PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Community-wide evaluation of methods for predicting the effect of mutations on protein–protein interactions

Rocco Moretti,[1] Sarel J. Fleishman,[2] Rudi Agius,[3] Mieczyslaw Torchala,[3] Paul A. Bates,[3] Panagiotis L. Kastritis,[4] João P. G. L. M. Rodrigues,[4] Mikaël Trellet,[4] Alexandre M. J. J. Bonvin,[4] Meng Cui,[5] Marianne Rooman,[6] Dimitri Gillis,[6] Yves Dehouck,[6] Iain Moal,[7] Miguel Romero-Durana,[7] Laura Perez-Cano,[7] Chiara Pallara,[7] Brian Jimenez,[7] Juan Fernandez-Recio,[7] Samuel Flores,[8] Michael Pacella,[9] Krishna Praneeth Kilambi,[10] Jeffrey J. Gray,[10,11] Petr Popov,[12,13] Sergei Grudinin,[12,13] Juan Esquivel-Rodríguez,[14] Daisuke Kihara,[14,15] Nan Zhao,[16] Dmitry Korkin,[16] Xiaolei Zhu,[17] Omar N. A. Demerdash,[17] Julie C. Mitchell,[17] Eiji Kanamori,[18] Yuko Tsuchiya,[19] Haruki Nakamura,[20] Hasup Lee,[21] Hahnbeom Park,[21] Chaok Seok,[21] Jamica Sarmiento,[22] Shide Liang,[22] Shusuke Teraguchi,[22] Daron M. Standley,[22] Hiromitsu Shimoyama,[23] Genki Terashi,[23] Mayuko Takeda-Shitaka,[23] Mitsuo Iwadate,[24] Hideaki Umeyama,[24] Dmitri Beglov,[25] David R. Hall,[25] Dima Kozakov,[25] Sandor Vajda,[25] Brian G. Pierce,[26] Howook Hwang,[26] Thom Vreven,[26] Zhiping Weng,[26] Yangyu Huang,[27] Haotian Li,[27] Xiufeng Yang,[27] Xiaofeng Ji,[27] Shiyong Liu,[27] Yi Xiao,[27] Martin Zacharias,[28] Sanbo Qin,[29] Huan-Xiang Zhou,[29] Sheng-You Huang,[30,31] Xiaoqin Zou,[30,31] Sameer Velankar,[32] Joël Janin,[33] Shoshana J. Wodak,[34,35] and David Baker[1,36]*

[1] Department of Biochemistry, University of Washington, Seattle, Washington 98195

[2] Department of Biological Chemistry, Weizmann Institute of Science, Rehovot 76100, Israel

[3] Biomolecular Modelling Laboratory, Cancer Research UK London Research Institute, London, WC2A 3LY, United Kingdom

[4] Bijvoet Center for Biomolecular Research, Faculty of Science - Chemistry, Utrecht University, Padualaan 8, 3584 CG, Utrecht, The Netherlands

[5] Department of Physiology and Biophysics, Virginia Commonwealth University, Richmond, Virginia 23298

[6] Department of BioModelling, BioInformatics and BioProcesses, Université Libre de Bruxelles (ULB), 1050 Brussels, Belgium

[7] Joint BSC-IRB Research Program in Computational Biology, Life Sciences Department, Barcelona Supercomputing Center, C/Jordi Girona 29, 08034 Barcelona, Spain

[8] Department of Cell and Molecular Biology, Uppsala University, Uppsala, 75124, Sweden

[9] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, Maryland 21218

[10] Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, Maryland 21218

[11] Program in Molecular Biophysics, Johns Hopkins University, Baltimore, Maryland 21218

[12] NANO-D, INRIA Grenoble-Rhone-Alpes Research Center, 38334 Saint Ismier Cedex, Montbonnot, France

[13] CNRS, Laboratoire Jean Kuntzmann, BP 53, Grenoble Cedex 9, France

[14] Department of Computer Science, Purdue Universit, West Lafayette, Indiana 47907

[15] Department of Biological Sciences, Purdue University, West Lafayette, Indiana 47907

[16] Informatics Institute and Department of Computer Science, University of Missouri-Columbia, MO, 65211

[17] Departments of Mathematics and Biochemistry, University of Wisconsin, Madison, Wisconsin 53706

[18] Japan Biological Informatics Consortium, Tokyo 135-0064, Japan

[19] Division of Life Sciences, Graduate School of Humanities and Sciences, Ochanomizu University, Tokyo 122-8610, Japan

[20] Institute for Protein Research, Osaka University, Osaka 565-0871, Japan

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: National Science Foundation; grant number: DBI-0845196 to D.K; Grant sponsor: National Science Foundation; grant number: IOS-1126992 to N.Z.; Grant sponsor: Spanish Ministry of Science; grant number: BIO2010–22324 to J. F.-R; Grant sponsor: Defense Threat Reduction Agency; grant number: HDTRA1-10-0040 to D.B..

*Correspondence to: Prof. David Baker, Molecular Engineering and Sciences, Box 351655, 4000 15th Ave NE, Seattle, WA 98195, USA. E-mail dabaker@uw.edu

Received 4 March 2013; Revised 13 June 2013; Accepted 18 June 2013
Published online 10 July 2013 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24356

[21] Department of Chemistry, Seoul National University, Seoul 151–747, Korea
[22] Systems Immunology Lab, WPI Immunology Frontier Research Center (IFReC), Osaka University, 3-1 Yamadaoka, Suita, Osaka 565–0871, Japan
[23] School of Pharmacy, Kitasato University, 108-8641 Japan
[24] Department of Biological Sciences, Faculty of Science and Engineering, Chuo University, Tokyo, 192-0393 Japan
[25] Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215
[26] Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, Massachusetts 01605
[27] Huazhong University of Science and Technology, 430074, Wuhan, China
[28] Physics Department, Technical University Munich, 85748, Garching, Germany
[29] Department of Physics and Institute of Molecular Biophysics, Florida State University, Tallahassee, Florida 32306
[30] Department of Physics and Astronomy, Dalton Cardiovascular Research Center, Informatics Institute, University of Missouri-Columbia; Columbia, Missouri 65211
[31] Department of Biochemistry, Dalton Cardiovascular Research Center, Informatics Institute, University of Missouri-Columbia; Columbia, Missouri 65211
[32] European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, United Kingdom
[33] IBBMC, Université Paris-Sud, 91405, Orsay, France
[34] Department of Biochemistry, University of Toronto, Ontario, M5S 1A8, Canada
[35] Hospital for Sick Children, 555 University Avenue, Toronto, Ontario M5K 1X8, Canada
[36] Howard Hughes Medical Institute, University of Washington, Seattle, Washington 98195

ABSTRACT

Community-wide blind prediction experiments such as CAPRI and CASP provide an objective measure of the current state of predictive methodology. Here we describe a community-wide assessment of methods to predict the effects of mutations on protein–protein interactions. Twenty-two groups predicted the effects of comprehensive saturation mutagenesis for two designed influenza hemagglutinin binders and the results were compared with experimental yeast display enrichment data obtained using deep sequencing. The most successful methods explicitly considered the effects of mutation on monomer stability in addition to binding affinity, carried out explicit side-chain sampling and backbone relaxation, evaluated packing, electrostatic, and solvation effects, and correctly identified around a third of the beneficial mutations. Much room for improvement remains for even the best techniques, and large-scale fitness landscapes should continue to provide an excellent test bed for continued evaluation of both existing and new prediction methodologies.

Proteins 2013; 81:1980–1987.

Key words: CAPRI; hemagglutinin; binding; deep mutational scanning; yeast display.

INTRODUCTION

Protein–protein interactions are crucial in biology.[1–3] Understanding the thermodynamics of protein–protein interactions is important for quantitative understanding of biological function and for enabling the design of proteins, small molecules and other compounds to modulate these interactions.[4,5] A large number of computational methods have been developed to predict protein–protein binding affinity.[6–9]

Blind community-wide tests of computational methods provide a means to objectively assess the current state of the art and identify potentially promising approaches. CASP has actively evaluated protein structure prediction methodology, and CAPRI has evaluated protein–protein docking methodology,[10–13] but there has been no similar test of methods for predicting the effects of mutation on protein–protein interactions.

Here we describe the results of a community wide test of methods for evaluating the effect of mutations on protein–protein interaction affinity. This test employed two comprehensive datasets on the effects of every point mutant on the enrichment under yeast display selection of two designed protein binders of influenza hemagglutinin (HA).

MATERIALS AND METHODS

Description of data

Enrichment data were derived from experiments described previously.[14] Briefly, single-point mutant variants were created, corresponding to all 20 amino acids at each of 53 and 45 positions for the computationally designed influenza binders HB36.4 and HB80.3, respectively. These were expressed as yeast cell surface-conjugates, and subjected to a nonpurifying selection for hemaglutannin binders using FACS (Fluorescence-Activated Cell Sorting) by using concentrations of HA roughly at the K_D of the respective interaction. The presort and enriched libraries were subjected to high-throughput sequencing on an Illumina GA-II sequencer,

and the enrichment value for each sequence was calculated as the base-2 logarithm of the ratio of the number of times the sequence was seen in the enriched library to the number seen in the naïve library.

Prediction

Participants in CAPRI round 26 exercise for targets T55 (HB36) and T56 (HB80) were asked to predict both the ranking (on an arbitrary 0–1 scale) and the mutational class (beneficial/neutral/deleterious) of each of mutation. A full description of the methods for each group is included in the Supporting Information. Predictions were completed prior to the public release of Whitehead *et al.*[14]

For the initial prediction round, participants were provided with a description of how the experimental data were derived, the starting sequences (Supporting Information Table S3), the positions at which mutations were made, and structures for HB36.3 (PDBID 3R2X)[15] and HB80.4 (provided as a prerelease structure, further refined and submitted as PDBID 4EEF)[14] complexes. (The structures for the HB36.4 and HB80.3 complexes were not provided, as they have not been crystallized.) HB36.3 differs from HB36.4 by a K64N mutation, and HB80.4 from HB80.3 by G12K, L17I, L21I, A35K, and S42K. Additionally, in the prereleased structure, the first HB80 chain, chain G, had been modeled with an additional K28A mutation.

To see if more specific knowledge of deep mutational scanning experimental data would help prediction, a second round of prediction was run. In addition to the information available from the first round, participants were also provided with the enrichment values of one half of the mutations, randomly selected (9 aa at each of the mutated positions plus the starting identity). Participants were free to modify their procedure how they saw fit to account for the additional information—details on how each group used the additional data are provided in the Supporting Information.

For classification purposes, mutations with a $\log_2$(enrichment ratio) greater than +2 were considered beneficial, and those with values less than −2 were considered deleterious. For the BLOSUM model, each mutation was assigned the BLOSUM62[16] matrix value (an integer in the range of −4 to 11) for the wild type/mutation amino acid pair. BLOSUM values greater than or equal to +2 were considered beneficial, and those less than or equal to −2 were considered deleterious. These cutoffs were chosen to most closely approximate the distribution of beneficial/neutral/deleterious mutations that was observed experimentally.

Continuous predictions were evaluated by the Kendall tau-b correlation to the $\log_2$(enrichment ratio) values, as calculated by the cor() function of R.[17] Kendall's tau examines the experimental and predicted values for the exhaustive list of mutation pairs, considering them concordant (e.g., $x_1 < x_2$ & $p_1 < p_2$) or discordant (e.g., $x_1 < x_2$ & $p_1 > p_2$). The tau-b metric is then $(C–D)/\sqrt{(N_x N_p)}$, where C is the number of concordant pairs, D the number of discordant pairs, and N_x and N_p the number of total pairs not tied on experimental and predicted values, respectively. To evaluate the correlation of mutants for a single position, a derivative of Kendall's tau-b was used, where pairs were evaluated only between mutations at the same position, but summed across all positions. AUC values were calculated with the ROCR package in R.[18] Predictions were evaluated on recall ([number of correctly predicted mutations for a class]/[total number of mutations in that class experimentally]) and precision ([number of correctly predicted mutations for a class]/[total number of mutations predicted to be in that class]).

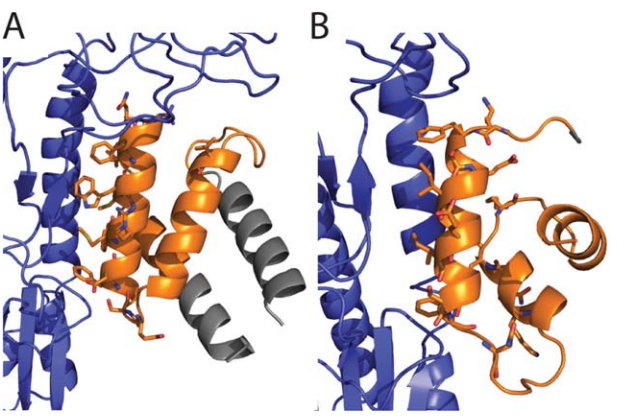

Figure 1

The structures of (**A**) HB36 (**B**) HB80 in complex with HA (blue) which were provided to participants. Residues probed in the deep sequencing enrichment experiment are in orange; the remainder are in grey. Residues at the interface are represented as sticks.

RESULTS

The HA binders HB36.4 and HB80.3 were designed previously using Rosetta.[15] Starting with these base designs, exhaustive single point mutant libraries were made and subjected to yeast display enrichment for binding to HA using nonpurifying FACS (Fluorescence-Activated Cell Sorting) selection.[14] By comparing the frequency of mutations in the enriched and unenriched libraries, an estimate of the effect of each point mutant on binding was obtained (the enrichment value, the $\log_2$ of the ratio of amino acid frequencies in the enriched library to that of the unenriched library).

Using crystal structures of design variants of HB36 and HB80 bound to HA as a guide (Fig. 1), participants made predictions of the effects of mutation on HA binding. These predictions were then compared to the

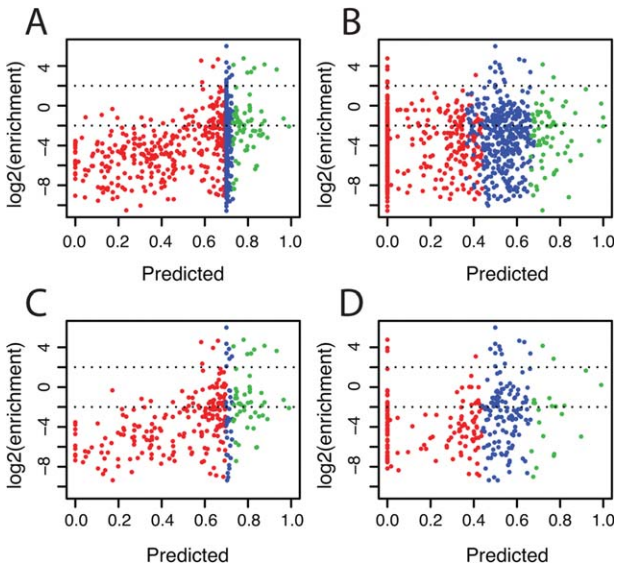

Figure 2

Predictor performance. Participant predictions (provided as a value between 0 and 1) are plotted versus the experimental enrichment value, with predictions of beneficial/neutral/deleterious colored green/blue/red. Enrichment ratios of −2 and 2, which defines the range of mutation considered experimentally neutral, are plotted with dotted lines. (**A**) Plot of all submitted predictions for HB80 for one of the top performing groups (G15, Weng). (**B**) Plot of all submitted predictions for HB80 for an average performing group. (**C**, **D**) As in (A) and (B), but only mutations at those positions in the interface are plotted.

experimental enrichment values (Supporting Information Fig. S1). The 22 groups that made submissions varied considerably in their ability to distinguish beneficial and deleterious mutations (Fig. 2A and B). Only two groups (G15, Weng, and G21s, Dehouck) had Kendall correlations above those of the BLOSUM62 model for both HB36 and HB80, although a few others (including G47, Flores, who only submitted predictions for HB36) were improved for a single protein (Supporting Information Table S1).

Of particular interest for applications of the prediction methods are the recall—the fraction of the experimentally beneficial mutations which are identified as such—and the precision—the fraction of predicted beneficial mutations which actually are. For HB36, 3.4% of the substitutions are experimentally beneficial, and for HB80, 2.4%. The precision of a method that selected randomly would hence be ∼3%; the BLOSUM model is roughly at this level. Three groups had precision better than 10% for both proteins (G05s, Bates; G15, Weng; G21, Fernandez-Recio), two of which (Weng and Fernandez-Recio) had recalls in the 25–40% range for both proteins (Supporting Information Table S1).

One limitation of the participant submitted classifications is that their performance is dependent on a somewhat arbitrary choice of threshold separating beneficial and nonbeneficial mutations. Another approach is to examine the performance of the ranking across all choices of thresholds with the area under the receiver operator characteristic (ROC) curve (AUC), which can also be interpreted as the probability that a randomly chosen positive item will be ranked appropriately against a randomly chosen negative item.[19] Several groups exceed the performance of the simple BLOSUM model for predicting both true beneficial (against neutral and deleterious) and true deleterious (against neutral and beneficial) mutations, with G15 (Weng; HB36 beneficial, deleterious AUC; HB80 beneficial, deleterious AUC: 0.667, 0.657; 0.705, 0.668) and G21 (Fernandez-Recio; 0.610, 0.726; 0.743, 0.651) on the Pareto front (Supporting Information Fig. S4).

The groups showing good performance were particularly successful in predicting deleterious mutations: low-ranking predictions were generally observed to be deleterious, whereas only a subset of the high-ranking predictions were beneficial (Supporting Information Fig. S1).

Mutations can influence binding if they disrupt the folded state, an effect particularly relevant for mutations away from the interface. To focus more on the ability of the methods to model-binding affinity independent of monomer stability, we also compared results on the subset of residues at the protein–protein interface (Fig. 2C, D, Supporting Information Fig. S1 and Table S1). The overall ranking of the groups did not change significantly on this subset.

It is instructive to break the results down based on the polarity of the initial and substituted residue. While the best groups did well predicting the effects of apolar to polar mutations, they overestimated the affinity of polar to polar and polar to apolar mutations (Fig. 3). This could be due to inaccuracies in treating electrostatics in the interfaces, as five of the six polar residues in the starting sequence for HB36 and three of the nine for HB80 are charged.

To test whether participants would be able to do better if they had additional data, in a second round nine mutations were randomly selected at each position of the two designed binders, and the experimental enrichment values for those mutations and for the starting amino acid were provided to participants. Fourteen groups submitted updated results, with improved results in most cases (Fig. 4A, B, Supporting Information Figs. S2, S3, and Table S1). Groups using machine learning techniques showed the greatest gains, though others using simpler reweighting strategies also improved performance. The top performing of these groups (G05s, Bates, and G21, Fernandez-Recio) included information from position/site specific models derived from the unblinded portion of the data, which, while potentially useful for evaluating combinations of mutations or modeling from sparse experimental data, would not be generalizable to other binding systems lacking experimental enrichment data.

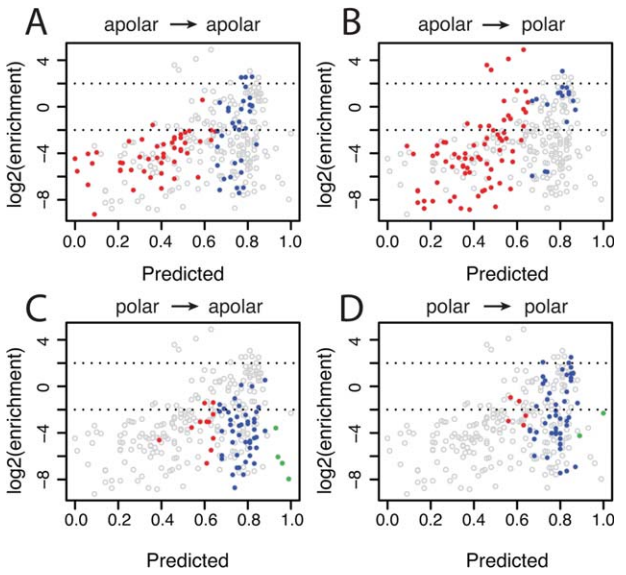

Figure 3

Mutations involving polar residues are more difficult to model. Breakdown by mutation polarity for a representative top performing group (G21s, Dehouck). The subset of HB36 interface mutations which are classed as (**A**) apolar to apolar, (**B**) apolar to polar, (**C**) polar to apolar, and (**D**) polar to polar for a representative top-performing group are displayed. For this analysis, residues D, E, H, K, N, Q, R, S, T, and Y are considered polar, and A, C, F, G, I, L, M, P, V, and W apolar. Green/blue/red correspond to participant's prediction of beneficial/neutral/deleterious. As a reference, the remaining interface mutations (those from the other three polarity groups for each graph) are plotted in grey.

Features contributing to good predictions

We used three approaches to identify factors which contributed to good predictions. First, to identify overall trends we evaluated the scoring and methodological features used by high performing groups. Second, we evaluated individual scoring terms used by several of the top-performing groups. Third, we released all of the experimental data to predictors, and asked groups to retrospectively identify which terms contributed to their performance.

The various protocols differ in how the mutant complexes are modeled. Some groups used coarse-grained models which do not require side-chain modeling, others kept all side chains other than the mutated one fixed, and others carried out various combinations of side-chain rotamer optimization, off-rotamer sampling, and backbone optimization. Many of the top performing groups optimized surrounding residues with off-rotamer sampling and backbone flexibility (Table I). Groups which normalized the score of the optimized mutant based on that of a similarly optimized reference structure also did somewhat better than average.

Groups which explicitly accounted for the effect of the mutant on structural stability generally performed better (Table I). Mutations which disrupt folding will necessarily

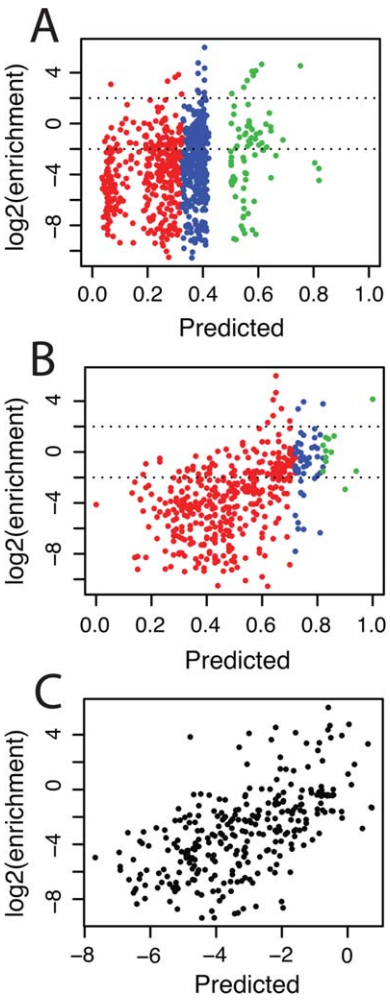

Figure 4

Improved performance upon refitting. Comparison of prediction results for group 21 (Fernandez-Recio) for (**A**) all round 1 HB80 interface predictions and (**B**) the reserved HB80 interface mutations for the round 2 predictions. (**C**) Prediction results for all interface positions when refit to the completely unblinded data without the position/site specific model.

Table I

Common features of top performing methods. Tallies of (# of groups using item in best performing half)/(# in other half)

	All Positions	Interface
Structural stability	5/3	6/2
Comparison to re-optimized starting structure	5/2	4/3
Entropy metric	4/1	3/2
Off-rotamer sampling	5/3	5/3
Statistical contact/distance metric	9/6	9/6
Lennard-Jones-style van der Waals	7/6	7/6
Other packing metrics	6/2	6/2
Optimization of surrounding residues	6/5	8/3
Backbone flexibility	5/3	6/2
Amino acid identity metric	2/3	1/4

Table II
Kendall correlation of individual metrics against experimental enrichment values.

	HB36 All Residues	HB80 All Residues	HB36 Interface	HB80 Interface
PoPMuSiC Packing Defect (D[a])[23]	0.300	0.288	0.294	0.260
Tobi T2 AP[b] (F)[22]	0.162	0.110	0.270	0.254
Tobi T1 AP (F)[22]	0.135	0.094	0.268	0.225
OPUS PSP (F)[21]	0.134	0.077	0.228	0.223
Tobi TSC CP (F)[22]	0.135	0.069	0.217	0.230
Skolnick SJKG CP (F)[30]	0.116	0.118	0.219	0.209
Floudas RMFCA CP (F)[31]	0.078	0.045	0.209	0.208
DComplex (F)[32]	0.140	0.071	0.256	0.206
FoldX hydrophob solv (B)[26]	nc[c]	nc	0.204	0.212
Park-Levitt HLPL CP (F)[33]	0.121	0.082	0.235	0.201
Li & Liang GEOMETRIC (F)	0.119	0.026	0.270	0.131
Boniecki Qp CP(F)[34]	0.182	0.062	0.265	0.155
Vendruscolo BFKV CP (F)[35]	0.166	0.057	0.242	0.185
Skolnick SKOa CP (F)[36]	0.155	0.087	0.237	0.160
FoldX bb_hbond (B)[26]	nc	nc	0.234	0.005
Miyazawa-Jernigan MJ2h CP (F)[37]	0.092	0.131	0.198	0.265
DFIRE2 (B)[38]	0.212	0.216	0.196	0.261
ACE (W)[25]	0.105	0.171	0.145	0.252
Tobi TB CP (F)[22]	0.109	0.070	0.111	0.233
Tanaka-Scheraga TS CP (F)[39]	0.050	0.096	0.153	0.223

[a]Calculated by (D)ehouck, (F)ernandez-Recio, (B)aker, (W)eng groups.
[b]AP—atomistic statistical potential; CP—coarse-grain statistical potential.
[c]Not calculated.

disrupt binding: P(binding) = P(folding)P(binding|folded), and mutations can affect either term. Methods which assume a stably folded protein will miss the effects of mutation on the first term. Accounting for stability is likely to be of particular importance for proteins with low starting stability.[20]

The highest performing groups employed packing metrics such as Lennard–Jones potentials (Table I). For example, the attractive portion of the van der Waals potential term was identified as one of the important terms by the Weng group (Table III), and statistical contact and distance scores, such as the OPUS_PSP group potential[21] and the Tobi coarse-grained potentials,[22] were among the single terms with the highest correlation to the enrichment data (Table II).

Other measures of packing such as convoluted fit and volume delta also correlated with improved performance (Table I). Of particular note is the PoPMuSiC packing defect term[23] from the Dehouck group, which correlated well with experimental results in both all residue and interface-only contexts (Table II), and was identified as the most influential term in the Dehouck group models (Table III). This coarse-grained metric measures the difference in residue volume between the starting and mutated residues, weighted for solvent accessibility.

Top groups also explicitly modeled electrostatics and solvation. Short range electrostatics were important for HB36, and Lazaridis-Karplus solvation[24] for HB80, according to the Weng group's analysis (Table III). While the ACE solvation term[25] by itself was correlated with the HB80 experimental results (Table II), adding it to a model with other terms had no appreciable benefit (Table III). The FoldX hydrophobic solvation term[26] correlates with interface enrichment values in both proteins (Table II), and the FoldX electrostatic terms ranked high in model feature importance (Supporting Information Table S2). Poisson–Boltzmann electrostatics have previously been shown to improve modeling this enrichment data.[14]

DISCUSSION

In the community wide test of methods for predicting the effects of point mutations on protein interaction reported here, the best groups are able to identify one-third of the beneficial mutations with less than a 10×

Table III
Evaluation of contribution of individual terms to prediction performance. Difference in Pearson correlation on omitting terms from all-data linear refits.

Dehouck Group[23,40,41]	HB36	HB80
packing defect	**0.167**	**0.075**
Solvent accessibility	0.018	0.005
Pairwise interactions	0.000	0.004
Backbone conformational preference	0.018	0.000
Weng Group[42]	**HB36**	**HB80**
vdW attractive	**0.055**	**0.056**
vdW repulsive	0.000	0.000
Solvation	0.012	**0.050**
Short range elec	**0.052**	0.020
Long range elec	0.015	0.029
Hydrogen bond	0.001	0.001
ACE	0.000	0.000

excess of mispredicted mutations. This is better than the performance of a simple model based on BLOSUM scores, and over three times the value expected from a random assignment. Accurate modeling clearly requires explicit consideration of the effects of mutations on stability, as methods which did not take this into account did not do as well. The best performing groups also modeled packing—either using a Lennard Jones model or considering volume changes—and electrostatics and solvation. The best methods used diverse overall approaches: machine learning (G21, Fernandez-Recio, and G05s, Bates), atom-level energy functions (G15, Weng), or coarse-grained models (G21s, Dehouck). The community wide experiment also reveals that there is considerable room for improvement in current methods; predicting the effect of mutations on polar starting positions appears to be a particular challenge.

We anticipate that many more comprehensive single-site scanning datasets should become available over the next several years as next generation sequencing methodology is increasingly applied to problems in biophysics. When modeling these data sets, it is important to recognize that there are a number of factors beyond binding affinity, such as stability, which contribute to the observed enrichment ratios in these experiments, and must be accounted for. Although enrichment results do not directly represent binding $\Delta\Delta G$ values, consideration of stability effects in making predictions is generally useful, as a theoretically tight binder is not useful if it is difficult or impossible to produce a folded protein. For those proteins which are stably folded, the values from deep mutational scanning experiments have been shown to match binding affinities.[27, 28] In particular, McLaughlin *et al.* found good correlation with the measured enrichment value and the $\Delta\Delta G$ of binding for 85 selected mutants (ref. 29, Supporting Information Fig. S2d).

The thousands of mutations which can be analyzed in parallel under identical conditions should compensate for many of the limitations of the high-throughput binding assays. For example, the prediction of small molecule-binding affinity to proteins is confounded by the fact that the available datasets consist of a small number of mutations on many different scaffolds with affinities measured by different groups using different techniques. As more comprehensive scanning datasets become available, further community wide experiments should continue to be useful for assessing methods and determining how best to model the effects of mutations on protein–protein interactions. The development of improved energy functions would also of course benefit from additional data from lower throughput but more accurate direct K_D measurements.

ACKNOWLEDGMENTS

R.A., M.T. and P.A.B. acknowledge funding from Cancer Research UK

Y.D. and M.R. are Postdoctoral Researcher and Research Director, respectively, at the Belgian F.R.S.-FNRS.

S.F. acknowledges funds from Uppsala University and eSSENCE (essenceofescience.se).

N.Z. and D.K. acknowledge funding from National Science Foundation (DBI-0845196 to D.K.). N.Z. is supported by National Science Foundation (IOS-1126992) and Paul K. and Diane Shumaker Fellowship

I.M., M.R.-D., L. P.-C., C. P., B. J., and J. F.-R. were supported by Grant # BIO2010-22324 from the Spanish Ministry of Science

R.M. and D.B. acknowledge funding from the Defense threat Reduction Agency (HDTRA1-10-0040 to D.B.). We thank T. A. Whitehead, A. Chevalier, and C. Bryan for assistance in interpreting the yeast display deep sequencing results.

REFERENCES

1. Jones S, Thornton JM. Principles of protein-protein interactions. Proc Natl Acad Sci USA 1996;93:13–20.
2. Nooren IMA, Thornton JM. Diversity of protein-protein interactions. EMBO J 2003;22:3486–3492.
3. Janin J, Bahadur RP, Chakrabarti P. Protein-protein interaction and quaternary structure. Q Rev Biophys 2008;41:133–180.
4. Arkin MR, Wells JA. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. Nat Rev Drug Discovery 2004;3:301–317.
5. Zhao L, Chmielewski J. Inhibiting protein-protein interactions using designed molecules. Curr Opin Struct Biol 2005;15:31–34.
6. Kastritis PL, Bonvin AMJJ. Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. J Proteome Res 2010;9:2216–2225.
7. Moal IH, Agius R, Bates PA. Protein-protein binding affinity prediction on a diverse set of structures. Bioinformatics 2011;27:3002–3009.
8. Vreven T, Hwang H, Pierce BG, Weng Z. Prediction of protein-protein binding free energies. Protein Sci 2012;21:396–404.
9. Kastritis PL, Bonvin AMJJ. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. J R Soc Interface 2012;10:20120835–20120835.
10. Janin J, Henrick K, Moult J, Eyck L Ten, Sternberg MJE, Vajda S, Vakser I, Wodak SJ. CAPRI: a Critical Assessment of PRedicted Interactions. Proteins 2003;52:2–9.
11. Moult J. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 2005;15:285–289.
12. Janin J. Protein-protein docking tested in blind predictions: the CAPRI experiment. Mol Biosyst 2010;6:2351–2362.
13. Wodak SJ, Méndez R. Prediction of protein-protein interactions: the CAPRI experiment, its evaluation and implications. Curr Opin Struct Biol 2004;14:242–249.
14. Whitehead TA, Chevalier A, Song Y, Dreyfus C, Fleishman SJ, Mattos C De, Myers CA, Kamisetty H, Blair P, Wilson IA, Baker D. Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. Nat Biotechnol 2012;30:543–548.
15. Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch E-M, Wilson IA, Baker D. Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science 2011;332:816–821.
16. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.

17. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2011.
18. Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics 2005;21:3940–3941.
19. Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett 2006;27:861–874.
20. Bloom JD, Labthavikul ST, Otey CR, Arnold FH. Protein stability promotes evolvability. Proc Natl Acad Sci USA 2006;103:5869–5874.
21. Lu M, Dousis AD, Ma J. OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. J Mol Biol 2008;376:288–301.
22. Tobi D. Designing coarse grained-and atom based-potentials for protein-protein docking. BMC Struct Biol 2010;10:40.
23. Dehouck Y, Grosfils A, Folch B, Gilis D, Bogaerts P, Rooman M. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. Bioinformatics 2009;25:2537–2543.
24. Lazaridis T, Karplus M. Effective energy function for proteins in solution. Proteins 1999;35:133–152.
25. Zhang C, Vasmatzis G, Cornette JL, DeLisi C. Determination of atomic desolvation energies from the structures of crystallized proteins. J Mol Biol 1997;267:707–726.
26. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 2002;320:369–387.
27. Fowler DM, Araya CL, Fleishman SJ, Kellogg EH, Stephany JJ, Baker D, Fields S. High-resolution mapping of protein sequence-function relationships. Nat Methods 2010;7:741–746.
28. Pál G, Kouadio J-LK, Artis DR, Kossiakoff AA, Sidhu SS. Comprehensive and quantitative mapping of energy landscapes for protein-protein interactions by rapid combinatorial scanning. J Biol Chem 2006;281:22378–22385.
29. McLaughlin RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. The spatial architecture of protein function and adaptation. Nature 2012;491:138–142.
30. Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding. When is the quasi-chemical approximation correct? Protein Sci 1997;6:676–688.
31. Rajgaria R, McAllister SR, Floudas CA. A novel high resolution Calpha–Calpha distance dependent force field based on a high quality decoy set. Proteins 2006;65:726–741.
32. Liu S, Zhang C, Zhou H, Zhou Y. A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. Proteins 2004;56:93–101.
33. Park B, Levitt M. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. J Mol Biol 1996;258:367–392.
34. Boniecki M, Rotkiewicz P, Skolnick J, Kolinski A. Protein fragment reconstruction using various modeling techniques. J Comput-Aided Mol Des 2003;17:725–738.
35. Bastolla U, Farwer J, Knapp EW, Vendruscolo M. How to guarantee optimal stability for most representative structures in the Protein Data Bank. Proteins 2001;44:79–96.
36. Skolnick J, Kolinski A, Ortiz A. Derivation of protein-specific pair potentials based on weak sequence fragment similarity. Proteins 2000;38:3–16.
37. Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 1996;256:623–644.
38. Yang Y, Zhou Y. Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. Protein Sci 2008;17:1212–1219.
39. Tanaka S, Scheraga HA. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules 1976;9:945–950.
40. Dehouck Y, Gilis D, Rooman M. A new generation of statistical potentials for proteins. Biophys J 2006;90:4010–4017.
41. Dehouck Y, Kwasigroch JM, Gilis D, Rooman M. PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. BMC bioinformatics 2011;12:151.
42. Haidar JN, Pierce B, Yu Y, Tong W, Li M, Weng Z. Structure-based design of a T-cell receptor leads to nearly 100-fold improvement in binding affinity for pepMHC. Proteins 2009;74:948–960.