CAGI 5

The Fifth International Experiment

Critical Assessment of Genome Interpretation

CHICAGO, JUL 5 - JUL 7

PLAZA BALLROOM A AND PLAZA BALLROOM B - LOBBY LEVEL (EAST TOWER)

HYATT REGENCY CHICAGO

HYATT REGENCY CHICAGO, 151 East Wacker Drive, Chicago, Illinois 60601, USA, T +1 312 565 1234, F +1 312 239 4541

Location: PLAZA BALLROOM A and PLAZA BALLROOM B - Lobby Level (East Tower)					
Date	**Start**	**End**	**Duration**	**Role**	**Speakers**
T Jul 5	14:00:00	14:15:00	0:15:00	Organizers	Steven Brenner, University of California Berkeley and John Moult, University of Maryland
Missense challenge: TPMT and PTEN					
T Jul 5	14:15:00	14:35:00	0:20:00	Data Provider	Kenneth Matreyek, University of Washington
T Jul 5	14:35:00	14:55:00	0:20:00	Assessor	Yana Bromberg, Rutgers University
T Jul 5	14:55:00	15:10:00	0:15:00	Predictor	Yizhou Yin, University of Maryland
Missense challenge: GAA					
T Jul 5	15:10:00	15:30:00	0:20:00	Data Provider + Assessor	Wyatt Clark, BioMarin
T Jul 5	15:30:00	15:45:00	0:15:00	Predictor	Emidio Capriotti, University of Bologna
				Break and poster session	
Missense challenge: Annotate all missense variants					
T Jul 5	16:15:00	16:45:00	0:30:00	Discussion	Sean Mooney, University of Washington
Missense challenge: CALM					
T Jul 5	16:45:00	17:05:00	0:20:00	Data Provider	Jochen Weile, University of Toronto
T Jul 5	17:05:00	17:25:00	0:20:00	Assessor	Nick Grishin, Howard Hughes Medical Institute, University of Texas Southwestern Medical Center
T Jul 5	17:25:00	17:40:00	0:15:00	Predictor	Panagiotis Katsonis, Baylor College of Medicine
T Jul 5	17:40:00	17:55:00	0:15:00	Discussion	Missense challenges initial discussion
T Jul 5	17:55:00	21:00:00	3:05:00		*Reception*
F Jul 6	8:45:00	8:50:00	0:05:00	Introduction	Gaia Andreoletti, University of California Berkeley
Missense challenge: Frataxin					
F Jul 6	8:50:00	9:05:00	0:15:00	Data Provider	Roberta Chiaraluce and Valerio Consalvi, Sapienza University of Rome
F Jul 6	9:05:00	9:25:00	0:20:00	Assessor	Emidio Capriotti, University of Bologna
F Jul 6	9:25:00	9:40:00	0:15:00	Predictor	Alexey Strokach, University of Toronto
Missense challenge: PCM1					
F Jul 6	9:40:00	10:00:00	0:20:00	Assessor	Marco Carraro, University of Padova
F Jul 6	10:00:00	10:15:00	0:15:00	Predictor	Rita Casadio, University of Bologna
F Jul 6	10:15:00	10:45:00	0:30:00	Discussion	Missense challenges discussion
				Break and poster session	
The CAGI Ethics forum					
F Jul 6	11:15:00	11:35:00	0:20:00	Ethics Forum	Barbara Koenig, University California San Francisco
F Jul 6	11:35:00	11:50:00	0:15:00	Ethics Forum Discussion	Ethics Forum Discussion
Splicing challenges: MapSy and VexSeq					
F Jul 6	11:50:00	12:10:00	0:20:00	Data Provider	William Fairbrother, Brown University
F Jul 6	12:10:00	12:30:00	0:20:00	Data Provider	Scott Adamson, UConn Health
F Jul 6	12:30:00	13:00:00	0:30:00	Assessor	Steve Mount, University of Maryland
				Lunch & poster session	
F Jul 6	14:30:00	14:45:00	0:15:00	Predictor	Tatsuhiko Naito, The University of Tokyo

F Jul 6	14:45:00	15:00:00	0:15:00	Predictor	Jun Cheng, Technical University of Munich
F Jul 6	15:00:00	15:15:00	0:15:00	Predictor	Ron Unger, Bar-Ilan University
F Jul 6	15:15:00	15:40:00	0:25:00	Discussion	Discussion on splicing challenges
F Jul 6	15:40:00	16:05:00	0:25:00	Discussion	Clinical implication of CAGI results
				Break and poster session	
Non-coding challenge: Regulation Saturation					
F Jul 6	16:35:00	16:50:00	0:15:00	Data Provider	Max Schubach, Berlin Institute of Health
F Jul 6	16:50:00	17:10:00	0:20:00	Assessor	Michael Beer, Johns Hopkins University
F Jul 6	17:10:00	17:25:00	0:15:00	Predictor	Alan Boyle, University of Michigan
F Jul 6	17:25:00	17:40:00	0:15:00	Predictor	Zhongxia Yan, University of California Berkeley
F Jul 6	17:40:00	18:10:00	0:30:00	Discussion	Discussion on Regulaiton Saturation challenge
S Jul 7	9:55:00	10:00:00	0:05:00	Introduction	Gaia Andreoletti, University of California Berkeley
Gene panel: Intellectual Disability					
S Jul 7	10:00:00	10:15:00	0:15:00	Data Provider	Emanuela Leonardi, University of Padova
S Jul 7	10:15:00	10:35:00	0:20:00	Assessor	Marco Carraro, University of Padova
S Jul 7	10:35:00	10:50:00	0:15:00	Predictor	Jingqi Chen, University of California Berkeley
S Jul 7	10:50:00	11:10:00	0:20:00	Discussion	ID Discussion
				Break and poster session	
Germline cancer challenge: CHEK2					
S Jul 7	11:40:00	12:00:00	0:20:00	Data Provider	Elad Ziv, University California San Francisco
S Jul 7	12:00:00	12:20:00	0:20:00	Assessor	Alin Voskanian and Maricel Kann, University of Maryland, Baltimore County
S Jul 7	12:20:00	12:50:00	0:30:00	Discussion	Discussion on *CHEK2* challenge
				Lunch & poster session	
Germline cancer challenge: ENIGMA					
S Jul 7	14:15:00	14:45:00	0:30:00	Data Provider + Assessor	Melissa Cline, University California Santa Cruz
S Jul 7	15:05:00	15:20:00	0:15:00	Predictor	Yang Shen, Texas A&M University
S Jul 7	15:20:00	15:50:00	0:30:00	Discussion	Discussion on cancer challenges
				Break and poster session	
Clinical genome challenge: Sick Kids challenge					
S Jul 7	16:20:00	16:50:00	0:30:00	Data Provider + Assessor	Stephen Meyn, University of Wisconsin
S Jul 7	16:50:00	17:20:00	0:30:00	SickKids panel	Discussion, led by Stephen Meyn
Complex trait challenge: Clotting disease (DVT or PE) exomes challenge					
S Jul 7	17:20:00	17:40:00	0:20:00	Data Provider + Assessor	Greg McInnes, University of Stanford
S Jul 7	17:40:00	17:55:00	0:15:00	Predictor	Yanran Wang, Rutgers University
S Jul 7	17:55:00	18:15:00	0:20:00	Discussion	Discussion on complex trait challenge
Closing remarks					
S Jul 7	18:15:00	18:45:00	0:30:00	Discussion	Future of CAGI

CAGI5: The Fifth International Experiment of the Critical Assessment of Genome Interpretation
Plaza ballroom A and plaza ballroom A - lobby level (east tower) - HYATT REGENCY CHICAGO, 151 East Wacker Drive, Chicago, Illinois 60601, USA, T +1 312 565 1234, F +1 312 239 4541

Abstracts:

Scott Adamson, UConn Health
Alan Boyle, University of Michigan
Emidio Capriotti, University of Alabama Birmingham
James Han, Yale University
Kunal Kundu, University Of Maryland
Kymberleigh Pagel, Indiana University Bloomington
Vikas Pejaver, Indiana University Bloomington
Lipika Ray, IBBR, University of Maryland
Castrense Savojardo, University of Bologna
Yang Shen, Texas A&M University
Prashanth Suravajhala Aarhus University
Robert Wang, University of California Berkeley
Jochen Weile, University of Toronto
Yizhou Yin, University of Maryland College Park

Vex-seq: High-Throughput Identification of the Impact of Genetic Variation on pre-mRNA Splicing Efficiency

Scott I. Adamson, Lijun Zhan, Brenton R. Graveley*

Presenting Author email: adamson@uchc.edu

Corresponding Author email: graveley@uchc.edu

Department of Genetics and Genome Sciences, Institute for Systems Genomics, UConn Health, Farmington, CT

Understanding the functional impact of genomic variants is one of the major goals of modern genetics and the underpinning of personalized medicine. To some extent, it is relatively easy to understand how non-synonymous protein coding variants exert their effects. Many synonymous and non-coding variants are known to act by altering the efficiency of pre-mRNA splicing. However, in most cases, it is exceedingly difficult to predict how these variants impact pre-mRNA splicing. Thus, a method that could simultaneously measure the splicing efficiency of thousands of exons and their variants would have a tremendous impact on the field and our understanding of genome function. We have developed a massively parallel approach using a novel barcoding design to test the impact of variants on pre-mRNA splicing, called variant exon sequencing (vex-seq). Using this approach we have tested the impact of 2,059 human genetic variants spanning 110 alternative exons. We interrogate the impact of exonic splicing regulatory features and splice site strength on the splicing efficiency of vex-seq test exons. Vex-seq yields data that reinforces known mechanisms of pre-mRNA splicing and can rapidly identify genomic variants that impact pre-mRNA splicing.

SEMpl: predicting the effects of single nucleotide polymorphisms on transcription factor binding affinity

Abstract
Background: One of the most surprising results to emerge from genome-wide association studies (GWAS) is 95% of all disease associated single nucleotide polymorphisms (SNPs) identified by this method reside in non-coding regions of the genome. Despite this finding, non-coding SNPs remain hugely understudied, due in part to the uncertain functional consequences of such mutations. However, a large proportion of these SNPs reside within regulatory regions of the genome, such as transcription factor binding sites (TFBSs). TFBSs only cover 8.1% of the genome, yet they contain 31% of GWAS SNPs. SNPs in these binding sites may alter the binding affinity of transcription factors, leading to changes in downstream gene expression, and ultimately human disease. Here, we propose a novel screening tool, SEMpl, which estimates transcription factor (TF) binding affinity to better predict disease causing SNPs in TFBSs.
Methods: SEMpl generates its predictions through observation of existing variants in TFBSs genome-wide using publically available data from the ENCODE database to generate SNP effect matrixes (SEMs). SEM scores represent the predicted change in binding affinity from average binding of the target TF.
Results: SEMpl has demonstrated a better correlation with experimental estimates of TF binding affinity than the current standard, position weight matrices (PWMs).
Significance: We hypothesize that SEMpl scores will allow researchers to better predict disease causing SNPs in TFBSs genome wide.

Predicting the impact of genetic variants with BioFoID tools

*Emidio Capriotti**
Department of Pharmacy and Biotechnology (FaBiT), University of Bologna.
Via F. Selmi 3. 40126 Bologna (Italy)
email: emidio.capriotti@unibo.it

During the last few years we developed several tools for predicting the impact of genetic variants at protein and nucleotide levels. The implemented methods are characterized by the types and number used for discriminating between pathogenic and benign variants. The simplest algorithm is PhD-SNP (Capriotti, et al., 2006), which is a support vector machine based method that takes in input only sequence-based extracted from the protein sequence profile. The most complex tool is SNPs&GO (Capriotti, et al., 2013b) which includes in the input features functional information encoded by Gene Ontology terms and, when available, protein structure features. More recent algorithms such as Meta-SNP (Capriotti, et al., 2013a) implements a meta prediction method combining 4 well-establish methods while PhD-SNP^{g} (Capriotti and Fariselli, 2017) uses the information retrieved on the UCSC genome browser to predict the impact of variants in non coding regions.
During the last edition of the CAGI we used modified version of these methods to predict the impact of the variants released for four challenges, namely the Cell-Cycle-Checkpoint Kinase 2 (CHEK2), the Acid Alpha-Glucosidase **(**GAA), the Calmodulin 1 (CALM1) and the Pericentriolar Material 1 (PCM1). Among these challenges we verified that PhD-SNP reached a good level of performances in the prediction of the fraction of tumor cases associated to a set of variants in the CHEK2 protein.
In particular on a set composed by 34 coding CHEK2 variants PhD-SNP achieved a balanced accuracy of 0.71 a Matthews' Correlation Coefficient of 0.41 and an Area Under the Curve (AUC) of 0.72. All the tools used for the CAGI challenges are available at http://snps.biofold.org/

References

Capriotti, E., Altman, R.B. and Bromberg, Y. Collective judgment predicts disease-associated single nucleotide variants. BMC Genomics 2013a;14 (Suppl. 3):S2.
Capriotti, E., Calabrese, R. and Casadio, R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. Bioinformatics 2006;22(22):2729-2734.
Capriotti, E., et al. WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation. BMC Genomics 2013b;14 Suppl 3:S6.
Capriotti, E. and Fariselli, P. PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. Nucleic Acids Res 2017;45(W1):W247-W252.

Prediction of the Effects of Missense Mutations in Pericentriolar Material 1

James Han[1], Kivilcim Ozturk[2], Hannah Carter*[2]

[1]Great Oak High School, Temecula, CA 92592 [2]Department of Medicine, University of California, San Diego, La Jolla, CA 92093

hcarte10@gmail.com

The aim of the PCM1 challenge was to classify 38 provided missense mutations into three categories: benign, pathogenic, or hypomorphic. In addition to classifying these mutations, the challenge involved finding the probabilities that each mutation caused the protein to resemble loss of function (MO p-value) or normal function (MO+WT p-value), standard deviations for each p-value, and a confidence score for each mutation assessed. Our approach centered around the use of the Variant Effect Scoring Tool (VEST) to predict the effects of the given missense mutations on the function of PCM1. VEST is a supervised machine learning-based classifier that provides the likelihood of a missense variant's involvement in a disease. The 38 missense mutations were initially analyzed by VEST, which assigned each mutation a score ranging from 0 to 1 and a p-value. We also curated 526 known disease-causing mutations from the literature and scored them with VEST to generate a distribution of pathogenic VEST scores. The p-value provided by VEST for each mutation was used as the MO+WT p-value, and the proportion of VEST scores in the pathogenic distribution as or more extreme than the given mutation's score was used as the MO p-value. The standard deviation of each calculated p-value was determined by finding the standard deviation of the set of VEST scores of the possible amino acid substitutions at the residue position of the given variant. For example, the standard deviation of the p-value of the variant G6D was calculated by finding the standard deviation of the p-values for the variants G6R, G6V, G6A, G6C, and G6S. Each of the 38 given variants were then classified by their MO and MO+WT p-values and their standard deviations. Finally, confidence values were assigned to each variant as functions of the VEST scores based on each variant's classification. Using our approach, 12 of the mutations (or 32%) were implicated as benign, 22 of the mutations (or 58%) were implicated as hypomorphic, and 4 of the mutations (or 10%) were implicated as pathogenic.

Predicting Missense Mutational Effects on Protein Functions and Cancer Pathogenicity

Mostafa Karimi, Yuanfei Sun, Yue Cao, Haoran Chen, Oluwaseyi Moronfoye, Yang Shen*
Department of Electrical and Computer Engineering
Center for Bioinformatics and Genomic Systems Engineering
Texas A&M University, College Station, Texas, USA

* Contact: yshen@tamu.edu

To predict missense mutational effect on proteins with known structure data (such as the frataxin challenge), we have developed a novel multistate protein design algorithm, named iCFN (Karimi and Shen, 2018), to derive optimal structures and energetics of individual proteins or protein-protein complexes upon mutation. The algorithm is exact in the sense that it guarantees the optimal solutions for given biophysical models. We have applied iCFN to explain missense mutations to pan-cancer DNA damage and repair (DDR) genes. We have found that, out of 547 mutations occurred to 15 DDR core genes with structural data, 354 (64.7%) such mutations were predicted to strongly destabilize ($\Delta\Delta G_{folc} \geq 3k_BT$) complexes formed by protein-protein interactions (PPIs). Our case study of the BRCA1-BARD1 RING-RING domain interaction predicted that the BRCA1 R7H mutation significantly affects intra- and inter-molecular interactions with BRCA1 E10 and BARD1 E117. Following the reciprocity principle, we predicted that a mutation at either site, potentially into a positively charged residue, would have a similar functional effect as BRCA1 R7H. Notably, BRCA1 E10K was previously found in Indian breast and breast-ovarian cancer families, which supports the link between mutational effects to protein functions and those to clinical outcomes.

To further anticipate mutations of certain effects from the protein sequence-structure-function relationship, we have developed an inverse protein design approach and applied iCFN to model *ESR1* activating mutations found in metastatic breast cancer patients. Our study has supported the underlying mechanistic hypotheses that these mutations stabilize estrogen receptor's agonist conformations relative to its antagonist conformations. Furthermore, we have predicted new mutations that are potentially activating, some of which are validated *in vitro*.

Going beyond individual protein-level mutational effects, we have developed novel machine learning algorithm to predict cancer pathogenicity (such as the ENIGMA challenge) from predicted molecular-level impacts. Considering that the required pathogenicity score is continuous whereas known pathogenicity classes are discrete, we have formulated the problem as ordinal regression and developed continuous loss functions tailored to discrete class labels. The linear regression model (which can be easily extended to the nonlinear case) was trained on ClinVar mutation data for breast cancer-related oncogenes and used to predict pathogenicity score for BRCA1 and BRCA2 missense mutations. For the missense mutations in Class 1 ("Not pathogenic), we correctly predicted 52 (true positives) of 62 (sensitivity: 84%). We also made 196 false positives but 186 of them are actually Class 2 ("Likely not pathogenic"). For those mutations in Class 5 ("Pathogenic"), we correctly predicted 5 (true positives) of 10 (sensitivity: 50%) but made 23 false positives including 2 from Class 4 ("Likely pathogenic").

Funding: This project is in part supported by NIGMS/NIH under R35GM124952.

Reference: Mostafa Karimi and Yang Shen, "iCFN: an efficient exact algorithm to multistate protein design", Bioinformatics, forthcoming.

Prediction of patient's clinical description and pathogenic variants from Intellectual Disability gene panel sequencing data.

Kunal Kundu [1, 2], Lipika R. Pal [1], Yizhou Yin [1, 2], John Moult [1, 3]*

[1] Institute for Bioscience and Biotechnology Research, University of Maryland, 9600 Gudelsky Drive, Rockville, MD 20850, [2] Computational Biology, Bioinformatics and Genomics, Biological Sciences Graduate Program, University of Maryland, College Park, MD 20742, USA, [3] Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742.

*Corresponding author
jmoult@umd.edu
Phone: (240) 314-6241
FAX: (240) 314-6255

Targeted gene sequencing panels are widely used for identifying putative causative variants in a set of genes or gene regions known to be relevant to a specific disease. However, it is often challenging to identify the causative variants, and a number of carefully controlled procedures are required for assessing the quality of data, accurate variant annotation, handling unphased genotypes, and devising an appropriate probability model that can prioritize primary and secondary disease findings. With these considerations in mind, we developed a new method (implemented in Python) and applied it to the Intellectual Disability panel challenge dataset. The method has five steps – 1. Gene – Phenotype Mapping 2. Variant Annotation, 3. QC analysis, 4. Variant Prioritization and 5. Probability scoring for the phenotypes. The inheritance pattern of the genes with respect to disease phenotypes are obtained from OMIM and HPO. The variants were annotated with the region of occurrence in the human genome, allele frequency, predicted impact on protein function, and previously reported disease association using the Varant tool (http://compbio.berkeley.edu/proj/varant/). The QC analysis assessed ti/tv ratio, heterozygous/homozygous ratio, common/rare/novel variant counts with respect to the control (1000 Genomes dataset). The variant prioritization algorithm is based on the variant annotations, QC parameters (Genotype Quality, Strand Bias, Read Depth and share count with other samples) and the known inheritance pattern for each gene. It uses consecutive search criteria (SC), starting with criteria deemed most reliable for finding causative variant, and progressing to those considered less reliable. We predefined eight SC for this purpose, primarily a combination of variant quality, shared count with other samples and predicted impact. A probability score for a sample to have a particular disease is computed, based on the type of filtered variants, quality of the variant, the prediction from *in-silico* tools, and inheritance pattern. Using this method we were able to assign 150 patients to one or more of the seven phenotypes in the challenge, although a substantial fraction of assignments is of low confidence. A total of six submissions were made based on which database (OMIM or OMIM+HPO) was used for gene-phenotype mapping, and the probability scoring model.

Clotting Exomes - CAGI

Kymberleigh A. Pagel, Moses Stamboulian, Yuxiang Jiang, and Predrag Radivojac*

School of Informatics, Computing, and Engineering, Indiana University, Bloomington, Indiana, USA
Corresponding Author: predrag@indiana.edu

The submitted predictions for individuals in the Clotting disease exomes challenge are derived utilizing a combination of variant pathogenicity scores in relevant genes and the relationship of each clinical covariate to the phenotype. Annotation of the protein coding variation in the raw VCF files was performed using ANNOVAR. We assign pathogenicity prediction scores to missense and stop gain variants with Mutpred2 and Mutpred-LOF, respectively. Per individual exome, we include only the variant with the highest pathogenicity prediction score within each gene in further analyses. Confirmed risk genes are used as “seed” genes on the human protein-protein interaction network for running a network propagation algorithm. The propagation algorithm are performed in a 5-fold cross validation manner to get an initial score between [0, 1] for all the genes. We then use the AlphaMax algorithm to estimate the positive proportion of the risk genes and calibrate those initial scores to be proper probability scores measuring the likelihood of a gene being associated with the disease.

We generate a beta distribution based upon the Mutpred scores of variants within the top one hundred highest scored genes for each phenotype. For each individual exome, we utilize the distribution to determine the p-value for the highest Mutpred scored variant within each gene. Next, we sought to incorporate the clinical covariates within a similar framework. For each clinical covariate, we search the published literature to find the mean and standard deviation values of the trait described in case/control studies. We utilize these variables from the literature to derive value distributions (binomial for gender and aspirin, Gaussian otherwise) that were used to derive p-values for each individual based upon their particular value for that covariate. The unnormalized score is the product of all gene and covariate scores, where each individual has scores for both VTE and atrial fibrillation. We then combine the VTE and atrial fibrillation score rankings using geometric mean, then transform with min-max normalization so that the values range between zero and one. The procedure is repeated one hundred times with differing amounts of seed genes (from 200 to 300), where the score for an individual is the mean score of the one hundred iterations.

SickKids5: Prediction of patient's clinical descriptions and pathogenic variants from their whole genome sequences

Lipika R. Pal [1], Kunal Kundu [1, 2], Yizhou Yin [1, 2], John Moult [1, 3]*

[1] Institute for Bioscience and Biotechnology Research, University of Maryland, 9600 Gudelsky Drive, Rockville, MD 20850, [2] Computational Biology, Bioinformatics and Genomics, Biological Sciences Graduate Program, University of Maryland, College Park, MD 20742, USA, [3] Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742. *Corresponding author: jmoult@umd.edu

The Sickkids5 challenge is to match phenotypic profiles with genotypic profiles for 24 undiagnosed cases. These children mainly have eye disorders, epilepsy, and connective tissue disorders (including Ehler-Danlos Syndrome).

A total of 213 clinical phenotype descriptions were extracted from the CAGI phenotype data provided for 24 children. These phenotypes were used to extract a total of 6239 potentially relevant genes from the Human Phenotype Ontology-based database (HPO) and the dbNSFP database. We also used the list of 319 genes from RetNet database for searching for eye disorder related variants. The gene list for secondary variants was taken from the table in the 2017 ACMG guidelines. The whole genome VCF files were annotated using the Varant and Annovar tools. Chromosome M was annotated and searched for pathogenic variants using the MSeqDR mv tool. Unlike the previous Sickkids challenge, where the genotype data were from Complete Genomics, this time the data are from Illumina Hiseq, including only SNVs and Indels.

A hierarchical scheme was used for identification of diagnostic variants, based on the strength of the evidence for disease relevance. All accepted high quality variants in the selected gene list with population frequency <1% (maximum frequency in any of the GnomAD database, 1000 genome data and in the ExAC database) were first categorized into ordered tiers: Category 1 – clinically relevant pathogenic variants from HGMD and ClinVar; Category 2 – nonsense, direct splice sites, frameshift and nonsynonymous deleterious (at least 60% of 7 methods - SNPs3D profile, SNPs3D stability, Polyphen2, Sift, Vest, REVEL and CADD, have predicted deleterious) mutations; Category 3 – other nonsynonymous mutations where deleteriousness agreement < 0.6 by these 7 methods; Category 4 – predicted benign nonsynonymous SNPs; Category 5 – variants close to a splice acceptor or donor site; Category 6 – variants annotated as UTR and intronic, where pathogenicity of these noncoding variants are based on any one pathogenic score of CADD, Eigen and GERP++; Categorized variants were further filtered for an appropriate inheritance model using the OMIM inheritance pattern.

For each phenotypic profile, each phenotips term was assigned a subjective value from 0 to 1, according to its importance. For example, if a connective tissue disorder is the most serious and definitive term in the profile, it was scored the highest. If seizure is also part of that profile with borderline occurrence, then that was assigned a lower value than would be the case if the term occurred in a profile where seizure is the most serious phenotype. We then calculated a weighted matching score between the phenotypic profile and each variant-carrying gene in a genome. For each genome, we examined the evidence supporting the top five scoring variants, considering gender match, inheritance pattern and correspondence with the OMIM disease description.

The clinical interpretability of MutPred2 predictions: the ENIGMA challenge as a case study

Vikas Pejaver[1,2], Kymberleigh A. Pagel[3], Predrag Radivojac[3,*], Sean D. Mooney[1,*]

1. Department of Biomedical Informatics and Medical Education, University of Washington, Seattle
2. The eScience Institute, University of Washington, Seattle
3. Department of Computer Science, School of Informatics, Computing, and Engineering, Indiana University, Bloomington

*Corresponding author email: predrag@indiana.edu, sdmooney@uw.edu

MutPred2 is a neural network-based model that predicts disease-associated variants and infers molecular mechanisms of disease. Using data sets from previous CAGI challenges, we recently evaluated the direct utility of MutPred2 scores in predicting broader notions of impact, operationalized in different ways. Remarkably, we found that, despite being trained as a general-purpose pathogenicity predictor using a large data set with simple binary labels ("Disease" or "Not Disease"), MutPred2 predictions capture molecular, cellular and organismal phenotypic effects of variants in protein-specific data sets. In this study, we further evaluate the direct interpretability of MutPred2 scores in a different context, the clinical setting, using variants classified by the ENIGMA Consortium as an exemplary data set. Similar to our previous observations, we found that MutPred2 generalized beyond the binary classification problem that it was trained for and predicted scores that aligned with the ENIGMA's 5-tier classification system (Figure 1). The overall area under the ROC curve (AUC) was 0.857 when "Uncertain" variants were excluded and all classes other than "Pathogenic" were treated as "Benign". When each gene was considered individually, the AUCs were 0.861 and 0.952 for BRCA1 and BRCA2, respectively. When "Likely Pathogenic" variants were included in the "Pathogenic" class, these values were 0.870, 0.924 and 0.787, respectively. Taken together with our observations from previous CAGI challenges, these results highlight the need to develop methods that are not only accurate but are interpretable in multiple contexts, and suggest that methods optimized to output scores that emphasize ranking of variants are more preferable than those that emphasize classification towards the extremes of pathogenicity or benignity.

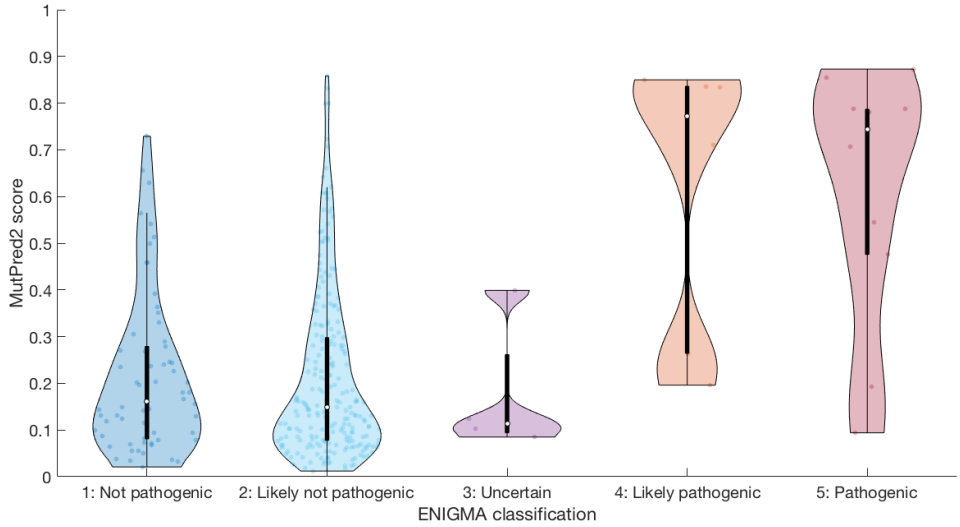

Figure 1. Violin plot showing MutPred2 score distributions over the five classes assigned by the ENIGMA Consortium. Lower MutPred2 scores are expected to be enriched among benign variants and higher scores are expected to be enriched among pathogenic variants.

INPS and INPS-3D: sequence- and structure-based prediction of protein stability change upon single-point variations

Castrense Savojardo[1], Pier Luigi Martelli[1*], Piero Fariselli[2], Giulia Babbi[1], Rita Casadio[1]
[1] Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Italy
[2] Dept. of Comparative Biomedicine and Food Science, University of Padova, Italy

* Correspondence: pierluigi.martelli@unibo.it

Abstract

Reliable tools for assessing the impact of non-synonymous Single Nucleotide Polymorphisms (nsSNPs) on protein stability are of prominent importance for elucidating the effect of variants in the genome. As soon as new protein sequences are produced by sequencing studies and new variants are discovered, methods that are able to characterize protein variants starting from protein sequence alone become extremely valuable, given the gap existing between the amount of sequence and structural information.

Here, we present INPS (Impact of Non-synonymous mutations on Protein Stability) [1], a method for predicting the impact of nsSNPs on protein stability, starting from sequence. INPS is based on a Support Vector Regression (SVR) approach and it is trained to predict the thermodynamic free energy change ($\Delta\Delta G$) upon single-point variations in protein sequences. INPS performance are comparable to those achieved by state-of-the-art methods based on protein structure, as assessed using a rigorous cross-validation procedure on a non-redundant dataset of protein variants (2,648 variants, correlation coefficient of 0.53). INPS performs very well also on a benchmarking dataset collecting 42 variations occurring in the tumor suppressor protein p53 (achieving a correlation coefficient of 0.71). Our results suggest that INPS is a tool suited for screening variants when the protein structure is not available.

INPS-3D [2] extends INPS by also including, when available, features extracted from the protein 3D structure. Structural information further improves the performance of the method reporting correlations coefficients of 0.58 and 0.76 on training and p53 datasets, respectively.

Finally, we report about the adoption of INPS and INPS-3D for two Critical Assessment of Genome Interpretation (CAGI) challenges: TPMT-PTEN and Frataxin.

References

[1] Fariselli P *et al.* (2015) INPS: predicting the impact of non-synonymous variations on protein stability from sequence. *Bioinformatics*, **31**(17):2816-28.

[2] Savojardo C *et al.* (2016) INPS-MD: a web server to predict stability of variants from sequence and structure. *Bioinformatics*, **32**(16):2542-2544.

Predicting changes in alternative splicing induced by genetic variants

Robert Wang[1], Yaqiong Wang[1]*, Zhiqiang Hu[1], Steven E. Brenner[1]

[1]University of California, Berkeley

*yqw@berkeley.edu

Accurate interpretation of genomic variants that alter RNA splicing is critical to precision medicine given that one-third of disease-causing variants might impact splicing (Lim et al., 2011). However, the complex regulation of RNA splicing renders identification of splicing variants to be difficult. Previous computational studies on predicting the impact of variants on splicing mainly focus on splicing regulatory motifs. However, these methods may be limited in predicting splicing changes in specific cells.

Recent studies also showed functions of chromatin modifications in regulating splicing. This suggests that integrating epigenetic annotations with RNA sequence-specific features may improve the prediction of variants that impact splicing.

We developed a tool named PEPSI (Predicting the Effects of variants on Percent Spliced In) for the Vex-seq challenge in CAGI 5. PEPSI can take in exonic or intronic variants and predict their impacts on changes in the percent spliced in (PSI) of the test exon with respect to the wild-type PSI. The training model uses random forests and integrates multiple layers of features including population allele frequencies, sequence conservation, RNA folding, and splicing associated sequence elements (e.g., splicing sites, splicing factor binding sites, branch-points) that are identified by *in silico* prediction tools and experiments. Additionally, our model also integrates cell-specific annotations related to the expression of trans-acting splicing factors and epigenetic features, providing a contextualized assessment of how variants affect the PSI of a given exon in a specified cell type. PEPSI was trained using variants on chromosomes 2 to 8 from the Vex-seq training set, and the trained model was used to predict the variant PSI of variants on chromosome 1. We observed good agreement (R^2 = 0.78) between model-predicted variant PSI and the experimental variant PSI for exons on chromosome 1. However, removal of epigenetic features from the training model did not significantly affect prediction accuracies for variant PSI (R^2 = 0.79). This suggests that epigenetics may not play a significant role in regulating splicing in the mini-gene system. The mini-gene system might not fully mimic the *in vivo* system in terms of the epigenetic regulation of splicing. PEPSI will be further trained on other experiment datasets to achieve a better evaluation on importance of different features. PEPSI can be used to study the splicing consequences of both exonic and intronic variants that may be involved in various Mendelian disorders.

Lim KH, Ferraris L, Filloux ME, Raphael BJ, Fairbrother WG. 2011. Using positional distribution to identify splicing elements and predict pre-mrna processing defects in human genes. *Proc Natl Acad Sci U S A 108*:11093-11098. PMCID:PMC3131313. doi:10.1073/pnas.1101135108

AN ATLAS OF VARIANT IMPACT MAPS FOR HUMAN DISEASE GENES

Jochen Weile[1,2,3,4], Song Sun[1,2,3,4,5], Atina G Cote[1,2,3], Jennifer Knapp[1,2,3], Marta Verby[1,2,3], Yingzhou Wu[1,2,3,4], Alan F Rubin[6], Carles Pons[7], Joseph C Mellor[8], Cassandra Wong[1,2], Natascha van Lieshout[1], Fan Yang[1,2,3,4], Murat Tasan[1,2,3,4], Guihong Tan[2,3], Shan Yang[9], Douglas M Fowler[10], Robert Nussbaum[9], Jesse D Bloom[11], Marc Vidal[12,13], David E Hill[12], Patrick Aloy[7,14], Frederick P Roth[1,2,3,4,15,*]

[1]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, ON, Canada
[2]The Donnelly Centre, University of Toronto, Toronto, ON, Canada
[3]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada
[4]Department of Computer Science, University of Toronto, Toronto, ON, Canada
[5]Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden
[6]Walter and Elizabeth Hall Institute for Medical Research, Parkville, Australia
[7]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute for Science and Technology, Barcelona, Catalonia, Spain
[8]SeqWell Inc, Boston, MA, USA
[9]Invitae Corp., San Francisco, CA, USA
[10]Department of Genome Sciences, University of Washington, Seattle, WA, USA
[11]Fred Hutchinson Research Center, Seattle, WA, USA
[12]Center for Cancer Systems Biology (CCSB), Dana-Farber Cancer Institute, Boston, MA, USA
[13]Department of Genetics, Harvard Medical School, Boston, MA, USA
[14]Institució Catalana de Recerca I Estudis Avançats (ICREA), Barcelona, Catalonia, Spain
[15]Canadian Institute for Advanced Research, Toronto, ON, Canada
* Corresponding Author: fritz.roth@utoronto.ca

Although we now routinely sequence human genomes, we cannot yet confidently identify functional variants. We recently developed a Deep Mutational Scanning framework that combines random codon-mutagenesis and multiplexed functional variation assays with computational imputation and refinement to yield exhaustive functional maps for human missense variants. We have applied this framework to seven disease-relevant human genes including Calmodulin, which formed the basis for one of last year’s CAGI challenges. The functional impact scores in these maps correspond to known protein features, and serve to confidently identify pathogenic variation and predict patient phenotypes. As we move to generate more variant impact maps for human disease genes towards a comprehensive atlas, we are developing a public database, MaveDB, and an accompanying ecosystem of apps to make our data available to the public.

Predict the effect of missense mutations on PTEN and TPMT protein stability

Yizhou Yin [1, 2], Kunal Kundu [1, 2], Lipika R. Pal [1], John Moult [1, 3]*

[1] Institute for Bioscience and Biotechnology Research, University of Maryland, 9600 Gudelsky Drive, Rockville, MD 20850, [2] Computational Biology, Bioinformatics and Genomics, Biological Sciences Graduate Program, University of Maryland, College Park, MD 20742, USA, [3] Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742.*Corresponding author: jmoult@umd.edu

In the PTEN and TPMT challenge in CAGI 5, participants were asked to predict the effect of mutations on the experimentally measured protein abundance relative to wild-type. Under the assumption that the protein abundance is closely related to the relative thermodynamic stability of proteins or to effects on specific biochemical mechanisms affecting half-life such as the ubiquitination [1], we tested both sequence-based methods and structure-based methods in this challenge.

Previously, in the NAGLU and SUMO ligase challenges in CAGI 4, we developed a sequence-based SVR ensemble method [2] and showed its effectiveness in estimating total protein activity for missense mutations. Here, we applied a similar strategy using the model trained on the NAGLU data. Next, we tested the use of two different structure-based methods, Rosetta and SNPs3D Stability. Rosetta estimates ddG for a mutation while SNPs3D Stability returns a binary estimation of whether ddG is likely greater or less than a threshold related to pathogenicity in monogenic disease. In additional submissions, we also applied the aforementioned methods to different subsets of mutations, e.g. surface vs. core.

As a consequence of the short time line for this challenge, the approach we used is crude. Nevertheless, we submitted it for the educational value. The initial analysis on the released answer keys showed a performance (RMSE 0.38 and 0.39, Pearson's r 0.49 and 0.47, Spearman's rho 0.49 and 0.49 for the best prediction on PTEN and TPMT respectively) comparable to that on the previous NAGLU and SUMO ligase challenges. It is interesting to see that the ensemble method is robust across different systems. It is also interesting to notice that most of our different strategies have very similar performance. Is this similarity masking some unique performance patterns on different subsets of the data by different approaches? Why didn't the state-of-the-art structure-based methods like Rosetta outperform sequence-based methods in these protein stability-related problems? Further analysis may reveal more valuable information to help better understand the underlying biology and to improve prediction models.

1. Gupta, A. and Leslie, N. R. Controlling PTEN (Phosphatase and Tensin Homolog) stability: A dominant role for Lysine 66. J. Biol. Chem. 2016; 291:18465-73.

2. Yin, Y., Kundu, K., Pal, L. R. and Moult, J. Ensemble variant interpretation methods to predict enzyme activity and assign pathogenicity in the CAGI4 NAGLU (Human N-Acetyle Glucosaminidase) and UBE2I (Human SUMO-ligase) challenges. Human Mutation. 2017; 38:1109-22.

CAGI5: The Fifth International Experiment of the Critical Assessment of Genome Interpretation
Plaza ballroom A and plaza ballroom A - lobby level (east tower) - HYATT REGENCY CHICAGO, 151 East Wacker Drive, Chicago, Illinois 60601, USA, T +1 312 565 1234, F +1 312 239 4541

Attendee information

Attendee Last Name	Attendee First Name	Company/Org	Country	Email
Adamson	Scott	UConn Health	USA	adamson@uchc.edu
Adhikari	Aashish	UC Berkeley	USA	aashishna@gmail.com
Andreoletti	Gaia	UC Berkeley	USA	gandreoletti@berkeley.edu
Beer	Michael	Johns Hopkins University	USA	mbeer@jhu.edu
Boyle	Alan	University of Michigan	USA	apboyle@umich.edu
Brenner	Steven	University of California, Berkeley	USA	brenner@compbio.berkeley.edu
Bromberg	Yana	Rutgers University	USA	yana@bromberglab.org
Capriotti	Emidio	University of Bologna	Italy	emidio.capriotti@unibo.it
Carraro	Marco	University of Padova	Italy	carraromarco3569@gmail.com
Carter	Hannah	UCSD	USA	hkcarter@ucsd.edu
Casadio	Rita	University of Bologna	Italy	rita.casadio@unibo.it
Chandonia	John-Marc	Berkeley National Lab	USA	jmchandonia@lbl.gov
Chen	Jingqi	UC Berkeley	USA	jingqi.chen@berkeley.edu
Cline	Melissa	UC Santa Cruz Genomics Institute	USA	cline@soe.ucsc.edu
Daneshjou	Roxana	Stanford University	USA	roxanad@stanford.edu
Dong	Shengcheng	University of Michigan	USA	apboyle@umich.edu
Furutsuki	Mabel	UC Berkeley	USA	mfurutsuki@berkeley.edu
Grishin	Nick	UTSW/HHMI	USA	grishin@chop.swmed.edu
Gupta	Sonal	Stanford University	USA	gupta.sonal1990@gmail.com
Han	James	UCSD	USA	hkcarter@ucsd.edu
Hubbard	Tim	King's College London / Genomics England	United Kingdom	tim.hubbard@kcl.ac.uk
Jun	Cheng	Technical University of Munich	Germany	chengju@in.tum.de
Karchin	Rachel	Johns Hopkins	USA	rachel.karchin@gmail.com
Katsonis	Panagiotis	Baylor College of Medicine	Greece	katsonis@bcm.edu
Koenig	Barbara	UCSF	USA	barbara.koenig@ucsf.edu
Kundu	Kunal	University of Maryland	USA	kkundu@umd.edu
Martelli	Pier Luigi	University of Bologna	Italy	pierluigi.martelli@unibo.it
McInnes	Greg	Stanford University	USA	gmcinnes@stanford.edu
Mooney	Sean	University of Washington	USA	sdmooney@uw.edu
Moult	John	University of Maryland	Israel	jmoult@yahoo.com
Mount	Steve	University of Maryland	USA	smount@umd.edu
Nishizaki	Sierra	University of Michigan	USA	apboyle@umich.edu
Pagel	Kym	Indiana University	USA	kpagel@iu.edu
Pejaver	Vikas	University of Washington	USA	vpejaver@uw.edu
Radivojac	Predrag	Indiana University	USA	predrag@indiana.edu
Ramola	Rashika	Indiana University	USA	rramola@iu.edu
Ray	Lipika	University of Maryland	USA	rlipika@gmail.com
Savojardo	Castrense	University of Bologna	Italy	castrense.savojardo2@unibo.it
Schubach	Max	Berlin Institute of Health (BIH)	Germany	max.schubach@bihealth.de
Shen	Yang	Texas A&M University	USA	yshen@tamu.edu
Shi	Fang-Yuan	Peking University	China	shify@mail.cbi.pku.edu.cn
Tatsuhiko	Naito	The University of Tokyo	Japan	tnaito0315@gmail.com
Unger	Ron	Bar-Ilan Univ	Israel	ubronron@yahoo.com

CAGI5: The Fifth International Experiment of the Critical Assessment of Genome Interpretation
Plaza ballroom A and plaza ballroom A - lobby level (east tower) - HYATT REGENCY CHICAGO, 151 East Wacker Drive, Chicago, Illinois 60601, USA, T +1 312 565 1234, F +1 312 239 4541

Voskanian	Alin	UMBC	USA	alinvoskanian@gmail.com
Wang	Robert	University of California Berkeley	USA	rwang916@berkeley.edu
Wang	Yaqiong	UC Berkeley	USA	yqw@berkeley.edu
Wang	Yanran	Rutgers University	USA	wang.yr89@gmail.com
Wang	Yu	Peking University	China	wangy@mail.cbi.pku.edu.cn
Weile	Jochen	University of Toronto	Canada	jochenweile@gmail.com
Yan	Zhongxia	UC Berkeley	USA	zeexyan@gmail.com
Yin	Yizhou	Institute for Bioscience and Biotechnology Research, University of Maryland	USA	yzysdc@gmail.com
Zhang	Jing	UTSW	USA	jingzhang.first@gmail.com