[10]: Combined Annotation Dependent Depletion “CADD is a tool for scoring the deleteriousness of single nucleotide variants as well as insertion/deletions variants in the human genome.”
- GERP[11]: Genomic Evolutionary Rate Profiling "GERP identifies constrained elements in multiple alignments by quantifying substitution deficits.”
- Mutation Assessor[12]: “predicts the functional impact of amino-acid substitutions in proteins.”
- Mutation Taster[13]: “evaluates the pathogenic potential of DNA sequence alterations.”
- PolyPhen2[14]: “is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations”
- SIFT[15]: “predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids”

Methodology

We used Perl to collect data from a variety of genomic databases. OMIM, HGMD, UniProtKB/SwissProt, and ClinVar were used as the positive set and common variants from dbSNP were used as a negative set. For features, we reformatted the data in order to annotate each mutation with all other mutations located at the same protein domain position, but in a different gene from these datasets as well as the 1,000 genomes project and Wellderly. Then we scored these mutations using the DS-Score. Next, we selected only variants from the positive and negative set located on genes that mapped to any variants from other genes, removing genes that have no domain family-level information. We then performed machine learning using a Random Forest classifier and compared to the performance of other classifiers and to the combination of other classifiers and our feature set.

ROC Curve



False Positive Rate

Interpretation of Findings

- Our results demonstrate that genomic data from model organisms and from paralogous human genes can predict if a human mutation is deleterious. In future projects, we will use more data from model organisms for greater predictive power.
- Our current method of classifing mutations has some predictive capability, but our data shows that it could be best implemented along with other predictors in order for more accurate mutation classification. These predictions could help diagnose disease, alert patients of dispositions to disease, and help facilitate targeted gene therapy.

References

[1] Hamosh, Ada et al. “Online Mendelian Inheritance in Man (OMIM), a Knowledgebase of Human Genes and Genetic Disorders.” Nucleic Acids Research 33.Database issue (2005): D514–D517
[2] Stenson PD, et al.. The Human Gene Mutation Database (HGMD) and its exploitation in the fields of personalized genomics and molecular evolution. 2012/09/06 edit Curr Protoc Bioinformatics 2012 [Chapter 1].
[3] Boeckmann B, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31:365–70.
[4] About ClinVar, www.ncbi.nlm.nih.gov/clinvar/, Vol. 2013.
[5] Sherry, S.T., *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* 29, 308-11.
[6] Abecasis, G. R.A. AutonL. D. BrooksM. A. DePristoR. M. DurbinR. E. HandsakerH. M. KangG. T. MarthG. A. McVean (2012). "An integrated map of genetic variation from 1,092 human genomes." Nature 491(7422): 56-65.
[7] https://www.scripps.org/news_items/4757-scripps-wellderly-genome-resource-now-available-to-researchers
[8] Eppig JT, Blake JA, Bult CJ, Kadin JA, Richardson JE;; The Mouse Genome Database Group. 2015. The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. Nucleic Acids Res. 2015 Jan 28;43(Database issue):D726-36.
[9] Cherry, J.M., E.L. Hong, C. Amundsen, R. Balakrishnan, G. Binkley, E.T. Chan, K.R. Christie, M.C. Costanzo, S.S. Dwight, S.R. Engel, et al., Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res, 2012. 40(Database issue): p. D700-5.*
[10] Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet. 2014 Feb 2. doi: 10.1038/ng.2892. PubMed PMID: 24487276.
[11] Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, et al. (2010) Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLoS Comput Biol 6(12): e1001025. doi:10.1371/journal.pcbi.1001025
[12] Boris Reva, Yevgeniy Antipin, and Chris Sander Predicting the functional impact of protein mutations: application to cancer genomics Nucl. Acids Res. first published online July 3, 2011 doi:10.1093/nar/gkr407
[13] Jana Marie Schwarz, David N Cooper, Markus Schuelke & Dominik Seelow. “MutationTaster2: mutation prediction for the deep-sequencing age” Nature Methods 11, 361–362 (2014) doi:10.1038/nmeth.2890 Published online 28 March 2014
[14] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. Nat Methods **7**(4):248-249 (2010).
[15] Ng PC, Henikoff S. “SIFT: Predicting amino acid changes that affect protein function.” Nucleic Acids Res. 2003 Jul 1;31(13):3812-4.