CAGI 6

The Sixth International Experiment

Critical Assessment of Genome Interpretation

14 – 16 May 2022

David Brower Center

Berkeley

Conference Program

(all times in PT)

Saturday 14th May

Registration opens

8:00	Register by 8:40 am to receive COVID test in time for start of meeting	

Welcome and overview

9:00	Welcome	CAGI Organizers

Missense challenge: HMBS

9:45	Session chair	Yana Bromberg
9:50	Data provider	Roth group
10:05	Assessor	Grishin group
10:25	Predictor	Yun Song
10:35	Predictor	Alexey Strokach
10:45	Discussion	
	Coffee break	

Missense cancer challenges: MAPK1, MAPK3

11:25	Data provider	Valerio Consalvi
11:40	Assessor	Emidio Capriotti
11:55	Predictor	Fabrizio Pucci
12:05	Predictor	Emil Alexov
12:15	Discussion	
	Lunch	

Missense challenge: Calmodulin

13:25	Session chair	Yana Bromberg
13:30	Data provider	Giuditta Dal Cortivo
13:45	Assessor	Emidio Capriotti
14:00	Predictor	Carlos Rodrigues
14:10	Predictor	Yang Shen
14:20	Discussion	

Abstract talks - part 1

14:30	Gabriel Cia, Giulia Babbi, Nurdan Kuru, Céline Marquet	
	Tea break	

Abstract talks - part 2

15:40	Muttaqi Alladin, Shen group, Milind Jagota	

Discussion of missense prediction methods and challenges

16:05	Discussion	All attendees

CAGI data for benchmarks discussion

16:35	Discussion	All attendees
	Cocktail reception	

Sunday 15th May

Registration opens

8:00		

Welcome

9:00	Welcome	CAGI Organizers

Polygenic Risk Score challenge

9:15	Session chair	
9:20	Data provider & assessor	Sung Chun, Shamil Sunyaev
9:50	Predictor	Doug Speed
10:00	Predictor	Zhao group
10:10	Discussion	
	Coffee break	

Clinical challenge: SickKids

11:20	Session chair	Alex Colavin
11:25	Data provider & assessor	Kyoko Yuki, Huayun Hu
11:55	Predictor	Kyoungyeul Lee
12:05	Predictor	Vicente Yepez
12:15	Discussion	
	Lunch	

Clinical challenge: Rare Genomes Project

13:25	Data provider & assessor	Sarah Stenton, Anne O'Donnell-Luria
13:55	Predictor	Panos Katsonis
14:05	Predictor	Susanna Zucca, Ivan Limongelli
14:15	Predictor	Jules Jacobsen
14:25	Discussion	

The CAGI Ethics Forum

14:35	Ethics Forum leader	Malia Fullerton
15:05	Discussion	
	Tea break	

CAGI & clinical recommendations

16:00	Introduction	Steven Brenner
16:10	Evidence-based calibration of computational tools for the clinical classification of missense variants	Vikas Pejaver
16:35	Review of the first decade of CAGI	Pedja Radivojac
17:00	Discussion	

CAGI6 feedback, leads and ideas for CAGI 7

17:10	Discussion	All participants

Assessor office hour (HMBS, PRS, RGP)

17:30		

Monday 16th May

Registration opens		
8:00		
Welcome		
9:00	Welcome	CAGI Organizers
Clinical challenges: ID panel		
9:10	Session chair	
9:15	Data provider & assessor	Emanuela Leonardi
9:45	Predictor	Raj Srinivasan
9:55	Predictor	Yexian Zhang
10:05	Discussion	
	Coffee break	
Splicing challenge		
10:45	Data provider & assessor	Carolina Jaramillo Oquendo, Diana Baralle
11:15	Predictor	Raphael Leman
11:25	Predictor	Yaqiong Wang
11:35	Predictor	Steve Mount
11:45	Discussion	
Abstract talks - part 3		
12:00	Brynja Matthíasardóttir, Yixuan Ye, Azza Althagafi, Chi Zhang	
	Lunch	
Keynote lecture		
13:30	Deep learning oracles for genomic discovery	Anshul Kundaje
Clinical challenge: Sherloc		
14:20	Data provider & assessor	Rachel Hovde, Peter Combs
14:50	Predictor	Ken Chen
15:00	Predictor	Joe Wu
15:10	Discussion	
Discussion of clinical prediction methods and challenges		
15:20		All attendees
Closing remarks		
15:50	CAGI organizers	

Conference venue

The CAGI 6 conference will be held in the Goldman Auditorium at the David Brower Center. The address of the conference venue can be found below. All scheduled conference events will take place at the conference venue.

Goldman Auditorium
David Brower Center
2159 Allston Way,
Berkeley, CA 94704

Contact information

CAGI 6 conference organizers:

Tina Bakolitsa
Email: bakolitsa@berkeley.edu
Phone: 510-990-1813

Wireless Internet

To access Wi-Fi in the Brower Center, join the "DBC Public" network. No password is required.

Due to data from human research participants and unpublished results, all CAGI materials are governed by the CAGI Data Use Policy. Tweeting and similar dissemination is encouraged only when explicitly permitted by presenters. The CAGI6 hashtag is #CAGI6.

Code of Conduct and CAGI DUA

CAGI maintains a simple list of core principles that provides the foundation for strong communities of scientists.

Code of Conduct: CAGI follows the ISMB code of conduct https://www.iscb.org/codeofconduct

CAGI Data Use Agreement: Essential information about how you can use CAGI data. Please read carefully because it might contain data restrictions that you did not expect.

CAGI aims to advance phenotypic interpretation of genomic variation. The CAGI experiments depend on the interrogation of data from people whose information has been collected as part of clinical care, following participation in a research project or biorepository, or from healthy volunteers. Some of these data -incorporating both genotypes and phenotypes- are highly sensitive and personal, and therefore must be handled with the utmost respect, integrity, and care including being maintained with the highest standards of data security and confidentiality. The success of CAGI also hinges critically on the generous contribution of pre-publication datasets and the participation of predictors and assessors. Many datasets affect individuals' careers.

To protect unpublished and sensitive data that have been shared with CAGI, and as a condition of participation in CAGI, CAGI participants must agree to the following dataset dissemination rules. We define CAGI "participants" as those who have any role in the CAGI experiment including predictors, assessors, data set providers, organizers and advisors.

- All datasets (including genotypes and phenotypes) are confidential until released by the dataset provider. Release may take the form of (a) datasets that are posted on the CAGI website and explicitly labeled as open public access, (b) explicit written permission from dataset provider to use the data for a limited set of applications, and/or (c) publication of the full contents of the dataset for unrestricted public use (publication of partial or restricted datasets constitutes release of only that partial or access-restricted dataset).
- CAGI participants agree not to share unreleased datasets with anyone except other registered and approved predictors who have agreed to these terms.
- CAGI participants agree to be responsible for maintaining the privacy and security of unreleased datasets, which they obtained from CAGI. As one example, CAGI participants must keep the files on secure systems and may not submit confidential data for predictions on third-party webservers.
- CAGI participants agree not to use an unreleased dataset for any other purposes than those described in the CAGI challenge for the dataset.
- CAGI participants agree not to use unreleased datasets for any commercial purpose.
- CAGI participants agree not to use any unreleased datasets in any publication, for example as a test case (even if the identity of the data is not disclosed) or for reporting a discovery that the CAGI participant might have made when analyzing the data.
- Following dataset release, CAGI participants may use the data with the same freedom and constraints as others who obtained the data without participation in CAGI via public mechanisms. Even after release, dataset use may be constrained (e.g., due to privacy issues) and participation in CAGI does not release CAGI participants from those constraints.
- Any requests for early release of dataset contents for a specific purpose must be submitted via the CAGI organizers, rather than directly to the dataset provider.

- In order to register for the CAGI dataset access, you must read, understand, and agree to these data use rules. If you agree to these rules, please register by providing your initials.

Dissemination of Slides Policy: CAGI follows the guidelines indicated below for disseminating slides of the CAGI conference.

- CAGI participants may use or publish any content of these slides only in compliance with the CAGI Data Use Agreement
- CAGI participants may not cite or publish any content of these slides except with the written permission of the originating author or in the primary CAGI publications
- CAGI participants must acknowledge the original authors of these slides and CAGI 6. Include the acknowledgement banner across the bottom
- CAGI participants must include the CAGI credits slide in their presentations
- Slides with a red slash (no remix) may not be included in a presentation unless all attendees are registered CAGI participants who signed the data use agreement
- Slides with a red X over the entire slide may not be used in any circumstances
- Slides with a blue slash (embargoed) may not be used unless the embargo is explicitly lifted
- Slides without slash or X may be tweeted unless flagged with the “No tweet” icon.
- CAGI participants using these slides must explain the “No tweet” icon, which means that the slide content should not be disseminated outside the presenting conference hall. If the presentation venue permits pervasive tweeting, you may not include these slides in your talk
- Note for slide authors: we expect all CAGI participants to abide by these restrictions, and will make best effort to ensure they are followed. However, but bear in mind that we have limited means of enforcing them, and therefore the restrictions cannot be guaranteed.

CAGI 6 Meeting Participants (in-person)

Gaia Andreoletti
Astellas Gene Therapies, South San Francisco, CA

Constantina Bakolitsa
University of California, Berkeley, Berkeley, CA

Gabriel Beriain
Université Libre de Bruxelles, Brussels, Belgium

Steven Brenner
University of California, Berkeley, Berkeley, CA

Yana Bromberg
Rutgers University, New Brunswick, NJ

Lawrence Carr
Patient advocate

Sung Chun
Harvard Medical School, Boston, MA

Pieter Jan Coenen
Invitae, Leuven, Belgium

Alexandre Colavin
Invitae, San Francisco, CA

Peter Combs
Invitae, San Francisco, CA

Qiang Cong
UT Southwestern Medical Center, Dallas, TX

Malia Fullerton
University of Washington, Seattle, WA

Reece Hart
MyOme Inc, Palo Alto, CA

Cindy Ho
University of California, Berkeley, Berkeley, CA

Roger Hoskins
University of California, Berkeley, Berkeley, CA

Rachel Hovde
Invitae, San Francisco, CA

Zhiqiang Hu
University of California, Berkeley, Berkeley, CA

Jules Jacobsen
Queen Mary University of London, London, UK

Milind Jagota
University of California, Berkeley, Berkeley, CA

Panagiotis Katsonis
Baylor College of Medicine, Houston, TX

Cyrielle Kint
Invitae, Leuven, Belgium

Anshul Kundaje
Stanford University, Stanford, CA

Kyle (Kyoungyeul) Lee
3billion, Seoul, South Korea

Ivan Limongelli

EnGenome, Pavia, Italy

Jennifer (Yu-Jen) Lin
University of California, Berkeley, Berkeley, CA

Selena Martinez
Patient advocate

M. Stephen Meyn
University of Wisconsin, Madison, WI

Reet Mishra
University of Berkeley, Berkeley, CA

Sean Mooney
University of Washington, Seattle, WA

Steve Mount
University of Maryland, College Park, MD

Anne O'Donnell-Luria
Broad Institute of MIT and Harvard, Cambridge, MA

Fabrizio Pucci
Université Libre de Bruxelles, Brussels, Belgium

Predrag Radivojac
Northeastern University, Boston, MA

Michael Snyder
Stanford University, Stanford, CA

Yun Song
University of California, Berkeley, Berkeley, CA

Sarah Stenton
Broad Institute of MIT and Harvard, Cambridge, MA

Qidi (Lily) Sun
University of California, Berkeley, Berkeley, CA

Shamil Sunyaev
Harvard Medical School, Boston, MA

Amanda Williams
Baylor College of Medicine, Houston, TX

Junwoo Woo
3billion, Seoul, South Korea

Yixuan Ye
Yale University, New Haven, CT

Chi Zhang
Yale University, New Haven, CT

Jing Zhang
UT Southwestern Medical Center, Dallas, TX

Susanna Zucca
EnGenome, Pavia, Italy

CAGI 6 Meeting Participants (remote)

Muttaqi Alladin
Indian Institute of Science, Bengaluru, India
Azza Althagafi
King Abdullah University of Science and Technology, Thuwal, Saudi Arabia
Emil Alexov
Clemson University, Clemson, NC
Maria Christina Aspromonte
University of Padova, Padova, Italy
Giulia Babbi
University of Bologna, Bologna, Italy
Diana Baralle
University of Southampton, Southampton, UK
Roberta Chiaraluce
Sapienza University, Rome, Italy
Valerio Consalvi
Sapienza University, Rome, Italy
Emidio Capriotti
University of Bologna, Bologna, Italy
Flavia Chen
Harvard Medical School, Boston, MA
Ken Chen
Sun Yat-sen University, Guangzhou, China
Giuditta Dal Cortivo
University of Verona, Verona, Italy
Danielle Dell' Orco
University of Verona, Verona, Italy
Piero Fariselli
University of Torino, Torino, Italy
Huayun Hou
SickKids Genome Clinic, Toronto, Canada
Tim Hubbard
King's College London, London, UK
Nurdan Kuru
Sabanci University, Istanbul, Turkey
Emanuela Leonardi
University of Padova, Padova, Italy
Raphael Leman
Centre François Baclesse, Caen, France
Olivier Lichtarge
Baylor College of Medicine, Houston, TX
Chang Lu
MRC Laboratory of Molecular Biology, Cambridge, UK
Céline Marquet
Technical University of Munich, Munich, Germany
Brynja Matthíasardóttir
University of Maryland, College Park, MD

Christian Mertes
Technical University of Munich, Munich, Germany
Carolina Jaramillo Oquendo
University of Southampton, Southampton, UK
Shailesh Panday
Clemson University, Clemson, NC
Vikas Pejaver
Ikahn School of Medicine at Mount Sinai, New York, NY
Carlos Rodrigues
University of Melbourne, Melbourne, Australia
Yang Shen
Texas A&M University, College Station, TX
Damian Smedley
Queen Mary University of London, London, UK
Douglas Speed
Aarhus University, Aarhus, Denmark
Rajgopal Srinivasan
TATA Consultancy Services, Hyderabad, India
Alexey Strokach
University of Toronto, Toronto, Canada Yuanfei Sun
Uma Sunderam
TATA Consultancy Services, Chennai, India
Wuwei Tan
Texas A&M University, College Station, TX
Warren van Loggerenberg
University of Toronto, Toronto, Canada
Yaqiong Wang
Fudan University, Shanghai, China
Joe (Yingzhou) Wu
University of Toronto, Toronto, Canada
Vicente Yepez
Technical University of Munich, Munich, Germany
Rujie Yin
Texas A&M University, College Station, TX
Kyoko Yuki
SickKids Genome Clinic, Toronto, Canada
Geyu Zhou
Yale University, New Haven, CT
Shaowen Zhu
Texas A&M University, College Station, TX

Conference Abstracts

Predicting the effects of missense variations and the case of MTHFR deficiency

Giulia Babbi[1], Castrense Savojardo[1], Samuele Bovo[1,2], Davide Baldazzi[1,3], Pier Luigi Martelli[1]*and Rita Casadio[1,4]

[1] Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna,40126 Bologna, Italy;

[2] Department of Agricultural and Food Sciences, University of Bologna, 40127 Bologna, Italy;

[3] Unit of Oncogenetics and Functional Oncogenomics, CRO Aviano, National Cancer Institute,IRCCS, 33081 Aviano, Italy.

[4] Institute of Biomembranes, Bioenergetics and Molecular Biotechnologies (IBIOM), ItalianNational Research Council (CNR), 70126 Bari, Italy
* Correspondence: pierluigi.martelli@unibo.it
giulia.babbi3@unibo.it (G.B.); castrense.savojardo2@unibo.it (C.S.); samuele.bovo@unibo.it (S.B.); davide.baldazzi8@unibo.it (D.B.); rita.casadio@unibo.it (R.C.)

Predicting the effect of variations on protein functional activity and disease associations is a strategic task of growing relevance – and evaluating these predictions is a goal of CAGI experiment itself. The Bologna Biocomputing Group provides resources and expertise for bothcurating databases and implementing tools to endow biological data with structural and functional annotations. Here we present some of our in-house computational tools and approaches that may help in solving this task, along with a specific case study on MTHFR deficiency that was part of the CAGI6 challenges.

For understanding the possible effect of missense variations in protein-protein interactions or domain-domain interfaces, we adopt our machine-learning based DeepREX tool (http://deeprex.biocomp.unibo.it). When possible, we analyse variations at the structural level, interms of the location of the residue with respect to the protein surface, and its distance from the active site and/or important functional motifs.

We adopt a consensus method to investigate whether missense variations are related to proteininstability. We use three state-of-the-art methods that predict the change of folding Gibbs free energy (ΔΔG) associated with a specific variation: our INPS/INPS-MD, based on machine- learning (https://inpsmd.biocomp.unibo.it), FoldX, based on statistical potentials, and PoPMuSiC2, based on statistical potentials and machine-learning.

To classify missense variations leading to disease insurgence we computed their likelihood of being disease-related with SNPs&GO (https://snps-and-go.biocomp.unibo.it). When the task isto select possible causative variations on more genes, we screened disease-related genes through our in-house resources eDGAR (http://edgar.biocomp.unibo.it) and PhenPath (http://phenpath.biocomp.unibo.it), both relying on NETGE-PLUS algorithm (http://net- ge2.biocomp.unibo.it) for enriching disease/phenotype set for functional annotations.

As a study case, we present the analysis of 72 missense variations of human MTHFR protein (Methylenetetrahydrofolate reductase) known to be associated with the disease “MTHFR deficiency”. By estimating the thermodynamic ΔΔG change according to the proposed consensus method, we find that 61% of the disease-related variations destabilize the protein, are present both in the catalytic and regulatory domain and correspond to known biochemical deficiencies. The propensity of solvent-accessible residues to be involved in protein-protein interaction sites, as predicted with DeepREX, indicates that most of the interaction sites are located in the regulatory domain. We show that patterns of disease-associated, physicochemical variation types, both in the catalytic and regulatory domains, are unique for theMTHFR deficiency when mapped into the protein architecture.

Mutational Effect Prediction with Protein Language Modeling andStructural Information

Yuanfei Sun, Yang Shen*
Texas A&M University, College Station, TX
77843yshen@tamu.edu

The workhorse molecules of life, proteins play a central role in cellular functions and are major drug targets. Their variations in humans and pathogens often lead to genetic diseases and therapeutic resistance. The ability to decipher the association between protein variations and resulting effects would facilitate disease prognostics and biologics design, e.g. antibodies.

Although multiplexed assays of variant effects (MAVE), such as deep mutational scanning (DMS) and massively parallel reporter assay (MPRA) experiments, are generating data about variant effects ranging from protein stability to cell viability, their speed and applicability are dwarfed by the amount of variants and effects to characterize. Therefore, there is a critical need to develop high-throughput and accurate computational tools that can overcome the lack or thelimitation of experimentally labeled variant data.

Our algorithmic answer for such “zero-shot” or “few-shot” variant effect prediction lies in the protein language models (PLMs). Attributed to advancements in sequencing technology, there are abundant “unlabeled” data of functional sequences (sampled by nature through evolution) across species. These primary sequences of amino acids resemble texts in natural languages and their distributions have been effectively learned by various pretrained LMs. We pretrained our transformer-based PLMs using highly sparse domain sequences from Pfam RepresentativeProteomes set (at 15% and 75% levels), and finetuned them with UniRef100 homology domain sequences within the given family. Using the modeled ratio of likelihood between variant and wild-type sequences as zero-shot predictors, our unsupervised models without the use of labeled data showed comparable performance over benchmark datasets against both alignment-based and alignment-free state-of-the-art methods (average difference in rank correlation: 0.007 to DeepSequence, 0.017 to Shin’s autoregressive model). We also compared two mainstream approaches for likelihood factorizations: autoregression and denoising. We found that the autoregressive model showed an edge for mutations at terminal regions while the denoising model outperformed for middle-region mutations. We further examined the performances over different order of missense mutations spanning from single to 24-site on HIS7_Yeast. Our results showed evidence that PLMs are epistasis aware. In an experiment to anticipate spike-protein mutations in COVID, our PLMs correctly discovered all single mutationsof the five WHO-defined Variants of Concern within top 5 predictions. Lastly, beyond modeling sequences only, we further embed structures as an extra modality information for PLMs through spatial message passing and multi-task learning. Our numerical results verify that the structure-aware sequence embeddings improve fitness prediction.

Assessment of the CAGI 6 HMBS challenge

Jing Zhang
UT Southwestern Medical Center, Dallas, TX

We will present the evaluation of predictions for the HMBS challenge. In the challenge, the participants were asked to predict effects on yeast growth caused by missense variants of human hydroxymethylbilane synthase, a protein involved in the third step of heme biosynthesis. For the evaluation, the disparate distributions and scaling between predictions and experimental scores remain the critical hurdle for impartial assessment. The performance of predictors implementing different algorithms and methods is similar. The Kendall tau ranges from 0.1 to 0.3 for 8 out of 10 groups. Most predictors are able to identify the highly deleterious (experimental score less than 0.3) or benign (experimental score more than 0.8) variants with modest accuracy with highest AUC above 0.7 respectively. However, for variants which slightly harm the growth of yeast with experimental growth score from 0.3 to 0.8, the performance of predictors is nearly random with maximum MCC less than 0.09. This pattern suggests that predictors work mostly like binary classifiers rather than predicting continuous scale scores. Furthermore, variants showing benefits on yeast growth are also poorly predicted. Meanwhile, the baseline predictor which is based purely on multiple sequence alignment outperforms most predictors with only three groups surpassing its performance. Nevertheless, the assessment scores of positive control and baseline predictor suggest substantial improvements in accuracy of predictions in the future should be possible, likely by methods not currently explored by predictors, which seem to be saturated at what they can achieve.

CalVEIR performance in the CAGI 6 HMBS challenge

Milind Jagota

Deep learning models of protein sequences such as AlphaFold have enabled recent breakthroughs in molecular biology. There has been interest in developing such models for variant effect prediction of coding regions of the genome. Models such as DeepSequence, EVE, and ESM-1v have approached missense variant effect prediction in an unsupervised manner with success. We developed a method for variant effect prediction that expands on these and conventional methods and tested our method on the HMBS prediction challenge in CAGI6. Our method was one of the top performers and combined novel structural features of protein function together with existing variant effect predictors. We trained a supervised regression on data from proteins that are distant from the target, demonstrating successful transfer across proteins. Our success provides new insight into representations of protein function and robust approaches to supervised learning for variant effect prediction.

Exploring the fitness landscape through structural andevolutionary models

by Fabrizio Pucci[1,2], Gabriel Cia[1,2], Marianne Rooman[1,2,*]

[1] Computational Biology and Bioinformatics, Université Libre de Bruxelles
[2] Interuniversity Institute of Bioinformatics in Brussels
* Correspondence: marianne.rooman@ulb.be

Despite the bioinformatics advances of the past decades, accurately predicting the impact of mutations on protein fitness remains a challenging goal. Indeed, the correlations between predicted and experimental protein fitness values are still quite limited, as was also found in the previous CAGI competition [1].

In the context of the current CAGI6 challenge, we present three different computational models that combine structural and coevolutionary information, which we applied to perform fitness predictions on the HBMS protein target:

- Model 1: As stability is one of the key ingredients of protein fitness, we used a rescaled version of our in-house structure-based PoPMuSiC predictor [2] that estimates the impact of variants on protein thermodynamic stability using a simplified representation of protein structures and statistical potentials.

- Model 2: This model uses a rescaled version of our deleteriousness prediction tool SNPMuSiC [3], which combines predictions based on protein structure and statistical potentials with the evolutionary score of the PROVEAN predictor [4].

- Model 3: In this last model, several prediction scores are combined through a simple linear regression model: the structure-based scores of PoPMuSiC [2] and MAESTRO[5], the residue solvent accessibility, and the (co)evolutionary scores such as PROVEAN [4] and EVCoupling [6].

We first applied these three models to the proteins CALM1, TPK, UBE2 and SUMO to identify the parameters of the models, and then blindly to HMBS. We conclude by comparing the experimental and predicted results and by discussing the advantages and limitations of the three approaches.

[1] Andreoletti, Gaia, et al. Human mutation 40.9 (2019): 1197-1201.
[2] Dehouck, Yves, et al. BMC Bioinformatics 12.1 (2011): 1-12
[3] Ancien, François, et al. Scientific Reports 8.1 (2018): 1-11.
[4] Choi, Yongwook, and Agnes P. Chan. Bioinformatics 31.16 (2015): 2745-2747.
[5] Laimer, Josef, et al., BMC Bioinformatics 16.1 (2015): 1-13.
[6] Hopf, Thomas A., et al. Bioinformatics 35.9 (2019): 1582-1584.

Assessing protein contribution to phenotypic change using short,coarse grain molecular dynamics simulations

Muttaqi A. Alladin*, Debnath Pal
Department of Computational and Data Sciences, Indian Institute of Science, Bengaluru 560012, India.
Muttaqi A. Alladin, Email: muttaqiahmad@iisc.ac.in

Debnath Pal, Email dpal@iisc.ac.in

Since the advent of high-throughput sequencing technologies, a large amount of data about proteins and protein variants has been generated, and interpreting the effect of these variants on the phenotype has been an enormous challenge. Although various methods exist that try to make a functional mapping between phenotype and genotype, many of these methods, like machine learning methods, are often computationally expensive to train and difficult to interpret.We use a relatively straightforward approach to create a functional mapping between the proteinvariants and the phenotype by using short, coarse grain molecular dynamics simulations. In our method, we carry out short coarse-grained molecular dynamics simulations (<10ns) for two different structures of the same protein. The two different structures could be where the same protein is in complex with different ligands/cofactors or where one structure is free, and the other structure is in complex with a ligand/cofactor. We have used this method in CAGI6 challenges for HMBS, MAPK1, and MAPK3 as targets. We have also used this method on Calmodulin. For MTHFR, SUMO1, and UBE2I, we explored getting a functional mapping between the phenotype and the variants using one single protein structure. As we are using a single structure instead of two, this method is not identical to the one used for CAGI6 Challenges or Calmodulin but is comparable as we are still using short coarse-grained molecular dynamics simulations to get a functional mapping. Although we don’t have the resultsof HMBS, MAPK1, and MAPK3 as the results of CAGI6 Challenges haven’t been publicly declared yet, this method has performed reasonably well on Calmodulin. For Calmodulin, we got a correlation of 0.6 with phenotype change for two-thirds of the data. Similarly, our functionalmapping of MTHFR also yielded a correlation coefficient of 0.6 for over two-thirds data. For SUMO1, we got a correlation coefficient of 0.7 for about 70% of the data, while UBE2I gave us acorrelation coefficient of 0.6 for 60% data. We hope that our method will open new avenues to rationally improve genome interpretation.

Impact of MAPK1/MAPK3 missense variants found in cancer:structural, function and stability experimental analysis

Maria Petrosino[1], Leonore Novak[1], Alessandra Pasquo[2], Emidio Capriotti[3], Roberta Chiaraluce[1],Valerio Consalvi*[1]

[1]Dipartimento di Scienze Biochimiche “A. Rossi Fanelli”, Sapienza University of Rome, Rome

(Italy)

[2]ENEA CR Frascati, Diagnostics and Metrology Laboratory FSN-TECFIS-DIM,Frascati (Italy)

*valerio.consalvi@uniroma1.it

[3]Department of Pharmacy and Biotechnology (FaBiT), University of Bologna. Bologna (Italy)

MAPK1(ERK2) and MAPK3 (ERK1) are serine/threonine kinase in the Ras-Raf-MEK-ERK signal transduction cascade that regulates cell proliferation, transcription, differentiation, and cell cycle progression. MAPK1 and MAPK3 are very similar in sequence (84%) and usually considered to be functionally redundant, although recent studies report evidence that they might play different roles. MAPK1/MAPK3 are activated by phosphorylation which occurs with strict specificity by MEK1/2 on Thr185/202 and Tyr187/204. Upon activation, MAPK1/MAPK3 translocate to the nucleus where they phosphorylate specific nuclear targets. Owing to their biological importance, they represent an important target of biomedical research and of a large part of drug discovery research.

A library of eleven MAPK1 and thirteen MAPK3 missense variants selected from the COSMICdatabase were analyzed by near and far-UV circular dichroism and intrinsic fluorescence spectra to determine thermodynamic stability. These are somatic variants detected in cancer tissues and are distributed along the protein sequence.

The thermodynamic stability was measured by monitoring the spectral changes (far-UV circular dichroism and intrinsic fluorescence emission) at increasing denaturant (guanidinium chloride) concentration. The variation of unfolding free energy (ΔG) is calculated by fitting the spectral changes at zero denaturant concentration (ΔGH2O). These data were used to calculate a $\Delta\Delta G$H2O value, the difference in unfolding free energy ΔGH2O between each variant and the wildtype protein, both in phosphorylated and unphosphorylated form. The catalytic efficiency $(k_{cat}/K_m)^{mut}/(k_{cat}/K_m)^{wt}$ of phosphorylated MAPK1 and MAPK3 missense variants was determined by a fluorescence assay based on Chelation-Enhanced Fluorescence upon substrate peptide phosphorylation by the kinase.

Assessing the predictions of the MAPK1/MAPK3 challenges

*Paola Turina[1], Maria Petrosino[2], Andrea Cicconardi[1], Leonore Novak[2], Alessandra Pasquo[3],Roberta Chiaraluce[2], Valerio Consalvi[2], Emidio Capriotti[1]**

[1] Department of Pharmacy and Biotechnology (FaBiT), University of Bologna. Bologna (Italy)
[2] Department of Biochemical Sciences, Sapienza University of Rome, Rome (Italy)
[3] ENEA CR Frascati, Diagnostics and Metrology Laboratory FSN-TECFIS-DIM, Frascati (Italy)

*email: emidio.capriotti@unibo.it

Data provider from the "Sapienza" University in Rome Verona characterized the impact ofmissense variants in MAPK1 and MAPK3. The experimental studies on 11 and 13 variants of MAPK1 and MAPK3 respectively, allowed to measure the free energy change (□G_{h2o}) and the enzymatic activity (k_{cat}/K_M) for the unphosphorylated and phosphorylated. The MAPK1/MAPK3 challenges were participated by 13 groups which submitted more than 40 predictions. Todetermine the effect of each mutant on protein stability and function, we calculated the variation of free energy change (□□G_{h2o}) and the variation of enzymatic activity (□k_{cat}/K_M). Comparing the prediction with experimental values of the □□G_{h2o} for MAPK3 we found that the best method fromTeam3 resulted in a Pearson Correlation Coefficient (PCC) of 0.64 and a Root-Mean-Square- Error (RMSE) of 1.9 kcal/mol for the unphosphorylated form, while a method from Team4 reacheda PCC of 0.68 and a RMSE of 1.2 kcal/mol for the phosphorylated form. For the prediction of the variation of activity of MAPK3 prediction from Team5 achieved and overall accuracy (Q2) of 0.75Matthews Correlation Coefficient 0.378 an Area Under the Receiver Operating Characteristic Curve of 0.80. More complex is the analysis of variants in MAPK1 for which a folding mechanism changes. Our analysis shows that Team 4 reached good performance for the Unphosphorylated form of MAPK1 while a method from Team2 predicts with good performance the (□k_{cat}/K_M). Overallfor the predictions for MAPK3 variants are more accurate than those achieved for the MAPK1 challenge.

Predicting changes in stability and catalytic efficiency of MAPK1 andMAPK3 variants using 3D structure information

by Fabrizio Pucci[1,2,*], Martin Schwersensky[1,2], Marianne Rooman[1,2,*],

[1] Computational Biology and Bioinformatics, Université Libre de Bruxelles
[2] Interuniversity Institute of Bioinformatics in Brussels
[*] Correspondence: fabrizio.pucci@ulb.be, marianne.rooman@ulb.be

A series of missense variants in mitogen-activated protein kinase (MAPK) of types 1 and 3, selected from the COSMIC database of somatic cancer mutations, were proposed as CAGI6 challenges. The goal was to predict the changes in stability and catalytic efficiency of both the unphosphorylated and phosphorylated forms of the proteins. For this purpose, we used the different X-ray structures available for these proteins in the Protein Data Bank, and modeled the phosphorylated forms by replacing the phosphorylated residues by negatively charged amino acids[1].

To predict the stability changes, we applied several models which exploit the 3D structure of the target protein: (1) our in-house predictor of folding free energy changes upon mutations ($\Delta\Delta G$), called PoPMuSiC[2], which uses several statistical potentials as input features of an artificial neural network; (2) an unbiased version of PoPMuSiC[3] ensuring that the $\Delta\Delta G$ value of every mutation is equal to minus the $\Delta\Delta G$ value of the reverse mutation; and (3) an average of PoPMuSiC and another $\Delta\Delta G$ predictor, MAESTRO[4].

To predict the change in catalytic efficiency of the variants, we used two rescaled versions of the PoPMuSiC $\Delta\Delta G$ predictor. As a third model, we developed a fitness predictor based on a linear regression model integrating four evolutionary features (sequence variation and covariation) and four structure-based features (PoPMuSiC, MAESTRO, SNPMuSiC[5], solvent accessibility), of which we identified the coefficients on the basis of experimental fitness values.

[1] Pearlman et al. (2011). Cell 147(4), 934-946.
[2] Dehouck Y et al. (2009). Bioinformatics, 25(19), 2537-2543.
[3] Pucci F et al. (2018). Bioinformatics, 34(21), 3659-3665.
[4] Laimer J et al. (2015). BMC bioinformatics, 16(1), 1-13.
[5] Ancien F et al. (2018). Scientific reports, 8(1), 1-11.

Application of HoTMuSiC to the calmodulin CAGI6 challenge

by Gabriel Cia[1,2], Marianne Rooman[1,2,*], Fabrizio Pucci[1,2]

[1] Computational Biology and Bioinformatics, Université Libre de Bruxelles
[2] Interuniversity Institute of Bioinformatics in Brussels
[*] Correspondence: marianne.rooman@ulb.be

Reliably estimating the change in thermal stability of proteins upon mutation is an important objective for the rational optimization of enzymes, especially in bioprocesses requiring unusual temperature conditions. We have recently developed a tool that predicts the change in melting temperature ΔT_m upon point mutations, which uses as input the wild-type protein structure and, when available, the wild-type melting temperature T_m. Our model, called HotMuSiC [1], relies on a combination of standard and temperature-dependent statistical potentials which were used as input features to train a neural network. The model was trained on a dataset of over 1,600 manually curated mutations with experimentally measured ΔT_m. It achieves a correlation of 0.61and a root mean square deviation of 4.2 °C between the predicted and experimental ΔT_m in 5-fold cross validation, which increases to 0.75 and 2.9 °C when ignoring the top 10% outliers.

To apply our HotMuSiC predictor to the calmodulin challenge, we used a modeled 3D structure for the apo form and an experimental structure for the holo form. Since the apo form follows a classical two-state folding transition, we simply applied our model to predict the changes in melting temperature, ΔT_m, of the variants. For the holo form, which follows a three-state transition, we predicted separately the ΔT_m of mutations situated in the C- and N-terminal domains; this means that we assumed that the melting temperature of the non-mutated domain remains unchanged. Furthermore, to predict the percentage of unfolding of each variant, we combined the predicted change in T_m with an experimentally determined enthalpy value ΔH_m and used a purposefully derived function relating these values with the percentage of unfolding.Finally, to predict the destabilization score of the variant, we simply calculated the change in the percentage of unfolding between the wild-type and the variant.

[1] Pucci, Fabrizio, et al. Scientific reports 6.1 (2016): 1-9.

Computational Tools to Predict Protein Stability Changes Upon MissenseMutations

Carlos H. M. Rodrigues[1], Stephanie Portelli[3], Douglas E. V. Pires[2], David B. Ascher[1, 3, *]

1 Computational Biology and Clinical Informatics, Baker Heart and Diabetes Institute,Melbourne, Victoria, Australia
2 School of Computing and Information Systems, University of Melbourne, Melbourne, Victoria,Australia
3 School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane,Queensland, Australia

*To whom correspondence should be addressed. D.B.A. Email: d.ascher@uq.edu.au

On-going technological advances have led to dramatic increases in the amounts of biological data being generated over the years. Along with the evolution of high performance computing and computational tools, this has provided us with a wealth of information, analytical power andthe opportunity to investigate fundamental health and biotechnological problems of different magnitude and kind, complementary to and able to guide conventional approaches.

Our group is interested in developing and experimentally validating novel computational methods to exploit this data, enhancing the impact of genome sequencing, structural genomics,and functional genomics on biology and medicine. One of our main areas of interest is in the development of predictive and analytical tools and databases to investigate and understand the relationship between protein sequence, structure and function and phenotype. These methods allow us to gain unique insights into the molecular basis of genetic diseases, as well as a betterunderstanding of the molecular mechanisms behind drug resistance, which has direct implications into guiding personalised patient treatment, the development of resistance resistantdrugs, and to aid the design of novel drugs.

For the Calmodulin Challenge in CAGI6, our team has submitted 6 different predictions for eachmissense mutation. These are derived from our well established suite of methods (mCSM, DUET, ENCoM, SDM, DynaMut and DynaMut2), which leverage physicochemical properties and distance pattern signatures extracted from protein structure data. Our tools are freely available to the scientific community and have been widely used in industry and academia all over the world with over 1 million hits/year.

Here we considered the APO structure of CaM as entry 1DMO on the Protein Data Bank, and entry 1CLL as the protein under Ca2+-saturating conditions. Each mutation and PDB structure were then input into our webservers and results were compiled accordingly. Our methods predict the effects of mutations in terms of changes in the Gibbs Free Energy of folding (ΔΔG),which were then used as a direct measure of melting temperature values requested for this challenge. As one of the selected best-performing predictors for this challenge we look forwardto the opportunity to present our findings and overall methodology used for our submissions.

Variant interpretation with BioFolD tools

*Andrea Cicconardi, Riccardo Ottalevi, Anna Benedetti, Paola Turina, Emidio Capriotti**
BioFolD Unit, 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, the BioFolD unit 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 of features used for detecting pathogenic variants and predicting the variation of protein stability. The tools for predicting pathogenic variants include PhD-SNP (Capriotti, et al., 2006), which is a support vector machine based approach based on sequence information extracted from the protein sequence profile, SNPs&GO (Capriotti, et al., 2013b) which relies on functional information encoded by Gene Ontology terms and, when available, protein structure features and Meta-SNP (Capriotti, et al., 2013a) a meta prediction toolcombining 4 well-establish methods. More recently, PhD-SNP^{g} (Capriotti and Fariselli, 2017) usesthe information retrieved on the UCSC genome browser to predict the impact of variants in noncoding regions and DDGun (Montanucci, et al., 2022) which predicts the variation of protein stability upon mutation.
During the last edition of the CAGI, we participated in five challenges using modified version of our methods to predict the functional effect of variants on hydroxymethylbilane synthase (HMBS), Serine/Threonine Kinase (STK11), methylenetetrahydrofolate reductase (MTHFR) and their clinical impact (Splicing VUS, Sherloc clinical classification). All the tools used for the CAGI challenges are available at https://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 diseasesassociated 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.
Montanucci L, Capriotti E, Birolo G, Benevenuta S, Pancotti C, Lal D, Fariselli P (2022). DDGun:an untrained predictor of protein stability changes upon amino acid variants. Nucleic Acids Research. DOI:10.1093/nar/gkac325.

Phylogeny-aware computing of tolerance for missense mutations

Nurdan Kuru[1], Onur Dereli[1], Emrah Akkoyun[1], Aylin Bircan[1], Oznur Tastan[1], Ogun Adebali[1*] [1] Faculty of Engineering and Natural Sciences, Sabanci University, Istanbul 34956, Turkey

* To whom correspondence should be addressed: oadebali@sabanciuniv.edu

With the advancement in high throughput sequencing technologies, our ability to detect genetic variation and predict the effect of a variant in clinical diagnosis has been revolutionized.

Understanding the effect of these missense mutations is critical for diagnosing rare diseases. Here, we propose a novel phylogeny-dependent probabilistic approach to predict the functionaleffects of missense mutations. Our approach exploits independent evolutionary events and phylogenetic relationships among species to measure the deleteriousness of a given variant.

We estimate the probability of observing any amino acid at the queried position of the protein inquestion by traveling through the phylogenetic tree. This process helps us analyze substitutionswithin the context of the phylogenic relations, and we use this information to assess substitutions' effects over the queried sequence. We assess the predictive performance of our approach (PHACT) on various subsets of a dataset, which contains 3023 proteins and 61662 variants obtained from Clinvar, Humsavar, and Gnomad. The experiments demonstrate that our method outperforms widely used pathogenicity prediction tools (i.e., SIFT and PolyPhen-2) and achieves similar or better predictive performance compared to existing conventional statistical approaches presented in dbNSFP (Figure 1). We now extend this approach to a machine learning-based approach. We use PHACT scores with other phylogenetic tree-related features in a gradient-boosting tree-based classifier. Our preliminary results show that the model outperforms other machine-learning-based algorithms.

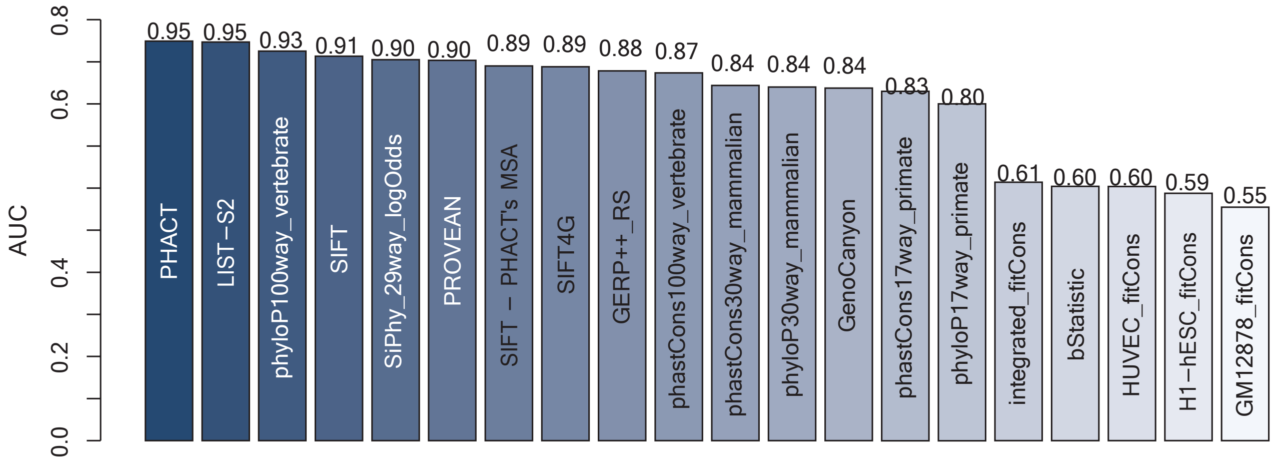

Figure 1: AUC comparison of PHACT against statistical pathogenicity predictionalgorithms presented in dbNSFP

Progress in complex phenotype prediction: the CAGI6 PRS challenge

Sung Chun
Harvard Medical School

Complex traits, which include common genetic disorders, are highly polygenic with numerous alleles of small effects contributing to the genetic risk. A proportion of this risk can be predicted by statistical methods that rely on results of Genome-Wide Association Studies (GWAS). Currently, most popular methods leverage GWAS summary statistics in the form of Polygenic Risk Scores (PRS). PRS has potential clinical utility for risk surveillance, prevention and personalized medicine. The CAGI6 PRS challenge assessed the performance of PRS algorithms using data on four phenotypes (Type 2 Diabetes, Breast Cancer, Inflammatory Bowel Disease and Coronary Artery Disease) representing disease areas that could benefit from PRS because of the availability of screening or early intervention options. We also assessed algorithms on a range of simulated data to get insight into the way algorithms perform in various parametric regimes. The accuracy of submitted algorithms was compared against a set of published baseline methods. The assessment revealed that one method outperformed state-of-the-art PRS algorithms in IBD (Nagelkerke's R^2=0.173 compared to 0.157) and across wide ranges of parameter values in simulated genetic architecture, highlighting that the current linear prediction models can be further improved. Algorithms that leverage functional annotations of genetic variants underperformed in comparison. Also, machine learning-based prediction models did not perform well. While this may be due to the paucity of non-linear genetic effects in complex traits, the current challenge was not structured to evaluate the full potential of these approaches. Due to privacy issues, it was not possible to openly share the full training data, and only a limited set of clinical covariates were available to use in prediction. Another limitation involved restricting this challenge to European ancestry individuals. Despite these limitations we hope to address in future challenges, we find the current PRS challenge valuable to assess the current state of the field. Despite these limitations which we hope to address in future challenges, we find the current PRS challenge valuable to assess the current state of the field.

MegaPRS Prediction of Complex Traits

Douglas Speed
Aarhus University, Aarhus, Denmark

I will explain MegaPRS, my tool for constructing polygenic risk scores (PRS) from summary statistics. MegaPRS improves upon existing PRS tools by allowing the user to specify the heritability model (how heritability is expected to be distributed across the genome). When applied to UK Biobank data, MegaPRS out-performs existing tools (e.g, SBLUP, lassosum, LDpred and SBayesR) for 223 out of 225 phenotypes. The average improvement in accuracy is 14% (SD 1), equivalent to increasing the sample size by a quarter. Furthermore, MegaPRS is computationally efficient, taking less than an hour to construct genome-wide PRS. MegaPRS is freely available within the software package LDAK (www.ldak.org).

Zhao group CAGI6-PRS challenge

Chi Zhang[1], Yixuan Ye[2], Geyu Zhou[2], Hongyu Zhao[1,2,*]

1. Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA
2. Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA

[*] Correspondence: hongyu.zhao@yale.edu

Polygenic risk score (PRS) has great promise for disease prevention, monitoring, and treatment.We participated in the CAGI6-PRS challenge with two PRS methods developed by our group: Annopred and SDPR and compared with several other state-of-the-art PRS methods (P+T, LDpred, and PRS-CS). More specifically, Annopred is a robust Bayesian framework that leverages diverse types of genomic and epigenomic functional annotations in genetic risk prediction, whereas SDPR is an efficient PRS method that does not rely on parametric assumptions of the effect size distributions nor validation datasets for parameter tuning.

For Annopred, we generated 88 candidate PRSs for each disease (CAD, BC, T2D, and IBD) under different tuning parameters. In particular, we estimated the effect sizes of all candidate SNPs using external GWAS summary statistics and computed PRS for all individuals in the UKBB training dataset. For each disease, the optimal tuning parameters were selected and thePRS model built using the “optimal” tuning parameter(s) was then evaluated and compared in the testing dataset. As parameter tuning is also required for P+T and LDpred, we followed the same procedure to evaluate their prediction performance. For SDPR, we generated one candidate PRS for each simulation and real dataset using provided summary statistics as input.We did not use the provided validation dataset since SDPR does not require parameter tuning.

Taken together, our results suggest that both AnnoPred and SDPR can significantly increase the accuracy of polygenic risk prediction and risk population stratification compared to the otherstate-of-the-art methods. Please refer to other abstracts of our group for an expanded discussion about Annopred, SDPR, and xPred—a novel method that improves cross-populationprediction of PRS.

Comparison of PRS methods for predicting Alzheimer's' Disease

Chi Zhang[1], Yixuan Ye[1], Hongyu Zhao[1,2,*]

1 Department of Biostatistics, Yale School of Public Health, New Haven, CT;

2 Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT;

* Correspondence: hongyu.zhao@yale.edu

Previous genome-wide association studies (GWAS) have revealed 38 susceptibility loci for Alzheimer's disease.And the Polygenic Risk Scores (PRS) have been widely created in multiple earlier studies to discriminate patients with AD from cognitively normal individuals AD GWAS data from the International Genomics of Alzheimer's Project (IGAP). Because several advanced PRS approaches have been developed in recent years, purpose of this study is to examine different PRS methodologies and to develop a robust PRS for AD. We investigated four PRS approaches (P+T, LDpred, PRScs, and AnnoPred) using results from two separate GWASs (IGAP GWAS: 21,982 cases and 41,944 controls; meta GWAS: 71,880 cases and 383,378 controls). We found that AnnoPred consistently increased prediction accuracy (AUC 0.675 using IGAP GWAS; AUC 0.696 using meta GWAS). Furthermore, as there is evidence for sex differences in AD symptomatology, progression, biomarkers, risk factor profiles, and treatment we derived sex-specific PRSs using ADGC data. These sex-specific PRSs are based on variousPRS models (PRScs, LDpred, AnnoPred, PRScsx, PleioPred and XPASS) to see if incorporating more information from the opposite sex could improve prediction performance.

Overall, AnnoPred based on sex-agnostic GWAS data provided the best prediction accuracy. This suggests that when sample size is limited, the benefits of larger sample size exceed the benefits of sex-specificity where the sample sizes of the ADGC data we used to create sex- specific AD GWASs are 4207 for female and 5702 for male. However, sex-specific PRS for ADis still worth investigating when there are more sex-specific GWAS results with larger sample sizes become available in the future.

Joint modeling of multi-population data and functional annotations to increase accuracy of polygenic risk prediction

Yixuan Ye[1], Yiliang Zhang[2], Leqi Xu[2], Wei Jiang[2], Chi Zhang[2], Geyu Zhou[1], Hongyu Zhao[1,2,*]

[1]Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA

[2]Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA

*Correspondence email: hongyu.zhao@yale.edu

The clinical use of polygenic risk scores (PRS) in non-European populations is hindered by the Eurocentric biases in genetic studies and the poor transferability of genetic results across populations. Here we propose a novel Bayesian PRS framework, xPred, which leverages GWAS summary statistics from multiple populations to boost the predictive power of PRS in under-represented populations. We also propose its extension, xPred-anno, to integrate functional annotations to upweight the genetic variants likely to be functional. Both xPred and xPred-anno employ a four-component mixture prior to model the effect sizes of genetic variants, where the genetic effects are coupled across populations via a shared proportion of causal SNPs. Through simulations and real data analyses on several quantitative traits and type 2 diabetes, we demonstrate that our approaches can substantially increase the accuracy of polygenic risk prediction and risk population stratification compared to the existing methods.

A fast and robust Bayesian nonparametric method for prediction ofcomplex traits using summary statistics

Geyu Zhou[1], Hongyu Zhao[1,2,*]

1. Program of Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA
2. Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA

[*]Correspondence: hongyu.zhao@yale.edu

Genetic prediction of complex traits has great promise for disease prevention, monitoring, and treatment. The development of accurate risk prediction models is hindered by the wide diversityof genetic architecture across different traits, limited access to individual level data for training and parameter tuning, and the demand for computational resources. To overcome the limitations of the most existing methods that make explicit assumptions on the underlying genetic architecture and need a separate validation data set for parameter tuning, we develop aSummary-statistics based Dirichlet Process Regression method SDPR that does not need to tune parameters. In our implementation, we refine the commonly used likelihood assumption todeal with the discrepancy between summary statistics and external reference panel. Through simulations, we show that SDPR is adaptive to different genetic architectures and robust to heterogeneity of per SNP sample sizes. In real data analysis, we compared the performance ofSDPR with 7 recently developed or current state of art methods (PRS-CS, SBayesR, LDpred, P+T, LDpred2, lassosum and DBSLMM) on 6 quantitative (height, BMI, HDL, LDL, total cholesterol and triglycerides) and 6 binary traits (coronary artery disease, breast cancer, IBD, type 2 diabetes, schizophrenia and bipolar). SDPR achieved the best performance for 6 traits (height, BMI, HDL, LDL, total cholesterol, breast cancer), and top tier performance for 4 additional traits (IBD, type 2 diabetes, schizophrenia, bipolar; within 0.003 of AUC difference compared with the top method). Furthermore, SDPR is able to fit the model in 15 minutes whenexecuted in parallel, significantly faster compared with 2-5 hours for PRS-CS, LDpred and LDpred2. Taken together, we believe that SDPR has the potential to be widely used given its competent performance on real traits, easiness to use and excellent computational efficiency.
SDPR is freely available at https://github.com/eldronzhou/SDPR.

Predicting diagnostic variants in the CAGI6 SickKids challenge

Junwoo Woo, Kyoungyeul Lee
3billion, Seoul, South Korea

Our mission at 3billion is to help end the diagnostic odyssey for genetic disease patients around the globe. Our method of prediction involves the use of ACMG Bayesian scores, semantic similarity scores, and scores generated using 3Cnet, a sequence-based deep neural network trained using clinical, common, and conserved mutation data.

ACMG Bayesian scores are calculated using EVIDENCE, our in-house bioinformatics pipeline for variant annotation while semantic similarity scores quantify the relation between patient phenotypes and those of candidate diseases with known associations to patient variants. Variant sequences are passed through 3Cnet to generate predicted pathogenicity scores. We trained a supervised logistic regression model on our repository of in-house patient data, using the Bayesian, semantic, and 3Cnet scores of variants as features and clinically confirmed variants (interpreted in accordance with the ACMG guidelines) as labels. We internally evaluated the performance of our method using top-K recall at the variant level of resolution and observed that in many cases, confirmed variants were discovered at favorable, lower values of K. We anticipate that the integration of this method to 3billion’s current internal genetic testing workflow will, on average, shorten the time required per case of genetic testing.

Gene prioritization for rare diseases integrating genotype, RNA-seq and phenotype - lessons from a CAGI 6 challenger team

Vicente A. Yépez[1], Christian Mertes[1,2,4], Nicholas H. Smith[1], Ines F. Scheller[1,3], Julien Gagneur[1,2,3,4]

1. Department of Informatics, Technical University of Munich, Garching, Germany
2. Munich Data Science Institute, Technical University of Munich, Garching, Germany
3. Institute of Computational Biology, Helmholtz Zentrum München, Neuherberg, Germany
4. Institute of Human Genetics, School of Medicine, Technical University of Munich, Munich, Germany

RNA sequencing has emerged as a complementary tool to DNA sequencing for rare disease diagnostics. However, gene prioritization methods integrating genotype, RNA-seq and phenotypes have been lacking. To address this need, the SickKids Genome Clinic released a CAGI 6 diagnostics challenge with nearly 80 genomes and RNA-seq samples[1]. We developed a gene prioritization model integrating variant annotations, mono-allelic expression, gene expression[2] and splicing outliers[3] (through our workflow DROP[4]), together with HPO-encoded phenotypes. The model is a gradient boosting machine (implemented using XGboost) trained on a cohort of 209 mitochondrial disease patients[5] from which half are diagnosed. Our model prioritizes the causal gene first for almost half of the diagnosed cases, and among the top 5 in more than 70% of them. Application to the CAGI6 SickKids cohort revealed several promising candidates. Our approach and publicly available software[6] can help find and prioritize candidate genes found by DNA and RNA sequencing and can be especially useful to reduce the burden of manual inspection in cohorts of hundreds of samples.

References

1. http://genomeinterpretation.org/cagi6-sickkids.html
2. Brechtmann et al, AJHG (2018)
3. Mertes et al, Nat Commun (2021)
4. Yépez et al, Nat Protoc (2021)
5. Yépez, Gusic et al, Genome Med (2022)
6. https://github.com/gagneurlab/cagi6_sickkids

Performance of diagnostic methods in identifying disease-causing variants: assessment of the Rare Genomes Project CAGI challenge

Sarah L. Stenton[1,2], Stephanie DiTroia[1,2], Vijay S. Ganesh[2], Emily Groopman[1,2], Gabrielle Lemire[1,2], Emily O'Heir[1,2], Ikeoluwa Osei-Owusu[2], Lynn S. Pais[1,2], Grace E. VanNoy[2], Michael Wilson[2], Christina Austin-Tse[2,3], Melanie O'Leary[2], Heidi L. Rehm[2,3], Anne O'Donnell-Luria*[1,2]

1. Division of Genetics and Genomics, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA
2. Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA
3. Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA

odonnell@broadinstitute.org, sstenton@broadinstitute.org

One major obstacle facing rare disease patients is simply obtaining a genetic diagnosis. The average "diagnostic odyssey" lasts more than five years, and over 60% of patients still lack a genetic diagnosis. The Rare Genomes Project (RGP) is a direct-to-participant research study on the utility of genome sequencing for rare disease diagnosis and gene discovery, led by genomics experts and clinicians at the Broad Institute of MIT and Harvard. Research subjects are consented for genomic sequencing and the sharing of their sequence and phenotype information with researchers working to understand the molecular causes of rare disease. In the RGP CAGI challenge, whole genome sequence and phenotype data from 30 RGP families were provided, consisting of both "solved" and "unsolved" cases. The challenge tasked participants with identifying the causative variant(s) in as many cases as possible. Participants submitted ranked causal variant predictions (limit 100 per proband, single or biallelic) with associated estimated probability of causal relationship values. Sixteen teams participated in the challenge, submitting variant predictions from a total of 52 different models. Model performance was determined by two independent methods and the mean rank determined those that were top-performing. The first method calculated maximum F-measure based on the submitted estimated probability of causal relationship values, a harmonic mean between the precision and recall of causal variant(s) in the solved cases. The second method allocated points to the prioritization of causal variant(s) within the first five, 10, 20, 50, and 100 ranked variants submitted for each proband in a weighted manner (100, 50, 25, 10, and 5 points, respectively). The top performing teams were able to recall a significant fraction of causal variants (in up to 13/14 solved cases), while in the unsolved cases one *de novo* near splice variant was deemed diagnostic, two credible leads are undergoing functional validation, and six candidates are being pursued as potential novel disease genes by entry into MatchMaker Exchange. In one such example, RNA-sequencing is underway to confirm the functional consequence of a deep intronic indel in *ASNS*, identified in *trans* with a frameshift variant in an unsolved case with a strong phenotypic match to asparagine synthetase deficiency. The identification of further potentially diagnostic variants illustrates promotion of synergy between researchers with clinical and computational expertise as a means of advancing the field of clinical genome interpretation.

Suggesting diagnoses for RGP patients with the eVai MachineLearning approach

Susanna Zucca[1], Ivan Limongelli[1], Ettore Rizzo[1], Federica De Paoli[1], Giovanna Nicora[1], MariaGiulia Carta[2], Riccardo Bellazzi[1,2], Paolo Magni[1,2]

1 enGenome srl, via Ferrata 5, Pavia, Italy

2 BMS Lab, University of Pavia, via Ferrata 5, Pavia, Italy

Corresponding authors: szucca@engenome.com, ilimongelli@engenome.com

Background:
The eVai platform (www.engenome.com) enables precise and early diagnosis of rare diseasesand supports geneticists with interpretation of genomic variants.

By combining Artificial Intelligence and International Guidelines, eVai classifies and prioritizesvariants for pathogenicity, suggesting the related genetic diagnoses.

Its ML-based approach was applied to Rare Genome Project (RGP) patients provided by theCAGI6 Challenge.

Materials and methods:
The eVai ML approach to suggest diagnoses was adapted to CAGI RGP data.
Training and test VCF files were analyzed through eVai and dataset features considering variantpathogenicity, variant quality, family segregation and phenotypic similarity were computed for single or compound heterozygous variants.

Different ML models were evaluated with a "Leave-one-proband-out" cross-validation on trainingset.
Selected models were trained on the CAGI RGP training set and used to predict test set.

Results: Among different models, two of them were selected according to their prioritization performances. Both models prioritize the causative variant in the first position for more than 74% of cases. Moreover, in more than 97% of cases, the causative variant was in the top 10 listfor both models.

Discussion: The eVai ML approach to suggest diagnoses goes beyond pathogenicity-based variant classification and mimics the geneticist manual review of candidate variants, matching patient's clinical phenotypes, family history and checking experimental data quality.

Applying Exomiser to the CAGI 6 Rare Genomes Project challenge

Jules Jacobsen
Queen Mary College London, London, UK

The Exomiser is a free, open source Java application for filtering and prioritisation of variants likely to be causative of Mendelian rare disease. It has been developed as part of the Monarch Initiative (monarchinitiative.org) since 2012 and is widely used in diagnostic pipelines around the globe.

The Exomiser uses patient phenotypic features encoded using the Human Phenotype Ontology (HPO) and a Variant Call Format (VCF) file of the patients exome / genome. It applies a variety of 'fuzzy' phenotype profile matching algorithms to prioritise segregating and de novo filtered variants likely to be causative of the patient's phenotype. The patient phenotype profile matching is run over known human rare disease, mouse and zebrafish knockout models which enables prioritisation of variants both in known disease-causing genes and discovery of new gene-disease associations.

For the Rare Genomes Project challenge, we ran Exomiser in various configurations (models) of varying degrees of permissiveness to allow detection of variants in known disease genes, incompletely penetrant, animal model and non-coding regions. The two most precise models were able to detect the diagnosed variant in the top 1, 3, and 10 prioritised variants in 31 (89%), 33 (94%) and 35 (100%) of the Rare Genomes training data cases.

AI-based Identification of Diagnostic Variants from Genotype and Phenotype

Azza Althagafi[1,2,3], Marwa Abdelhakim[1], and Robert Hoehndorf[1,2,*]
[1] Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology (KAUST), Thuwal 23955, Saudi Arabia. [2] Computer, Electrical and Mathematical Sciences Engineering Division (CEMSE), KAUST, Thuwal, Saudi Arabia. [3] Computer Science Department, College of Computers and Information Technology, Taif University, Taif, Saudi Arabia

emails: azza.althagafi,marwa.abdelhakim,robert.hoehndorf@kaust.edu.sa

* Correspondence: robert.hoehndorf@kaust.edu.sa

Family-based Filtering

To choose the most suitable mode of inheritance for each case, we studied both the training set and the actual testing set for all possibilities while considering the ethnicity. Consequently, in some cases, we prioritized some mode of inheritance filters over others. We use ethnicity to prioritize a family filter based on recessive mode of inheritance when we suspect likely consanginuity (i.e., we use the amount of consanguinuity within an ethnic group as a prior when selecting the mode of inheritance filter to apply). For the family-based filtering, we utilized the recently published method Slivar [11]. The method explores practical guidelines for variant (SNPand INDEL) filtering and reports the expected number of candidates for de novo dominant, recessive, and autosomal dominant modes of inheritance. We evaluated different settings and configurations based on the family pedigrees. Using Slivar, for the trios or quads, and duos: we use segregating denovo, segregating recessive, and compound heterozygous compound-hets filtering. For the proband only cases: we used segregating dominant and segregating recessive filtering for the variants.

Causal variants prediction

After filtering variants, our approach for predicting the causative variant(s) is by combining two main sources of information; the first utilizes the genomic features and pathogenicity predictionusing CADD [12], and the second is based on the phenotype annotations for the affected families combined with the ontology-based machine learning method DL2vec [3]. We made four submissions based on the different gene-phenotype representations using DL2vec. We mainly utilize three types of gene annotation features for supervised learning as they perform best in our previous experiments [3]: Gene Ontology (GO) [2], Mammalian Phenotype Ontology (MP) [15] , and the Human Phenotype Ontology (HPO) [13]. Specifically, we obtain the annotations of human genes with functions and cellular locations encoded by the GO, and the phenotypes of their mouse orthologs from the Mouse Genome Informatics (MGI) database and characterized using the MP, and the phenotypes of the human genes using HPO.Furthermore, we obtain phenotype annotations of human diseases with the Human Phenotype Ontology (HPO), in addition to the phenotypes obtained from the training set. To combine the annotations using the different ontologies, we use the integrated PhenomeNET ontology [14].

We jointly embed the gene and disease, their ontology-based annotations, and the ontologies used in the annotations in a vector space. We generate embeddings individually using GO, MP, and HP annotations, and their union. We then use a pointwise learning-to-rank model to prioritize gene–disease pairs based on gene–disease associations in the Online Mendelian Inheritance in Men (OMIM) database [1], and the phenotypes in our training set. Our model is based on neural networks; given a pair of embedding vectors G and D as input, the model independently transforms the embeddings into a lower-dimensional representations using two fully-connected hidden layers, and then computes the inner product followed by a sigmoid function that outputs a value between 0 and 1, and which we use as the prediction score for an association between G and D. We combine DL2Vec predictions with CADD predictions, and use weighted prediction scores as the final predication score for the variants.

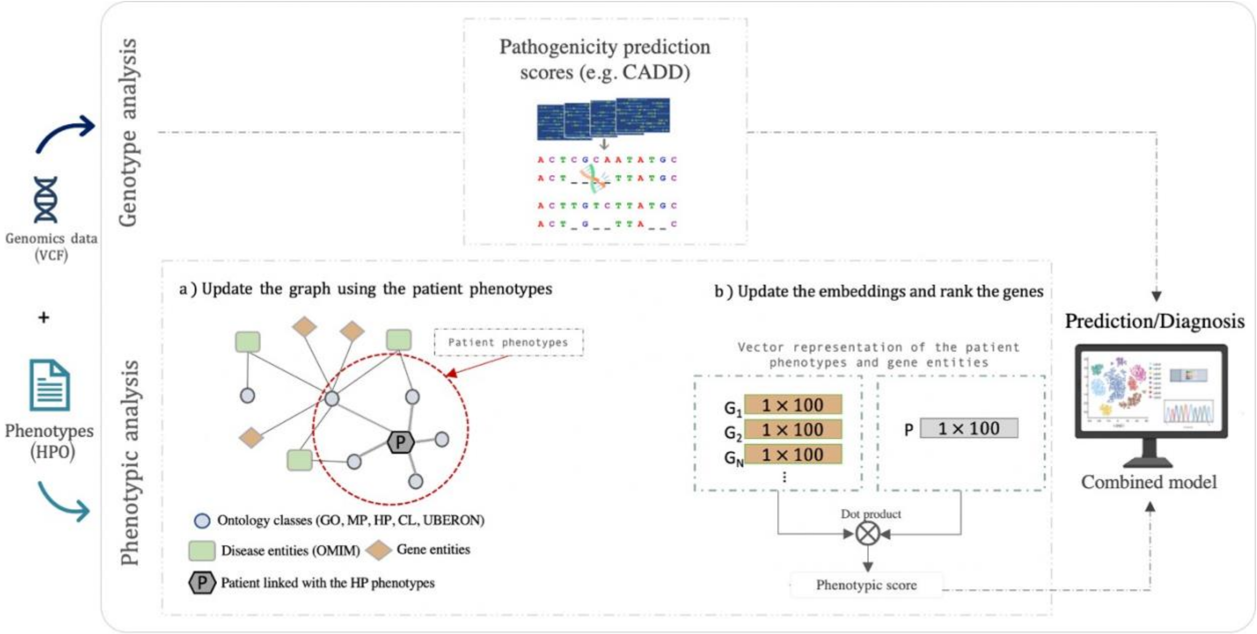

Figure 1: Model Workflow

References

1. Amberger, J., Bocchini, C., Hamosh, A.: A new face and new challenges for online mendelian inheritance in man (OMIM). Human mutation 32(5), 564–567 (2011)

2. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.: Gene ontology: tool for the unification of biology. Nature genetics 25(1), 25–29 (2000)

3. Chen, J., Althagafi, A., Hoehndorf, R.: Predicting candidate genes from phenotypes, functions and anatomical site of expression. Bioinformatics 37(6), 853–860 (2021)

4. Collins, R.L., , Brand, H., Karczewski, K.J., Zhao, X., Alf˙oldi, J., Francioli, L.C., Khera, A.V., Lowther, C., Gauthier, L.D., Wang, H., Watts, N.A., Solomonson, M., O'Donnell-Luria, A., Baumann, A., Munshi, R., Walker, M., Whelan, C.W., Huang, Y., Brookings, T., Sharpe, T., Stone, M.R., Valkanas, E., Fu, J., Tiao, G., Laricchia, K.M., Ruano-Rubio, V., Stevens, C., Gupta, N., Cusick, C., Margolin, L., Taylor, K.D., Lin, H.J., Rich, S.S., Post, W.S., Chen, Y.D.I., Rotter, J.I., Nusbaum, C., Philippakis, A., Lander, E., Gabriel, S., Neale, B.M., Kathiresan, S., Daly, M.J., Banks, E., MacArthur, D.G., and, M.E.T.: A structural variation reference for medical and population genetics. Nature 581(7809), 444–451 (May 2020). https://doi.org/10.1038/s41586-020-2287-8,

https://doi.org/10.1038/s41586-020-2287-8

5. Consortium, .G.P., et al.: An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422), 56 (2012)

6. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al.: The variant call format and vcftools. Bioinformatics 27(15), 2156–2158 (2011)

7. Karczewski, K.J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D.M., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K.E., Cummings, B.B., et al.: The exac browser: displaying reference data information from over 60 000 exomes. Nucleic acids research 45(D1), D840–D845 (2017)

8. K¨ohler, S., Carmody, L., Vasilevsky, N., Jacobsen, J.O.B., Danis, D., Gourdine, J.P., Gargano, M., Harris, N.L., Matentzoglu, N., McMurry, J.A., et al.: Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic acids research 47(D1), D1018– D1027 (2019)

9. McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., Cunningham, F.: Deriving the consequences of genomic variants with the ensembl api and snp effect predictor. Bioinformatics 26(16), 2069–2070 (2010)

10. Narasimhan, V., Danecek, P., Scally, A., Xue, Y., Tyler-Smith, C., Durbin, R.: Bcftools/roh: a hidden markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics 32(11), 1749–1751 (2016)

11. Pedersen, B.S., Brown, J.M., Dashnow, H., Wallace, A.D., Velinder, M., Tristani-Firouzi, M., Schiffman, J.D., Tvrdik, T., Mao, R., Best, D.H., et al.: Effective variant filtering and expected candidate variant yield in studies of rare human disease. NPJ Genomic Medicine 6(1), 1–8 (2021)

12. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J., Kircher, M.: Cadd: predicting the deleteriousness of variants throughout the human genome. Nucleic acids research 47(D1), D886–D894 (2019)

13. Robinson, P.N., K¨ohler, S., Bauer, S., Seelow, D., Horn, D., Mundlos, S.: The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. The American Journal of Human Genetics 83(5), 610–615 (2008)

14. Rodr´ıguez-Garc´ıa, M.A., Gkoutos, G.V., Schofield, P.N., Hoehndorf, R.: Integrating phenotype ontolo- ´ gies with PhenomeNET. Journal of biomedical semantics 8(1), 58 (2017)

15. Smith, C.L., Eppig, J.T.: The mammalian phenotype ontology: enabling robust annotation and comparative analysis. Wiley Interdisciplinary Reviews: Systems Biology and Medicine 1(3), 390–399 (2009)

16. Wang, K., Li, M., Hakonarson, H.: Annovar: functional annotation of genetic variants from highthroughput sequencing data. Nucleic acids research 38(16), e164–e164 (2010).

Prediction of Neurodevelopmental Disorders and Pathogenic Variants from Gene Panel Sequences

Yexian Zhang[1,2], Qi Li[1,2], Maggie Haitian Wang[1,2*]
[1] CUHK Shenzhen Research Institute, Shenzhen, China;[2] JC School of Public Health and Primary Care, Chinese University of Hong Kong, Hong Kong SAR, China.

Correspondence: maggiew@cuhk.edu.hk

Gene panel sequencing analysis is widely used for identifying causative genetic variations and the genetic diagnosis of inherited disease. In the Critical Assessment of Genome Interpretation 6 (CAGI6) Intellectual Disability panel challenge, we proposed a novel method for identifying pathogenic variants, and performing the polygenic risk scores (PRS) to predict disease phenotypes. The Ensembl Variant Effect Predictor (VEP) tool and the precalculated REVEL scores were used for annotating the raw VCF files. For variants reported with multiple REVEL scores, we selected the highest score. Variants were filtered based on the following four criteria: (1) absent or with a MAF < 5% in 1000 Genomes Project data, (2) not homozygous reference alleles, (3) present in only one sample, (4) protein-altering variants. The ClinVar databases, phenotype-specific Phenolyzer score, and REVEL score were used to prioritize variants. The variant reported as pathogenic/likely pathogenic in ClinVar or the top-ranked variant that combines the ranking of Phenolyzer score and REVEL score was identified as putative causative variants. The PRS is constructed using the effect sizes extracted from published GWAS summary statistics or derived from training dataset by logistic regression analysis.

Variant impact estimation using Evolutionary Action in the CAGI 6 Intellectual Disability Panel, SickKids6 and Rare Genomes Project challenges

Amanda Williams
Baylor College of Medicine

The difficulty of identifying causal variants makes diagnosis and treatment hard. Computational methods can prioritize variants and could provide doctors with diagnosis and treatment options. CAGI aims to assess computational methods ability to assist in clinical settings. We used variant impact scores and allele frequency to address the Intellectual Disability Panel, Rare Genomes Project, and Sickkids6 challenges in CAGI6. Variant impact was estimated using the Evolutionary Action (EA) method, and allele frequency was estimated according to GnomAD, the UK Biobank, the test data sets, and the training data sets. We then investigated genes for genotype-to-phenotype relationships according to ClinVar, DisGeNet, Human Phenotype Ontology (HPO), and GeneCards. For the Rare Genomes Project and Sickkids6 challenge, we prioritized variants through a combination of variant impact, allele frequency, and genotype-to-phenotype relationship. Using these features, we matched neurodevelopmental phenotypes to patients in the Intellectual Disability Panel. Overall, we indicated casual genes, driver variants, and matched phenotypes to patients.

Evaluating the impact of variants of unknown significance on splicing in CAGI6

Brynja Matthíasardóttir[1,2], Chiao-Feng Lin[3], and Stephen M. Mount[1]*

[1] University of Maryland, College Park, MD;

[2] National Human Genome Research Institute, NIH, Bethesda, MD;

[3] DNAnexus, Mountain View, CA;

Contact: brynmatt@umd.edu, chiaofeng.lin@gmail.com, smount@umd.edu*

Introduction: The degree to which genomic variants that affect splicing are responsible for disease-causing mutations remains unknown, and estimates of the fraction of disease causing mutations impacting splicing range from 10% to over 50%. A closely related problem is our ability to identify variants of unknown significance with impact on gene expression through splicing. Over the years, methods have evolved from consensus methods, information content, maximum entropy and hidden Markov models, to deep neural networks.

Methods: In the SplicingVUS challenge, participants were asked to predict splicing disruption from variants of unknown significance. We used direct application of the deep neural network SpliceAI, with and without supplementation by a detailed expert consideration of individual cases. For the first method, a threshold of 0.21 was selected based on the optimal accuracy of training data from the challenge (accuracy of 0.88). Variants with a maximum SpliceAI score of 0.21 or greater were assigned a classification of 1, indicating that splicing is altered relative to controls. The second method included a more in-depth multi-layered approach. Variant annotation was performed considering SpliceAI scores (threshold 0.21), gnomAD variant count, MaxEnt scores, 100 vertebrate conservation score (UCSC) and CADD scores. Factors considered during variant classification included the assessment of mutations in the 5' splice site core and mutations that create AG dinucleotides upstream of a 3' splice site. Missense mutations with a high CADD score were assumed to affect protein function rather than splicing.

Results, Conclusions and Next Steps: In the future, we will attempt to formalize domain knowledge by applying various machine learning methods to this data set and to additional data sets.

Assessing the CAGI 6 Sherloc challenge
Rachel Hovde, Peter Combs, Yuya Kobayashi

Invitae is a large-scale medical genetic testing company with clinical and sequence data from over 3 million patients. This rich dataset allows our internal researchers to interpret the clinical impact of previously unknown variants with high confidence using a semi-quantitative system called Sherloc. We regularly submit genetic variant interpretations to ClinVar, but this year, we held back a set of ~70 thousand newly-interpreted variants to use as a test set for the Sherloc Clinical Prediction Challenge. We invited participants to predict the clinical impact of each variant in the test set, as well as a score indicating their confidence in that prediction, and we evaluated contestants based on how closely they matched our own high-confidence interpretations.

Evaluation metrics were developed to assess clinical applicability of models, as well as scientific contribution. There was a broad range of performance among the participants, but several methods achieved high recall and precision, showing potential clinical applicability.   Gradient-boosted trees were the most popular method, and both winning teams used this algorithm.

Teams were given the opportunity to specialize in biologically meaningful subcategories of variants, and there were teams that seemed to have particular insight into missense variants and intronic inDels. In general, for variants that were hardest for teams to call , our prediction was based on private clinical patient data and other highly manual analyses.

We have been developing a number of internal software tools for securely sharing models and for retraining and evaluating them over time. We hope to share these in future CAGI competitions.

Prioritizing pathogenic variants using functional genomics data and DNA, protein sequence-derived features

Ken Chen[1], Maolin Ding[1], Huiying Zhao[2] and Yuedong Yang[1,3,*]
[1] School of Computer Science and Engineering, Sun Yat-sen University, 510000, Guangzhou, China,
[2] Sun Yat-sen Memorial Hospital, Sun Yat-sen University, 510000, Guangzhou, China,
[3] Key Laboratory of Machine Intelligence and Advanced Computing (Sun Yat-sen University), Ministry of Education, China.
* To whom correspondence should be addressed. Email: yangyd25@mail.sysu.edu.cn

Predicting the pathogenicity of genetic variants has always been a challenge in genomics. However, existing methods usually only focus on certain kinds of mutations or can not make accurate predictions at a genome-wide scale. Here, we reported a novel model, FASVAR, which could be applied to single nucleotide variants (SNVs) and short indels in coding and non-coding regions. FASVAR incorporates a variety of features, including gene annotation, splicing, conservation scores, transcription factor binding, chromatin accessibility, and histone marks. For coding variants, we additionally compiled protein features based on multiple sequence alignment (MSA) and predicted disorder. Considering that these features may have distinct contributions to prioritizing the variants of different types, we trained separate models for SNVs and indels in coding and non-coding regions. Then we built an ensemble model based on the separated models. To avoid overfitting, we split the training data into five folds by chromosomes to adopt the cross-validation strategy for training and fine-tuning the model. Our model outperforms state-of-the-art methods on independent test samples released by the Clinvar database recently. The analysis of feature importance suggests that MSA-derived features and RNA splicing features are the most important for coding and non-coding variants, respectively. Meanwhile, conservation scores and histone marks make critical contributions to the predictions for all kinds of variations. Notably, FASVAR significantly improves the predictions for non-coding SNVs, which may facilitate the study of non-coding regions.

Quantification of genotype-phenotype relationships in Pompe disease: a patient-derived model predicting age of onset, diseaseseverity and progression

Reet Mishra[1], Zhiqiang Hu[1], Dona Kanavy, Jennifer L. Goldstein, Deeksha S. Bali, Yuanbin Ru, G. Karen Yu, Jonathan H. LeBowitz, Wyatt T. Clark, Constantina Bakolitsa

Pompe disease (PD) is a rare autosomal recessive disorder caused by acid -glucosidase (GAA) pathogenic variants. Two different clinical forms have been described, depending on the age of onset: infantile (IOPD, <1 year of age) and late (LOPD, juvenile or adult). Early diagnosis of PD is critical, and initiation of enzyme replacement therapy can improve motor and respiratory function as well as survival. However, several factors can complicate and delay diagnosis, especially for LOPD, from PD's broad clinical spectrum and overlap with many other neuromuscular disorders, to variable diagnostic approaches in different countries, insufficient awareness of PD clinical manifestations, and a large number of GAA variants of unknown significance.

We generated the largest publicly available PD database, including the genotypes (two causal variants) and disease severity (infantile, juvenile, or adult) for 1,750 patients. Integrative analysisof variants observed in IOPD and LOPD patients and their gnomAD frequencies suggested that people with two less severe variants could be healthy. For example, most people with homozygous -32-13T>G variants, the top causal variants of PD, do not develop PD in their lifetime. This challenges the binary annotation of pathogenicity (pathogenic or benign) in popular variant impact databases, such as ClinVar or HGMD. We built a linear model with the GAA enzyme activity as an intermediate, accurately predicting the disease severity from a patient’s genotype. An independent test suggests the model can accurately distinguish between IOPD and LOPD patients with an AUC of 0.95. Our prediction provides finer gradations of variants’ pathogenicity. To extend our model to unobserved pathogenic variants, we experimentally measured the *in vitro* enzyme activities for 357 low-frequency GAA variants in ExAC database with unknown significance and applied a clinical evaluation to determine their pathogenicity. Ourobservations strongly suggest that PD does not follow a classic autosomal-recessive model. Wepropose a refined model integrating genotype pathogenicity and other genetic, environmental and age-associated modifiers.

Our analysis can improve future diagnostic screening for PD, and could be similarly applied to other monogenic diseases characterized by graded severity.

Acknowledgements

We would like to thank the CAGI community for helping us improve the sixth edition of CAGI challenges. Your feedback, at all stages of this experiment, made for better data quality and challenge design, thereby also improving assessment and presentation of results at the CAGI 6 conference. We hope this collaborative spirit continues and inspires a new generation of researchers in genome interpretation.

Funding for this conference was made possible (in part) by U24 HG007346 from the National Human Genome Research Institute. The CAGI experiment has been primarily supported by U24 HG007346 from the NHGRI. Support for CAGI has also come through a research agreement from TATA Consultancy Services.

The views expressed in written conference materials or publications and by speakers and moderators do not necessarily reflect the official policies of the Department of Health and Human Services; nor does mention by trade names, commercial practices, or organizations imply endorsement by the U.S. Government.