Human Mutation
Variation, Informatics, and Disease

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY

Assessing Computational Predictions of the Phenotypic Effect of Cystathionine-beta-Synthase Variants

Journal:	*Human Mutation*
Manuscript ID	Draft
Wiley - Manuscript type:	Special Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Kasak, Laura; University of California Berkeley, Department of Plant and Microbial Biology; University of Tartu, Institute of Biomedicine and Translational Medicine Bakolitsa, Constantina ; University of California Berkeley, Department of Plant and Microbial Biology Hu, Zhiqiang; University of California Berkeley, Department of Plant and Microbial Biology Rine, Jasper; University of California Berkeley, California Institute for Quantitative Biosciences Dimster-Denk, Dago; University of California Berkeley, California Institute for Quantitative Biosciences; Pionyr Immunotherapeutics Pandey, Gaurav; University of California Berkeley, Department of Plant and Microbial Biology; Icahn School of Medicine at Mount Sinai, Department of Genetics and Genomic Sciences and Icahn Institute for Data Science and Genomic Technology Bromberg, Yana; Rutgers University, Department of Biochemistry and Microbiology Cao, Chen; University of Maryland, Institute for Bioscience and Biotechnology Research; University of Maryland at College Park, Computational Biology, Bioinformatics and Genomics, Biological Sciences Graduate Program Capriotti, Emidio; Stanford University, Department of Bioengineering; University of Bologna, BioFolD Group, Department of Pharmacy and Biotechnology (FaBiT) Casadio, Rita; University of Bologna, Biocomputing Group, Department of Pharmacy and Biotechnology Giollo, Manuel; University of Padua, Department of Biomedical Sciences Katsonis, Panagiotis; Baylor College of Medicine, Department of Molecular and Human Genetics Leonardi, Emanuela; University of Padua, Department for Woman and Child Health; Pediatric Research Institute , Lichtarge, Olivier; Baylor College of Medicine, Department of Molecular and Human Genetics Martelli, Pier Luigi; University of Bologna, Department of Pharmacy and Biotechnology Mooney, Sean; Buck Institute for Research on Aging; University of Washington, Department of Biomedical Informatics and Medical Education Pal, Lipika; University of Maryland, Institute for Bioscience and

	Biotechnology Research Radivojac, Predrag; Indiana University Bloomington, School of Informatics and Computing; Northeastern University, Khoury College of Computer Sciences SAVOJARDO, CASTRENSE; University of Bologna, Department of Pharmacy and Biotechnology Thusberg, Janita; Buck Institute for Research on Aging; Invitae Tosatto, Silvio; University of Padua, Department of Biomedical Sciences Vihinen, Mauno; University of Tampere, Institute of Medical Technology; Lund University, Department of Experimental Medical Science Väliaho, Jouni; University of Tampere, Institute of Medical Technology Repo, Susanna; University of California Berkeley, Department of Plant and Microbial Biology; Wellcome Genome Campus, ELIXIR Moult, John; University of Maryland, Institute for Bioscience and Biotechnology Research; University of Maryland at College Park, Department of Cell Biology and Molecular Genetics Brenner, Steven; University of California Berkeley, Department of Plant and Microbial Biology Friedberg, Iddo; Miami University, Department of Microbiology; Iowa State University, Department of Veterinary Microbiology and Preventive Medicine
Key Words:	CAGI challenge, cystathionine-beta-synthase, single amino acid substitution, critical assessment, phenotype prediction, machine learning, homocystinuria

Assessing Computational Predictions of the Phenotypic Effect of Cystathionine-beta-Synthase Variants

Laura Kasak[1,2], Constantina Bakolitsa[1], Zhiqiang Hu[1], Jasper Rine[3], Dago F. Dimster-Denk[3,*], Gaurav Pandey[1,*], Yana Bromberg[4], Chen Cao[5,6], Emidio Capriotti[7,*], Rita Casadio[8], Manuel Giollo[9], Panagiotis Katsonis[10], Emanuela Leonardi[11,*], Oliver Lichtarge[10], Pier Luigi Martelli[8], Sean D. Mooney[12,*], Lipika R. Pal[5], Predrag Radivojac[13], Castrense Savojardo[8], Janita Thusberg[12,*], Silvio C.E. Tosatto[9], Mauno Vihinen[14,*], Jouni Väliaho[14], Susanna Repo[1,*], John Moult[5,15], Steven E. Brenner[1], Iddo Friedberg[16,17,*]

[1]Department of Plant and Microbial Biology, University of California, Berkeley, CA, USA

[2]Institute of Biomedicine and Translational Medicine, University of Tartu, Tartu, Estonia

[3]California Institute for Quantitative Biosciences, University of California, Berkeley, CA, USA

[4]Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ, USA

[5]Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, MD, USA

[6]Computational Biology, Bioinformatics and Genomics, Biological Sciences Graduate Program, University of Maryland, College Park, MD, USA

[7]Department of Bioengineering, Stanford University, Stanford, CA, USA

[8]Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy

[9]Department of Biomedical Sciences, University of Padua, Padua, Italy

[10]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

[11]Department for Woman and Child Health, University of Padua, Italy

[12]Buck Institute for Research on Aging, Novato, CA, USA

[13]School of Informatics and Computing, Indiana University, Bloomington, IN, USA

[14]Institute of Medical Technology, University of Tampere, Tampere, Finland

[15]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD, USA

[16]Department of Microbiology, Miami University, Oxford, OH, USA

[17]Department of Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, IA USA

Correspondence

Iddo Friedberg, Department of Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, IA USA

Email: idoerg@iastate.edu

[*]Present address: Dago F. Dimster-Denk, Pionyr Immunotherapeutics, San Francisco, CA, USA; Gaurav Pandey, Department of Genetics and Genomic Sciences and Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, NY, USA; Emidio Capriotti, BioFolD Group, Department of Pharmacy and Biotechnology (FaBiT), University of Bologna, Bologna, Italy; Emanuela Leonardi, Department for Woman and Child Health, University of Padua, Italy; Pediatric Research Institute, Citta della Speranza, Padua, Italy; Sean D. Mooney, Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, USA; Predrag Radivojac, Khoury College of Computer Sciences, Northeastern University, Boston MA, USA; Mauno Vihinen, Department of Experimental Medical Science, Lund University, Lund, Sweden; Janita Thusberg, Invitae, San Francisco, CA, USA; Susanna Repo, ELIXIR, Wellcome Genome Campus, Hinxton, Cambridge, UK; Iddo Friedberg, Department of Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, IA USA

Contract grant sponsor: NIH (U41 HG007346, R13 HG006650, GM120374, GM079656, GM066099, R01GM114434, 1 U01 GM115486 01); Ministry of Education, Universities and Research (FFABR grant); Italian Ministry of Health (GR-2011-02347754); Fondazione Istituto di Ricerca Pediatrica - Città della Speranza (18-04); NSF (ABI 1458359); NLM (R01LM009722); Marie Curie International Outgoing Fellowship (PIOF-GA-2009-237751).

ORCID IDs: Laura Kasak, 0000-0003-4182-2396; Constantina Bakolitsa, 0000-0002-6980-9831; Zhiqiang Hu, 0000-0001-8854-3410; Emidio Capriotti, 0000-0002-2323-0963; Panagiotis Katsonis, 0000-0002-7172-1644; Emanuela Leonardi, 0000-0001-8486-8461; Olivier Lichtarge, 0000-0003-4057-7122; Sean D. Mooney, 0000-0003-2654-0833; John Moult, 0000-0002-3012-2282; Lipika R. Pal, 0000-0002-3390-110X; Gaurav Pandey, 0000-0003-1939-679X; Susanna Repo, 0000-0003-3488-4767; Jasper Rine, 0000-0003-2297-9814; Janita Thusberg, 0000-0001-8028-4426; Silvio C.E. Tosatto, 0000-0003-4525-7793; Mauno Vihinen, 0000-0002-9614-7976; Steven E. Brenner, 0000-0001-7559-6185; Iddo Friedberg, 0000-0002-1789-8000

Abstract

Accurate prediction of the impact of genomic variation on phenotype is a major goal of computational biology and an important contributor to personalized medicine. Computational predictions can lead to a better understanding of the mechanisms underlying genetic diseases, including cancer, but their adoption requires thorough and unbiased assessment. Cystathionine-beta-synthase (CBS) is an enzyme that catalyzes the first step of the transsulfuration pathway, from homocysteine to cystathionine, and in which variations are associated with human hyperhomocysteinemia and homocystinuria. We have created a computational challenge under the CAGI framework to evaluate how well different methods can predict the phenotypic effect(s) of CBS single amino acid substitutions using a blinded experimental data set. CAGI participants were asked to predict yeast growth based on the identity of the mutations. The performance of the methods was evaluated using several metrics. The CBS challenge highlighted the difficulty of predicting the phenotype of an *ex vivo* system in a model organism when classification models were trained on human disease data. We also discuss the variations in difficulty of prediction for known benign and deleterious variants, as well as identify methodological and experimental constraints with lessons to be learned for future challenges.

Keywords: CAGI challenge, cystathionine-beta-synthase, single amino acid substitution, critical assessment, phenotype prediction, machine learning

INTRODUCTION

One of the central challenges in biology is to determine the impact of genomic variation on the phenotype(s) of an organism. As the amount of genomic data increases and accumulates at an exponential rate, comprehensive and accurate prediction algorithms are needed when biological experiments are expensive or difficult to execute (Fernald, Capriotti, Daneshjou, Karczewski, & Altman, 2011). Missense mutations are the most studied class of protein-altering variants; however, even today the algorithms often disagree on the pathogenicity prediction of the variants (Ioannidis et al., 2016). To determine the optimal use of each algorithm in different tasks, a thorough and unbiased methodological assessment is required. The ultimate aim is to attain a better understanding of the complex genotype-phenotype relationship, and, most importantly, provide the basis for clinical application to improve human health (Rost, Radivojac, & Bromberg, 2016).

Since 2010, the Critical Assessment of Genome Interpretation (CAGI) experiment has been evaluating bioinformatics tools developed for phenotype prediction from genomic variation data (Hoskins et al., 2017). Here, we present an objective assessment of predictions on the effects of single amino acid substitutions in the function of cystathionine-beta-synthase (*CBS*, MIM# 613381). CBS is an extensively studied vitamin-dependent enzyme involved in cysteine biosynthesis via the transsulfuration pathway. The molecular architecture of human CBS comprises an N-terminal heme-binding domain (residues 1-70), followed by the catalytic domain (residues 71-381) and a C-terminal regulatory domain (residues 412-551) (Majtan et al., 2018). The heme domain, which is found only in mammalian forms of CBS, lacks any significant structural elements and exhibits significant sequence diversity. Changes in the heme's coordination environment can be transmitted to the active site ~20 Å away, leading to inhibition of CBS activity (Weeks, Singh, Madzelan, Banerjee, & Spiro, 2009). The central domain belongs to the family of pyridoxal-5'-phosphate (PLP)-dependent enzymes, with the PLP cofactor bound via a Schiff base to K119 in the CBS active site. The C-terminal domain, also known as the Bateman module, contains two consecutive CBS-motifs (residues 412-471 and 477-551) that form distinct binding sites for S-adenosyl-methionine (AdoMet) and enable CBS homotetramerization. Two high-affinity and four low-affinity AdoMet binding sites have been identified per CBS tetramer, with distinct roles proposed in the regulation and activation (Pey, Majtan, Sanchez-Ruiz, & Kraus, 2013). Catalytic and regulatory domains are joined by a flexible linker (residues 382-411) that is sensitive to proteolysis. Targeted proteolysis of CBS results in a truncated, dimeric and more active form of the enzyme, adding another layer of CBS regulation (Skovby, Kraus, & Rosenberg, 1984; Zou & Banerjee, 2003).

Homocystinuria due to CBS deficiency (MIM# 236200) is an autosomal recessive inborn error of sulfur amino acid metabolism, characterized by increased levels of homocysteine in the urine (Mudd, Levy, & Kraus, 2001). The estimated worldwide prevalence of homocystinuria is approximately 1 in 100,000 (Moorthie, Cameron, Sagoo, Bonham, & Burton, 2014). More than 160 different disease-associated variants have been identified in the *CBS* gene (http://cbs.lf1.cuni.cz/index.php). The majority of these are substitutions that do not involve catalytic residues, suggesting that their effect resides in structural or conformational perturbations leading to a misfolded protein (Majtan et al., 2018). About one-half of homocystinuric patients respond to high doses of pyridoxine, the soluble form of PLP (Mudd et al., 2001) and several mutations are clearly pyridoxine remediable (B6-responsive homocystinuria): A114V, R266K, R369H, K384E, L539S, and the most common substitution I278T, which accounts for ~20% of all homocystinuric alleles (Dimster-Denk, Tripp, Marini, Marqusee, & Rine, 2013; Moat et al., 2004; Skovby, Gaustadnes, & Mudd, 2010).

Since CBS is well studied and its *ex-vivo* mutation effects are easily quantified, it is a tractable system for investigating phenotype - genotype relationships, making it an attractive target for the CAGI challenges. CAGI1 (2010) included the first CBS challenge, where participants were asked to predict yeast growth rates when compared with wild-type yeast based on amino acid substitution information. This dataset involved 51 synthetic single amino acid substitutions within the human CBS coding region. The second CBS challenge (part of CAGI 2 in 2011) concerned the properties of a larger set of variants (78 amino acid substitutions) that had been observed in patients with homocystinuria. For both challenges, participants were asked to submit predictions on the effect of the variants in the function of CBS at both high (400 ng/ml) and low (2 ng/ml) cofactor (pyridoxine) concentration. Altogether, 38 predictions from CAGI1 and CAGI2 were assessed. Methods employed varied from the purely structural to ones combining both structural and sequence conservation information or annotation, and from meta-predictors (models that use the predictions of other methods as features) to methods driven mainly by sequence, while one submission was a set of random predictions.

In general, deleterious mutations were better predicted than variants with minor or no effects on phenotype, with the hardest to predict effects often involving variants with weak sequence and structural signals. The use of distinct assessment criteria revealed differences in performance between methods, with methods integrating sequence, structure and functional features performing best overall. To improve the predictive potential of these types of studies, both experimental and computational methods need to be better tailored to the biology under investigation.

METHODS

Dataset

The CAGI1 CBS challenge included 51 synthetic single amino acid substitutions within the human CBS (Table 1) (Dimster-Denk et al., 2013), while the CAGI2 CBS challenge comprised 78 single amino acid variants that had been observed in patients with homocystinuria (Mayfield et al., 2012). The variant nomenclature refers to the human CBS cDNA (mRNA reference sequence NM_000071, protein reference sequence NP_000062).

The functionality of the variants was tested in an *in vivo* yeast complementation assay, where the human *CBS* allele is expressed in yeast and functionally complements the yeast ortholog, *CYS4*, which was removed from the chromosome. In this assay, growth is dependent upon the level of mutant human CBS function, and growth rates are expressed as a percentage relative to wild type grown with the same amount of exogenous pyridoxine supplementation. An experimental standard deviation is also available based on 3-4 repeated assays. The assay was performed in the presence of high (400 ng/ml) and low (2 ng/ml) pyridoxine concentrations. For a detailed description of this assay, see Mayfield et al. (2012) and Dimster-Denk et al. (2013). Participants were asked to submit predictions on the effect of the variants on the function of CBS both in high and low pyridoxine concentrations. The submitted prediction was requested as a numeric value, i.e., percent of growth when compared with wild-type yeast, with a standard deviation. The predictions were assessed against the percent of growth values actually measured for each substitution in the yeast assay.

Prediction assessment

The correlation between the predicted effect of the mutations and the actual effects serves as an initially simple but powerful measure to assess the accuracy of the prediction methods. Because the mutation data are not derived from a normal distribution, nonparametric tests were used to assess the methods. Both Spearman's rank correlation and Kendall's Tau correlation (KCC) were used to assess each algorithm's predictions against the observed growth rates. The root-mean-square deviation (RMSD) was also calculated to estimate the difference between experimental and predicted values. In order to assess the accuracy of the algorithms in a clinical sense, evaluation was also conducted against a binary version of the experimental growth rates. A threshold of 0 was chosen for the binary version and performance was evaluated in terms of area under the ROC curve (AUC), sensitivity, specificity, accuracy, and positive/negative predictive value (Lever, Krzywinski, & Altman, 2016). For experimental data, the total number of negatives (no growth substitutions) were defined as N=TN+FP and the total number of experimental positives (growth detected) as P=TP+FN. All the performance indices are shown in Supp. Table 1. An overall ranking of the methods was defined as a mean ranking of three different measures (KCC, AUC, and RMSD) shown in Supp. Table 2. These three measures were chosen since they assess complementary aspects of prediction performance.

Characterization of easy and hard to predict variants

We examined the mutation effects that were easiest and hardest to predict in order to determine whether they shared any common features. We first identified these mutations by summing the binary predictions (0 for no growth, >0 for growth) at low pyridoxine conditions across all methods for each variant. The variants with the lowest and highest summed scores were individually examined in terms of sequence, solvent-accessibility and location within the CBS structure.

To determine the likelihood of a variant being benign or deleterious through sequence analysis, scores for amino acid substitutions were taken from the BLOSUM90 substitution matrix (Henikoff & Henikoff, 1992). Scores of -1, 0 or 1 were classified as moderate substitutions, i.e. substitutions with a likelihood of arising by chance in terms of evolution and therefore of unknown effect on CBS function. Scores >1 were classified as conservative substitutions with a projected benign effect on CBS, while scores <-1 were classified as non-conservative, indicating a potentially deleterious effect on CBS function. Solvent accessible surface area (SASA) was calculated for the human CBS monomer (PDB id 4COO) using GetArea (Fraczkiewicz & Braun, 1998) and when different, dimer SASA results were noted. Secondary structure assignments and analysis were according to PDB id 4COO and visual inspection of the structure.

Method uniqueness in prediction results

For CAGI2, evaluation of the specific contribution of each prediction to the variance with experimental results was addressed using a multiple linear regression model. First, a multiple linear regression model was built with the best methods from each group. The top method from each group was chosen based on the highest adjusted R^2 values of every single method, to exclude predictions using modified versions of the same methods. The final methods included in the model were SID#16, 23, 26, 27, 29, 34, 36, and 41. Subsequently, methods were removed one at a time, and the linear regression equation was recalculated. The contribution of each method to the model was estimated from the delta adjusted R^2 values. SID#25 was excluded from the model as it lacked predictions for 10 substitutions.

RESULTS

CAGI1 challenge

In the CAGI1 CBS challenge, participants were asked to submit predictions to assess the impact of 51 single amino acid substitutions upon the function of the human CBS enzyme in both high (400 ng/ml) and low (2 ng/ml) pyridoxine concentrations. The function of the variants had been experimentally tested in an *in vivo* yeast complementation assay (Dimster-Denk et al., 2013). Twenty predictions from 13 groups were submitted to this challenge (Table 2), which were assessed blindly. A summary of each method is described in Supporting Information. Of the 13 participating groups, nine submitted one prediction, two contributed two, one submitted three and one provided four different submissions. Some methods used sequence-only or structure-only information, some employed meta-predictors, and others combined sequence, structural and annotation data. SID#17 (submission identification number) submitted only raw data without predictions and was excluded from the assessment. Most participants (18/19) provided predictions for both high and low pyridoxine concentration; however, seven predictions did not distinguish between the different cofactor levels. The majority of the predictions did not include standard deviations (13/19), and most of the methods that included estimates of reliability for each prediction appeared to be arbitrary (constant values like σ=5, 10, 15; n=5/7).

1. Assessment across different performance metrics

No single evaluation measure can capture a method's performance; thus, various measures were used to assess the phenotype prediction programs, including Kendall's Tau correlation coefficient (KCC), precision/recall, accuracy, and root-mean-square deviation (RMSD), *inter alia* (Supp. Table 1). For KCC, most of the prediction methods display statistically significant correlation with experimental data at both pyridoxine concentrations (Figure 1). Methods SID#2, SID#5 and SID#20 showed strong correlation at both high and low cofactor concentrations. SID#7 was the second best at high pyridoxine concentration (τ=0.50, p=4.87x10^{-6}); however, it showed little correlation at low cofactor concentration (τ=0.25, p=0.034). For RMSD, SID#5, which is a meta-predictor, was the best and second best at low and high cofactor concentration, respectively. The highest accuracy of 82.4% was achieved by three methods SID#1, 6 and 20 (Supp. Table 1). The first two combined sequence and structure information, while SID#20 is a meta-predictor, integrating several prediction methods. SID #2, 3, 9 and 11 achieved 100% sensitivity, whereas other methods had the highest specificity (94-100%, SID#7, SID#14) at both cofactor concentrations. The most sensitive models were mostly structure-based. The majority of the methods had good results for PPV, where the values varied from 65-100%; however, for NPV, the values displayed a much wider range (0-90%), meaning that the methods are better at predicting benign than deleterious variants.

2. Overall ranking

In order to carry out an overall performance assessment, ranks of the prediction methods based on KCC, AUC and RMSD were averaged to obtain the overall ranks of the methods (Supp. Table 2). This revealed SID#2 as the best performing method, having a mean rank of 2.2 across all measures, with SID#5 close behind (mean rank of 2.3). The first method is based on sequence information integrated with functional and structural annotations, while the other is a meta-predictor. One of the methods that did not perform as well, SID#16, was biased toward the prediction of low growth variants, whereas SID#13 was more conservative, with moderate to high growth predicted for most of the substitutions. More than half of the predictions actually did better than the baseline method, SIFT (SID#15), which ranked 11th overall.

3. Easy and hard to predict variants

We examined variants that were the easiest or hardest to predict based on the consensus output of all methods to determine whether they shared any common features (Figure 2B, Supp. Fig. 1). At low cofactor concentration, there were overall 12% (2/17) deleterious variants and 18% (6/34) benign variants whose effects were predicted incorrectly by more than half of the methods, our definition of *consensus*.

The deleterious mutations that were easiest to predict at a low cofactor concentration were L154P, N228K, G258R and G457E. The majority of these are non-conservative substitutions and are located within helices in the CBS structure (Supp. Table 3). The easiest to predict benign variants (low pyridoxine) were K271E, V356A and T383S, with moderate conservation scores and were again mostly located within helices. Among the better predicted benign variants, the majority were partially or fully exposed to the solvent.

The hardest to predict deleterious variants at both high and low pyridoxine concentrations were H65L and F385Q. H65 is located in the H1-H2 loop and axially coordinates the iron atom on one side of the heme plane, with C52 on the other, and mutation of either of these residues results in low catalytic activity (Ojha, Wu, LoBrutto, & Banerjee, 2002). Interestingly, the functional impact of these variants was not easy to assess from sequence and structure comparisons. Although H65 and the sequence flanking it are locally conserved, the heme-binding domain itself (comprising approximately the first 70 N-terminal residues), with the exception of a short 5-residue helix, has no secondary structure elements and does not resemble other heme proteins in either primary sequence or tertiary structure (Kumar et al., 2018). F385Q is located in the H17-H18 loop that forms part of the linker connecting the N-terminal catalytic domain with the C-terminal regulatory domain, and lies within an aromatic cluster of residues Y381, F332, F334, F385, W390, F396 enclosed by salt bridges R336-D388, K394-E302, and K384-E302 connecting helices H12-H14, H17, and H18. Erroneous coordination between aromatic residues can disrupt the extended **π-π** networks formed by aromatic clusters, thereby affecting protein stability and folding (Madhusudan Makwana & Mahalakshmi, 2015). Additionally, both these variants involve non-conservative substitutions, and thus could be expected to have a deleterious effect on CBS function.

The hardest to predict benign variants (low pyridoxine) were surprising in that they involved non-conserved substitutions, so they could be expected to disrupt CBS function. Additionally, within the structure, some were implicated in functionally relevant regions of CBS, such as the dimer interface (L345P) and the active site (V118G, adjacent to the PLP-ligating K119). All inaccurate consensus predictions of benign variants at low pyridoxine were for variants that confer sensitivity to reduced pyridoxine levels relative to the major allele (Dimster-Denk et al., 2013). The methods that correctly predicted all of these variants (SID#2, SID#3 and SID#9) displayed a broad spread both in features used (sequence, structure and thermodynamics respectively) and in overall performance (Table 2). Additionally, SID#2 and SID#9 did not distinguish between high and low pyridoxine.

CAGI2 challenge

In the CAGI2 CBS challenge, 84 single amino acid variants that had been observed in patients with homocystinuria were collected and functionally tested in an *in vivo* yeast complementation assay (Mayfield et al., 2012). 78 had experimental values for both pyridoxine concentrations; 6 'hem1 rescue' variants were left out from the assessment due to absent/conflicting data (Table 1). Participants were again asked to submit predictions of the effect of the variants on the function of CBS both in high and low cofactor concentration. This challenge attracted 20 submissions from 9 groups (Table 2) that were assessed without knowledge of the identity of the predictors. An overview of the methods is provided in Supporting Information. Four groups submitted one submission each, three groups submitted two each, and one group each contributed four and six predictions respectively. Four groups participated in both CAGI1 and CAGI2 CBS challenges. As in CAGI1, features used to generate the predictions ranged from sequence- or structure-only information to meta-predictors and methods combining sequence, structural and functional annotation data. SID#31 was excluded from the assessment due to its constant growth rate prediction of 100 for all substitutions. Almost all groups (17/19) provided distinct values for high/low cofactor concentrations. For this challenge, most submitters also provided standard deviations (13/19). Only one of the methods had arbitrary standard deviation values for all predictions (SID#26, σ=10). In addition to prediction programs, reference results were obtained by submitting the mutations to the SNAP (SID#50) and SIFT (SID#51) public servers (Bromberg, Yachdav, & Rost, 2008; P. Kumar, Henikoff, & Ng, 2009).

1. Assessment across different performance metrics

The same assessment measures as in CAGI1 were used in CAGI2 (Methods). Looking at KCC, over half of the predictions were highly significant (Figure 3); however, even the best deviated substantially from experimental values. At both high and low pyridoxine concentrations, methods SID#16 and SID#26 showed the strongest correlation with the experimental data and had also high AUC values (Figure 3). The latter was also the top predictor in terms of RMSD. In terms of accuracy, a structure-based method (SID#25) had the highest value (72%) at high and low pyridoxine concentration (Supp. Table 2). At both cofactor concentrations, methods SID#16, 26, 34, and 41 had the highest sensitivity (100%). Most of these methods employed integrated sequence and structure information. SID#23 was the top method (high pyridoxine) with a 75% specificity, SID#27 and SID#28 scored 83% (low pyridoxine). SID#23 used structural data, whereas the other two are meta-predictors. In contrast to the CAGI1 CBS challenge, NPV showed higher median values than PPV (65 vs 57%), implying that the probability of loss of function was slightly better predicted than the probability of having no or minor effects on the phenotype.

2. Overall ranking

As in the CAGI1 CBS challenge, we computed the overall ranks representing the performance of the submissions. Based on this criterion, the top methods in the CAGI2 CBS challenge were SID#26 and SID#16 with overall ranks of 1.8 and 2.3 respectively. Both methods utilized combined evolutionary information and structural features and ranked higher than the best baseline method (SID#50). Almost all of the methods performed better than the random predictor (SID#24), only SID#32 ranked even worse according to all assessment measures at both cofactor concentrations (Supp. Table 2). SID#32 is a random forest classifier that predicted no growth or growth (only values 0 or 100 for both concentrations). Two of the baseline methods, SNAP and SIFT (SID#50 and SID#51, respectively), ranked overall third and sixth, respectively.

3. Easy and hard to predict variants

The hardest and easiest to predict variants (Figure 2C, Supp. Fig. 2) showed similar trends to those observed in the CAGI1 CBS challenge. As observed before, the inaccurately predicted benign variants (low pyridoxine) involved non-conservative substitutions and were located in loop regions of the CBS structure (Supp. Table 3). The most accurately predicted deleterious variants (low pyridoxine) involved a majority of non-conservative mutations of residues located within stable secondary structure elements (helices), while the hardest to predict deleterious predictions involved residues with moderate conservation scores mostly located within loops.

Within this last group, a number of mutants have been indirectly implicated in CBS function through involvement in homooligomerization, redox sensing and regulation. A355 lies within helix H15, which is in turn sandwiched between H4 and strand beta3 at the CBS homodimerization interface. By introducing a kink in H15, the A355P mutant could potentially disrupt the folding in this region of the protein thereby impacting CBS function. Similarly, V168 is positioned at the homodimer interface while M391 is located within helix H18, a region of putative involvement in CBS homotetramer formation (Ereno-Orbea, Majtan, Oyenarte, Kraus, & Martinez-Cruz, 2013). A288 packs against W323 on strand beta6, and next to it, F324 packs against A360 in helix H15. A361 lies within interacting distance of C370, a residue that has been implicated in homocystinuria (Kraus et al., 1999) and proposed to modulate CBS function through interaction with an endogenous regulator such as nitric oxide (Eto & Kimura, 2002). The A361T mutant could therefore potentially interfere with a functionally relevant modification (e.g. S-nitrosylation) of C370. Similarly, modification of A288 could disrupt the pairing or orientation between beta5 and beta6, thereby potentially impacting the 272-CxxC-275 oxygen sensing motif of CBS, a redox active disulfide bond that allosterically controls CBS activity (Niu et al., 2018).

The most inaccurately predicted deleterious mutant, E302K, lies within interacting distance of one of the two active site loops (situated between helices H6 and H7). Recent studies have highlighted the importance of conformational flexibility of the loops defining the entrance to the catalytic site (Majtan et al., 2018).

4. Correlation between methods and unique contribution of different methods

To have a better understanding of the strengths and weaknesses of the different methods, we investigated the correlation between their predictions (Figure 4). Correlation heatmaps for high and low pyridoxine concentration had negligible differences. The strongest predictor for method correlation appeared to be the relation to a single group (Table 2). For example, SID#27-28, SID#33&41 and all (SID#19-22) except one method (SID#23) were highly correlated among each other. However, SID#29-32 and 35-36 had higher correlations with other methods than among their own group. SID#29-32 were the only predictors that used simply two states (growth rates of 0 or 100). Interestingly, two of the best ranking methods (SID#16 and SID#26) were strongly correlated. In addition, SID# 27, 28 and 34 showed a strong correlation with the top prediction methods, although were based on different features.

Baseline methods SNAP (SID#50) and SIFT (SID#51) showed strong correlation with the best performing predictions, which is partly expected as SID#16 was based on a version of the SNAP algorithm.

In order to assess the specific contribution of each method to the variance with experimental results in CAGI2, we applied a multiple linear regression model as described in Methods above. For high pyridoxine concentration, this revealed SID#16 and SID#36 as the most significant contributors (Δ adjusted R^2 values of 0.053 and 0.041, respectively). At the same time, for low pyridoxine concentration, SID#36 and SID#27 contributed the most (0.054 and 0.053, respectively). SID#36 is based on protein structure, sequence homology and included functional information, whereas SID#16 combined evolutionary information with structural features, and SID#27 is a meta-predictor (Figure 5).

DISCUSSION

Prediction features in relation to performance

In terms of prediction features, different methods performed well in distinct assessment measures. We observed that methods integrating sequence and structural information performed the best overall, ranking first or second (SID#2 in CAGI1, SID#16 and SID#26 in CAGI2). Methods that used only structural information (SID#9 and 11 in CAGI1, SID#19-20 in CAGI2) did not perform as well as those combining additional features. However, in terms of individual evaluation metrics, a structure-based method showed the highest accuracy of 72% (SID#25) in CAGI2. In CAGI1, SID#1 and 20 were the best at both cofactor concentrations, reaching accuracy of 82%. The first one combined structure and sequence data while SID#20 is a meta-predictor. In terms of sensitivity, most of the top-performing methods were structure-based in CAGI1 (SID#3, SID#9, and SID#11), while in CAGI2, the most sensitive algorithms combined structure with sequence data (SID#16, SID#26, SID#34). At the same time, for specificity, almost all of the top methods were meta-predictors in CAGI2 (SID#27-28). These observations suggest that combining different features and methods would yield the best results, as has been indicated previously (Grimm et al., 2015; Tang & Thomas, 2016). Some methods are tailored to predict whether a variant affects the function of the protein in hand and others are optimized to determine whether a variant is pathogenic or benign in the clinical sense (Grimm et al., 2015; Katsonis et al., 2014; Pejaver, Mooney, & Radivojac, 2017).

The importance of integrating information from different sources is reflected in the most inaccurately predicted mutants that tended towards non-conserved substitutions, structural uncertainty, or both. The power of combining structural, sequence and functional information was visible in CAGI2, where the overall performance of a structure and sequence combined method (SID#35) was improved significantly (by five ranks) with the inclusion of functional annotation data (SID#36). The latter was also the method that uniquely contributed the most predictive power of all methods at low pyridoxine concentration. Another structure-based method (SID#25) that incorporated functional information (trained on the CAGI1 dataset) also performed strongly. Methods trained on HGMD (Human Gene Mutation Database) mutations (SID#35-36) would be expected to perform well (Dong et al., 2015; Ioannidis et al., 2016; Pejaver et al., 2017). Interestingly, however, the best methods in CAGI2 CBS challenge showed variable training data, from no training to training on PMD (the Protein Mutant Database), HGMD, and CAGI1 CBS variant data (Supporting Information).

Limitations of the challenges

In terms of methodological limitations, most methods were developed to predict pathogenicity in humans or enzyme activity, not yeast growth or the effect of cofactor concentration on growth rate, something that could at least in part explain the difficulties they encountered in identifying the remediable class of variants (Supporting Information). So, while a meaningful distinction could be made in these challenges between growth and no growth under low pyridoxine conditions, this was not the case for distinguishing between rescue (high pyridoxine) and no rescue (low pyridoxine) variants. Similarly, only qualitative comparisons could be made between CAGI1 and CAGI2, since the datasets differed in size and type, and only four groups participated in both challenges. Among these, one group used different versions of their method (SID#1 in CAGI1, and SID#26 in CAGI2), while two others did not make use of the CBS training data.

Another limitation of the assessment involved requesting standard deviation as estimates of reliability from predictors, as opposed to the more commonly employed confidence levels that most prediction methods provide. Consequently, some predictors did not provide these values, chose them arbitrarily, or provided large values with the result that they could not be reasonably used in these assessments.

These challenges also revealed a number of experimental limitations. Yeast CBS lacks the heme domain and is not regulated by AdoMet (Jhee, McPhie, & Miles, 2000), thereby engaging different pathways in the enzyme's regulation and physiological roles. Additionally, over-expression can result in non-physiological effects, including protein aggregation. These differences could help explain some of the inconsistencies observed in the experimental study in which yeast growth phenotypes did not match the clinical data (Mayfield et al., 2012). In a similar study, several variants identical to the ones used in these experiments resulted in contrasting yeast growth phenotypes (Wei, Wang, Wang, Kruger, & Dunbrack, 2010).

In addition, the clinical assessment of the majority of variants explored in this study has since changed. Of the 78 alleles described in CAGI2 as having been observed in patients with homocystinuria, only 30 are currently classified as pathogenic or likely pathogenic in ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/, accessed March 25, 2019), with an additional 16 annotated as being of uncertain significance or with conflicting interpretations of pathogenicity (Table 1). Eight substitutions (P78R, K102N, D234N, R266K, V320A, T353M, V371M, D444N) out of 22 that showed experimental growth rates of ≥85% in CAGI2 are currently annotated as pathogenic or likely pathogenic. Most of the participants made accurate predictions for these 'benign' variants (Figure 2). It is important to mention here that ClinVar had not yet been launched during the first two CAGI challenges. Also, not all the (likely) pathogenic *CBS* variants currently present in ClinVar have been collected from clinical testing, some are based on literature with no assertion criteria provided. Ideally, different functional assays should be applied, in order to increase the confidence in the observed phenotypic effect of the studied variant, because the function of a gene can differ in distinct organisms. Finally, mutations in *cis* with the ability to either suppress other pathogenic missense mutations or increase the severity of the clinical phenotype continue to be reported (de Franchis, Kraus, Kozich, Sebastio, & Kraus, 1999; Shan, Dunbrack, Christopher, & Kruger, 2001), raising the possibility that the incidence of double mutant alleles may be underestimated in homocystinuric patients.

Conclusion

CBS is a multi-functional enzyme with complex biology and intricate regulation that remains the object of much study. Our assessment of the CAGI1 and CAGI2 CBS challenges highlighted the strengths and weaknesses of different prediction features and approaches, as well as the need to address issues of methodological and experimental limitations. Both computational and experimental methods need to be tailored to the particular biological question under investigation in order to improve the predictive potential of the variant effect. It is hoped that future iterations of CAGI will see improvements on all these fronts.

Acknowledgements

The authors are grateful to the CAGI organizers, data providers and participants.

Disclosure statement: The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available to registered users from the CAGI web site https://genomeinterpretation.org/content/cagi-2011-results.

References

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., . . . Sunyaev, S. R. (2010). A method and server for predicting damaging missense mutations. *Nat Methods, 7*(4), 248-249. doi:10.1038/nmeth0410-248

Bromberg, Y., & Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res, 35*(11), 3823-3835. doi:10.1093/nar/gkm238

Capriotti, E., Fariselli, P., Rossi, I., & Casadio, R. (2008). A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics, 9 Suppl 2*, S6. doi:10.1186/1471-2105-9-S2-S6

de Franchis, R., Kraus, E., Kozich, V., Sebastio, G., & Kraus, J. P. (1999). Four novel mutations in the cystathionine beta-synthase gene: effect of a second linked mutation on the severity of the homocystinuric phenotype. *Hum Mutat, 13*(6), 453-457. doi:10.1002/(SICI)1098-1004(1999)13:6<453::AID-HUMU4>3.0.CO;2-K

Dimster-Denk, D., Tripp, K. W., Marini, N. J., Marqusee, S., & Rine, J. (2013). Mono and dual cofactor dependence of human cystathionine beta-synthase enzyme variants in vivo and in vitro. *G3 (Bethesda), 3*(10), 1619-1628. doi:10.1534/g3.113.006916

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., & Liu, X. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet, 24*(8), 2125-2137. doi:10.1093/hmg/ddu733

Ereno-Orbea, J., Majtan, T., Oyenarte, I., Kraus, J. P., & Martinez-Cruz, L. A. (2013). Structural basis of regulation and oligomerization of human cystathionine beta-synthase, the central enzyme of transsulfuration. *Proc Natl Acad Sci U S A, 110*(40), E3790-3799. doi:10.1073/pnas.1313683110

Eto, K., & Kimura, H. (2002). A novel enhancing mechanism for hydrogen sulfide-producing activity of cystathionine beta-synthase. *J Biol Chem, 277*(45), 42680-42685. doi:10.1074/jbc.M205835200

Fernald, G. H., Capriotti, E., Daneshjou, R., Karczewski, K. J., & Altman, R. B. (2011). Bioinformatics challenges for personalized medicine. *Bioinformatics, 27*(13), 1741-1748. doi:10.1093/bioinformatics/btr295

Fraczkiewicz, R., & Braun, W. (1998). Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *Journal of Computational Chemistry, 19*(3), 319-333. doi:10.1002/(sici)1096-987x(199802)19:3<319::aid-jcc6>3.0.co;2-w

Grimm, D. G., Azencott, C. A., Aicheler, F., Gieraths, U., MacArthur, D. G., Samocha, K. E., . . . Borgwardt, K. M. (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Hum Mutat, 36*(5), 513-523. doi:10.1002/humu.22768

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A, 89*(22), 10915-10919.

Hoskins, R. A., Repo, S., Barsky, D., Andreoletti, G., Moult, J., & Brenner, S. E. (2017). Reports from CAGI: The Critical Assessment of Genome Interpretation. *Hum Mutat, 38*(9), 1039-1041. doi:10.1002/humu.23290

Ioannidis, N. M., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., . . . Sieh, W. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet, 99*(4), 877-885. doi:10.1016/j.ajhg.2016.08.016

Jhee, K. H., McPhie, P., & Miles, E. W. (2000). Yeast cystathionine beta-synthase is a pyridoxal phosphate enzyme but, unlike the human enzyme, is not a heme protein. *J Biol Chem, 275*(16), 11541-11544.

Katsonis, P., Koire, A., Wilson, S. J., Hsu, T. K., Lua, R. C., Wilkins, A. D., & Lichtarge, O. (2014). Single nucleotide variations: biological impact and theoretical interpretation. *Protein Sci, 23*(12), 1650-1666. doi:10.1002/pro.2552

Katsonis, P., & Lichtarge, O. (2014). A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness. *Genome Res, 24*(12), 2050-2058. doi:10.1101/gr.176214.114

Kraus, J. P., Janosik, M., Kozich, V., Mandell, R., Shih, V., Sperandeo, M. P., . . . Gaustadnes, M. (1999). Cystathionine beta-synthase mutations in homocystinuria. *Hum Mutat, 13*(5), 362-375. doi:10.1002/(SICI)1098-1004(1999)13:5<362::AID-HUMU4>3.0.CO;2-K

Kumar, A., Wissbrock, A., Goradia, N., Bellstedt, P., Ramachandran, R., Imhof, D., & Ohlenschlager, O. (2018). Heme interaction of the intrinsically disordered N-terminal peptide segment of human cystathionine-beta-synthase. *Sci Rep, 8*(1), 2474. doi:10.1038/s41598-018-20841-z

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc, 4*(7), 1073-1081. doi:10.1038/nprot.2009.86

Lever, J., Krzywinski, M., & Altman, N. (2016). Classification evaluation. *Nature Methods, 13*, 603. doi:10.1038/nmeth.3945

Madhusudan Makwana, K., & Mahalakshmi, R. (2015). Implications of aromatic-aromatic interactions: From protein structures to peptide models. *Protein Sci, 24*(12), 1920-1933. doi:10.1002/pro.2814

Majtan, T., Pey, A. L., Gimenez-Mascarell, P., Martinez-Cruz, L. A., Szabo, C., Kozich, V., & Kraus, J. P. (2018). Potential Pharmacological Chaperones for Cystathionine Beta-Synthase-Deficient Homocystinuria. *Handb Exp Pharmacol, 245*, 345-383. doi:10.1007/164_2017_72

Mayfield, J. A., Davies, M. W., Dimster-Denk, D., Pleskac, N., McCarthy, S., Boydston, E. A., . . . Rine, J. (2012). Surrogate genetics and metabolic profiling for characterization of human disease alleles. *Genetics, 190*(4), 1309-1323. doi:10.1534/genetics.111.137471

Moat, S. J., Bao, L., Fowler, B., Bonham, J. R., Walter, J. H., & Kraus, J. P. (2004). The molecular basis of cystathionine beta-synthase (CBS) deficiency in UK and US patients with homocystinuria. *Hum Mutat, 23*(2), 206. doi:10.1002/humu.9214

Moorthie, S., Cameron, L., Sagoo, G. S., Bonham, J. R., & Burton, H. (2014). Systematic review and meta-analysis to estimate the birth prevalence of five inherited metabolic diseases. *J Inherit Metab Dis, 37*(6), 889-898. doi:10.1007/s10545-014-9729-0

Mudd, S. H., Levy, H. L., & Kraus, J. P. (2001). Disorders of transsulfuration. In C. R. Scriver, A. Beaudet, W. Sly, & D. Valle (Eds.), *The Metabolic Basis of Inherited Disease* (pp. 2007–2056): McGraw-Hill, New York.

Niu, W., Wang, J., Qian, J., Wang, M., Wu, P., Chen, F., & Yan, S. (2018). Allosteric control of human cystathionine beta-synthase activity by a redox active disulfide bond. *J Biol Chem, 293*(7), 2523-2533. doi:10.1074/jbc.RA117.000103

Ojha, S., Wu, J., LoBrutto, R., & Banerjee, R. (2002). Effects of heme ligand mutations including a pathogenic variant, H65R, on the properties of human cystathionine beta-synthase. *Biochemistry, 41*(14), 4649-4654.

Olatubosun, A., Valiaho, J., Harkonen, J., Thusberg, J., & Vihinen, M. (2012). PON-P: integrated predictor for pathogenicity of missense variants. *Hum Mutat, 33*(8), 1166-1174. doi:10.1002/humu.22102

Pejaver, V., Mooney, S. D., & Radivojac, P. (2017). Missense variant pathogenicity predictors generalize well across a range of function-specific prediction challenges. *Hum Mutat, 38*(9), 1092-1108. doi:10.1002/humu.23258

Pey, A. L., Majtan, T., Sanchez-Ruiz, J. M., & Kraus, J. P. (2013). Human cystathionine beta-synthase (CBS) contains two classes of binding sites for S-adenosylmethionine (SAM): complex regulation of CBS activity and stability by SAM. *Biochem J, 449*(1), 109-121. doi:10.1042/BJ20120731

Rost, B., Radivojac, P., & Bromberg, Y. (2016). Protein function in precision medicine: deep understanding with machine learning. *FEBS Lett, 590*(15), 2327-2341. doi:10.1002/1873-3468.12307

Shan, X., Dunbrack, R. L., Jr., Christopher, S. A., & Kruger, W. D. (2001). Mutations in the regulatory domain of cystathionine beta synthase can functionally suppress patient-derived mutations in cis. *Hum Mol Genet, 10*(6), 635-643.

Skovby, F., Gaustadnes, M., & Mudd, S. H. (2010). A revisit to the natural history of homocystinuria due to cystathionine beta-synthase deficiency. *Mol Genet Metab, 99*(1), 1-3. doi:10.1016/j.ymgme.2009.09.009

Skovby, F., Kraus, J. P., & Rosenberg, L. E. (1984). Biosynthesis and proteolytic activation of cystathionine beta-synthase in rat liver. *J Biol Chem, 259*(1), 588-593.

Zou, C. G., & Banerjee, R. (2003). Tumor necrosis factor-alpha-induced targeted proteolysis of cystathionine beta-synthase modulates redox homeostasis. *J Biol Chem, 278*(19), 16802-16808. doi:10.1074/jbc.M212376200

Tang, H., & Thomas, P. D. (2016). Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation. *Genetics, 203*(2), 635-647. doi:10.1534/genetics.116.190033

Tavtigian, S. V., Byrnes, G. B., Goldgar, D. E., & Thomas, A. (2008). Classification of rare missense substitutions, using risk surfaces, with genetic- and molecular-epidemiology applications. *Hum Mutat, 29*(11), 1342-1354. doi:10.1002/humu.20896

Thomas, P. D., Kejariwal, A., Guo, N., Mi, H., Campbell, M. J., Muruganujan, A., & Lazareva-Ulitsky, B. (2006). Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools. *Nucleic Acids Res, 34*(Web Server issue), W645-650.

Weeks, C. L., Singh, S., Madzelan, P., Banerjee, R., & Spiro, T. G. (2009). Heme regulation of human cystathionine beta-synthase activity: insights from fluorescence and Raman spectroscopy. *J Am Chem Soc, 131*(35), 12809-12816. doi:10.1021/ja904468w

Wei, Q., Wang, L., Wang, Q., Kruger, W. D., & Dunbrack, R. L., Jr. (2010). Testing computational prediction of missense mutation phenotypes: functional characterization of 204 mutations of human cystathionine beta synthase. *Proteins, 78*(9), 2058-2074. doi:10.1002/prot.22722

Ye, Z. Q., Zhao, S. Q., Gao, G., Liu, X. Q., Langlois, R. E., Lu, H., & Wei, L. (2007). Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). *Bioinformatics, 23*(12), 1444-1450. doi:10.1093/bioinformatics/btm119

Yoo, J., Lee, Y., Kim, Y., Rha, S. Y., & Kim, Y. (2008). SNPAnalyzer 2.0: a web-based integrated workbench for linkage disequilibrium analysis and association analysis. *BMC Bioinformatics, 9*, 290. doi:10.1186/1471-2105-9-290

Yue, P., & Moult, J. (2006). Identification and analysis of deleterious human SNPs. *J Mol Biol, 356*(5), 1263-1274. doi:10.1016/j.jmb.2005.12.025

[dataset] CAGI; 2010/2011; CBS challenge dataset; Dataset available to registered users from the CAGI web site: https://genomeinterpretation.org/content/cagi-2011-results.

Table 1. CAGI1 prediction dataset: 51 single amino acid substitutions within the human CBS coding region. CAGI2 prediction dataset: 78 single amino acid variants that had been observed in patients with homocystinuria. Current (2019) ClinVar pathogenic (P) or likely pathogenic (LP) status of the CAGI2 variants in relation to homocystinuria are shown. C denotes conflicting interpretations of pathogenicity, U uncertain significance.

CAGI1		CAGI2		
Nucleotide variant	Protein variant	Nucleotide variant	Protein variant	ClinVar (2019)
194A>T	H65L	194A>G	H65R	
250A>G	I84V	304G>C	A69P	
353T>G	V118G	209C>T	P70L	U
370C>G	L124V	233C>G	P78R	LP
370,371CT>GC	L124A	253G>C	G85R	
379A>G	I127V	260C>A	T87N	
424A>G	I142V	262C>T	P88S	
424,425AT>GC	I142A	302T>C	L101P	P/LP
425T>A	I142N	304A>C	K102Q	
460461CT>GC	L154A	306G>C	K102N	LP
461T>C	L154P	325T>C	C109R	P/LP
529A>G	K177E	341C>T	A114V	P/LP
541,542CT>GC	L181A	346G>A	G116R	LP
562A>G	I188V	361C>T	R121C	C
566T>C	V189A	362G>A	R121H	LP
629T>A	L210Q	362G>T	R121L	
640A>G	I214V	376A>G	M126V	
659T>G	L220R	384G>T	E128D	
684C>A	N228K	393G>C	E131D	U
718A>G	I240V	415G>A	G139R	P
718719AT>GC	I240A	429C>G	I143M	
721C>G	L241V	430G>A	E144K	P/LP
742,743CT>GC	L248A	434C>T	P145L	P
755T>C	V252A	442G>C	G148R	LP
772G>C	G258R	451G>A	G151R	
800A>T	K267M	457G>C	G153R	U
799A>G	K267E	461T>A	L154Q	
811A>G	K271E	463G>A	A155T	
829,830AT>CC	I277P	473C>T	A158V	
839T>C	V280A	494G>A	C165Y	P
856,857AT>GC	I286A	502G>A	V168M	C
877C>G	L293V	539T>C	V180A	U
931A>G	I311V	572C>T	T191M	P
1012,1013CT>GC	L338A	593A>T	D198V	
1023A>T	Q341H	671G>A	R224H	

1034T>C	L345P	676G>A	A226T	LP
1061T>G	V354G	683A>G	N228S	U
1067T>C	V356A	684C>A	N228K	
1073T>C	V358A	700G>A	D234N	P
1112T>C	V371A	715G>A	E239K	
1115T>C	V372A	770C>T	T257M	P/LP
1120,1121CT>GC	L374A	775G>A	G259S	U
1147A>T	T383S	785C>T	T262M	P/LP
1153,1154,1155TTC>CAA	F385Q	796A>G	R266G	
1153T>C	F385L	797G>A	R266K	P/LP
1223G>T	W408L	824G>A	C275Y	
1268T>C	L423P	833T>C	I278T	P
1298A>T	H433L	862G>A	A288T	U
1370G>A	G457E	862G>C	A288P	
1468A>C	I490L	904G>A	E302K	LP
1646A>G	D549G	919G>A	G307S	P
		959T>C	V320A	LP
		992C>A	A331E	LP
		992C>T	A331V	
		1007G>A	R336H	LP
		1039G>A	G347S	LP
		1046G>A	S349N	
		1055G>A	S352N	
		1058C>T	T353M	P/LP
		1060G>A	V354M	
		1063G>C	A355P	
		1081G>A	A361T	
		1105C>T	R369C	U
		1106G>A	R369H	U
		1106G>C	R369P	
		1109G>A	C370Y	LP
		1111G>A	V371M	LP
		1126G>A	D376N	
		1150A>G	K384E	P
		1173G>A	M391I	
		1265C>T	P422L	U
		1301C>A	T434N	U
		1304T>C	I435T	U
		1316G>A	R439Q	C
		1330G>A	D444N	P/LP
		1367T>C	L456P	
		1397C>T	S466L	U
		1572C>A	Q526K	

Table 2. Overview of the phenotype prediction programs used to generate predictions for the CAGI1 and CAGI2 CBS challenges

Submission ID	Group ID	Program name	Program features	Reference
CAGI1				
SID#1	Lichtarge lab	Evolutionary Action working version	sequence, structure	
SID#2	Bromberg lab	SNAP	sequence, structure, annotation	Bromberg et al., 2007
SID#3	Wei lab	SAPRED	structure	Ye et al., 2007
SID#4	Switch lab		structure	
SID#5	Vihinen lab	PON-P	meta-predictor	Olatubosun, Valiaho, Harkonen, Thusberg, & Vihinen, 2012
SID#6	Vihinen lab	PolyPhen2	sequence, structure	Adzhubei et al., 2010
SID#7	Vihinen lab	SNPanalyzer	sequence	Yoo, Lee, Kim, Rha, & Kim, 2008
SID#8	Vihinen lab	Panther	sequence	Thomas et al., 2006
SID#9	Casadio lab	IMutant3	structure, thermal stability	Capriotti, Fariselli, Rossi, & Casadio, 2008
SID#10	Casadio lab	IMutant4	structure, thermal stability	
SID#11	Casadio lab	IMutant baseline	structure, thermal stability	
SID#12	Forman lab	SDM	sequence, structure	
SID#13	BioFolD Unit	IMutant3	sequence, structure	Capriotti et al., 2008
SID#14	Karchin lab		sequence, structure	
SID#16	Mooney lab		meta-predictor	
SID#18	Forman lab	SDM	sequence, structure	
SID#19	Tavtigian lab	AlignGVGD	sequence	Tavtigian, Byrnes, Goldgar, & Thomas, 2008
SID#20	Tavtigian lab	AlignGVGD	meta-predictor	
CAGI2				
SID#16	Bromberg lab	SNAP	sequence, structure	Bromberg et al., 2007
SID#19	Tosatto lab		structure	
SID#20	Tosatto lab		structure	
SID#21	Tosatto lab		structure	
SID#22	Tosatto lab		structure	
SID#23	Tosatto lab		structure	
SID#24	Tosatto lab	D100 roll	random	
SID#25	Switch lab		structure	
SID#26	Lichtarge Lab	Evolutionary Action	sequence, structure	Katsonis & Lichtarge, 2014
SID#27	Vihinen lab	PON-P	meta-predictor	Olatubosun et al., 2012

SID#28	Vihinen lab	PON-P	meta-predictor	Olatubosun et al., 2012
SID#29	Shatsky lab		meta-predictor	
SID#30	Shatsky lab		meta-predictor	
SID#32	Shatsky lab		meta-predictor	
SID#33	Mooney lab		meta-predictor	
SID#34	Sunyayev lab		sequence, structure	
SID#35	Moult lab	SNPs3D SVM	sequence, structure	Yue & Moult, 2006
SID#36	Moult lab	SNPs3D SVM	sequence, structure, annotation	
SID#41	Mooney lab		meta-predictor	

Figure legends

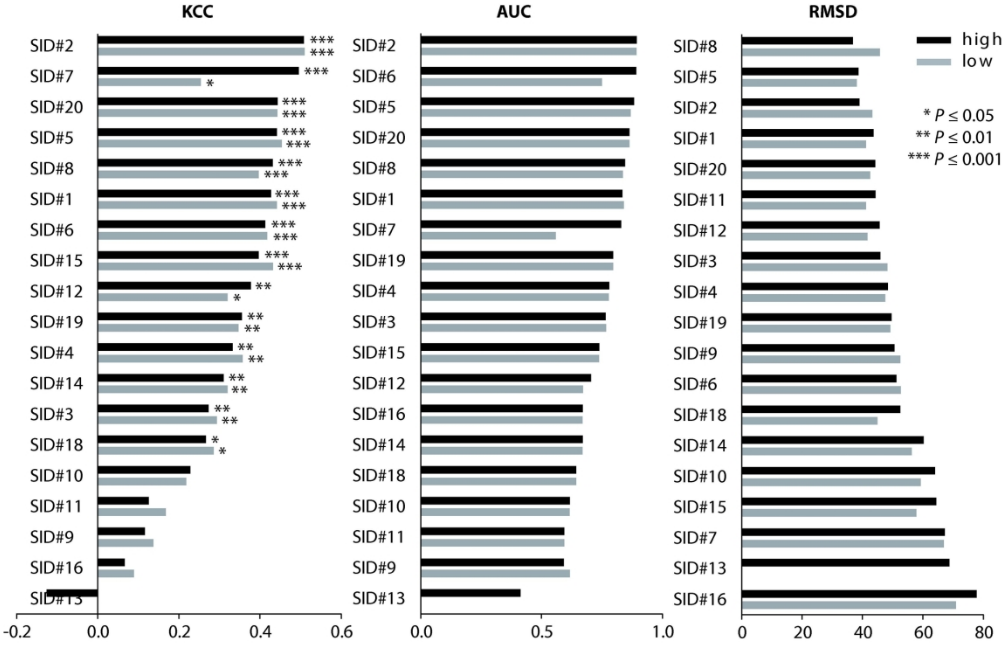
Figure 1. Kendall's Tau correlation coefficient (KCC), area under the ROC curve (AUC) and root-mean-square deviation (RMSD) for the phenotype prediction programs at high and low cofactor concentration in CAGI1. Statistical significance of correlation scores is indicated with asterisks.

Figure 2. Consensus predictions for CBS. (A) CBS domain diagram, (B) CAGI1, (C) CAGI2. The percentage of correct predictions for deleterious (red) and benign (blue) variants is shown for each experimentally determined variant at low pyridoxine concentration. Residues are shaded in the color of the corresponding domain, with the linker region highlighted in orange.

Figure 3. Kendall's Tau correlation coefficient (KCC), area under the ROC curve (AUC) and root-mean-square deviation (RMSD) for the phenotype predictions at high and low cofactor concentration in the CAGI2 CBS challenge. Statistical significance of correlation scores is indicated with asterisks.

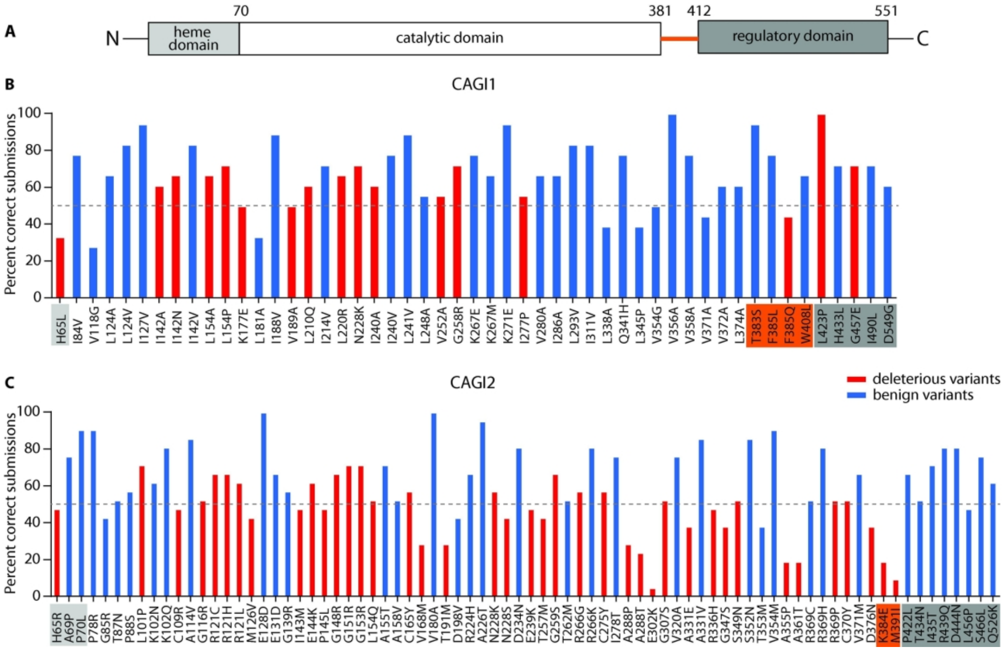
Figure 4. Spearman's rank correlation among methods and with experimental data (Exp) for high and low cofactor concentration. Each cell shows the correlation between two methods, with a color scale ranging from red (perfect correlation) to white (no correlation) and blue (perfect anti-correlation).

Figure 5. Δ adjusted R^2 values of the methods from the linear regression model for high and low cofactor concentration, quantifying the contribution of each method to the proportion of total variance explained.



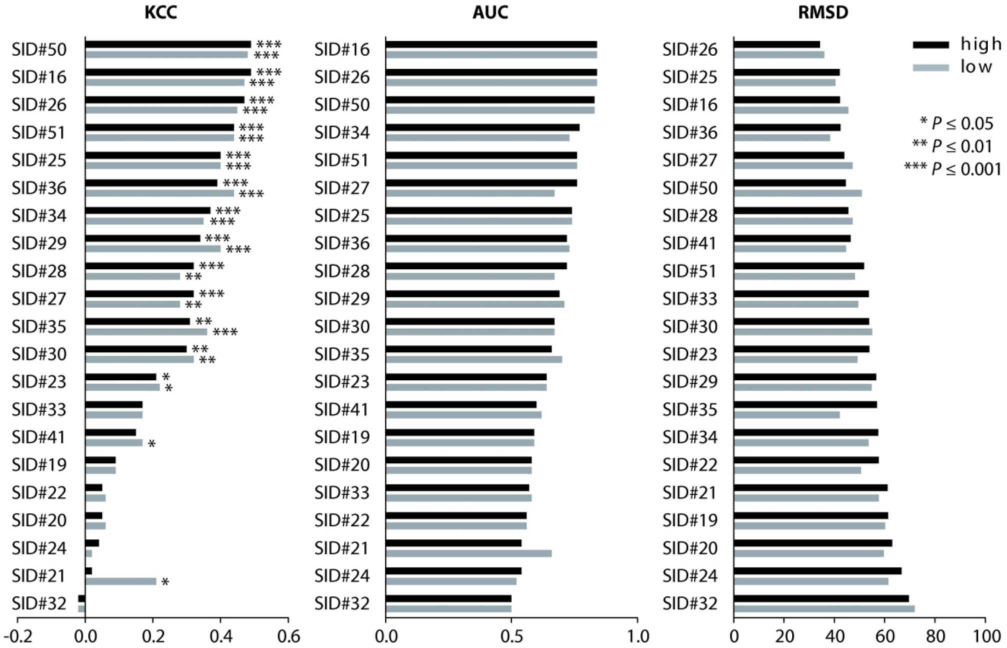
Kendall's Tau correlation coefficient (KCC), area under the ROC curve (AUC) and root-mean-square deviation (RMSD) for the phenotype prediction programs at high and low cofactor concentration in CAGI1. Statistical significance of correlation scores is indicated with asterisks.

174x111mm (300 x 300 DPI)



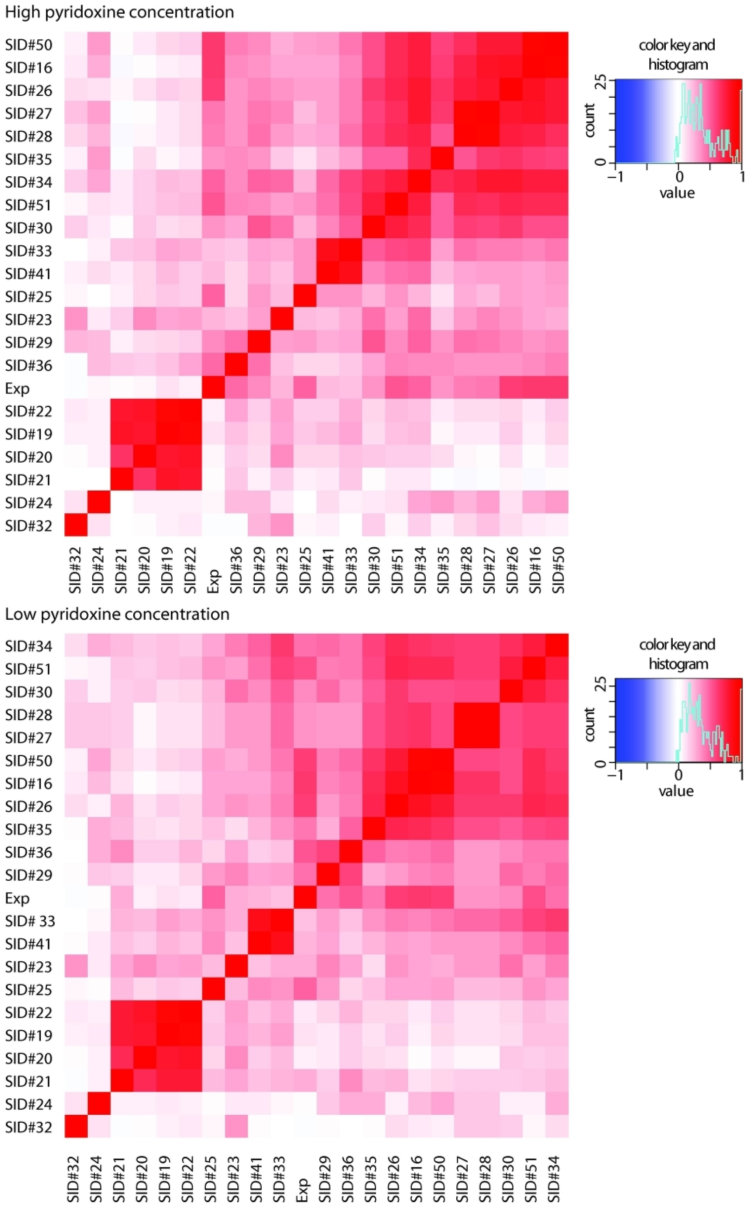
Consensus predictions for CBS. (A) CBS domain diagram, (B) CAGI1, (C) CAGI2. The percentage of correct predictions for deleterious (red) and benign (blue) variants is shown for each experimentally determined variant at low pyridoxine concentration. Residues are shaded in the color of the corresponding domain, with the linker region highlighted in orange.

180x116mm (300 x 300 DPI)



Kendall's Tau correlation coefficient (KCC), area under the ROC curve (AUC) and root-mean-square deviation (RMSD) for the phenotype predictions at high and low cofactor concentration in the CAGI2 CBS challenge. Statistical significance of correlation scores is indicated with asterisks.

167x107mm (300 x 300 DPI)

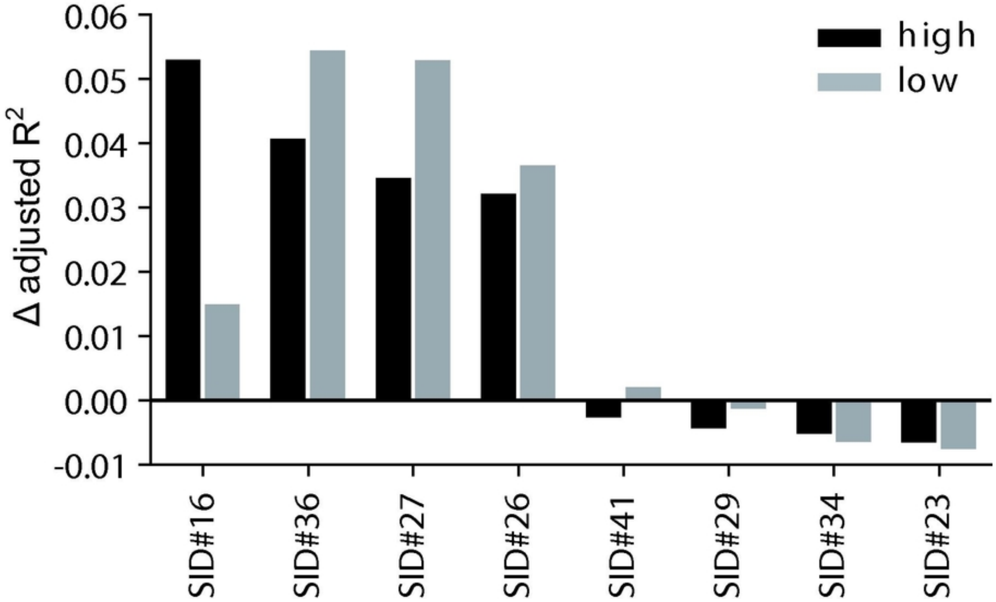


High pyridoxine concentration

Low pyridoxine concentration

Spearman’s rank correlation among methods and with experimental data (Exp) for high and low cofactor concentration. Each cell shows the correlation between two methods, with a color scale ranging from red (perfect correlation) to white (no correlation) and blue (perfect anti-correlation).

145x235mm (300 x 300 DPI)



Δ adjusted R^2 values of the methods from the linear regression model for high and low cofactor concentration, quantifying the contribution of each method to the proportion of total variance explained.

77x46mm (300 x 300 DPI)

Assessing Computational Predictions of the Phenotypic Effect of Cystathionine-beta-Synthase Variants

Laura Kasak[1,2], Constantina Bakolitsa[1], Zhiqiang Hu[1], Jasper Rine[3], Dago F. Dimster-Denk[3,*], Gaurav Pandey[1,*], Yana Bromberg[4], Chen Cao[5,6], Emidio Capriotti[7,*], Rita Casadio[8], Manuel Giollo[9], Panagiotis Katsonis[10], Emanuela Leonardi[11,*], Oliver Lichtarge[10], Pier Luigi Martelli[8], Sean D. Mooney[12,*], Lipika R. Pal[5], Predrag Radivojac[13], Castrense Savojardo[8], Janita Thusberg[12,*], Silvio C.E. Tosatto[9], Mauno Vihinen[14,*], Jouni Väliaho[14], Susanna Repo[1,*], John Moult[5,15], Steven E. Brenner[1], Iddo Friedberg[16,17,*]

[1]Department of Plant and Microbial Biology, University of California, Berkeley, CA, USA

[2]Institute of Biomedicine and Translational Medicine, University of Tartu, Tartu, Estonia

[3]California Institute for Quantitative Biosciences, University of California, Berkeley, CA, USA

[4]Department of Biochemistry and Microbiology, Rutgers University, New Brunswick, NJ, USA

[5]Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, MD, USA

[6]Computational Biology, Bioinformatics and Genomics, Biological Sciences Graduate Program, University of Maryland, College Park, MD, USA

[7]Department of Bioengineering, Stanford University, Stanford, CA, USA

[8]Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Bologna, Italy

[9]Department of Biomedical Sciences, University of Padua, Padua, Italy

[10]Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA

[11]Department for Woman and Child Health, University of Padua, Italy

[12]Buck Institute for Research on Aging, Novato, CA, USA

[13]School of Informatics and Computing, Indiana University, Bloomington, IN, USA

[14]Institute of Medical Technology, University of Tampere, Tampere, Finland

[15]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD, USA

[16]Department of Microbiology, Miami University, Oxford, OH, USA

[17]Department of Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, IA USA

Correspondence

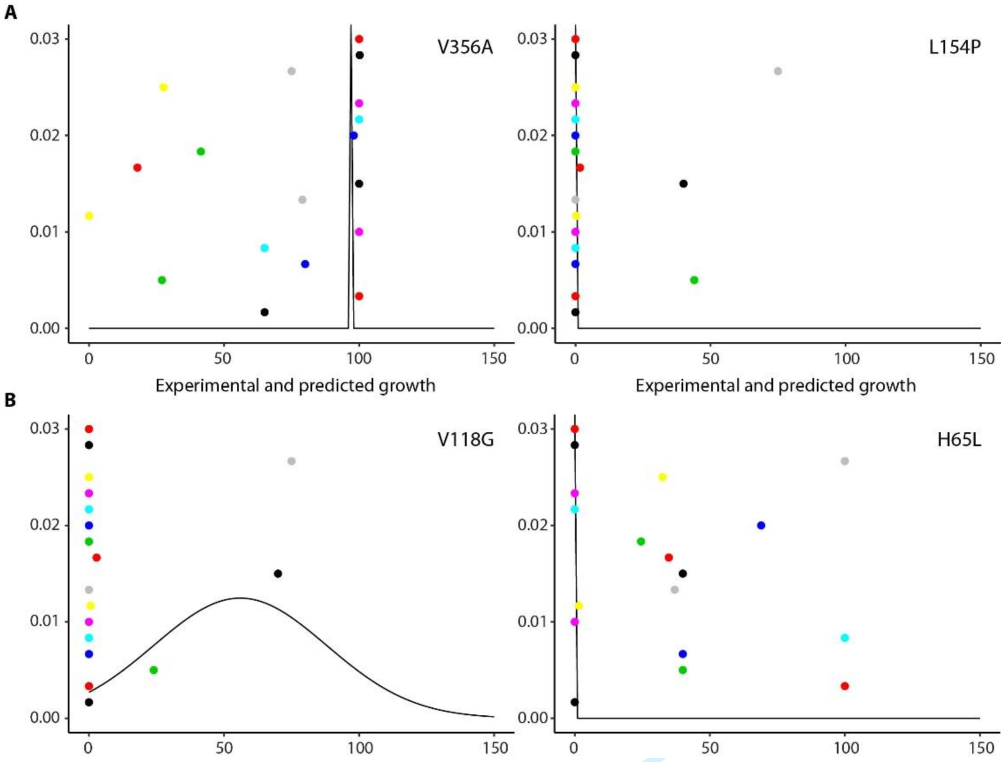
Iddo Friedberg, Department of Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, IA USA

Email: idoerg@iastate.edu

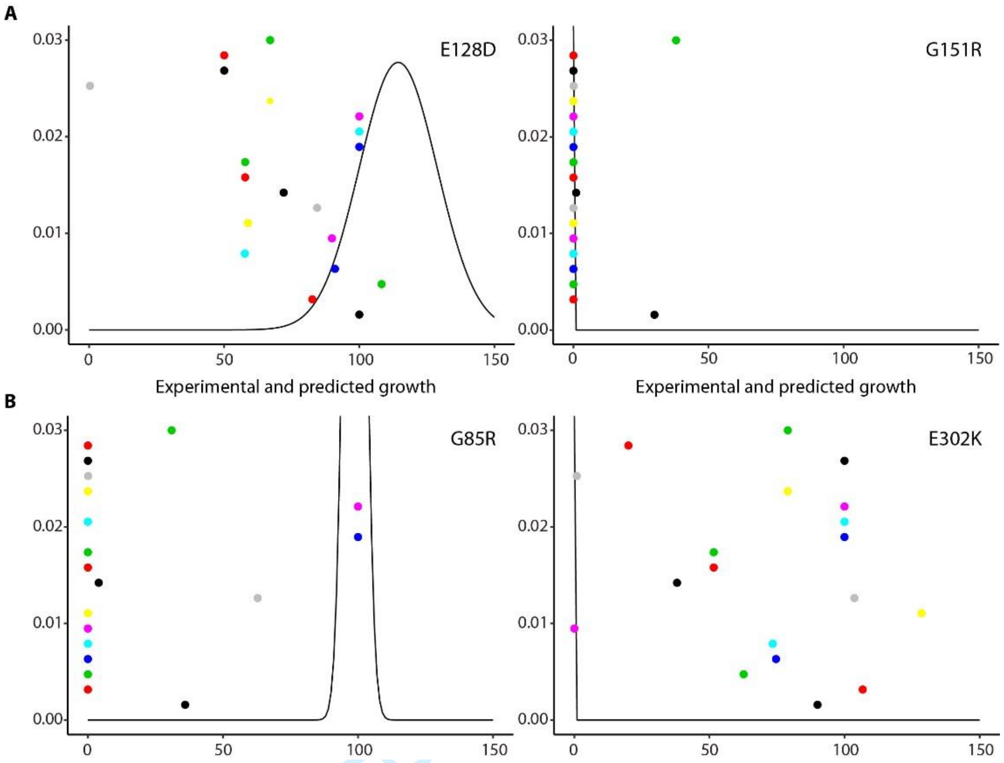
[*]Present address: Dago F. Dimster-Denk, Pionyr Immunotherapeutics, San Francisco, CA, USA; Gaurav Pandey, Department of Genetics and Genomic Sciences and Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, NY, USA; Emidio Capriotti, BioFolD Group, Department of Pharmacy and Biotechnology (FaBiT), University of Bologna, Bologna, Italy; Emanuela Leonardi, Department for Woman and Child Health, University of Padua, Italy; Pediatric Research Institute, Citta della Speranza, Padua, Italy; Sean D. Mooney, Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, USA; Predrag Radivojac, Khoury College of Computer Sciences, Northeastern University, Boston MA, USA; Mauno Vihinen, Department of Experimental Medical Science, Lund University, Lund, Sweden; Janita Thusberg, Invitae, San Francisco, CA, USA; Susanna Repo, ELIXIR, Wellcome Genome Campus, Hinxton, Cambridge, UK; Iddo Friedberg, Department of Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, IA USA

Supplementary Material

Supplementary Figures



Supplementary Figure 1. Individual predictions versus experimental data at low pyridoxine concentration in the CAGI1 CBS challenge. (A) Examples of easy to predict benign (V356A) and deleterious (L154P) variants. (B) Examples of hard to predict benign (V118G) and deleterious (H65L) variants. Curve is the experimentally measured distribution; absence of curve means that the yeast did not grow at all. Each dot corresponds to a prediction, with submissions 1 through 18 distributed vertically. X-axis is % yeast growth of the mutated protein relative to wild-type. Y axis – 18 submissions 1-18.



Supplemental Figure 2. Individual substitution predictions versus experimental data at low pyridoxine concentration in the CAGI2 CBS challenge. (A) Examples of easy to predict benign (E128D) and deleterious (G151R) variants. (B) Examples of hard to predict benign (G85R) and deleterious (E302K) variants. Curve is the experimentally measured distribution; absence of curve means that the yeast did not grow at all. Each dot corresponds to a prediction, with submissions 1 through 19 distributed vertically. X-axis is % yeast growth of the mutated protein relative to wild-type. Y axis – 19 submissions 1-19.

Supplementary Tables

Supplementary Table 1. Performance measures considered in the assessment. KCC, Kendall's correlation coefficient; AUC, area under the ROC curve; ACC, accuracy; PPV, positive predictive value; NPV, negative predictive value; RMSD, root mean square deviation.

	High pyridoxine concentration									Low pyridoxine concentration								
Submission ID	Spearman's ρ	KCC	AUC	recall	specificity	ACC	PPV	NPV	RMSD	Spearman's ρ	KCC	AUC	recall	specificity	ACC	PPV	NPV	RMSD
CAGI1																		
SID#1	0.54	0.43	0.83	0.85	0.76	0.82	0.88	0.72	43.6	0.60	0.44	0.84	0.85	0.76	0.82	0.88	0.72	41.1
SID#2	0.63	0.51	0.89	1.00	0.00	0.67	0.67	NA	39.0	0.65	0.51	0.89	1.00	0.00	0.67	0.67	NA	43.2
SID#3	0.38	0.27	0.77	1.00	0.00	0.67	0.67	NA	45.9	0.42	0.29	0.77	1.00	0.00	0.67	0.67	NA	48.3
SID#4	0.43	0.33	0.78	0.73	0.73	0.73	0.85	0.58	48.4	0.47	0.36	0.78	0.73	0.73	0.73	0.85	0.58	47.5
SID#5	0.58	0.44	0.88	0.97	0.53	0.82	0.80	0.90	38.7	0.59	0.45	0.87	0.91	0.59	0.80	0.81	0.77	38.2
SID#6	0.55	0.41	0.89	0.97	0.53	0.82	0.80	0.90	51.2	0.51	0.42	0.75	0.50	1.00	0.67	1.00	0.50	52.7
SID#7	0.62	0.50	0.83	0.74	0.94	0.80	0.96	0.64	67.3	0.30	0.25	0.56	0.12	1.00	0.41	1.00	0.36	67.0
SID#8	0.59	0.43	0.85	1.00	0.00	0.67	0.67	NA	36.8	0.55	0.40	0.84	0.71	0.93	0.78	0.96	0.61	45.8
SID#9	0.16	0.12	0.59	1.00	0.00	0.67	0.67	NA	50.6	0.18	0.14	0.62	1.00	0.00	0.67	0.67	NA	52.5
SID#10	0.27	0.23	0.62	0.30	0.93	0.51	0.90	0.40	64.0	0.25	0.22	0.62	0.30	0.93	0.51	0.90	0.40	59.3
SID#11	0.15	0.13	0.59	1.00	0.00	0.70	0.70	NA	44.3	0.23	0.17	0.59	1.00	0.00	0.70	0.70	NA	41.2
SID#12	0.45	0.38	0.70	0.93	0.40	0.74	0.74	0.75	45.7	0.39	0.32	0.67	0.93	0.40	0.74	0.74	0.75	41.7
SID#13	-0.18	-0.13	0.41	0.94	0.00	0.63	0.65	0.00	68.8	NA	NA	NA	NA	NA	NA	NA	NA	NA
SID#14	0.38	0.31	0.67	0.41	0.94	0.59	0.93	0.44	60.3	0.39	0.32	0.67	0.41	0.94	0.59	0.93	0.44	56.3
SID#15	0.48	0.4	0.74	0.58	0.88	0.67	0.90	0.50	64.5	0.53	0.43	0.74	0.58	0.88	0.67	0.90	0.50	57.8

SID#16	0.08	0.07	0.67	0.85	0.12	0.61	0.66	0.29	77.9	0.12	0.09	0.67	0.85	0.12	0.61	0.66	0.29	71.0
SID#18	0.32	0.27	0.64	0.83	0.40	0.69	0.74	0.55	52.4	0.34	0.29	0.64	0.83	0.40	0.69	0.74	0.55	44.9
SID#19	0.45	0.36	0.80	0.74	0.82	0.76	0.89	0.61	49.6	0.44	0.35	0.80	0.74	0.82	0.76	0.89	0.61	49.2
SID#20	0.55	0.44	0.86	0.85	0.76	0.82	0.88	0.72	44.3	0.57	0.44	0.86	0.85	0.76	0.82	0.88	0.72	42.6
	High pyridoxine concentration									Low pyridoxine concentration								
Submission ID	Spearman's ρ	KCC	AUC	recall	specificity	ACC	PPV	NPV	RMSD	Spearman's ρ	KCC	AUC	recall	specificity	ACC	PPV	NPV	RMSD
CAGI2																		
SID#16	0.64	0.49	0.84	1.00	0.00	0.49	0.49	NA	42.3	0.62	0.47	0.84	1.00	0.00	0.49	0.49	NA	45.6
SID#19	0.11	0.09	0.59	0.68	0.50	0.59	0.57	0.63	61.4	0.12	0.09	0.59	0.68	0.50	0.59	0.57	0.63	60.2
SID#20	0.06	0.05	0.58	0.53	0.58	0.55	0.54	0.56	63.0	0.07	0.06	0.58	0.53	0.58	0.55	0.54	0.56	59.6
SID#21	0.03	0.02	0.54	0.68	0.50	0.59	0.57	0.63	61.2	0.27	0.21	0.66	0.68	0.50	0.59	0.57	0.63	57.7
SID#22	0.07	0.05	0.56	0.68	0.50	0.59	0.57	0.63	57.6	0.08	0.06	0.56	0.68	0.50	0.59	0.57	0.63	50.7
SID#23	0.26	0.21	0.64	0.47	0.75	0.62	0.64	0.60	54.0	0.27	0.22	0.64	0.47	0.75	0.62	0.64	0.60	49.3
SID#24	0.06	0.04	0.54	0.61	0.45	0.53	0.51	0.55	66.7	0.02	0.02	0.52	0.61	0.45	0.53	0.51	0.55	61.5
SID#25	0.49	0.40	0.74	0.83	0.64	0.72	0.63	0.83	42.2	0.49	0.40	0.74	0.83	0.64	0.72	0.63	0.83	40.5
SID#26	0.62	0.47	0.84	1.00	0.00	0.49	0.49	NA	34.2	0.60	0.45	0.84	1.00	0.03	0.50	0.49	1.00	36.0
SID#27	0.44	0.32	0.76	1.00	0.28	0.63	0.57	1.00	44.1	0.34	0.28	0.67	0.50	0.83	0.67	0.73	0.63	47.3
SID#28	0.40	0.32	0.72	0.68	0.73	0.71	0.70	0.71	45.5	0.34	0.28	0.67	0.50	0.83	0.67	0.73	0.63	47.3
SID#29	0.38	0.34	0.69	0.79	0.60	0.69	0.65	0.75	56.6	0.45	0.40	0.71	0.74	0.68	0.71	0.68	0.73	54.9
SID#30	0.34	0.30	0.67	0.61	0.73	0.67	0.68	0.66	53.9	0.36	0.32	0.67	0.61	0.73	0.67	0.68	0.66	55.1
SID#32	-0.02	-0.02	0.50	0.53	0.48	0.50	0.49	0.51	69.6	-0.03	-0.02	0.50	0.53	0.48	0.50	0.49	0.51	71.9
SID#33	0.21	0.17	0.57	0.39	0.68	0.54	0.54	0.54	53.8	0.22	0.17	0.58	0.39	0.68	0.54	0.54	0.54	49.5
SID#34	0.48	0.37	0.77	1.00	0.00	0.49	0.49	NA	57.4	0.44	0.35	0.73	1.00	0.00	0.49	0.49	NA	53.6

SID#35	0.35	0.31	0.66	0.74	0.58	0.65	0.62	0.70	57.0	0.43	0.36	0.70	0.74	0.58	0.65	0.62	0.70	42.1
SID#36	0.47	0.39	0.72	0.68	0.68	0.68	0.67	0.69	42.5	0.52	0.44	0.73	0.68	0.68	0.68	0.67	0.69	38.3
SID#41	0.22	0.15	0.60	1.00	0.08	0.53	0.51	1.00	46.5	0.24	0.17	0.62	1.00	0.08	0.53	0.51	1.00	44.7
SID#50	0.62	0.49	0.83	1.00	0.00	0.49	0.49	NA	44.5	0.61	0.48	0.83	1.00	0.00	0.49	0.49	NA	50.9
SID#51	0.54	0.44	0.76	0.68	0.78	0.73	0.74	0.72	51.7	0.55	0.44	0.76	0.68	0.78	0.73	0.74	0.72	48.3

Supplementary Table 2. Ranking of the phenotype prediction programs. KCC, Kendall's correlation coefficient; AUC, area under the ROC curve; RMSD, root mean square deviation.

	High pyridoxine concentration			Low pyridoxine concentration				
Submission ID	KCC	AUC	RMSD	KCC	AUC	RMSD	mean	overall
CAGI1								
SID#1	6	6	4	4	4	2	4.3	4
SID#2	1	1	3	1	1	6	2.2	1
SID#3	13	10	8	12	8	10	10.2	10
SID#4	11	9	9	8	7	9	8.8	7
SID#5	4	3	2	2	2	1	2.3	2
SID#6	7	2	12	6	9	13	8.2	6
SID#7	2	7	17	14	18	17	12.5	12
SID#8	5	5	1	7	5	8	5.2	5
SID#9	17	18	11	17	15	12	15	16
SID#10	15	16	15	15	16	16	15.5	17
SID#11	16	17	6	16	17	3	12.5	12
SID#12	9	12	7	10	11	4	8.8	7
SID#13	19	19	18				18.7	19
SID#14	12	13	14	11	12	14	12.7	14
SID#15	8	11	16	5	10	15	10.8	11
SID#16	18	13	19	18	12	18	16.3	18
SID#18	14	15	13	13	14	7	12.7	14
SID#19	10	8	10	9	6	11	9	9
SID#20	3	4	5	3	3	5	3.8	3
CAGI2								
SID#16	1	1	3	2	1	6	2.3	2
SID#19	16	15	18	17	16	19	16.8	18
SID#20	17	16	19	19	17	18	17.7	19
SID#21	20	19	17	14	13	17	16.7	16
SID#22	17	18	16	18	19	12	16.7	16
SID#23	13	13	12	13	14	10	12.5	14
SID#24	19	19	20	20	20	20	19.7	20
SID#25	5	7	2	7	5	3	4.8	4
SID#26	3	1	1	3	2	1	1.8	1

SID#27	9	5	5	11	10	7	7.8	7
SID#28	9	8	7	11	10	7	8.7	8
SID#29	8	10	13	6	8	15	10	11
SID#30	12	11	11	10	12	16	12	12
SID#32	21	21	21	21	21	21	21	21
SID#33	14	17	10	16	18	11	14.3	15
SID#34	7	4	15	9	6	14	9.2	9
SID#35	11	12	14	8	9	4	9.7	10
SID#36	6	8	4	5	7	2	5.3	5
SID#41	15	14	8	15	15	5	12	12
SID#50	1	3	6	1	3	13	4.5	3
SID#51	4	5	9	4	4	9	5.8	6

Supplementary Table 3. Features of easiest and hardest to predict variants in CAGI1 and CAGI2 at low pyridoxine concentration. Sequence substitutions scores were taken from the BLOSUM90 matrix. Solvent accessible surface area (SASA) was calculated for the monomer (PDB id 4COO) using GetArea and differences from dimer SASA results are noted. O=out (solvent-exposed), I=in, O/I=intermediate state.

CAGI1

Protein variant	Number of correct predictions	BLOSUM90	SASA	Location in structure	Comments
Easiest deleterious					
L154P	13	-4	I	H6	
N228K	13	0	I	H9	
G258R	13	-3	I	H11	
G457E	13	-3	I	H21-H22 loop	
Hardest deleterious					
H65L	6	-4	O/I	H1_H2 loop	
F385Q	8	-4	I	H17-H18 loop	
Easiest benign					
L124V	15	0	O/I	H5	
I142V	15	3	I	H5-H6 loop	
I188V	16	3	I	H7-H8 loop	
L241V	16	0	I	H10	
K271E	17	0	O	H11	
L293V	15	0	O/I	H12	
I311V	15	3	O/I	H12-H13 loop	
V356A	18	-1	I	H15	
T383S	17	1	O	H17-H18 loop	
Hardest benign					
V118G	5	-5	I	H5	
L181A	6	-2	I	H7	
L338A	7	-2	I	H14	
L345P	7	-4	O	H14-H15 loop	I in dimer
V354G	9	-5	I	H15	
V371A	8	-1	I	beta3	

CAGI2

Protein variant	Number of correct predictions	BLOSUM90	SASA	Location in structure	Comments
Easiest deleterious					
L101P	15	-4	I	H3-beta2 loop	
G151R	15	-3	I	H6	
G153R	15	-3	I	H6	
R121C	14	-5	I	H5	
R121H	14	0	I	H5	

G148R	14	-3	I	H6 start	
G259S	14	-1	I	H11	
Hardest deleterious					
E302K	1	0	O/I	H12-H13 loop	
M391I	2	1	I	H18	
A355P	4	-1	I	H15	
A361T	4	0	I	H15	
K384E	4	0	I	H17-H18 loop	
A288T	6	0	I	beta5-H12 loop	
V168M	6	0	I	H6-H7 loop	
A288P	6	-1	I	beta5-H12 loop	
T191M	6	-1	I	H7-H8 loop	
Easiest benign					
E128D	21	1	O	H5	O/I in dimer
V180A	21	-1	O/I	H7	I in dimer
A226T	19	0	O	H9	
P70L	19	-4	I	H1-H2 loop	
P78R	19	-3	O	H1-H2 loop	
V354M	19	0	I	H15	
A114V	18	-1	I	H4-H5 loop	
S352N	18	0	I	H15	
Hardest benign					
T353M	8	-1	I	H15	
D198V	9	-5	O	H7-H8 loop	O/I in dimer
G85R	9	-3	I	H2-beta1 loop	
L456P	10	-4	O/I	H21-H22 loop	

Predictor method descriptions

A short description of each method.

CAGI1

SID#1 – pre-mature version of the Evolutionary Action method

This submission was from the Lichtarge lab and it used a method that was under development (at the time of submission) to predict the impact of coding variants. Although the basic hypothesis was the same to the version that was published later as the Evolutionary Action (EA) method (Katsonis & Lichtarge 2014), the two versions differ drastically in technical and conceptual improvements that were added later. These differences may result in substantial difference in performance. Briefly, the hypothesis is that the impact of SNVs in protein function depends on two terms: the evolutionary importance of the mutated residue and the size of substitution. The evolutionary importance was measured by the Evolutionary Trace (ET) method and the size of substitution was approximated by ET-dependent log-odds. The variant impact was calculated as the product of the two terms, which was normalized between 0 (benign) and 100 (pathogenic).

The example phenotypes were used to guide how to convert the variant impact into CBS function activity. This limited data set led to the following conclusions:

-Most mutants have binary activity, with few exceptions of partial function.

-A variant impact score of 40 roughly separates functional from non-functional mutants.

-For variants with partial function, the growth rate at 2 ng/ml Pyridoxine is about half of the growth rate at 400 ng/ml Pyridoxine.

Considering the above, the submitted answers were estimated as:

For 400 ng/ml: 0% if the raw score is less than 30, 100% if the raw score is more than 50 and linearly extrapolated between 0 and 100 if the raw score is from 30 to 50, respectively.

For 2 ng/ml: 0% if the raw score is less than 30, 100% if the raw score is more than 70 and linearly extrapolated between 0 and 100 if the raw score is from 30 to 70, respectively.

Katsonis, P. & Lichtarge, O. (2014) ‘A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness’, *Genome Research* **24**(12):2050-8.

SID#2 - SNAP

Neural-network based method for the prediction of functional effects of single amino acid substitutions. By default, for each submitted substitution, SNAP reports a single binary prediction (Neutral/Non-neutral), which is associated with an RI (reliability index, range 0-9) and a value of “expected accuracy” (0- 100%; in testing, the accuracy of SNAP predictions at the given RI). In the present evaluation of neutral predictions, no change (100% growth rate) was reported for both cofactor concentrations (the “standard deviation” was set arbitrarily to 10%). The predictors reported observing that increased RI scores correlated fairly well with the severities of protein function change. Thus, for non-neutral predictions a decrease of 10% in growth rate was assumed for each 1 point drop in RI. These scores were reported for both cofactor concentrations.

SID#3 - SAPRED

A description of the method can be found in “Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism “SAP” (Ye et al. 2007). From the 51 mutations used in this challenge, 45 could be mapped onto the CBS structure. The predictors used the homology modeling software Modeller to build a structure model of the 45 mutants. This was used as partial input to SAPRED. For the other 6 mutations, sequence only data were used in SAPRED_SEQ. SAPRED outputs for each mutation, ranging from 0 for benign to 1 for deleterious, were linearly converted to growth rates as follows:

In the high pyridoxine medium:

SARPED score	Growth rate	Classification
0.95 – 1.00	0 - 20	Severely impaired
0.75 – 0.95	20 - 80	Impaired
0 – 0.75	80 - 125	No effect

In the low pyridoxine medium:

SARPED score	Growth rate	Classification
0.90 – 1.00	0 - 20	Severely impaired
0.70 – 0.90	20 - 80	Impaired
0 – 0.70	80 - 125	No effect

Ye, Z. Q.; Zhao, S. Q.; Gao, G.; Liu, X. Q.; Langlois, R. E.; Lu, H. & Wei, L. (2007) ‘Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP)’, *Bioinformatics* **15**(23), 1444-1450.

SID#4 – FoldX3.b4

Predictions were based on the CBS structure (PDB id 1JBQ). Using the FoldX3.b4 algorithm, the structure was first repaired (RepairPDB command) and all mutants were subsequently generated using the BuildModel command. FoldX reports the change in thermodynamic stability of the protein as a result of mutation as a DDG (change in free energy difference between wild type and mutant, measured in kcal.mol^{-1}). The AnalyseComplex command was additionally used to obtain the effect of the mutation on the free energy of dimmer formation. The largest DDG value was taken to report the most severe effect. Moreover, an artificial penalty of 3 and 3 kcal.mol^{-1} was applied when the wild type residue was making a contact with the pyridoxal 5-phosphate or heme groups. Based on the training data, the destabilization of the mutants were converted into yeast growth rates as follows:

ΔΔG	growth_rate_400ng/ml_pyr	growth_rate_2ng/ml
>3	severely_impaired (0)	severely_impaired
2-3	impaired (70)	impaired (37)
1.3-2	none (103)	impaired (79)
<1.3	none (105)	none (100)

SID#5 – PON-P

PON-P (Olatubosun et al. 2012) is a machine learning -based predictor for variation pathogenicity analysis that combines results from a number of existing tolerance and protein stability prediction tools and makes an overall prediction based on all the data. For methodological details see SID#27.It has been trained with about 31,000 known variation cases of pathogenic and neutral missense variants. The machine learning method applied is random forests, comprising of 800 trees. The system calculates pathogenicity score and expected accuracy. Conversion from pathogenicity scores to growth rates (ranging from 0 to 105) was made with pathogenic cases given a value of 0 and neutral ones a value of 105. Only cases with a pathogenicity score below 0.5 were considered as having activity in the low (2 ng/ml) pyridoxine experiment.

Olatubosun, A.; Väliaho, J.; Härkönen, J., Thusberg; J. & Vihinen, M. (2012) 'PON-P: integrated predictor for pathogenicity of missense variants', *Human Mutation* **33**(8):1166-74.

SID#6 – PON-P-PolyPhen

Predictions were run with PolyPhen v. 2.0.22 (Adzhubei et al. 2010) by using Pathogenic-Or-Not-Pipeline (PON-P). Default values of the method were used. Pathogenic cases were given a value of 0 and neutral ones a value of 100. Only cases with a pathogenicity score below 5 were considered to have activity in the low (2 ng/ml) pyridoxine experiment.

Adzhubei, I. A.; Schmidt, S.; Peshkin, L.; Ramensky, V. E.; Gerasimona, A.; Bork, P.; Kondrashov, A. S. & Sunyaev, S. R. (2010) 'A method and server for predicting damaging missense mutations', *Nature Methods* **7**(4):248-249.

SID#7 – PON-P-SNPAanalyzer

Predictions were run with SNPAnalyzer (Yoo et al. 2008) by using Pathogenic-Or-Not-Pipeline (PON-P). Default values of the method were used. Pathogenic cases were given a value of 0 and neutral ones a value of 100. Only cases with a pathogenicity score below 5 were considered to have activity in the low (2 ng/ml) pyridoxine experiment.

Yoo, J.; Lee, Y.; Kim, Y.; Rha, S. Y. & Kim, Y. (2008) 'SNPAnalyzer 2.0: a web-based integrated workbench for linkage disequilibrium analysis and association analysis', *BMC Bioinformatics* **9**:290.

SID#8 – PON-P-PANTHER

Predictions were run with PANTHER SNV scoring tool v. 1.01 (Thomas et al. 2006) by using Pathogenic-Or-Not-Pipeline (PON-P). Default values of the method were used. Pathogenic cases were given a value of 0 and neutral ones a value of 100. Only cases with a pathogenicity score below 5 were considered to have activity in the low (2 ng/ml) pyridoxine experiment.

Thomas, P. D.; Kejariwal, A.; Guo, N.; Mi, H.; Campbell, M. J.; Muruganujan, A. & Lazareva-Ulitshy, B. (2006) ‘Applications for protein sequence-function evolution data: mRNA/protein expression analysis and coding SNP scoring tools’, *Nucleic Acids Research* **34**:W645-650.

SID#9 – I-Mutant3

Predictions were based on the DDG of unfolding predicted with IMutant3 (Capriotti et al. 2008). The SVM based predictor analyzed the mutation type, the residue composition of the portion of the sequence flanking the mutation site, and, when available, the composition of the three-dimensional environment of the mutated residue and its relative solvent accessibility. The relative growth rates were computed starting from the predicted DDG values on the basis of the following conversion table:

|DDG |< 1 : 100% growth rate
1<|DDG |<2 : 75%

2<|DDG |<3 : 50%
3<|DDG |<4 : 25%
|DDG |>4 : 0%

The raw data column reported the predicted DDG and indicated whether the prediction was based on structural information (STR) or sequence information (SEQ). Growth rates at high and low pyridoxine concentration were assumed to be equal. Standard deviation was not computed.

Capriotti, E.; Fariselli, P.; Rossi, I. & Casadio, R. (2008) ‘A three-state prediction of single point mutations on protein stability changes’, *BMC Bioinformatics* **9** Suppl 2:S6.

SID#10 – I-Mutant4

Predictions were based on I-Mutant4, a method employing SVM with an RBF kernel. Input involves the protein mutation, the residue composition of the portion of the sequence flanking the mutation site, the composition of the three-dimensional environment of the mutated residue, the composition of the sequence profile in a 3-residue-long window centered in the mutated residue, and the conservation of the mutated position in multiple sequence alignment. Protein stability was considered perturbed if |DDG|>1, otherwise the mutation was considered conservative. From the binary output, relative growth rates were predicted as follows:

Perturbed stability → 0% relative growth rate

Non Perturbed stability → 100% relative growth rate

Growth rates at high and low pyridoxine concentration were assumed to be equal. Standard deviation was not computed.

SID#11 – ProTherm

Predictions were based on a statistical analysis of the correlation between mutation type and protein stability perturbation computed from the data of the current release of the ProTherm database.

Protein stability was considered perturbed if the |DDG|> 1, otherwise the mutation was considered a conservative mutation (in term of protein stability).

The raw probability for a mutation X-->Y to perturb the protein stability was computed as the ratio between the number of mutations X-->Y conducive to protein perturbation and the total number of X-->Y mutations in the data set, as derived from ProTherm, and were reported in the raw data column. The associated standard errors were evaluated with a binomial approximation and were reported in the raw data column between brackets.

The percentage relative growth rate was computed starting from the perturbing probabilities (Pp) using the following equations:

Relative growth rate (%) = 100 * (1-Pp)

StandardError[Relative growth rate] = 100* StandardError[Pp]

For some mutation types, the data reported in the ProTherm database was not sufficient to compute the perturbation probability.

SID#12 - SDM

This analysis of CBS was driven by the assumption that deleterious protein mutations act through two mechanisms: (1) by destabilizing the protein structure, or (2) by disrupting functional residues, e.g. the active site(s) or binding site(s). To predict destabilization, "SDM" (http://mordred.bioc.cam.ac.uk/~sdm/sdm.php) was used with PDB id 1JBQ. The results were given in the "raw data" column.

Functional residues were identified as those described in the publication of the crystal structure (Meier et al., EMBO J., vol 20, p 3910, 2001) and classified accordingly as: Active site, Heme binding, Dimer interface, Oxidoreductase active site (ORAS) motif, or "non-functional".

Subsequently, rules were developed to determine the effects on the catalytic rates in both low and high pyridoxine concentration, of structural destabilization and/or alteration of functional residues. For this data on CBS mutations was collected from three sources: (1) Kraus et al., Human Mut., vol 13, p 362, 1999, Table 2 and Table 5; (2) "CBS mutation database of genotype and vitamin B6 responsiveness in CBS deficiency" at http://cbs.lf1.cuni.cz/cbsdata/b6response.htm; (3) the CAGI CBS "Example data set". From these, prediction rules were defined as follows:

Rules	Catalytic rate in high [pyridoxine]	Catalytic rate in low [pyridoxine]
Severely destabilising mutations (ABS(SDM) > 3.85)	0	0
Mutating an "Active site" residue	0	0
For ABS(SDM)<3.85 and "non-functional" residue	100	80
Mutating an "ORAS motif" residue	100	100
Mutating a "Heme binding" residue	70	40
Mutations at residue 267, by analogy with residue 266	100	40
Mutations at residues 383/385, by analogy with residue 384	80	20

SID#13 – I-Mutant3.0

Predictions were performed by calculating the mean of two SVM predictors implemented in I-Mutant3.0 (http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi).

SID#14 – SwissModel, Rosetta, HMMER + PHYLIP

Cystathionine beta-synthase (CBS) mutants were scored using a sum of scores derived from structure-based energy calculations and phylogenetic considerations. Mutant scores were converted to percentage yeast growth (relative to yeast harboring wild-type human CBS) using cutoffs derived from empirical evidence (i.e., the six experimental results that were provided). All calculations were assumed to model a scenario where catalysis is not limited by pyridoxine concentration (i.e., >= 400 ng/ml). Predicted growth was then extrapolated to a low0concentration scenario based on observations from the provided experimental results.

Structure-based energy calculations were performed using Rosetta (Rohl et al. 2001) using both the ligand-bound and ligand-free dimer to model the interactions between individual components. The starting structure was a homology model of the monomeric CBS PALP domain retrieved from SwissModel (Schwede et al. 2004). The homology model was used because, unlike the crystal structure, it did not contain any chain breaks. The RMSD of the homology model relative to the crystal structure was <1.0 Å. Bond lengths and dihedral angles were idealized using Rosetta – this procedure was carried out to ensure that good (low) energies could be obtained in subsequent calculations. To prepare the dimer CBS structure, the X-ray crystal structure was used to align two CBS monomers from the homology model.

Three separate scores were derived based on Rosetta energy estimates (weighted sums of semi-empirical energies). For the wild type and each mutant, the ligand-bound monomer was used to calculate enthalpic and pseudo-entropic energies and the ligand-free dimer to calculate an enthalpic interfacial energy.

For all calculations, an ensemble of 10 structures was created for wild-type and mutant CBS. For ligand-bound monomeric CBS, Rosetta calculations were run with an initial 25 rounds of coupled backbone minimization and side-chain repacking, using conjugate gradient minimization. Next, a second 25 rounds were ran using steepest descent minimization. For ligand-free CBS dimers, the energy was calculated using coupled rigid-body, side-chain, and backbone minimization.

Calculations were performed on IBM iDataPex cluster with 500 Quad Core Intel Xeon E5472 processors, located at Johns Hopkins University's Institute for Computational Medicine. This protocol yielded four structural ensembles for each mutation: a wild-type ligand-bound monomeric ensemble, a mutant ligand-bound monomeric ensemble, a wild-type ligand-free dimeric ensemble and a mutant ligand-free dimeric ensemble. For each ensemble, the mean and standard deviation of energies was computed and converted into the scoring function given below:

DE_{res} is an enthalpic term that represents the estimated difference in mean energies of the wild type and mutant ligand-bound monomeric ensembles at the position where the mutation occurred. S.D. is the standard deviation of the wildtype and mutant ligand-bound monomeric ensembles at the position where the mutation occurred, which was interpreted as a pseudo-entropic term. DE_{int} is an enthalpic term that was estimated as the difference in the mean energies between the wild type and mutant ligand-free dimeric ensembles measured at all interfacial residues. The standard deviations of the interfacial residue energies were not used because they were all close to 0.

For phylogenetic analysis, CBS orthologs were collected from several databases: UCSC Genome Browser 46-way genome alignment, GeneCards (Rebhan et al. 1998), OMA group 45361 (Schneider et al. 2007), OrthoDB (Kriventseva et al. 2008), and PhylomeDB(Huerta-Cepas et al. 2008). Prokaryotes were not included because they mostly contained CBS paralogs, such as cysteine synthases, rather than orthologs. The HMMER hmmscan program (Eddy 1998) was used to identify the location of the pyridoxal-phosphate binding domain, known as PALP in the Pfam database (Bateman et al. 2002), in human CBS and its ortholog sequences. The PALP domain sequences were then aligned with CLUSTALW (Larkin et al. 2007) (gap-opening penalty 15.0, gap extension penalty 0.3). PALP sequences with fewer than 250 aligned residues were identified by eye and dropped from the alignment. The final alignment contained 45 sequences.

Next, the maximum likelihood tree-building method *promlk* in the PHYLIP suite (Felsenstein 1989) was used, with default parameters, to build a rooted phylogenetic tree of the aligned sequences. The sequences were sorted by the ordering of leaves in the phylogenetic tree, with the human sequence positioned at the top of the list. Inspired by the analysis of Marini *et al.* (Marini et al. 2010), each alignment column was visually inspected to see how far down the list the human reference amino acid residue was conserved. While the bread mold CBS sequence appeared to be at the most optimal depth for thresholding conservation based on agreement with the given experimental examples, each column was reviewed at length and the importance of conservation at some positions was downweighted in certain putative outlier species, including insects (as suggested by Marini et al.) and rhesus macaque (in which likely deleterious prolines are observed in positions 118, 124 and 140). Each mutation was assigned a binary conservation score of 1 if it occurred in what was decided to be a sufficiently conserved alignment column and 0 otherwise.

The final raw scoring function for each mutation was:

Mutations were ordered based on their raw scores, which were partitioned into no activity, intermediate activity and wild-type activity. Thresholds were chosen based on the experimental results of six mutations provided. Forty-one mutations were predicted to have either no activity or wild-type activity, and these were predicted to not be distinguishable on the basis of the yeast complementation assay. These mutations were assigned standard deviations of zero. Based on the experimental results provided, the remaining 10 intermediate-activity mutations were predicted to have activities ranging from 70 to 100%. These 10 mutations were therefore assigned a mean activity of 85% and standard deviation of 15%.

Eddy, S. R. (1998), 'Profile hidden Markov models.', *Bioinformatics* **14**(9), 755-763.

Felsenstein, J. (1989), 'PHYLIP – Phylogeny Inference Package (Version 3.2)', *Cladistics* **5**, 164-166.

Huerta-Cepas, J.; Bueno, A.; Dopazo, J. & Gabaldón, T. (2008), 'PhylomeDB: a database for genome-wide collections of gene phylogenies', *Nucleic Acids Research* **36**(suppl 1), D491-D496.

Kriventseva, E. V.; Rahman, N.; Espinosa, O. & Zdobnov, E. M. (2008), 'OrthoDB: the hierarchical catalog of eukaryotic orthologs', *Nucleic Acids Research* **36**(suppl 1), D271-D275.

Larkin, M.; Blackshields, G.; Brown, N.; Chenna, R.; McGettigan, P.; McWilliam, H.; Valentin, F.; Wallace, I.; Wilm, A.; Lopez, R.; Thompson, J.; Gibson, T. & Higgins, D. (2007), 'Clustal W and Clustal X version 2.0', *Bioinformatics* **23**(21), 2947-2948.

Marini, N. J.; Thomas, P. D. & Rine, J. (2010), 'The Use of Orthologous Sequences to Predict the Impact of Amino Acid Substitutions on Protein Function', *PLoS Genet* **6**(5), e1000968.

Rebhan, M.; Chalifa-Caspi, V.; Prilusky, J. & Lancet, D. (1998), 'GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support', *Bioinformatics* **14**(8), 656-664.

Rohl, C. A.; Strauss, C. E.; Misura, K. M. & Baker, D. (2004), ‘Protein Structure Prediction Using Rosetta’, *Methods in Enzymology* **383**, 66-93.

Schneider, A.; Dessimoz, C. & Gonnet, G. H. (2007), 'OMA Browser—Exploring orthologous relations across 352 complete genomes', *Bioinformatics* **23**(16), 2180-2182.

Schwede, T.; Kopp, J.; Guex, N. & Peitsch, M. C. (2003), 'SWISS-MODEL: an automated protein homology-modeling server', *Nucleic Acids Research* **31**(13), 3381-3385.

SID#15 - SIFT

Sequence submission, default settings.

SID#16 - Metapredictor

CBS variants were classified by linear regression. *E. coli* activity for 36 CBS mutations was collected from literature and used as the training set. The following features were used in the prediction:

(1) MutPred score (Li et al. 2009)

MutPred is a random forest -based computational method for predicting the effect of amino acid substitutions, based on protein sequence. It models changes of structural features, such as secondary structure and stability changes, and functional sites, such as catalytic residues and post-translational modifications,

between wild-type and mutant sequences. These changes are expressed as probabilities of gain or loss of structure and function, are used to classify disease mutations from HGMD and neutral mutations from SWISS-Prot database

by random forest. MutPred can provide insight into the specific molecular mechanism responsible for the disease state.

(2) SIFT score (Ng and Henikoff 2001)

(3) gain of catalytic residue predicted from sequence (Xin et al. 2010)

(4) loss of catalytic residue predicted from sequence

(5) gain of catalytic residue predicted from structure

(6) loss of catalytic residue predicted from structure

(7) predicted stability changes

(8) confidence index of (7)

Li, B.; Krishnan, V. G.; Mort, M. E.; Xin, F.; Kamati, K. K.; Cooper, D. N.; Mooney, S. D. & Radivojac, P. (2009) ‘Automated inference of molecular mechanisms of disease from amino acid substitutions’, *Bioinformatics* **25**(21), 2744-2750.

Ng, P. C. & Henikoff, S. (2001) ‘Predicting deleterious amino acid substitutions’, Genome Res. **11**(5) 863-874.

Xin, F.; Myers, S.; Li, Y. F.; Cooper, D. N.; Mooney, S. D. & Radivojac, P. (2010) ‘Structure-based kernels for the prediction of catalytic residues and their involvement in human inherited diseases’, *Bioinformatics* **26**(16), 1975-1982.

SID#18 – SDM

CBS residues that were potentially interacting with a ligand or the other protein chain in the PDB file 1JBQ were identified using two distance cut-offs: VDW radii + 0.5 Å, or 6.05 Å (maximum length of water-mediated hydrogen bond). Subsequently, a set of known mutations and known phenotypes from the literature ((1) Kraus et al., Human Mut., vol 13, p 362, 1999, Table 2 and Table 5; (2) CAGI example data).

SDM was ran on the "known" mutations and the CAGI prediction data set (where structure information available) using chain A of PDB file 1JBQ for input, to predict the stability change upon mutation.

Using the genotype, "proteotype" (i.e. SDM predictions and interaction information) and phenotype information from the "known" mutations, rules were developed for predicting the effects of novel mutations as follows:

Proteotype	High pyridoxine value	Low pyridoxine value
SDM < -3	0	0
SDM > 1.68	0	0
No interactions, SDM between -3 and 1.68	103	79
Interacts with Heme	70	37
Interacts with PLP	0	0
Interacts with other monomer, moderate SDM < -1	0	0

SID#19 – Align-GVGD http://agvgd.iarc.fr

A CBS multiple sequence alignment was prepared containing 14 metazoan sequences from human through starlet anemone. All of the missense substitutions were scored using Align-GVGD with this alignment. Data from the example variants were used in a linear regression of % growth vs Align-GVGD grade to get an idea of correlation and variability. Since Align-GVGD normally gives 7 grades, expected % growth was evenly spaced from 0 for the most severe grade to 100 for the least severe grade (steps of 16.7%). While there was some evidence for a difference in the regression for the example variants, this was not judged sufficient to be acted on. Therefore, no differentiation was made between low and high pyridoxine. For the standard deviation, a value of 0.6 x the 95% confidence interval of the Y-intercept of linear regression was used. This resulted in the same standard deviation for all variants. However, the lower bound of the confidence interval should be zero for grades with point estimates less than one SD from zero.

SID#20 – Align-GVGD + SIFT

A CBS multiple sequence alignment was prepared containing 14 metazoan sequences from human through starlet anemone. All of the missense substitutions were scored using Align-GVGD with this alignment. All of the missense substitutions were scored using SIFT under default parameters. The 7 Align-GVGD grades were converted to an ordered series of 0 (severe) to 6 (benign). SIFT scores were converted to -2 (SIFT=0.000), 0 (0.00<SIFT<0.055), +e (0.055<SIFT<1.00). The Align-GVGD grading and SIFT grading were added together. To this total, 2 was added so that the most severe grade would have a score of 7. Data from the example variants were used in a linear regression of % growth vs combined grade to get an idea of correlation and variability. Since the combined grade system gave 11 grades, expected % growth was evenly spaced from 0 for the most severe grade to 100 for the least severe grade in steps of 10%. Though not very different from this simple scoring, linear regression results were ignored. While there was some evidence for a difference in the regression for the example variants, this was not judged sufficient to be acted on. Therefore, no differentiation was made between low and high pyridoxine. For the standard deviation, a value of 0.5 x the 95% confidence interval of the Y-intercept of linear regression was used. This resulted in the same standard deviation for all variants. However, the lower bound of the confidence interval should be zero for grades with point estimates less than one SD from zero.

CAGI2

SID#16 – SNAP

Predictions were made using SNAP (Screening for Non-Acceptable Polymorphisms), a neural network based method for prediction of functional effects of single amino acid substitutions. By default, for each submitted substitution, SNAP reports a single binary prediction (Neutral/Non-neutral), which is associated with an RI (reliability index, range 0-9) and a value of "expected accuracy" (0-100%; in testing, the accuracy of SNAP predictions at the given RI). The SNAP RI scores are computed by taking a mean of ten neural nets trained on different data sets – the raw score. Unfortunately, none of these values map directly to percentage change in yeast growth rate (and there is no way to differentiate the circumstances such as co-factor concentrations). We have previously observed that increased scores correlate fairly well with the severities of protein function change. Thus, we used the raw score as reference and followed the steps below in making predictions (our "standard deviation" is computed from the variation in predictions on different networks and is the same for both co-factors):

1. For unreliable predictions (-10<raw score <=10) we report 90% growth rate.

2. For neutral predictions (raw score <= -10) we report no change (100% growth rate) for both cofactor concentrations.

3. For non-neutral predictions we assumed a decrease of 1% growth rate for each 1-point drop in the mean of the predictions. We assumed that lower concentrations of cofactors would have a more deleterious effect on already deleterious mutants, as was observed in previously reported CBS data. Thus, we report a 5% lower drop for lower concentrations.

Overall, it is important to note that while we can attempt to adapt SNAP output to the various prediction challenges, accounting for all possible variations on the theme is nearly impossible. Thus, it is more realistic to leave the interpretation of mutation neutrality to the expert experimentalists working with the protein at hand.

SID#19 – graph analysis + KNN

The algorithms used involved graphs analysis and K-Nearest Neighbor. The estimation of the protein's activity was performed in two steps, namely a *classification* and an *estimate*. During the classification we aimed to determine if the protein's mutant was active or not. This choice was performed assessing our novel metrics. Clearly, if protein is inactive, the activity percentage is set to 0, otherwise we start the estimation step. Exploiting the same novel metrics, we computed the protein activity using the most similar CBS mutants in a training set, according to K-NN definition. Our metrics were based on a graph representation of the protein, where nodes corresponds to amino acids, and edges are the chemical bond type (e.g. Hydrogen bond, pi-cation bond). Using statistical methods, we were able to estimate the *relevance* of each residue in the network. Therefore, a direct comparison of the wild-type nodes relevance with the mutant ones provided an index of similarity between the two graphs.

SID#20 – graph analysis, KNN, regression + majority voting

The algorithms used involved graphs analysis, K-Nearest Neighbor, regression, and majority method. The estimation of the protein's activity was performed in two steps, namely a *classification* and an *estimate*. During the classification we aimed to determine if the protein's mutant was active or not. This choice was performed assessing a set of five well known mutation analyzers plus a novel metric. With such majority voting we understand if a protein is inactive, and we can thus set to 0 its activity. Otherwise we started the estimation step. Exploiting the same novel metric, we computed the protein activity using the most similar CBS mutants in a training set, according to K-NN definition. Such estimate was improved in the end, by mapping its value on a regression curve. Our metric was based on a graph representation of the protein, where nodes correspond to amino acids, and edges are the chemical bond type (e.g. hydrogen bond, pi-cation bond). Using statistical methods, we were able to estimate the *relevance* of each residue in the network. Therefore, a direct comparison of the wild-type node's relevance with the mutant ones provided an index of similarity between the two graphs.

SID#21 – graph analysis + regression

The algorithms involved concern graphs analysis and regression. The estimation of the protein's activity was performed in two steps, namely a *classification* and an *estimate*. During the classification we aimed to determine if the protein's mutant was active or not. This choice was performed assessing our novel metric. Clearly, if the protein is inactive, the activity percentage is set to 0, otherwise we started the estimation step. Exploiting the same novel metrics, we computed the protein activity. Such estimate was improved in the end, mapping its value on a regression curve. Our metric was based on a graph representation of the protein, where nodes correspond to amino acids, and edges are the chemical bond type (e.g. Hydrogen bond, pi-cation bond). Using statistical methods, we were able to estimate the *relevance* of each residue in the network. Therefore, a direct comparison of the wild-type nodes relevance with the mutant ones provided an index of similarity between the two graphs.

SID#22 – graph analysis + KNN

The algorithms used involved graph analysis and K-Nearest Neighbor. The estimation of the protein's activity was performed in two steps, namely a *classification* and an *estimate.* During the classification we aimed to determine if the protein's mutant was active or not. This choice was performed assessing our novel metric. Clearly, if protein is inactive, the activity percentage is set to 0, otherwise we started the estimation step. Exploiting the same novel metric, we computed the protein activity using the most similar CBS mutants in a training set, according to 3-NN definition. Our metric was based on a graph representation of the protein, where nodes correspond to aminoacids, and edges are the chemical bond type (e.g. Hydrogen bond, pi-cation bond). Using statistical methods, we were able to estimate the *relevance* of each residue in the network. Therefore, a direct comparison of the wild-type nodes relevance with the mutant ones provided an index of similarity between the two graphs.

SID#23 – graph analysis, KNN, regression + majority voting

The algorithms used involved graphs analysis, K-Nearest Neighbor, regression, and majority method. The estimation of the protein's activity was performed in two steps, namely a *classification* and an *estimate*. During the classification we aimed to determine if the protein's mutant was active or not. This choice was performed assessing a set of five well known mutation analysers. With such majority voting we understand if a protein is inactive, and we can thus set to 0 its activity. Otherwise we started the estimation step. Exploiting the same novel metric, we computed the protein activity using the most similar CBS mutants in a training set, according to K-NN definition. Such estimate was improved in the end, mapping its value on a regression curve. Our metric was based on a graph representation of the protein, where nodes correspond to aminoacids, and edges are the chemical bond type (e.g. Hydrogen bond, pi-cation bond). Using statistical methods, we were able to estimate the *relevance* of each residue in the network. Therefore, a direct comparison of the wild-type nodes relevance with the mutant ones provided an index of similarity between the two graphs.

SID#24 – random (D100 roll)

The algorithm used was D100 roll. The estimation of the protein's activity was performed in two steps, namely a *classification* and an *estimate*. During the classification we aimed to determine if the protein's mutant was active or not. This choice was performed rolling a 100-

sided die. If the result was below 50, then the mutation was predicted pathogenic, and therefore the protein activity was 0.

During the estimate, we rolled again the dice, and we added the result to a minimum value M. In this way, the predicted activity will be bounded in [M, M+100].

SID#25 – FoldX3.b4

Predictions were based on the pdb structure 1jbq. Using the FoldX3.b4 algorithm, the structure was first repaired (RepairPDB command) and subsequently all the mutants were generated using the BuildModel command. FoldX reports the change in thermodynamic stability of the protein as a result of mutation as a so-called DDG (change in free energy difference between wild type and mutant, kcal.mol^{-1}). In addition, the AnalyseComplex command was used to obtain the effect of the mutation on the free energy of dimer formation. The largest DDG value was taken to report the most severe effect. Moreover, when the wild type residue was making a contact with the pyridoxal 5-phosphate or heme groups, an artificial penalty of 4 kcal.mol^{-1}was applied, since FoldX has not parameterisation for these groups. Based on the training data, the destabilisation of the mutants was converted into growth rates as shown in the table.

DDG	growth_rate_400ng/ml_pyr	growth_rate_2ng/ml
> 3	severely_impaired (0)	severely_impaired
2-3	impaired (70)	impaired(60)
1.5-2	impaired (80)	impaired (70)
1-1.5	none (101)	none (93)
<1	none (120)	none (110)

SID#26 – Evolutionary Action

This submission used the Evolutionary Action (EA) method prior to its publication (Katsonis & Lichtarge 2014). In contrast to the pre-mature version used in CAGI1, this version only had minor technical differences to the published EA version. Briefly, EA does not use any training since it relies on a formal equation of the genotype-phenotype relationship. This equation states that the fitness effect of a mutation equals the product of the sensitivity of the mutated position with the magnitude of the change. The sensitivity of the position is calculated by quantifying the correlation of the residue variations with phylogenetic branching within an alignment of homologous sequences (Lichtarge et al. 1996; Mihalek et al. 2004; Lichtarge and Wilkins 2010). The magnitude of the change is calculated from substitution likelihood according to numerous sequence alignments for the given context (strata of sensitivity of the position, and optionally additional stratification based on structural features). The product is then normalized to represent the percentile rank of each variant within the protein in the scale of 0 (benign) to 100 (pathogenic). The EA scores are available for all human variants at: http://mammoth.bcm.tmc.edu/EvolutionaryAction.

The loss of CBS activity at high co-factor concentration was set to 100−EA. At low co-factor concentration, the EA scores were scaled to yield lower CBS activities (as guided by the test data), such that an EA of 70 will yield 10% CBS activity instead of 30% (linear scaling without changing the extremes, so, EA of 0 and 100 will still yield 100% and 0% CBS activity, respectively).

Katsonis, P. & Lichtarge, O. (2014) 'A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness', *Genome Research* **24**(12):2050-8.

SID#27 – PON-P (metapredictor)

The Pathogenic-or-Not-Pipeline (PON-P) provides high quality prediction of the effect of amino acid substitutions, by: (1) aggregating predictions of five predictors and deducing a consensus prediction; (2) determining the reliability of the consensus prediction and based on that, classifying the cases as neutral, pathogenic, or unclassified variant (UV). It is a machine learning approach based on Random Forest.

For training the PON-P predictor, we built a pathogenic (positive) dataset of 14,610 amino acid substitutions obtained by manual curation from the PhenCode database[1] (downloaded in June 2009), IDbases[2] and 16 individual Locus Specific Databases (LSDBs). The selected cases were annotated as disease-causing either in SwissVar[3] or in the LSDB. The negative (neutral) dataset cases was obtained from dbSNP build 131 by filtering variations with population frequency > 0.1 and with chromosome count _50.

PON-P was tested on an independent dataset extracted from the Protein Mutant Database (PMD)[4]. A random forest is classification and regression trees (CARTs)[5].

To construct a random forest, CARTs are grown on a bootstrap set. At each node of each tree, a random subset of attributes is selected as candidate attributes for node splitting. The number of the attributes is predefined. Each tree is grown fully. Since the random sampling is done with replacement, approximately 33% of the data was not be used in tree construction. This data, referred to as the out-of-bag data, was passed down the tree, and used to obtain the out-of-bag estimate of the performance of the tree. This scheme implies that a specific case in the dataset will be used in performance evaluation in approximately 33% of the trees in the random forest. For each class, the out-of-bag error estimate was computed as follows: each case in the class is passed down those trees for which it is out of bag, and the class that gets the majority vote is the predicted class. The predictions are aggregated for the whole dataset, and the error rate is the generalization error of the random forest for the class.

Using the PON-P metaserver, we obtained predictions from four tolerance methods, PhD-SNP[7] (version 2.0.6), SIFT[8] (4.0.3), PolyPhen-2[9] (2.0.22) and SNAP[10] (1.0.8), as well as from a stability predictor, I-Mutant[11] (3.0.8). To select optimally relevant features we adopted greedy backward elimination approach. PON-P returns a probability value which indicates the likelihood of the protein being functionally impaired by the variation on which the prediction is made as well as an associated standard error. Prediction was made on the 84 missense variations in the CAGI CBS dataset using these two outputs. If we denote the probability output of PON-P by *x*, and the associated standard error by *se*, we converted these to growth rates as follows:

High pyridoxine conc.

Prediction (*hp*) = *x* x 110

SD = *se* if *x*>0, otherwise 0

Low pyridoxine conc.

Prediction (*lp*) = *hp* – (0.08 x *hp*) if *hp*>99, *hp* – (0.3 x *hp*) if *hp* >40, 0 otherwise

SD = *se* if *lp*>0, otherwise 0

1. Giardine, B., Riemer, C., Hefferon, T., Thomas, D., Hsu, F., Zielenski, J., Sang, Y., Elnitski, L., Cutting, G., Trumbower, H. et al. (2007). PhenCode: connecting ENCODE data with mutations and phenotype. *Human Mutation* **28**, 554-562.

2. Piirilä, H., Valiäho, J., Vihinen, M. (2006). Immunodeficiency mutation databases (IDbases). *Hum. Mutat.* **27**, 1200-1208.

3. Yip, Y.L., Scheib, H., Diemand, A.V., Gattiker, A., Famiglietti, L.M., Gasteiger, E., Bairoch, A. (2004). The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. *Hum. Mutat.* **23**, 464-470.

4. Nishikawa, K., Ishino, S., Takenaka, H., Norioka, N., Hirai, T., Yao, T., Seto, Y. (1994). Constructing a protein mutant database. *Protein Eng.* **7**, 733.

5. Breiman, L. (2001). Random Forests. *Machine Learning* **45**, 5-32.

6. Breiman, L., Friedman, J., Ohlsen, R., Stone, C. (1984). Classification and Regression Trees (Wadsworth, Belmont, CA: Chapman and Hall/CRC).

7. Capriotti, E., Calabrese, R., Casadio, R. (2006). Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* **22**, 2729-2734.

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

9. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248-249.

10. Bromberg, Y., Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.* **35**, 3823-3835.

11. Capriotti, E., Fariselli, P., Rossi, I., Casadio, R. (2008). A three-state prediction of single point mutations on protein stability changes. *BMC Bioinformatics* **9** Suppl 2, S6.

SID#28 – PON-P (metapredictor)

This is a modification of submission SID#27, thus only differences are indicated.

PON-P returns a probability value which indicates the likelihood of the protein being functionally impaired by the variation on which a prediction is made as well as an associated standard error.

Predictions were made on the 84 missense variations in the CAGI CBS dataset using these two outputs. If we denote the probability output of PON-P by *x*, and the associated standard error by *se*, we converted these to growth rates as follows:

High pyridoxine conc.

Prediction (*hp*) = *x* x 110 if x≥023, 0 otherwise

SD = *se* if *x*>0, otherwise 0

Low pyridoxine conc.

Prediction (*lp*) = *hp* – (0.08 x *hp*) if *hp*>99, *hp* – (0.3 x *hp*) if *hp* >40, 0 otherwise

SD = *se* if *lp*>0, otherwise 0

SID#29 – Metapredictor (SIFT + BLOSUM)

Random forest classifier. Trained by last year's 2010 CBS dataset.

SID#30 – Metapredictor (SIFT)

Random forest classifier. SIFT used as a threshold cutoff.

SID#32 – Metapredictor (SIFT + BLOSUM)

Random forest classifier. Trained by last year's 2010 CBS dataset.

SID#33 – Metapredictor (MutPred, SIFT, PolyPhen2)

The following features were used to represent the functional impact of each CBS mutation:

(1) MutPred score

(2) SIFT score

(3) PolyPhen 2 score

(4) predicted stability (prediction plus reliability index)

(5) predicted catalytic residue (gain and loss)

The relative growth rate was modeled from 56 training data points as a linear model of the above features. Two models were built for two experimental conditions, e.g., high pyridoxine and low pyridoxine. For the high pyridoxine, the linear model could explain 32 percent of the variance among observed growth rates, and for the low pyridoxine, the same measure reduced to 25 percent. In reporting results, negative predictions were bounded to zero to agree with the actual experimental results.

SID#34 – PolyPhen-2 v2.1.0

Predictions were obtained using Polyphen-2 v2.1.0 software available for download here: http://genetics.bwh.harvard.edu/pph2/dokuwiki/downloads

The method is described in detail in:

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. Nat Methods 7(4):248-249 (2010)

Copies of the paper and supplement can be downloaded here:

http://genetics.bwh.harvard.edu/pph2/dokuwiki/about

Version 2.1.0 of the software is described here:

http://genetics.bwh.harvard.edu/pph2/dokuwiki/news

Posterior probabilities obtained from the software were converted to enzyme activity estimates by fitting a linear regression to 23 variants with known qualitative effects from UniProtKB and using fitted model to derive activities for all 84 mutations.

HumVar model (for putatively strongly deleterious mutations) was used for predictions under high-pyridoxine concentrations; HumDuv model (highest sensitivity) was used for predictions under low-pyridoxine concentrations.

SID#35 – SNPs3D

The following method was used for prediction of the relative growth rate in yeast for the missense variants in CBS, at high co-factor (pyridoxine) concentration (400ng/ml) and at low co-factor concentration (2ng/ml).

Background:

In previous work, we have established that the majority of monogenic disease mutations involve significant destabilization of protein structure [1]. We use a support vector machine (SVM) to detect destabilization effects, and a second support vector machine to detect all types of effect on *in vivo* protein function [2]. Both SVMs were trained on monogenic disease missense mutations and a control set of interspecies differences [1], [2]. The structure method uses 15 features based on detailed structural effects, including reduction in hydrophobic area, overpacking, backbone strain, and loss of electrostatic interactions. The general method uses five conservation and residue substitution type features derived from the family multiple sequence alignment.

The standard versions of these SNPs3D methods were used in the CAGI 2010 CBS challenge. A high fraction of those mutations were predicted to be destabilizing, and predictions from the two methods were generally consistent. Both methods were effective in identifying mutations with zero growth, and most mutations with 100% growth, but there were more false positives than expected – mutations where growth is intermediate or near normal but which were assigned as having a high deleterious impact by the SVMs. The most likely interpretation of this observation is that relatively low levels of CBS activity are sufficient for near full growth rate in yeast, compared with levels typically causing disease for monogenic disease proteins as a whole. In general, the experimental data and the predictions were consistent with improved growth at higher cofactor concentration being a consequence of increased stability conferred by cofactor binding compensating for the destabilizing effect of mutations. Further, it appears that cofactor rescue is only possible when the original destabilization is relatively mild. These observations led to the two prediction methods used in CAGI 2011.

Recalibrated SNPs3D SVM method:

The examination of our results for CBS in CAGI 2010 led to three observations: (1) The effective threshold confidence score for zero growth mutations in the profile SVM is approximately -2.0, rather than the expected zero. (2) Confidence scores >0.5 were observed for mutations which had near normal growth rates in yeast at the lower cofactor concentration, and were therefore unaffected by additional cofactor. (3) Mutations where growth rate is affected by cofactor had a range of intermediate scores – our method is not sensitive enough to resolve this level of subtlety, nor would we expect it to be.

Based on these observations, we re-parameterized the method specifically for CBS, treating high negative scores as zero growth, high positive scores as 100% growth, and intermediate scores as 100% for high cofactor concentration and 50% for low cofactor concentration.

References:

[1] Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol. 2005 Oct 21;353(2):459-73

[2] Yue P, Moult J. Identification and analysis of deleterious human SNPs. J Mol Biol. 2006 Mar 10;356(5):1263-74.

SID#36 – SNPs3D + annotation

The following methods were used for prediction of the relative growth rate in yeast for the missense variants in CBS, at high co-factor (pyridoxine) concentration (400ng/ml) and at low co-factor concentration (2ng/ml).

Background:

In previous work, we have established that the majority of monogenic disease mutations involve significant destabilization of protein structure [1]. We use a support vector machine (SVM) to detect destabilization effects, and a second support vector machine to detect all types of effect on *in vivo* protein function [2]. Both SVMs were trained on monogenic disease missense mutations and a control set of interspecies differences [1], [2]. The structure method uses 15 features based on detailed structural effects, including reduction in hydrophobic area, overpacking, backbone strain, and loss of electrostatic interactions. The general method uses five conservation and residue substitution type features derived from the family multiple sequence alignment.

The standard versions of these SNPs3D methods were used in the CAGI 2010 CBS challenge. A high fraction of those mutations were predicted to be destabilizing, and predictions from the two methods were generally consistent. Both methods were effective in identifying mutations with zero growth, and most mutations with 100% growth, but there were more false positives than expected – mutations where growth is intermediate or near normal but which were assigned as having a high deleterious impact by the SVMs. The most likely interpretation of this observation is that relatively low levels of CBS activity are sufficient for near full growth rate in yeast, compared with levels typically causing disease for monogenic disease proteins as a whole. In general, the experimental data and the predictions were consistent with improved growth at higher cofactor concentration being a consequence of increased stability conferred by cofactor binding compensating for the destabilizing effect of mutations. Further, it appears that cofactor rescue is only possible when the original destabilization is relatively mild. These observations led to the two prediction methods used in CAGI 2011.

Methods using SNPs3D SVM scores together with other information:

Since these are disease mutations, there is extensive database and literature information available. We therefore also developed a method which combined the SVM outputs with annotation from these sources, in some instances augmented by inspection of the structure. While it would be possible to fully automate this method, in fact, the integration of information was performed manually. In addition to the SVMs, data sources included partial disease severity information and patient responsiveness to cofactor for some mutations, and very limited information on mutation growth in E.coli and yeast [3], [4]. There is considerable inconsistency between these data sources, so that definitive answers were generally not available.

References:

[1] Yue P, Li Z, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. J Mol Biol. 2005 Oct 21;353(2):459-73

[2] Yue P, Moult J. Identification and analysis of deleterious human SNPs. J Mol Biol. 2006 Mar 10;356(5):1263-74.

[3] CBS mutation database (http://cbs.lf1.cuni.cz/index.php)

[4] Kim CE, Gallagher PM, Guttormsen AB, Refsum H, Ueland PM, Ose L, Folling I, Whitehead AS, Tsai MY, Kruger WD. Functional modeling of vitamin responsiveness in yeast: a common pyridoxine-responsive cystathionine beta-synthase mutation in homocystinuria. Hum Mol Genet. 1997 Dec;6(13):2213-21.s

SID#50 – SNAP baseline

SID#51 – SIFT baseline