This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520931, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Predicting nsSNPs that disrupt protein-protein interactions using docking

Norman Goodacre[1], Nathan Edwards[1], Mark Danielsen[1], Peter Uetz[2], Cathy Wu[1,3]

[1] Department of Biochemistry and Molecular and Cellular Biology
Georgetown University Medical Center
Washington, DC 20007
{nfg4,nje5,dan}@georgetown.edu

[2] Center for the Study of Biological Complexity
Virginia Commonwealth University
Richmond, VA 23284
peter@uetz.us

[3] Center for Bioinformatics and Computational Biology and Protein Information Resource
University of Delaware
Newark, DE 19711
wuc@dbi.udel.edu

ABSTRACT

The human genome contains a large number of protein polymorphisms due to individual genome variation. How many of these polymorphisms lead to altered protein-protein interaction is unknown. We have developed a method to address this question. The intersection of the SKEMPI database (of affinity constants among interacting proteins) and CAPRI 4.0 docking benchmark was docked using HADDOCK, leading to a training set of 166 mutant pairs. A random forest classifier that uses the differences in resulting docking scores between the 166 mutant pairs and their wild-types was used, to distinguish between variants that have either completely or partially lost binding ability. 50% of non-binders were correctly predicted with a false discovery rate of only 2%. The model was tested on a set of 15 HIV-1 - human, as well as 7 human - human glioblastoma-related, mutant proteins pairs: 50% of combined non-binders were correctly predicted with a false discovery rate of 10%. The model was also used to identify 10 protein-protein interactions between human proteins and their HIV-1 partners that are likely to be abolished by rare non-synonymous single-nucleotide polymorphisms (nsSNPs). These nsSNPs may represent novel and potentially therapeutically-valuable targets for anti-viral therapy by disruption of viral binding.

Categories and Subject Descriptors

I.2.1 [**ARTIFICIAL INTELLIGENCE Applications and Expert Systems**]: Medicine and science
J.3 [**LIFE AND MEDICAL SCIENCES**]: Biology and genetics

BCB '14, September 20 - 23 2014, Newport Beach, CA,
USA ACM 978-1-4503-2894-4/14/09.
http://dx.doi.org/10.1145/2649387.2649397

General Terms

Algorithms, Performance, Design, Experimentation

Keywords

Non-synonymous polymorphism, PPI, protein docking, interface, machine learning, mutant

1. INTRODUCTION

Proteins in human populations display a wide array of sequence polymorphisms. However, these sequence changes do not always affect the biological activity of specific proteins. In fact, it remains difficult to determine *a priori* whether changes in protein sequence will affect a protein's activity such as protein-protein interactions (PPIs). Although a number of tools have been developed to predict the functional effect of SNPs [1], these tools mainly focus on protein stability, rather than protein interaction. Estimates of the proportion of total nsSNPs involved in disease via altered PPI ranges from 4% [2] to 10% [3]. Here we describe a novel tool that predicts whether a change in amino acid sequence leads to loss of protein-protein interaction (PPI). We developed this model using the SKEMPI database of kinetic mutants (with experimentally-determined kD) [4], and tested it using a set of relatively well characterized HIV-1 – human protein interactions, as well as a second set of human - human interactions known to play a role in glioblastoma.

HIV-1 was chosen because it provides one of the clearest examples of the effect of a single SNP on disease. Interestingly, certain individuals are completely resistant to HIV- 1. These individuals possess a truncated version of the HIV-1 surface receptor CCR5, which is found on the surface of CD4+ T- helper cells and macrophages. This non-functional CCR5Δ32 variant prevents HIV-1 entry [5]. Of particular interest is that resistance to HIV can be engineered by transplanting stem cells containing CCR5Δ32 into patients [6]. This raises the possibility that there are other loss of binding variants in the human population that confer HIV resistance. Indeed, HIV interacts with around 1000 human proteins, including the alternative receptor CXCR4, that have been shown to have one or more sequence variants. HIV-1 is also one of the best studied human viruses from a structural perspective. For instance, the Subramanian group has used cryo-electron microscopy to produce structures of HIV-1 proteins in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520931, IEEE/ACM Transactions on Computational Biology and Bioinformatics

multimeric form [7], investigated the strain-specificity of such complexes [8] and probed the dynamics of HIV – human protein complexes under a variety of environmental and cellular conditions [9]. The glioblastoma test set was chosen because many forms of cancer are cause by altered PPIs, and kinetic interaction data is available for glioblastoma in particular.

The relative lack of computational tools to predict altered interactions is due in part to the range of interactions in which proteins can be involved (long-term vs. transient, in complex or binary), and also ways in which these interactions can be altered (modulated by post-translational modification / cofactor binding, increased or decreased in affinity by mutation). However, recent surveys indicate that the physical distribution of disease-nsSNPs is unique. Disease-nsSNPs often cluster together at the protein surface [10]. Such clusters of disease-nsSNPs tend to be found at protein interaction interfaces [11][12] and to be involved in the same disease; neither is true for non-disease-causing nsSNPs. These studies suggest that a model of altered interaction based on direct physical changes at the interface could explain the mechanism of many disease-nsSNPs.

It is unlikely, however, that such a model would succeed by using only static structures of proteins. As revealed by the instrumental evolutionary trace (ET) method of Lichtarge, interfaces are highly modular: families of related proteins often contain an ancestral core of interface residues about which additional functional clusters have arisen over the course of evolution [13]. In fact, the majority of observed contacts (pairs of residues within binding distance, typically 6Å) in co-crystal complexes present in the Protein Data Bank do not actually contribute to binding. Studies (e.g. [2]) also suggest that the majority of nsSNPs, even those present at interaction interfaces, are not likely to affect interaction. Flexible protein-protein docking tools such as HADDOCK are therefore necessary to more accurately capture the key residues responsible for binding kinetics, within the broader interface.

Although docking tools do not always produce accurate results, ongoing community benchmarking efforts such as Critical Assessment of Predicted Interactions (CAPRI), currently in its 4th iteration [14], are accelerating algorithm development e.g. through the development of advanced rescoring methods. Therefore, docking tools are likely to become increasingly relevant to predictive models of altered binding. CAPRI 4.0 consists of 144 structurally well-defined protein pairs (known crystal structure of proteins in both bound and unbound forms) of various functions encompassing most known binding modes. The recently-released SKEMPI database is also likely to hone the predictive capabilities of docking tools [4]. SKEMPI provides by far the largest publically-available resource to date of kinetic information for protein mutants, with over 3,000 mutants across 169 protein pairs [4]. Free energy of binding (ΔG) values are provided for all mutant-containing protein pairs, as well as all wild-type pairs. This allows mutant pairs to be sub-divided into classes, e.g. "binders" (unaffected or mildly weakened) vs. "non-binders" (severely weakened). Comparison of docking-derived energy and other physical features to experimentally-determined values is likely to improve the accuracy of docking tools in the near future.

We present a model that employs protein docking to predict complexes of a subset of mutant pairs in the SKEMPI database. The physical and energy scores from docking are used to train a machine-learning algorithm to differentiate between a class of binders and a class of non-binders, as described above. Because of its superior performance in recent rounds of the community wide Critical Assessment of Predicted Interactions competition (CAPRI) [15], especially when interface information is available, the HADDOCK docking tool [16] is used.

The proposed model also continues the development of "double delta" or "ΔΔ"energy scores (**Figure 1,** see Methods for details). Moal and Fernández-Recio (2013) used statistical pairwise amino acid potentials to predict ΔΔG of SKEMPI mutants [17], while Demerdash and Mitchell (2013) [18] developed a hybrid model containing energetic and non-energetic terms in order to re-rank docking results and select "native" poses from thousands of decoys. In comparison to these methods, the proposed model uses more extensive structural information, since it is based on entire docked complexes. SKEMPI mutant pairs as well as SKEMPI wild-type pairs are docked, Δ scores for energetic effects across the entire interface or complex are calculated in each case, and then Δ scores are compared to generate ΔΔ scores. The use of ΔΔ scores allows the proposed model to make predictions for a range of protein pairs (enzyme-inhibitor, ligand-receptor, or virus-host), especially when normalized as a proportion of the wild-type. Use of the entire complex allows for effects such as the strain imposed on bonds and angles underlying mutated residues to be measured, which is not possible for pairwise potentials. Future users would need only to submit a pair of wild-type proteins (with contact information), and one or more mutant-containing pairs of proteins to the HADDOCK webserver [19].

2. METHODS

2.1 Docking

39 wild-type systems (i.e. protein pairs) from CAPRI 4.0 and their 496 associated mutants located within the interaction interface from SKEMPI were submitted to the HADDOCK webserver for docking, although only a fraction of these (12 pairs and 166 associated mutants) passed quality control measures described below, and were ultimately used in the training set (**See section 2.2**). Surface contacts were used as ambiguous interaction restraints (AIRs). Contacts and interface residues were derived from the bound structures (co-crystal complexes) in the CAPRI 4.0 benchmark using CAPRI definitions (all residues ≤ 6.0 and 10.0 Å, respectively, from the opposite chain) (Janin, 2010). Surface residues were calculated in NACCESS [20] using a threshold of > 50% solvent accessibility of either the main or side chain in the unbound structure. The mutant proteins for docking were created in Chimera 1.8.1 [21] using the Dunbrack rotamer library [22] (no optimization of the global protein structure was performed). All hetero-atoms (non-protein atoms such as water or crystallization factors), and additional chains were removed prior to docking.

The performance of the classifier depends on the quality of docking results, and therefore stringent quality-control measures were taken. Wild-type protein pairs and all associated mutants were discarded if the wild-type could not successfully be docked. All wild-type results were compared to their co-crystal complexes to ensure their poses were biologically acceptable. The fraction of native contacts (fnc), the ligand RMSD (l-RMSD), and interface RMSD (i-RMSD) were calculated and star ratings were given according to standard CAPRI protocol [23]. Pairs with less than one star were discarded, along with associated mutants. Finally, all scores were averaged over the top 10 poses from the highest-ranking HADDOCK cluster, a common refinement step for docking algorithms [24] that has been reported to improve the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520931, IEEE/ACM Transactions on Computational Biology and Bioinformatics

quality of docking results [25][26].

2.2 Training

Mutants with k_D < 1/10 wild-type were labelled as "binders", while those with k_D > 1000x wild-type were labelled as "non-binders". These cut-offs were determined by observing natural peaks in the distribution of k_D-fold for the initial 496 mutants in the SKEMPI / CAPRI 4.0 overlap (not shown). Mutants with k_D between 10- and 1000-fold were also discarded, along with associated wild-type pairs. The final training set consisted of 12 wild-type protein pairs and 166 associated mutants (87 binders, 79 non-binders). Binders and non-binders were analyzed based on physicochemical class (hydrophobic, aromatic, etc.) as well as size change (**Figure 2**).

Initially, 21 features were calculated from the HADDOCK docking runs (**Figure 3**). Except for the conservation score, an external metric and common tool for predicting loss of binding, all scores were ΔΔ energy or ΔΔ physical scores. ΔΔ scores are based on Δ scores, which themselves are used by docking programs to score poses (change in energy / physical parameter upon binding). ΔΔ scores are the difference (mutant – wild-type) in Δ scores, comparing the "quality" of binding in the mutant, with reference to its original wild-type complex (**Figure 1,** right). All ΔΔ scores except for two residue-residue contact potentials (see paragraph below) were normalized as a proportion of the wild-type Δ score. Physical ΔΔ scores included differences in the buried surface area (BSA) and conformational rearrangement during binding. Energetic ΔΔ scores included differences in electrostatic, Van der Waals, or covalent bond energies at various sites in the complex, such as the interface, internal regions (core), entire complex, or entire complex plus water solvent (**Figure 3**).

Ultimately, 4 features were retained for the model. These were: "Conservation" or "Cons", "ΔΔBond", "ΔΔG", and "ΔΔBSA". Cons is an external metric that was found to increase the performance of the other 3 features, when combined in the model (its standalone performance is also compared to that of the model). This score approximates the disruptiveness of a mutation, and is defined as the inverse of the value from the 2008 Le and Gascuel amino-acid replacement matrix [27]. This replacement matrix estimates the probability of a substitution using the equation $P(t) = e^{Qt}$, where t is time, e is the natural log, and Q is mutation rate observed in seed sequences for Pfam families. Inverses were used because this matrix gives higher scores for more common substitutions, rather than rarer and more disruptive ones. For multiple and compound mutants (>1 mutation in one or both proteins, respectively), scores for individual mutations were summed. No difference in calculation was performed for multiple and compound mutants. ΔΔBond is the difference in mutant and wild-type Δ Bond, where ΔBond is the difference in docked and non-docked covalent bond energies. ΔΔG is the difference in mutant and wild-type ΔG. ΔΔBSA is the difference in mutant and wild-type buried surface area, or BSA. Because by definition the BSA for undocked protein pairs is 0, ΔBSA (the difference in docked and non-docked BSA) is equivalent to BSA and the two terms are used interchangeably in the present study. It is important to note that scoring functions for docking tools have been optimized for directing docking, rather than producing realistic energy values. Therefore, although HADDOCK is among the few docking tools with a realistic force field, the "energy" scores used in the present work should be interpreted as parameters of docking rather than physical values. For example, ΔG possesses an entropic component, which is typically evaluated using normal mode analysis (e.g. by molecular dynamics software). Because this is a computationally very costly analysis, docking tools generally use the number of rotatable bonds as an approximation of entropy [28].

As a second external metric commonly used to predict loss of binding, residue-residue (pairwise) contact potentials were calculated. These were calculated by combining ΔΔ HADDOCK Van der Waals and electrostatic scores at the level of individual contact residues (**Figure 3,** green bars), rather than for the whole interface. Finally, a combined external model (CEM) was created using both external metrics (conservation and pairwise contact potentials).

Weka is a flexible, Java-based environment for machine learning algorithm development [29]. In the current version (3.7), Weka supports a number of feature-refinement protocols, including CfsSubsetEval, which minimizes redundancy among features, and BestFirst, which maximizes the informativeness (predictive value) of features. 4 features, present at least 80% of the time during tenfold cross-validation, using the CfsSubsetEval Attribute Evaluator with the BestFirst Search Method in Weka 3.7, were kept in the final model. Random forests, formalized by Breiman (2001) [30], are a family of ensemble classification methods that are particularly suitable when a number of distinct combinations of features and threshold values may be predictive of the same class. Random forests were found to outperform other popular classifiers, including artificial neural networks (ANNs), Bayesian networks (BNs), and Support Vector Machines (SVMs), although overall performance was comparable for BNs. The model presented in this study consists of a random forest classifier (N=100) created in Weka 3.7 using a core of the 4 most informative features (**Figure 3,** blue bars). This classifier was trained according to the class labels of "binding" and "non-binding", defined as above, and tested in tenfold cross-validation. The following pseudo code summarizes the procedure used to create the model.

```
for wild-type protein pair in SKEMPI / CAPRI 4.0 overlap:
    calculate surface contacts from co-crystal structure
    redock in HADDOCK using unbound wild-type structures
    compare docking result to co-crystal structure
    if docking fails:
        discard protein pair and associated mutants
    if docking result < 1 star:
        discard protein pair and associated mutants
    extract Δ scores from docking files (score_bound^wt - score_unbound^wt)
    average Δ scores from top 10 decoys of top cluster
    for mutant protein pair:
        if kD < 10*wild-type kD:
            label as binder
        if kD > 1000*wild-type kD:
            label as non-binder
        if binder or non-binder:
            create mutant unbound structure(s) in Chimera v1.6
            dock in HADDOCK using same parameters as for wt
            extract Δ scores (score bound^mut - scoreunbound^mut)
            average Δ scores from top 10 decoys of top cluster
            calculate ΔΔ scores (Δ score^mut - Δ score^wt)
load ΔΔ scores (features) into Weka v3.7
refine non-redundant, highly-informative set of features
train RandomForest classifier using mutant class labels
```

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520931, IEEE/ACM Transactions on Computational Biology and Bioinformatics

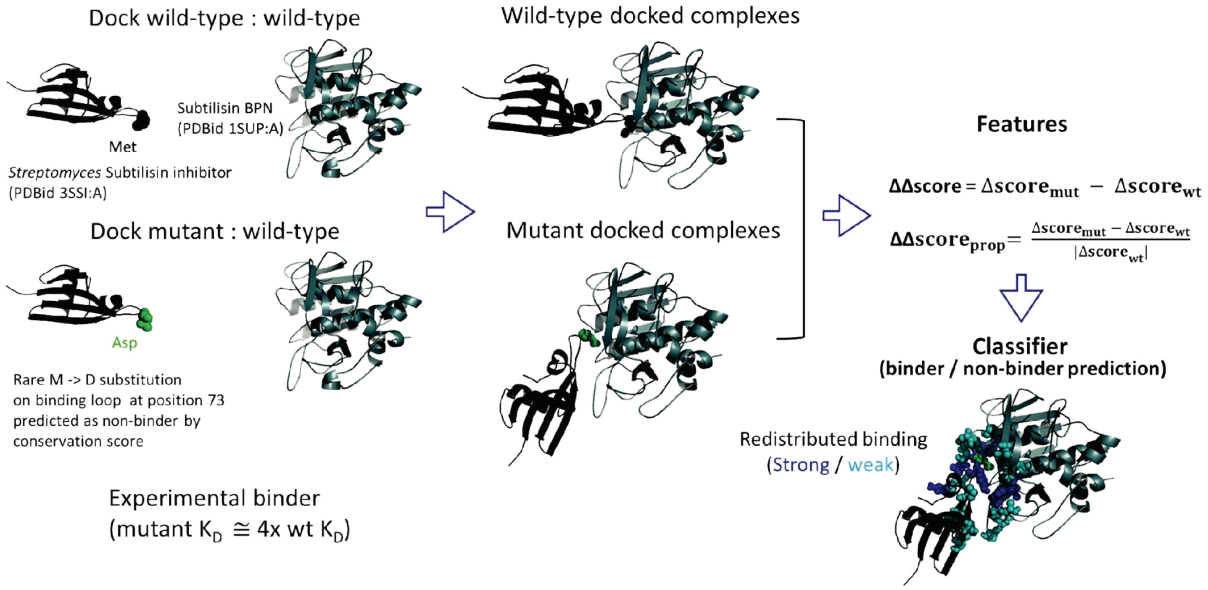

Figure 1. Procedure to distinguish binding from non-binding protein mutants. The experimental design is illustrated by Subtilisin BPN and its inhibitor in Streptomyces (PDB 1SUP and 3SSI, respectively; co-crystal complex: PDB 2SIC). One mutant (M73D) of the inhibitor (green spheres) was classified as a non-binder using the conservation score and contact energy. However, docking scores revealed significant residual binding, potentially explaining why the mutant has not completely lost binding, and also correctly predicts the mutant as a binder.

2.3 External test sets

In addition to tenfold cross-validation, the classifier was tested on a set of 15 mutant-containing HIV-1-human protein pairs (10 binders, 5 non-binders): 6 Capsid – Cyclophilin A mutants from SKEMPI [4], 3 Vpr – TFIIB mutants [31], and 7 integrase – LEDGF mutants [32]. For the HIV-1 test set, mutations were approximately evenly distributed among HIV-1 and human proteins. Finally, the classifier was tested on a set of 7 human mutant-containing protein pairs thought to play a role in the development of glioblastoma by losing interaction [33]. The glioblastoma set was used because, unlike the majority of the HIV-1 set, quantitative information ($\Delta\Delta G$) on loss of binding affinity was available.

2.4 Case study – predicting HIV-1 interaction – abolishing human nsSNPs

In order to demonstrate the utility of the classifier for addressing one of primary biological questions for which it was designed, predicting the effect of genetic variants on PPIs, a case study involving the known nsSNPs of biochemically well-characterized human-HIV-1 PPIs was conducted. These were PPIs for which the exact or approximate interaction interface had been experimentally determined. While 131 such interactions could be found in the literature, only around 20% of these had crystal structures in the Protein Data Bank encompassing the entire interface on both sides of the interaction (i.e. for both proteins). AIRs were calculated as during training, using NACCESS to predict surface residues. Predictions were made for a total of 58 nsSNPs (those contained by the crystal structures) involving 18 PPIs (18 human proteins and their 9 HIV-1 protein partners) by docking using the HADDOCK webserver and extracting scores as described above. Mutants were constructed using Chimera 1.8.1 as above, incorporating the rare forms of all nsSNPs that could be incorporated into the pdb structure. Originally, 23 PPIs were identified, but 5 could not be docked (CCR5-gp120, CCR2-gp120, PKR-Tat, SMUG1-Vpr, and p53-Nef). In order to ascertain which nsSNPs were most likely, from a biological perspective, to affect interaction, proximity to the interaction interface was calculated as any atom within 10.0 Å of an experimentally-determined interacting residue. 20 of the 58 nsSNPs were proximal to interfaces.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520931, IEEE/ACM Transactions on Computational Biology and Bioinformatics

3. RESULTS

3.1 Docking

7 of the 39 overlapping SKEMPI / CAPRI 4.0 pairs could not be docked and were discarded. Exactly half of the remaining docked wild-type pairs (16) received at least a one-star rating (6 one-star, 10 two-star - not shown) and were retained. Three wild-type docked pairs received a one-star rating but were omitted because the interaction partners were rotated or 180 degree around the interface, compared to the co-crystal complex. A further 4 wild-types pairs were discarded because associated mutants contained no binders or non-binders. The final training set of docking results consisted of 12 wild-type pairs and their associated 166 mutant-containing pairs (87 binders, 79 non-binders) (**Table 1**).

3.2 Training

57 binders (60%) contained a mutation from either a positive, polar, or hydrophobic residue to a residue of a different class. 66 non-binders (83%) contained a mutation from an aromatic residue to a residue of a different class, which in 48 cases (60%) was a mutation to a hydrophobic residue. Overall, non-binders had a greater tendency to contain substitutions with amino acids smaller than the originals, as evidenced by a predominantly negative distribution of Δ size (mutant – wild-type) (**Figure 2**). Often, these non-binding mutations consisted of an aromatic or other large residue replaced with a smaller residue. Multiple mutations were also more common among non-binders (not shown). The distribution for binders was centered around 0.

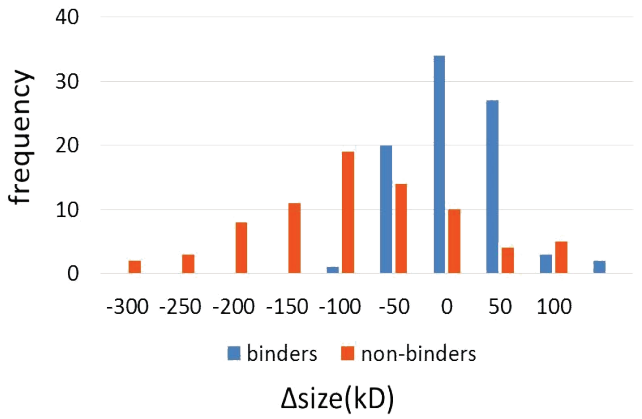

Figure 2. Size shifts are more negative for non-binders than for binders, in the training set. Mutations of aromatic to small hydrophobic residues, as well as mutations involving multiple residues, were common among non-binders (see text for details).

There was some redundancy among the final 12 protein pairs (**Table 1**). Ras and Rac, present in complexes 1LFD and 1E96, respectively, are both part of the Ras superfamily of small GTPases and share 30% sequence identity. Ras is also present in complex 1HE8, but in this case the partner is the activator PI-3 kinase rather than Ras interacting protein.

Of the 21 original features, 5 appeared to provide optimal performance according to both BestFirst exhaustive subset sampling in Weka and classifier precision and recall. These 5 were also the most informative, and appeared to measure distinct aspects of binding, including free energy, buried surface area, and improper-bond energy. However, the phi-psi angle feature was removed because it offered little extra performance when added to the other 4. The 4 features used in the final model, as well as features used to approximate residue-residue (pairwise) contact potentials, are shown in **Figure 3**. For the 4 final features, the difference in distributions for true binders and true non-binders is evident, with higher average values and proportionally even higher variance for non-binders (**Figure 4**), despite the presence of a number of positive outlier scores among binders (~10% binders for all features, excluding BSA). BSA was the only score that was higher on average for non-binders than for binders, and also the only one for which the p-value was > .05 using a one-tailed t-test with unequal variance.

The Q-value curve, which shows the # of positive (i.e. non-binding) predictions made for given false discovery rates, indicates significant improvement in predictive performance compared to either the pairwise contact potentials or the combined model (CEM) of pairwise contact potentials and conservation score (**Figure 5**). The model (solid line) is able to make 34 correct non-binder predictions without incurring a false positive, while the CEM (dashed line) is able to make only 23 such predictions. The model predicts half of true positives with a false discovery rate or FDR=2%, while the CEM predicts half of true positives with FDR=9% (**Figure 5,** red arrow).

A confidence threshold of c(nonbinder) > 0.60 was set by visual inspection. This corresponds to a FDR of 10% and 73 positive predictions (64/79 true positive predictions) (**Figure 5,** blue arrow). This confidence threshold was used during additional classification tasks. The same threshold appeared to be optimal for binder predictions, as well: c(binder) > 0.60. With this threshold, precision, recall, specificity, and F1 score for non-binders were: 0.89, 0.84, 0.91, and 0.86, respectively. For binders, these scores were: 0.89, 0.80, 0.89, and 0.84, respectively. The area under the receiver-operator curve was 0.93. The unlabeled set of mutant-containing protein pairs, those with kofold between 10 and 1000, were largely classified as binders. An example of one of the mutants is illustrated in **Figure 1** (a M73D substitution in subtilisin inhibitor, PDBid 2SIC, chain I). Although this substitution is extremely uncommon, and therefore would rank as a non-binder using the conservation score alone, it would be predicted correctly to be a binder.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520931, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Table 1. Docking results and training set. Of the 39 wild-type protein pairs that overlap between the CAPRI 4.0 docking benchmark and the SKEMPI database, 12 produced biologically-accurate structures when docked and contained at least one "binder" or "non-binder" mutant in SKEMPI (k_D < 10-fold of wild-type, k_D > 1000-fold of wild-type, respectively). The protein names, species, and PDB entries for these 12 protein pairs (columns 1-2), as well as the CAPRI docking ratings ("wild-type docking") are shown. The numbers of "binders" and "non-binders" are also shown ("mutant docking").

		wild-type docking				models docked		
Protein pair from CAPRI 4.0 / SKEMPI	**PDBID_chains**	**f_{nc} (prop contacts recaptured)**	**ligand RMSD**	**interface RMSD**	**stars**	**# SKEMPI binders**	**# SKEMPI non-binders**	**Total**
RAC1_NCF2 (*Homo sapiens*)	1E96_A_B	0.5	2.94	0.84	**	1	0	**1**
CHEY_CHEA (*Escherichia coli*)	1FFW_A_B	0.33	3.04	0.73	**	4	0	**4**
GRB2_VAV (*Mus musculus*)	1GCQ_B_C	0.12	3.5	1.09	*	3	0	**3**
PK3CG_RASH (*Homo sapiens*)	1HE8_A_B	0.24	3.48	0.94	*	2	0	**2**
BLAT_BLIP (*Escherichia. coli, Streptomyces clavuligerus*)	1JTG_A_B	0.12	3.58	0.91	*	4	48	**52**
TGFB3_TGFR2 (*Homo sapiens*)	1KTZ_A_B	0.9	2.2	0.67	**	2	2	**4**
GNDS_RASH (*Rattus norvegicus, Homo sapiens*)	1LFD_A_B	0.53	2.88	0.93	**	6	0	**6**
ACES_FAS2 (*Mus musculus, Dendroaspis angusticeps*)	1MAH_A_F	0.36	2.74	0.84	**	4	6	**10**
SUBT_IOVO (*Bacillus licheniformis, Meleagris. gallopavo*)	1R0R_E_I	0.9	1.18	0.71	**	47	23	**70**
ACTB_PROF1 (*Bos taurus*)	2BTF_A_P	0.53	2.1	1.1	**	2	0	**2**
UPA_UPAR (*Homo sapiens*)	2I9B_A_E	0.15	5.27	2.14	*	4	0	**4**
SUBT_SSI (*Bacillus amyloliquefaciens, Streptomyces albogriseolus*)	2SIC_E_I	0.76	2.62	0.74	**	8	0	**8**
						87	**79**	**166**

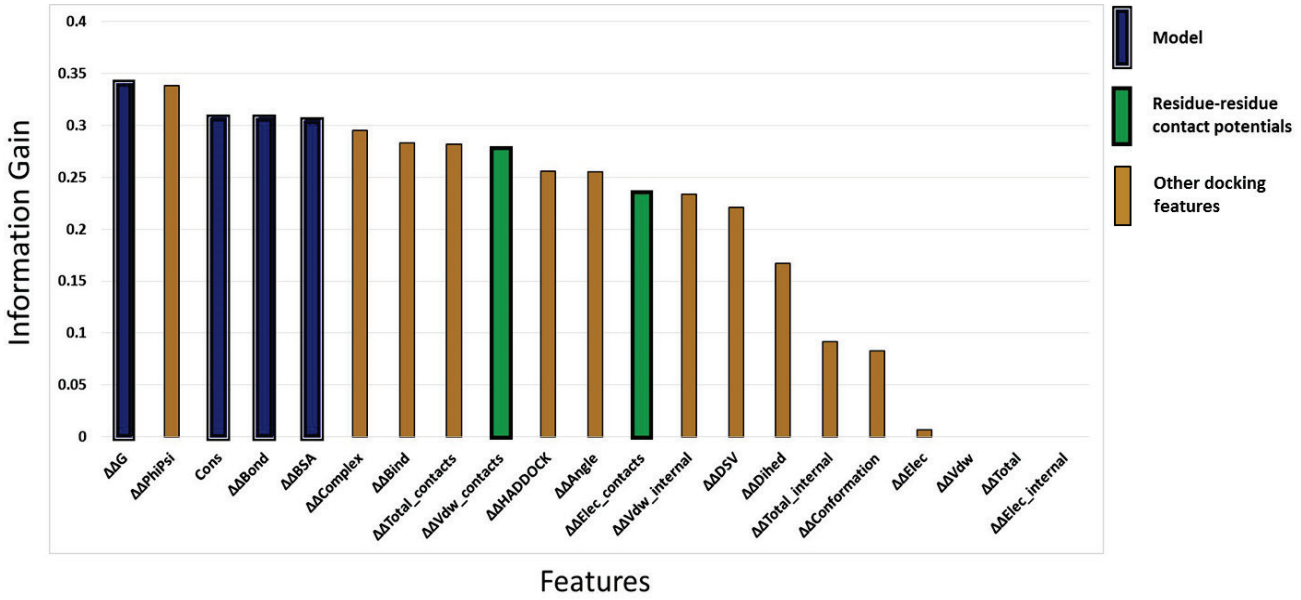

Figure 3. Docking-derived and conservation features for predicting loss of binding. In all, 21 features candidate features generated during docking were sampled in Weka (and one external amino-acid replacement score based on sequence conservation). The features selected for the final model are shown in blue with double-lined edges, while features used to capture residue-residue contact potentials are shown in green with bold edges. The conservation feature was also used in the combined external model (CEM) (cons + green bars).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520931, IEEE/ACM Transactions on Computational Biology and Bioinformatics

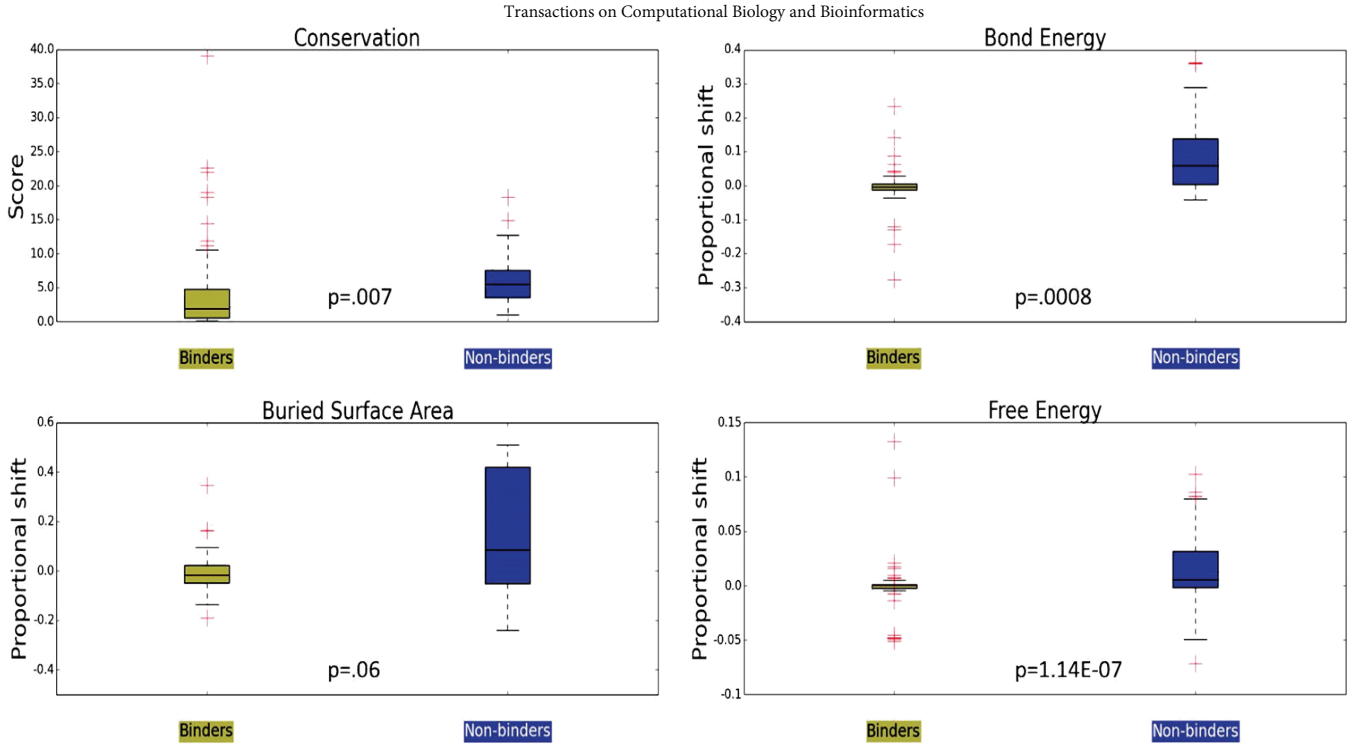

Figure 4. Boxplots of values for features used in model. Binders and non-binders are shown in tan and blue, respectively, while outliers are indicated by red plus signs.

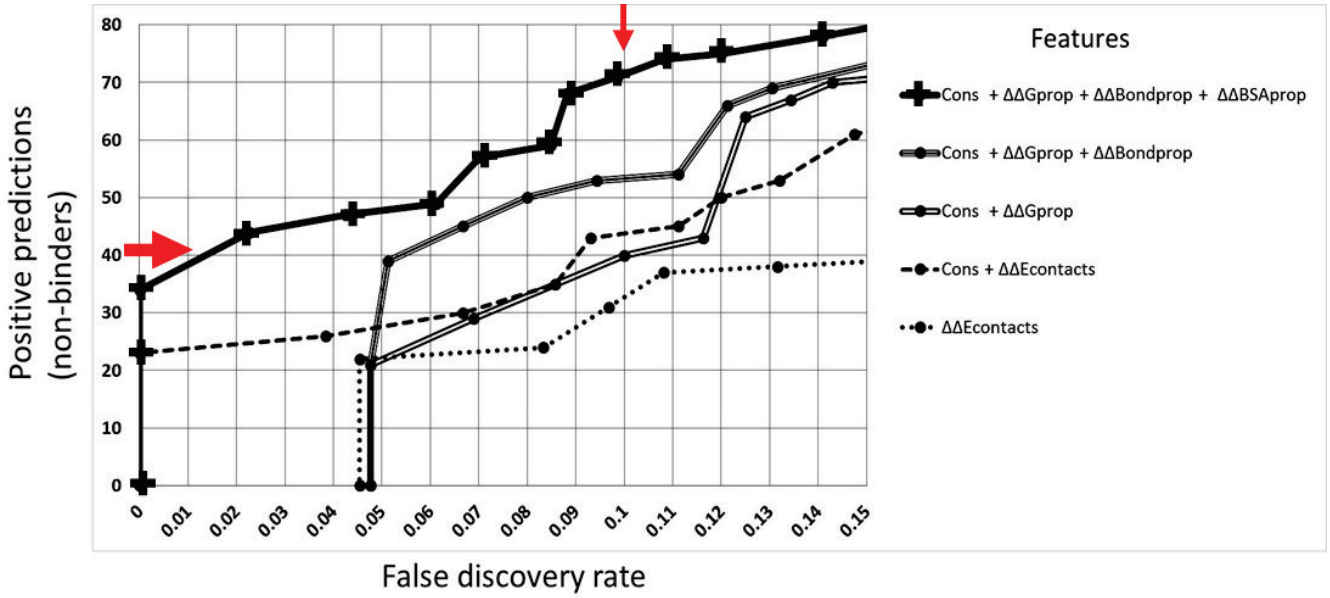

Figure 5. Combining multiple docking-derived features enhances predictive performance. The number of total positive (i.e. non-binding) predictions is plotted against the false discovery rate (FDR). The complete set of four features (solid line with crosses) shows improved performance over subsets of three (solid grey line) and two (compound line) features, even though the remaining features in the subsets have higher predictive performance when used in isolation. Combining features also enhances performance over external metrics that do not use full interface information, such as amino-acid replacement scores based on sequence conservation, pairwise amino-acid energy potentials (dotted line), or both (CEM - dashed line). The large red arrow indicates half of all positive predictions. The small red arrow indicates the FDR corresponding to a confidence threshold of c > 0.60.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520931, IEEE/ACM Transactions on Computational Biology and Bioinformatics

3.3 Cross-validation

HADDOCK docking is capable of optimizing backbone and side-chain conformations at the interface, and provides various physical and energy-based features (**Figure 3**). The model performs particularly well at predicting mutant protein pairs with strongly-diminished affinity of interaction (non-binders), distinguishing them from less disruptive mutants on the basis of characteristic patterns of redistributed binding at the interface. Specifically, less favorable changes in energy upon binding in mutants compared to their wild-type "parent" protein pairs (+ ΔΔ energy scores), as well as the presence of markedly fewer outliers, appear to define non-binders (**Figure 4**).

3.4 Comparison with pairwise potentials

Pairwise residue contact potentials derived from docking were among the most informative on an individual basis, but did not combine well with the most informative core features, being largely redundant with other features of the model. A model combining residue contact potentials with accessible surface area (ASA) [34] has been shown to be useful in predicting "hot spot" residues – those most essential for binding [35]. Therefore, we combined BSA with contact potentials to see if we could achieve a similar synergism at the level of the entire interface. Although we noted a modest improvement in performance, the improvement was less significant than when adding BSA to other 3 features of the present model. This may be due to the fact that, for the majority of non-binders, hot spot residues have been removed.

The model shows several advantages in performance terms, notably its low false discovery rate (FDR), with ~50% non-binders correctly predicted with an FDR of 2%, compared to an FDR of 9% for a combined model (CEM) based on conservation scores and pairwise residue contact potentials (**Figure 5**).

3.5 External test sets

For the HIV-1 test set of 5 binders and 10 non-binders, 4 and 7 predictions were made with $c > 0.60$, of which 3 and 4 were correct, respectively (**Table 2, top)**. The FDR was thus 50% for binders, and 20% for non-binders.

For the glioblastoma test set of 7 non-binders, 7 predictions were made with $c > 0.60$, of which 5 were correct. Increasing the confidence threshold to 0.80, 5 predictions remained, of which all 5 were correct (**Table 2, bottom**). Because there were no binders in this set, no FDR can be given.

Validation on external test sets of HIV-1 and human glioblastoma mutants showed results similar to those from tenfold cross-validation, in particular for the glioblastoma set, for which increasing the confidence threshold modestly eliminated all false binders without losing any true binders (not shown). This suggests that tightening the confidence threshold is an effective means of adjusting the model to eliminate false predictions, supporting results from cross-validation (**Figure 5**). There was a relatively higher rate of errors in the HIV-1 dataset, with 4 ambiguous predictions, and only 7 of the remaining 11 correct (64%). It should be noted, however, that the FDR for non-binders was fairly modest, at 20%.

Although the LEDGF-integrase pair accounted for fewer than half of the HIV-1 dataset (7/15 mutants), 3/4 non-predictions ($c < 0.60$ for either class) and 2/4 false predictions were found among its mutants. Omitting LEDGF-integrase predictions, the FDR for non-binders is 0%. These less accurate results for LEDGF-integrase may be due to the class labelling methodology or artefacts of the docking methodology, as elaborated in the **Discussion** (**section 5.2)**.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520931, IEEE/ACM Transactions on Computational Biology and Bioinformatics

Table 2. Non-synonymous SNPs predicted to cause loss of binding between human and HIV-1 proteins. Protein names and PDB structures used in docking are provided in columns 1-2 and 4-5. NsSNPs are listed in column 3 (those found at or near interface residues in bold, light blue). Amino acid positions are for Uniprot canonical sequences. Proximity to interface is defined as within 10 angstroms from a literature-reported interacting residue. Non-binding prediction status (Yes(c >= 0.60) or No), as well as confidence values for positive cases, is shown in the final column.

External test set 1 - HIV-1

Human protein name	PDB structure	mutation	HIV1 partner name	PDB structure	binding type	Non-binder prediction , confidence
LEDGF	2B4J_D	D366N	Integrase	2B4J_A	non-binding	N
LEDGF	2B4J_D	D366A	Integrase	2B4J_A	non-binding	N
LEDGF	2B4J_D	V370A	Integrase	2B4J_A	binding	Y, 0.81
LEDGF	2B4J_D	I365A	Integrase	2B4J_A	non-binding	Y, 0.87
LEDGF	2B4J_D	K360A	Integrase	2B4J_A	non-binding	N
LEDGF	2B4J_D	V408A	Integrase	2B4J_A	binding	N
LEDGF	2B4J_D	F406A	Integrase	2B4J_A	non-binding	N
TFIIB	1RLY_A	R53A_T54A	Vpr	1M8L_A	binding	N
TFIIB	1RLY_A	F55A	Vpr	1M8L_A	binding	N
TFIIB	1RLY_A	W52A	Vpr	1M8L_A	binding	N
CypA	1AK4_A	H487R	Capsid	1AK4_D	non-binding	N
CypA	1AK4_A	A488G	Capsid	1AK4_D	non-binding	N
CypA	1AK4_A	G489A	Capsid	1AK4_D	non-binding	Y, 0.95
CypA	1AK4_A	G489V	Capsid	1AK4_D	non-binding	Y, 0.84
CypA	1AK4_A	P490A	Capsid	1AK4_D	non-binding	Y, 0.88

External test set 2 - glioblastoma

Human protein name	PDB structure	mutation	partner name	PDB structure	binding type	Non-binder prediction , confidence
p53	1YCS_A	P177S	53BP2	1YCS_B	non-binding	Y, 0.72
p53	1YCS_A	R248H	53BP2	1YCS_B	non-binding	Y, 0.84
p53	1YCS_A	R248Q	53BP2	1YCS_B	non-binding	N
p53	1YCS_A	R248W	53BP2	1YCS_B	non-binding	N
p53	1YCS_A	R273C	53BP2	1YCS_B	non-binding	Y, 0.83
HRAS	1NVU_R	G12D	SOS1	1NVU_S	non-binding	Y, 0.95
RHOE	2V55_B	D67Y	ROCK1	2V55_A	non-binding	Y, 0.9

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520931, IEEE/ACM Transactions on Computational Biology and Bioinformatics

3.6 Case study – predicting nsSNPs that abolish human – HIV-1 PPIs

Of the 18 human – HIV-1 PPIs in the case study that could successfully be docked, 10 were predicted to be non-binders when nsSNP rare form variant(s) were included (c > 0.60) (**Table 3**, rightmost column). Of the 18 successfully-docked mutants, 10 were predicted to be non-binders, including 8 of the 10 successfully-docked cases with nsSNPs at the interface and 1 case of a single-nsSNP mutant at the interface. By contrast, the 8 predicted binder nsSNPs were primarily external to the interface (6/8 cases), although there were 3 cases of single-nsSNP mutants for this class. Thus, predicted non-binders generally fell within the interface and had multiple mutations, while predicted binders generally fell outside the interface and had fewer mutations. It must also be noted that the confidence of predictions was not particularly high for any of the non-binders, ranging between 0.60 (the threshold) and 0.77. The highest confidence prediction was for APOBEC3F – Vif. It is interesting that this highest prediction was for a member of a family of closely-related human proteins, with considerable redundancy in function. APOBEC3H nsSNPs were also predicted to abolish interaction, although the primary target [36] of HIV-1 Vif, APOBEC3G, did not have nsSNPs that prevent this interaction. AN evolutionary explanation of this finding is elaborated below, in the **Discussion (section 4.3**). A (relatively) high-confidence non-binding prediction (0.73) was also made for Alix-p6. This is likely a result of the availability of crystal contacts for this pair, and the presence of 7 nsSNPs overall with 2 at the interaction interface. The kinases (Lck, Hck, and Fyn) which are hijacked by HIV-1 Nef to orchestrate down-regulation of T-cell surface MHC I and II surface receptors, as well as the anti-lentiviral protein BST-2, all showed loss of interaction upon mutation to their nsSNP rare forms, even though the interacting residues information was only general (SH3 domain).

Table 3. Non-synonymous SNPs predicted to cause loss of binding between human and HIV-1 proteins. Protein names and PDB ids (columns 1-2, 4-5) are shown. NsSNPs are listed in column 3 (those found at or near interface residues in bold, light blue). Amino acid positions are for Uniprot canonical sequences. Proximity to interface is defined as within 10 angstroms from a literature-reported interacting residue. Non-binding prediction status (Yes(c >= 0.60) or No), as well as confidence values for positive cases, is shown in the final column.

Human protein	PDB structure	nsSNPs (@ interface)	HIV1 partner	PDB structure	Non-binder prediction, conf
CD4	4H8W:C	K191E, F227S, R265W	Gp120	4H8W:G	N
Lck	4D8K:A	**G201S**	Nef	4NEE:C	Y, 0.65
ß-TrCP	1P22:A	**A543S**, P592H	Vpu	1VPU:A	Y, 0.66
TRIM5α	4B3N:A (SMR)	G31S, H43Y ,C58Y, G110E, V112F, R136Q, G249D, H419Y, C467S, P479L	Capsid	1E6J:P	N
Dynamin2	3SNH:A (SMR)	P263L	Nef	4NEE:C	N
SIRT1	4KXQ:A	D3E, V484D	Tat	1JFW:A (SMR)	Y, 0.6
TFIIB	1RLY:A	P19S	Vpr	1M8L:A	N
APOBEC3G	3V4K:A	**H186R**, R256H, Q275E	Vif	4N9F:G	N
APOBEC3F	4IOU:A	R48P, Q61L, P97L, A108S, A178T, V231I, **Y307C**	Vif	4N9F:G	Y, 0.77
APOBEC3H	4J4J:A (SMR)	R18L, **G105R, K121E, K121N**, K140E, E178D	Vif	4N9F:G	Y, 0.63
APOBEC3B	3VM8:A (SMR)	**K62E**, P98L, S109A, **T146K**, R351H	Vif	4N9F:G	N
Hck kinase	1AD5:A	A44T, **M105L**, P502Q	Nef	4NEE:C	Y, 0.69
AP1G1	1W63:A	V195G, P685H	Nef	4NEE:C	N
Erk1	2ZOQ:A	E323K	Nef	4NEE:C	N
Fyn	1Y57:A (SMR)	I445F, D506E	Nef	4NEE:C	Y, 0.65
Importin-α	1IAL:A (SMR)	**A157V,** P165R, G365S, T430P, K453N	Vpr	1M8L:A:1-96	Y, 0.60
Alix	2XS1:A	V7M, A309T, **V378I, G429S**, N550S, K638E, S730L	P6	2R05:B	Y, 0.6
Alix	2XS1:A	V7M, A309T, **V378I, G429S**, N550S, K638E, S730L	Nucleo capsid	1A1T:A (SMR)	Y, 0.69

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCBB.2016.2520931, IEEE/ACM Transactions on Computational Biology and Bioinformatics

4. CONCLUSION

The model was found to perform optimally when using one conservation-based score and three docking-based scores for mutation (ΔΔG, ΔΔBond and ΔΔBSA). Many of the non-binder mutants used in the training set had aromatic or other large residues substituted with smaller residues. This may explain why two of the three docking features (ΔΔBond and ΔΔBSA) quantify redistributed binding across the interface. The model appears to outperform both sequence-conservation and a pairwise-potential – based predictive models. Specifically, the model generates predictions with a very low false discovery rate, provided the confidence threshold is set suitably high (at least c>0.60). This low false discovery rate was also found in external validation using HIV-1 – human and glioblastoma-related mutants. The model was used to discovery ten cases wherein an nsSNP in a human protein abolished interaction with an HIV-1 partner protein.

5. DISCUSSION

5.1 Comparison with existing models

An estimated 10,000 – 25,000 SNPs [37] that code for altered versions of 3,200 human proteins [38] (non-synonymous SNPs or nsSNPs) are believed to play a role in disease. It has been estimated that as much as 10% of these nsSNPs may exert this effect by altering protein-protein interactions [3], including with viral proteins [5].

However, existing techniques such as amino-acid conservation scores are insufficient for predicting mutations that disrupt interaction, particularly in a disease context. A recent structural SNPs survey by Das et al. (2014) [39] found that variants at interaction interfaces tend to disrupt interactions of greater biophysical strength, compared to variants outside the interface. However, variants at interaction interfaces do not fall upon more highly conserved residues, compared to those outside. Therefore, measuring the magnitude of binding energy disruption (ΔΔG or other $\Delta\Delta E_{score}$) seems to be a promising means of improving predictive capabilities.

The SKEMPI database of experimentally-defined kinetic mutants has already led to development of more refined pairwise potentials. A handful of recent studies have used SKEMPI either for training [40] or validation [41][42] of predictive models of protein interaction. These studies are encouraging, as they are among the first successful attempts to make binding predictions based on energy scores that are generalizable across proteins pairs. The novel predictive model presented here adds to such models, using the full structure of the protein interaction complex, in particular the interface, as depicted in **Figure 1**. The expansion of databases like SKEMPI is likely to accelerate the development of docking tools, as more compound and synergistic mutations are added.

The performance of HADDOCK depends in part upon the accuracy of active interface restraint information. Co-crystal complexes are not available for many of the more than 1,000 HIV-1 - human protein pairs that may be investigated in the future. However, considerable overlap exists between human-human and human-virus interfaces [43]. Therefore, human-human interfaces may be used. Interfaces can be obtained from databases such as 3DID [44] or iPfam [45]. In fact, a tool based on 3DID recently developed by Gonzalez, Liao and Wu (2013) [46] can provide a confidence score to rank interacting residues. Alternatively, interacting-residues prediction programs such as ProMate [47] or the consensus tool CPORT [48] can be used.

5.2 External test sets

The results suggest that the non-quantitative terms from literature used to assign class labels were ambiguous. The glioblastoma mutants all had experimentally-measured + ΔΔG values, and had more experimental evidence of binding loss. In addition, the use of a monomer of HIV-1 integrase for docking with human LEDGF rather than a dimer (the current model was only trained on binary complexes) may have produced incorrect poses. Lab data indicates that significantly more hinging occurs when the monomer, rather than the dimer, is docked (results not shown).

5.3 Case study of nsSNPs that abolish human – HIV-1 PPIs

The case study is valuable because it serves as further evidence that the predictive model can be applied to its original and primary purpose: predicting the effect of sequence variation on essential protein interactions of pathogens (with their host).

Equally importantly, these findings (nsSNPs with interaction-abolishing effects in 10 human proteins) have potential medical relevance, as they consist of mutations that could be cloned into T-cells that are then administered into AIDS patients to confer lasting immunity, following the overall methodological approach of Hutter et al. in their 2009 experimental therapy [6].

It is tempting to speculate that the APOBEC3 family of proteins has been in an evolutionary arms race with primate lentiviral Vif proteins for some time, and that the known nsSNPs have evolved as escape mutants for APOBEC3F, and APOBEC3H, but not yet APOBEC3G, proteins. It has been found that only a single amino acid differs between human and macaque APOBEC3G – the latter is not bound by lentiviral Vif [49]. APOBEC3B nsSNPs also were not predicted to lose interaction with HIV-1 Vif, but the B form is not a major player in HIV-1 infection.

Predictions of non-binding for 2 of the 8 mutants outside the interface must be interpreted only tentatively, as the model was not trained on mutants outside of interaction interfaces. Nevertheless, the finding that the majority of nsSNP-mutants predicted to cause loss of binding were mutants within the interface, and vice versa, supports the model.

Another important point to acknowledge is methodological in nature. NsSNPs were incorporated (for each protein) as a single ensemble during the *in silico* preparation of structural mutants. While the findings suggest that interface nsSNPs are the predominant causes of binding loss in this experiment, 3 human proteins had >1 interface nsSNP. Future studies should follow up on the present study with predictions of the effects of individual nsSNPs.

5.4 Comparison of Random Forest with other machine-learning classifiers

Bayesian networks (BNs) may be of some value in the development of future models. As indicated, the performance for BNs was close to that of Random Forests during cross - validation. By contrast, ANNs and especially SVMs displayed a high false-negative (FNR) rate for non-binders. Over 50% of actual non-binders were incorrectly classified by SVMs, although the FDR was lower than for Random Forests. BNs

classified HIV-1, but not glioblastoma, non-binding mutants slightly more accurately than did RandomForest.

5.5 Future directions

5.5.1 Features

The finding that greater buried surface area (positive ΔΔBSA) was characteristic of non-binders, yet is typically associated with higher binding affinity in experimental findings [42] also warrants deeper investigation (**Figure 4**). ΔΔBSA also contained the fewest outliers of any feature in the model (**Figure 4**, red plus marks), suggesting that binding redistribution is consistently different for the two classes (although p=.06). This may be due to the prevalence, among non-binders, of mutations converting aromatic to hydrophobic or other class of residue. Aromatic residues contain bulky side chains whose removal would allow the two proteins to come closer together, with an increase in Lennard-Jones potentials. In agreement with this explanation, non-binding mutants generally replaced larger amino acids with smaller amino acids, which was not found to be true of binding mutants (**Figure 3**). Aromatic residues are also well-known to be over-represented among hot spots, contributing substantially to binding affinity [50]. Alternatively, the removal of hot spot residues may force the docking software to introduce numerous weak compensatory interactions e.g. through rotation of hydrophobic side-chains. More thorough investigation of these possibilities would clarify the findings of the present study.

More realistic energy (and other) scoring functions could improve performance. For example, the type of energy driving interaction at the core of the interface, where hot spots predominate [50], is often distinct from that found at the periphery of the interface, where solvent interactions are involved ("O-ring" theory of Bogan and Thorn [51]). Geometric scores for scoring final docking poses, such as ZDOCK pairwise shape complementary (PSC) [52] might add value to the existing model. Additional structural elements such as fold or motif could be included, again bringing more structural "context" for energy scores. Recently, a docking affinity benchmark was published [53], wherein prediction of realistic ΔG was found to be particularly difficult in cases involving significant conformational rearrangement. Modeling conformational rearrangement also continues to hamper the performance of docking tools at the primary task of complex prediction [14]. To address this problem, one of the original 21 features generated for the present model was a score for conformational rearrangement: rmsd(ΔΔposition), where Δposition is a vector of residue displacements occurring during docking, for either the wild-type or mutant. This feature was not found to add significant predictive value to the model, therefore, more advanced (e.g. geometric) scores are required. The iAlign tool developed by Gao and Skolnick [54] scores similarity of interfaces between two pairs of proteins, and has been recommended for scoring docking predictions [55]. However, this tool was not found to add significant value in a pilot study (results not shown), perhaps due to the use of a single representative structure from each docking, rather than a consensus or averaged structure. Global docking tools such as ZDOCK could also be used to verify the accuracy of the wild-type docking by consensus, in cases where no co-crystal complex is available, so that future users of the model can be more assured of ΔΔ scores with predictive value. Alternatively, if a known non-binder exists, that mutant could be used as a positive control for non-binding (although this does not inform about the true wild-type binding conformation). A third way to verify docking results is to compare them with solved crystal structures of homologous complexes.

5.5.2 Classes of altered binding

The existing model could be expanded to include other classes of altered binding, such as "super-binders" with enhanced affinity. A recent study used SKEMPI mutants to train a classifier for nsSNPs that affect protein-interactions, using three classes – no effect, diminished binding, and enhanced binding [56]. However, the classifier did not define a class of "non-binders", as in the present study. There are many other classes of binding that could be defined, for instance enthalpy-driven vs. entropy-driven binding. Such a classifier could aid in the development of more sophisticated free energy (ΔG) scoring functions. There is preliminary evidence that disease-causing nsSNPs that alter protein interactions act through distinct mechanisms [56]. The same study also leverages the class of "undefined" (medium effect) mutants in SKEMPI to improve predictions for binding and non-binding mutants, using a technique known as semi-supervised learning. The functional insight that future tools such as the one in the present study might shed on interaction-altering human SNPs would prove invaluable to the current understanding of human genetic variation in disease.

6. REFERENCES

[1] Karchin, R. (2009). Next generation tools for the annotation of human SNPs. *Briefings in bioinformatics*, *10*(1), 35-52.

[2] Schuster-Böckler, B., & Bateman, A. (2008). Protein interactions in human genetic diseases. *Genome biology*, *9*(1), R9.

[3] Ferrer-Costa, C., Orozco, M., & de la Cruz, X. (2002). Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *Journal of molecular biology*, *315*(4), 771-786.

[4] Moal, I. H., & Fernández-Recio, J. (2012). SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics*, *28*(20), 2600-2607.

[5] Marmor, M., Sheppard, H. W., Donnell, D., Bozeman, S., & Celum, C. (2001). Homozygous and heterozygous CCR5-[DELTA] 32 genotypes are associated with resistance to HIV infection. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, *27*(5), 472-481.

[6] Hütter, G., Nowak, D., Mossner, M., Ganepola, S., Müßig, A., Allers, K., ... & Thiel, E. (2009). Long-term control of HIV by CCR5 Delta32/Delta32 stem-cell transplantation. *New England Journal of Medicine*, *360*(7), 692-698.

[7] Liu, J., Bartesaghi, A., Borgnia, M. J., Sapiro, G., & Subramaniam, S. (2008). Molecular architecture of native HIV-1 gp120 trimers. *Nature*, *455*(7209), 109-113.

[8] White, T. A., Bartesaghi, A., Borgnia, M. J., Meyerson, J. R., de la Cruz, M. J. V., Bess, J. W., ... & Subramaniam, S. (2010). Molecular architectures of trimeric SIV and HIV-1 envelope glycoproteins on intact viruses: strain-dependent variation in quaternary structure. *PLoS pathogens*, *6*(12), e1001249.

[9] Earl, L. A., Lifson, J. D., & Subramaniam, S. (2013). Catching HIV 'in the act'with 3D electron microscopy. *Trends in microbiology*, *21*(8), 397-404.

[10] Ye, Y., Li, Z., & Godzik, A. (2005, December). Modeling and analyzing three-dimensional structures of human disease proteins.

In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (pp. 439-450).

[11] Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S. M., & Yu, H. (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature biotechnology*, *30*(2), 159-164.

[12] David, A., Razali, R., Wass, M. N., & Sternberg, M. J. (2012). Protein–protein interaction sites are hot spots for disease associated nonsynonymous SNPs. *Human mutation*, *33*(2), 359-363.

[13] Lichtarge, O., Bourne, H. R., & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *Journal of molecular biology*, *257*(2), 342-358.

[14] Hwang, H., Vreven, T., Janin, J., & Weng, Z. (2010). Protein–protein docking benchmark version 4.0. *Proteins: Structure, Function, and Bioinformatics*,*78*(15), 3111-3114.

[15] Lensink, M. F., & Wodak, S. J. (2010). Docking and scoring protein interactions: CAPRI 2009. *Proteins: Structure, Function, and Bioinformatics*,*78*(15), 3073-3084.

[16] Dominguez, C., Boelens, R., & Bonvin, A. M. (2003). HADDOCK: a protein-protein docking approach based on biochemical or biophysical information.*Journal of the American Chemical Society*, *125*(7), 1731-1737.

[17] Moal, I. H., & Fernandez-Recio, J. (2013). Intermolecular contact potentials for protein–protein interactions extracted from binding free energy changes upon mutation. *Journal of Chemical Theory and Computation*, *9*(8), 3715-3727.

[18] Demerdash, O. N., & Mitchell, J. C. (2013). Using physical potentials and learned models to distinguish native binding interfaces from de novo designed interfaces that do not bind. *Proteins: Structure, Function, and Bioinformatics*,*81*(11), 1919-1930.

[19] De Vries, S. J., van Dijk, M., & Bonvin, A. M. (2010). The HADDOCK web server for data-driven biomolecular docking. *Nature protocols*, *5*(5), 883-897.

[20] Hubbard, S. J., & Thornton, J. M. (1993). Naccess. *Computer Program, Department of Biochemistry and Molecular Biology, University College London*,*2*(1).

[21] Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry*,*25*(13), 1605-1612.

[22] Dunbrack Jr, R. L., & Karplus, M. (1993). Backbone-dependent rotamer library for proteins application to side-chain prediction. *Journal of molecular biology*,*230*(2), 543-574.

[23] Janin, J. (2010). Protein–protein docking tested in blind predictions: the CAPRI experiment. *Molecular BioSystems*, *6*(12), 2351-2362.

[24] Vajda, S., & Kozakov, D. (2009). Convergence and combination of methods in protein–protein docking. *Current opinion in structural biology*, *19*(2), 164-170.

[25] Lorenzen, S., & Zhang, Y. (2007). Identification of near native structures by clustering protein docking conformations. *PROTEINS: Structure, Function, and Bioinformatics*, *68*(1), 187-194.

[26] Kozakov, D., Hall, D. R., Beglov, D., Brenke, R., Comeau, S. R., Shen, Y., ... & Vajda, S. (2010). Achieving reliability and high accuracy in automated protein docking: ClusPro, PIPER, SDU, and stability analysis in CAPRI rounds 13– 19.*Proteins: Structure, Function, and Bioinformatics*, *78*(15), 3124-3130.

[27] Le, S. Q., & Gascuel, O. (2008). An improved general amino acid replacement matrix. *Molecular biology and evolution*, *25*(7), 1307-1320.

[28] Moitessier, N., Englebienne, P., Lee, D., Lawandi, J., & Corbeil, A. C. (2008). Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. *British Journal of Pharmacology*, *153*(S1), S7-S26.

[29] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), 10-18.

[30] Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

[31] Agostini, I., Navarro, J. M., Bouhamdan, M., Willetts, K., Rey, F., Spire, B., ... & Sire, J. (1999). The HIV-1 Vpr co-activator induces a conformational change in TFIIB. *FEBS letters*, *450*(3), 235-239.

[32] Cherepanov, P., Ambrosio, A. L., Rahman, S., Ellenberger, T., & Engelman, A. (2005). Structural basis for the recognition between HIV-1 integrase and transcriptional coactivator p75. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(48), 17308-17313.

[33] Nishi, H., Tyagi, M., Teng, S., Shoemaker, B. A., Hashimoto, K., Alexov, E., ... & Panchenko, A. R. (2013). Cancer missense mutations alter binding properties of proteins and their interaction networks. *PloS one*, *8*(6), e66273.

[34] Tuncbag, N., Gursoy, A., & Keskin, O. (2009). Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics*, *25*(12), 1513-1520.

[35] Moreira, I. S., Fernandes, P. A., & Ramos, M. J. (2007). Hot spots—A review of the protein–protein interface determinant amino acid residues. *Proteins: Structure, Function, and Bioinformatics*, *68*(4), 803-812

[36] Marin, M., Rose, K. M., Kozak, S. L., & Kabat, D. (2003). HIV-1 Vif protein binds the editing enzyme APOBEC3G and induces its degradation. *Nature medicine*, *9*(11), 1398-1403.

[37] de Beer, T. A., Laskowski, R. A., Parks, S. L., Sipos, B., Goldman, N., & Thornton, J. M. (2013). Amino Acid Changes in Disease-Associated Variants Differ Radically from Variants Observed in the 1000 Genomes Project Dataset.*PLoS computational biology*, *9*(12), e1003382.

[38] Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), 5/12/2014. World Wide Web URL: http://omim.org/

[39] Das, J., Lee, H. R., Sagar, A., Fragoza, R., Liang, J., Wei, X., ... & Yu, H. (2014). Elucidating Common Structural Features of Human Pathogenic Variations Using Large Scale Atomic Resolution Protein Networks. *Human mutation*, *35*(5), 585-593.

[40] de Vries, S. J., & Bonvin, A. M. (2011). CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK.*PLoS One*, *6*(3), e17695.

[41] Moal, I. H., & Fernandez-Recio, J. (2013). Intermolecular contact potentials for protein–protein interactions extracted from binding free energy changes upon mutation. *Journal of Chemical Theory and Computation*, *9*(8), 3715-3727.

[42] Dehouck, Y., Kwasigroch, J. M., Rooman, M., & Gilis, D. (2013). BeAtMuSiC: prediction of changes in protein–protein binding affinity on mutations. *Nucleic acids research*, *41*(W1), W333-W339.

[43] Franzosa, E. A., & Xia, Y. (2011). Structural principles within the human-virus protein-protein interaction network. *Proceedings of the National Academy of Sciences*, *108*(26), 10538-10543.

[44] Stein, A., Russell, R. B., & Aloy, P. (2005). 3did: interacting protein domains of known three-dimensional structure. *Nucleic acids research*, *33*(suppl 1), D413-D417.

[45] Finn, R. D., Miller, B. L., Clements, J., & Bateman, A. (2014). iPfam: a database of protein family and domain interactions found in the Protein Data Bank. *Nucleic acids research*, *42*(D1), D364-D373

[46] González, A. J., Liao, L., & Wu, C. H. (2013). Prediction of contact matrix for protein–protein interaction. *Bioinformatics*, *29*(8), 1018-1025.

[47] Neuvirth, H., Raz, R., & Schreiber, G. (2004). ProMate: a structure based prediction program to identify the location of protein–protein binding sites. *Journal of molecular biology*, *338*(1), 181-199.

[48] de Vries, S. J., & Bonvin, A. M. (2011). CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. *PLoS One*, *6*(3), e17695.

[49] Schröfelbauer, B., Chen, D., & Landau, N. R. (2004). A single amino acid of APOBEC3G controls its species-specific interaction with virion infectivity factor (Vif). *Proceedings of the National Academy of Sciences of the United States of America*, *101*(11), 3927-3932.

[50] Moreira, I. S., Fernandes, P. A., & Ramos, M. J. (2007). Hot spots—A review of the protein–protein interface determinant amino acid residues. *Proteins: Structure, Function, and Bioinformatics*, *68*(4), 803-812

[51] Bogan, A. A., & Thorn, K. S. (1998). Anatomy of hot spots in protein interfaces.*Journal of molecular biology*, *280*(1), 1-9

[52] Pierce, B. G., Hourai, Y., & Weng, Z. (2011). Accelerating protein docking in ZDOCK using an advanced 3D convolution library. *PloS one*, *6*(9), e24657.

[53] Kastritis, P. L., Moal, I. H., Hwang, H., Weng, Z.,.Bates, P. A., Bonvin, A. M., & Janin, J. (2011). A structure based benchmark for protein–protein binding affinity. *Protein Science*, *20*(3), 482-491.

[54] Gao, M., & Skolnick, J. (2010). iAlign: a method for the structural comparison of protein–protein interfaces. *Bioinformatics*, *26*(18), 2259-2265

[55] Gao, M., & Skolnick, J. (2011). New benchmark metrics for protein protein docking methods. *Proteins: Structure, Function, and Bioinformatics*, *79*(5), 1623-1634.

[56] Zhao, N., Han, J. G., Shyu, C. R., & Korkin, D. (2014). Determining Effects of Non-synonymous SNPs on Protein-Protein Interactions using Supervised and Semi-supervised Learning. *PLoS computational biology*, *10*(5), e1003592

Author Biographies

Norman Goodacre received his PhD from Georgetown University in 2014. He is currently a postdoctoral fellow at the Center for Biologics Evaluation and Research (CBER) at the U.S. Food and Drug Administration. His research interests include protein modeling and synthetic biology.

Nathan J. Edwards received his PhD degree from Cornell University in 2001. He has been working in the area of proteomics informatics, especially peptide identification from tandem mass-spectra, for more than 10 years, first with Celera Genomics and Applied Biosystems, then at the University of Maryland, College Park and Georgetown University, where he is an Associate Professor. Other research interests include glycoproteomics and metabolomics informatics, and pathogen detection using DNA and proteomics based signatures. He is actively involved in the Clinical Proteomics Tumor Analysis Consortium in the Data Coordinating Center. He has published 49 peer reviewed manuscripts and given more than 45 invited talks.

Mark Danielsen received his Ph.D. from the University of South Carolina, Columbia in 1983. He became a faculty member at Georgetown University Medical School in 1989 after fellowships at Stanford University and the Laboratory of Molecular Biology, The Medical Research Council, Cambridge, England. He is director of the Bioinformatics Graduate Program in the department of Biochemistry and Molecular and Cellular Biology. His research expertise is in the areas of molecular diagnostics of infectious disease, and molecular endocrinology.

Peter Uetz received his PhD from EMBL and the University of Heidelberg in 1997. After having held faculty positions at the Karlsruhe Institute of Technology (KIT), The Institute for Genomic Research (TIGR), and the J Craig Venter Institute (JCVI), he has been an Associate Professor at Virginia Commonwealth University since 2011. Dr. Uetz has worked on vertebrate developmental biology as well as microbial functional genomics with a focus on protein interaction networks and protein function. He has published more than 120 papers that have been cited more than 10,000 times.

Cathy H. Wu received her PhD degree from Purdue University in 1984. She is the Edward G. Jefferson Chair and Director of Center for Bioinformatics and Computational Biology at the University of Delaware and the Director of the Protein Information Resource at UD and Georgetown University. She has conducted bioinformatics research for 25 years and is the PI/Co-PI on several consortium projects, including the UniProt and the Protein Ontology. She serves on several advisory boards, including the ACM SIGBio, and has served on over 50 international conference organizing committees. Her research encompasses protein structure-function, biomedical text mining and ontology, systems biology, and translational bioinformatics. She has published more than 220 peer-reviewed papers and 12 books, conference proceedings and journal special issues, as well as given more than 150 invited talks.