BIOINFORMATICS ORIGINAL PAPER
Vol. 28 no. 20 2012, pages 2600–2607
doi:10.1093/bioinformatics/bts489

Structural bioinformatics
Advance Access publication August 1, 2012

SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models

Iain H. Moal and Juan Fernández-Recio*

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

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Empirical models for the prediction of how changes in sequence alter protein–protein binding kinetics and thermodynamics can garner insights into many aspects of molecular biology. However, such models require empirical training data and proper validation before they can be widely applied. Previous databases contained few stabilizing mutations and no discussion of their inherent biases or how this impacts model construction or validation.

Results: We present SKEMPI, a database of 3047 binding free energy changes upon mutation assembled from the scientific literature, for protein–protein heterodimeric complexes with experimentally determined structures. This represents over four times more data than previously collected. Changes in 713 association and dissociation rates and 127 enthalpies and entropies were also recorded. The existence of biases towards specific mutations, residues, interfaces, proteins and protein families is discussed in the context of how the data can be used to construct predictive models. Finally, a cross-validation scheme is presented which is capable of estimating the efficacy of derived models on future data in which these biases are not present.

Availability: The database is available online at http://life.bsc.es/pid/mutation_database/

Contact: juanf@bsc.es

Received on June 26, 2012; revised on July 26, 2012; accepted on July 27, 2012

1 INTRODUCTION

Ascertaining how the alteration of protein sequence influences the thermodynamics and kinetics of binding is of fundamental importance to many disparate areas of biomedical research. Among others, its application is central to antibody engineering (Kuroda *et al.,* 2012), interaction design (Fleishman *et al*., 2011; Kortemme and Baker, 2004; Mandell and Kortemme, 2009), locating binding hotspots (Bogan and Thorn, 1998; Grosdidier and Fernandez-Recio, 2008; Grosdidier and Fernandez-Recio, 2012; Guerois *et al*., 2002; Kortemme and Baker, 2002; Moreira *et al*., 2007; Tuncbag *et al*., 2009; Xia *et al*., 2010), determining the energetics of specific moiety contacts (for instance, Anderson *et al*., 1990), uncovering binding mechanisms (Weikl and von Deuster, 2009), characterizing transitions states (for instance, Harel *et al*., 2007), determining the diversity of affinities and specificities of extant proteins (Empie and Laskowski, 1982) as well as ascertaining the functional consequences of potentially pathological mutations (for instance, Tischkowitz *et al*., 2008). Should it become possible to estimate how single or multiple mutations influence the kinetics and thermodynamics of protein binding on a much greater scale, many further investigations would materialize. For instance, determining the consequences of polymorphisms and mutations from initiatives such as the 1000 genomes project (Altshuler *et al.,* 2010), the Cancer Genome Project (http://www.sanger.ac.uk/genetics/CGP/) or the Cancer Genome Atlas (http://cancergenome.nih.gov/) promises to yield significant insights into the mechanisms of disease on a systemic level. Yet, the amount of data being produced is staggering. Applied to genetic data obtained in a clinical setting, such methods could find use in personalized medicine. Used to characterize sequence-function landscapes, these methods could test theoretical models regarding the relationship between mutational robustness and evolvability (Draghi *et al*., 2010). Applied to orthologs and paralogs, this would indicate how affinities and specificities vary across phylogenetic trees, giving insights in how organisms have adapted to their niches at the interaction network level. Furthermore, a more widespread application of *de novo* interface design could be used to rewire interaction networks.

Incisive and large-scale experimental studies have illuminated many aspects of protein binding (Ashkenazi *et al*., 1990; Bass *et al*., 1991; Farady *et al*., 2007; Keeble *et al*., 2008; Kelley *et al*., 1995; Lu *et al*., 1997, 2001; Reichmann *et al*., 2005, 2007; Schreiber and Fersht, 1995), whereby the binding of mutants is typically characterized using surface plasmon resonance, isothermal titration calorimetry or spectroscopic techniques. However, these approaches are resource intensive, which prohibits their use in the high-throughput characterization of sequence-function landscapes in many systems. Recourse can be made in combinatorial methods that circumvent the need for direct physical measurement (Ernst *et al*., 2010; Fowler *et al*., 2010; Pal *et al*., 2005; Weiss *et al*., 2000; Whitehead *et al*., 2012; Wu *et al*., 2011). However, these methods can suffer from challenges arising from library generation, display, selection and sequencing (Araya and Fowler, 2011). Currently, it is unlikely that any of these methods alone will be able to keep up with the abundance of data generated by modern genomics initiatives. In light of this and ever increasing computational resources, computational models are an attractive option where structural data are available. Methods such as MM-PBSA and MM-GBSA can be used to derive free energies from structural ensembles generated using Monte Carlo sampling or molecular dynamics simulation.

*To whom correspondence should be addressed.

However, errors can arise from force field approximations and insufficient conformational sampling, and the production of trajectories can be costly (Hou *et al.*, 2011). Promisingly, it is possible to produce empirically parameterized functions for the physical modelling or statistical inference of $\Delta\Delta G$ (Benedix *et al.*, 2009; Guerois *et al.*, 2002; Kamisetty *et al.*, 2011; Tong *et al.*, 2004) or even $\Delta\Delta H$, $\Delta\Delta S$, Δk_{on} and Δk_{off}. Such empirical functions have the potential to be highly efficient and, where the role of conformational flexibility is limited, highly accurate. These methods are limited by the availability of training data; with greater data, finer inferences can be made and a broader range of phenomena investigated (Fleishman and Baker, 2012). However, special care must be taken to avoid subtle ways of overfitting or deriving overly optimistic estimates of the generalisation error.

In this article, we present SKEMPI: Structural database of Kinetics and Energetics of Mutant Protein Interactions. SKEMPI is a large, manually curated database of experimentally measured changes in binding free energy, entropy, enthalpy and rate constants, upon mutation. The data may be used to investigate the structural principles governing the influence of mutations upon binding, for the evaluation of theoretical or experimental techniques, as well as for the training of empirical functions. As the kinetic and thermodynamic changes reported in the literature are not systematic and reflect the interests of the experimentalists who determine them, the data have specific biases towards certain residues, types of mutation, spatial locations, binding sites, proteins and protein families. These biases are reviewed and discussed in the context of how they can be accounted for in the construction and evaluation of empirical models. Specifically, a cross-validation framework is presented within which empirically trained models of arbitrary complexity can be evaluated, such that the resultant cross-validation error is not lowered by these biases and can thus be used to give an indication of how well the model is likely to perform when applied to mutations that are not necessarily of the same type as those overrepresented in the training set.

2 METHODS

The kinetic and thermodynamic data on mutational effects in protein–protein interactions were collected directly from the literature. Some of the data were found via the Alanine Scanning Energetics database (ASEdb) (Thorn and Bogan, 2001) and PINT (Kumar and Gromiha, 2006) databases, as well as during the literature search for a previously published binding free energy benchmark (Kastritis *et al.*, 2011). Where possible, K_d, ΔH, ΔS, ΔG, k_{on} and k_{off} values, which are not explicitly reported, were generated via the relationships $\Delta G = \Delta H - T\Delta S$, $\Delta G = RT \ln K_d$ and $K_d = k_{off}/k_{on} = 1/K_a$. All dissociation constants, association rates and dissociation rates were reported in units of M, $M^{-1} s^{-1}$ and s^{-1}, respectively. Enthalpies and entropies were either reported as, or converted to, kcal mol^{-1} and cal $mol^{-1} K^{-1}$. No mutations involving insertions, deletions or residues containing post-translational modification were included, nor were mutations in which the affinity of the mutant was beyond the detection threshold of the experimental method were used. Where numerical values were not explicitly included in the relevant publication, values were obtained directly from the authors, extracted from a figure or taken from the ASEdb. These cases are noted in the database. The classification of amino acid positions as surface, interior, rim, core or support is as described in Levy (2010). Where multiple structures were available, the highest resolution structures were selected. Where structures of mutant forms were available, the reverse mutation back to the wild type was also included in the database.

To determine which proteins are homologs, we used the GAP4 program (Bonfield *et al.*, 1995). Proteins were defined as homologs if they have a similarity score of greater than 50 and at least 30% sequence identity. Two proteins were deemed to bind to the same site on a shared binding partner or homologous binding partners when at least 70% of the interface residues on the binding partner, for the interaction with the fewest such residues, was shared by the second interaction. Interface residues were defined as those with a non-hydrogen atom within 10 Å of a non-hydrogen atom on the binding partner.

3 RESULTS

In total, 3047 $\Delta\Delta G$ measurements of 2792 unique mutations or sets of mutations were found for 158 structures of 85 protein–protein complexes. For comparison, the available databases ASEdb (Thorn and Bogan, 2001) and PINT (Kumar and Gromiha, 2006), respectively, contain 620 and 699 binding affinity data cross-referenced to dimer structures in the protein databank, covering 26 and 32 systems, respectively. The $\Delta\Delta G$ values collected range from −12.4 to 12.4 kcal mol^{-1}. These extreme values correspond to the conversion between the wild-type P1-Lys BPTI/trypsin complex, which has femtomolar affinity, and the micromolar P1-Gly variant, both of which have crystal structures available (Helland *et al.*, 1999). Kinetic rate constants were available for the wild-type and mutant forms of 713 mutations, covering a range of $\Delta \log10(k_{on})$ of −3.6 to 2.4 and $\Delta \log10(k_{off})$ of −8.5 to 6.8, with k_{on} and k_{off} units of $M^{-1} s^{-1}$ and s^{-1}, respectively. Furthermore, changes in entropies and enthalpies were available for 127 mutations, with $\Delta\Delta H$ ranging from −7.5 to 17.6 kcal mol^{-1} and $\Delta\Delta S$ from −31.5 to 33.2 cal $mol^{-1} K^{-1}$. In total, the database contains 2317 single mutants, 364 double mutants, 103 triple mutants and 263 values corresponding to between 4 and 27 mutations, with many examples of additive and non-additive effects. The data come from many studies, including site-directed experiments, systematic mutation scans, homolog scanning mutagenesis, cognate/non-cognate pairs and interface engineering studies. The majority of the data were measured using surface plasmon resonance, isothermal titration calorimetry and spectroscopic methods such as stopped-flow fluorescence. Some of the data were reported as inhibition constants, and thus enzyme mutants that reduce catalytic turnover could result in the overestimation of binding strength, such as may be the case for the tissue factor/FAB K165A/K166A mutant (Huang *et al.*, 1998). This is, however, not common in the dataset. In agreement with a previous work (Kastritis *et al.*, 2011), the reported standard errors in K_d for the wild-type and mutant affinities are typically up to around 50%, corresponding to around 0.25 kcal mol^{-1}, while binding energies for interactions that appear more than once tend to vary by around 0.5 kcal mol^{-1}, with differences rarely exceeding 1.5 kcal mol^{-1}. These figures, therefore, give a more accurate picture of the experimental uncertainties in the data.

Many different types of mutations are present, including mutations that destroy intramolecular disulfide bridges, mutations corresponding to structural changes (such as mutations to and from proline) and mutations accompanied by the recruitment and burial of ions to compensate for charges introduced at the interface. Due to the large number of mutations derived from alanine scanning experiments, around a third of the data correspond to mutations to alanine. When these are removed, and amino acids are classified as hydrophobic (A, V, I, L, M, F, Y and W), polar (S, T, N and Q), positive (R, K and H), negative (D and E) or other (C, G and P), 36% of the mutations are within the same category, 13% are from hydrophobic amino acids to polar or charged, 11% are from polar to hydrophobic or charged, 5% correspond to a charge swap, 16% to a mutation from a charged residue to a polar or hydrophobic and the remaining correspond to mutations to or from the ‘other’ category.

A summary of the database is shown in Table 1. The complete database is available at the website http://life.bsc.es/pid/mutation_database/. The crystal structures were checked for unknown and modified residues, most of which were away from the binding site (4CPA, 2BTF, 2FTL, 1DAN, 1S1Q) although a modified residue does covalently attach the subunits of the ubiquitin/UCH-L3 complex (1XD3). The complexes were also checked for co-factors at the binding interface, by searching for chemical species with at least one atom within 4 Å of both binding partners. These included sulphate ions, mostly for the BTPI/trypsin complexes (2FTL, 3BTQ, 3BTT, 3BTE, 3BTM, 3BTD, 3BTH, 3BTF, 3BTW, 3BTG, 2JEL), metal ions (1DVF, 2O3B, 4CPA, 1M8Q, 2J0T), sugars (1DAN, 1NMB), GDP or heme groups buried within one of the monomers (1GRN, 2B10, 2PCC) or agents associated with the buffer or protein precipitation/purification (2O3B, 1TMG, 3BN9, 2WPT, 3NPS). None of these species was found intercalated between the subunits of the complex.

4 DISCUSSION

One of the motivating forces behind assembling this dataset is that a major obstacle to gaining a deep understanding of the binding process and to efficiently engineering protein interfaces is the lack of structural and affinity data (Wodak, 2012). Such affinity data can be used to train appropriate ΔΔG models, ideally with some form of overfitting avoidance bias, complexity commensurate to the number of training examples and accounting for the biases in the training set. A number of ΔΔG functions have been developed previously (Benedix *et al.*, 2009; Guerois *et al.*, 2002; Kamisetty *et al.*, 2011; Tong *et al.*, 2004), although with very few stabilizing mutations present in the data used for parameterization. Furthermore, a large proportion of the data have come from alanine scanning experiments. As such, models have been trained to estimate the influence of clashes or the removal of stabilising contacts. The most common goal of interface engineering, however, is affinity optimization. The data presented here include 303 mutations that stabilize the interaction by at least 0.5 kcal mol^{-1}, and up to 12.4 kcal mol^{-1}. This is, in part, due to a number of crystal structures that have been solved for lower affinity mutant complexes (notably BPTI/trypsin (Helland *et al.*,1999; Pasternak *et al.*, 2001), SGB/OMTKY3 (Bateman *et al.*, 2000, 2001), Colicin E9/IM9 (Keeble *et al.*, 2008), barnase/barstar (Vaughan *et al.*, 1999), hemagglutinin/IgG1 (Fleury *et al.*, 1998), subtilisin BNP’/CI2 (Radisky *et al.*, 2005), Radisky *et al.*, 2004), cytochrome C/peroxidase (Kang and Crane, 2005), HEW lysozyme/antibody (Fields *et al.*, 1996), β-lactamase/BLIP (Reichmann *et al.*, 2005), cyclophilin A/HIV-1 capsid (Howard *et al.*, 2003) and Efb-C/C3d (Haspel *et al.*, 2008) which can be used to train the strengthening reverse mutation back to the wild type. Other sources of strengthening mutations include homolog scanning experiments (Kotzsch *et al.*, 2008; Li *et al.*, 1997), mutations that were initially discovered via phage display (Lang *et al.*, 2000), mutations found via computational interface redesign (Reynolds *et al.*, 2008; Selzer *et al.*, 2000; Kiel *et al.*, 2004) and from systematic scans (Krowarsch *et al.*, 1999; Lu *et al.*, 2001).

The ΔΔG models that have been developed to date combine a small number of weighted terms in linear combination, with few adjustable parameters. However, binding involves many different physical phenomena, and many of these have numerous different models available, rendering the *a priori* selection of terms difficult. In principle, terms could be selected based upon their predictive value, either using a feature selection algorithm or by manual trial and error. However, as related criteria are used to evaluate final models and the predictive value of terms or combinations of terms, both strategies run the risk of overestimating the accuracy of the final model, even if the number of selected terms is restricted by Akaike or Bayesian information criteria. Such concerns can be alleviated by reserving a subset of the data specifically for model validation or by performing cross-validation as an outer loop to the whole model construction process. As a further precaution, the model should not then be revised in light of the performance on the reserved data or the outer cross-validation performance. As long as this is done, then either of these strategies could be used not just for the evaluation of linear models with feature selection (such as Moal and Bates, 2012) but also for more complex machine learning models (such as Moal *et al.*, 2011). Indeed, we suggest that it is best to use both strategies. Final models can be evaluated on separate test sets not included in the presented benchmark, such as the enrichment ratios derived from indirect measurement of binding affinities (Ernst *et al.*, 2010; Weiss *et al.*, 2000) (with the hGH/hGHbp interaction removed from the training data when applicable). Additionally, the cross-validation scheme outlined below can be used.

There are a number of biases present in the benchmark, and the way in which such data are used can bias models towards specific residues, residue types, mutations, proteins, protein families and interfaces. Furthermore, attempts to estimate the generalization error using re-sampling (for instance, bootstrapping or cross-validation) will have similar biases unless accounted for. Table 1 shows an uneven distribution of mutations in specific spatial locations in and away from the interface. Non-interface residues are classified as either surface or interior, and interface residues as rim (partially buried upon binding), core (mostly exposed prior to binding and fully buried in complex) and support (mostly buried prior to binding and fully buried in complex) (Levy, 2010). While the number of interfacial residues in the support, rim and core of interfaces is roughly equal, the database contains twice as many mutations in the rim as in the support region, and twice as many again in the core. The data are also heavily biased towards alanine mutants, which account for around a third of the whole dataset. In model training, one could attempt to account for this using stratified sampling or, better, weighting training mutations against the more frequently occurring category. Another bias is towards

Table 1. A summary of the systems in the SKEMPI database

Protein 1	Protein 2	#muts	#kin	#ther	#PDBs	INT	SUR	COR	SUP	RIM	Ala
Turkey ovomucoid third domain	Streptomyces griseus proteinase B	307	0	0	17	3.6	15.6	50.1	0.3	30.5	3.4
Bovine alpha-chymotrypsin	Turkey ovomucoid third domain	291	0	0	1	3.8	21.4	46.5	9.6	18.7	3.5
Turkey ovomucoid third domain	Subtilisin Carlsberg	280	0	0	1	3.2	16.6	39	9.7	31.5	3.9
Human leukocyte elastase	Turkey ovomucoid third domain	255	0	0	1	3.6	20.2	47.8	0	28.4	4
hGH-binding protein	Human growth hormone	193	58	3	1	16.5	25.9	18.4	18.9	20.3	97.6
BLIP	TEM-1 beta-lactamase	176	100	0	3	6.9	2.9	48.8	31.7	9.6	92.3
Factor VIIa	Tissue factor	130	74	0	1	9.3	23	31.1	10.6	26.1	91.9
HyHEL-10	HEW Lysozyme	111	45	12	1	0	0	82.7	11.8	5.5	43.3
Barnase	Barstar	105	79	31	6	1.3	5.2	56.2	22.2	15	64.7
Interferon gamma receptor	mAbs A6	74	15	0	1	9.1	25.3	37.4	11.1	17.2	47.5
Colicin E9 immunity protein	Colicin E9 DNase	65	57	0	7	14	46.8	18.3	9.7	11.3	31.7
CD4	gp120	56	0	0	1	11.9	50.8	18.6	5.1	13.6	96.6
Colicin E9 DNase	Colicin E2 immunity protein	53	30	0	1	17.1	24	17.7	18.3	22.9	22.3
IgG1-kappa D1.3 Fv	HEW Lysozyme	53	0	3	4	0	1.5	64.2	11.9	22.4	89.6
IgG1-kappa D1.3 Fv	E5.2 Fv	46	0	0	1	0	0	64.4	8.5	27.1	98.3
Angiogenin	Ribonuclease inhibitor	45	45	0	1	1.6	4.8	57.1	27	9.5	84.1
Jel42 antibody	Histadine-containing protein HPr	43	0	0	1	7	39.5	20.9	11.6	20.9	14
Subtilisin BPN	Chymotrypsin inhibitor 2	40	10	0	15	8.9	0	73.3	15.6	2.2	40
HEW Lysozyme	HyHEL-63 Fab	39	0	0	1	0	0	75	21.2	3.8	100
BLIP	SHV-1 beta-lactamase	37	0	0	2	12.7	0	63.6	14.5	9.1	58.2
Interleukin-4	Interleukin-4 receptor	36	36	0	1	2.8	22.2	44.4	19.4	11.1	47.2
Bone morphogenetic protein-2	BMPR-IA receptor	35	29	0	4	12.7	28.7	21	11.5	26.1	29.9
Bovine alpha-chymotrypsin	BPTI	32	15	13	1	9.4	3.1	71.9	6.2	9.4	50
Bovine trypsin	BPTI	29	4	0	10	0	0	96.6	3.4	0	17.2
Membrane-type serine protease 1	S4 Fab	27	0	0	1	11.1	40.7	22.2	0	25.9	100
Membrane-type serine protease 1	BPTI	27	0	0	1	0	44.4	7.4	14.8	33.3	100
TGF-beta type II receptor	Transforming growth factor beta 3	27	2	0	1	22.2	3.7	29.6	11.1	33.3	77.8
Membrane-type serine protease 1	E2 Fab	25	0	0	1	8	32	32	4	24	100
RalGSD-RBD	H-Ras1	25	25	25	1	8.6	42.9	5.7	0	42.9	20
Metalloproteinase inhibitor 1	MMP1 Interstitial collagenase	22	0	0	1	6.7	0	60	3.3	30	26.7
RNase A	Ribonuclease inhibitor	21	17	0	1	0	0	50	34.6	15.4	92.3
UCH-L3	Ubiquitin	18	0	0	1	0	33.3	22.2	22.2	22.2	50
Chemotaxis protein CheY	Chemotaxis protein CheA	17	0	0	1	11.8	29.4	17.6	11.8	29.4	47.1
Cytochrome C peroxidase	Cytochrome C	16	0	12	5	0	0	36.4	9.1	54.5	68.2
HIV-1 capsid protein	Cyclophilin A	16	0	0	2	0	0	50	0	50	43.8
HL-A2-flu	JM22	15	13	11	2	0	0	66.7	20	13.3	60
Protein A/Z	IgG1 MO61 Fc	14	5	0	1	28.6	7.1	21.4	14.3	28.6	57.1
Bone morphogenetic protein-2	Crossveinless 2	13	6	0	1	0	0	46.2	7.7	46.2	38.5
Acetylcholinesterase	Fasciculin	13	6	0	1	11.1	0	44.4	44.4	0	5.6
ZipA	FtsZ fragment	12	0	0	1	0	0	33.3	0	66.7	83.3
Immunoglobulin FAB 5G9	Tissue factor	11	0	0	1	0	7.1	14.3	21.4	57.1	92.9
Phosphatidylinositol 3-kinase	H-Ras1	10	0	0	1	5.3	21.1	0	52.6	21.1	36.8
Subtilisin BPN	Streptomyces subtilisin inhibitor	10	8	0	1	0	0	100	0	0	10
Subtype N9 neuraminidase	Antibody NC10	8	7	0	1	0	0	87.5	0	12.5	12.5
IgG1 MO61 Fc	B domain of Protein G	8	0	0	1	0	0	66.7	0	33.3	100
Carboxypeptidase A	Potato carboxypeptidase inhibitor	8	0	0	1	0	0	87.5	12.5	0	12.5
NucA nuclease	NuiA nuclease inhibitor	8	0	0	1	0	0	55.6	0	44.4	55.6
Fibrinogen-binding protein Efb-C	Complement C3d	8	4	4	3	0	0	100	0	0	50
Integrin alpha-L	Intercellular adhesion molecule I	7	7	0	1	54.5	45.5	0	0	0	0
VavS	Growth factor receptor-bound protein 2	7	0	0	1	0	0	71.4	28.6	0	57.1
Beta-chain of 14.3.d	Staphylococcal enterotoxin C3	7	0	0	1	0	0	57.1	28.6	14.3	100
AML1 Runx1 Runt domain	Core-binding factor beta	6	0	0	1	0	0	50	33.3	16.7	100
Ubiquitin	Tumor susceptibility gene 101 protein	6	0	0	1	0	0	16.7	33.3	50	100
Cytochrome C peroxidase	Non-cognate Cytochrome C	6	0	6	1	0	16.7	16.7	0	66.7	0
Rac-1	p67phox	6	0	0	1	0	66.7	16.7	0	16.7	0
Bovine alpha-chymotrypsin	Eglin c	6	0	0	1	0	0	100	0	0	0
Eglin c	Subtilisin Carlsberg	6	0	0	1	0	0	100	0	0	0
Inhibitor of vertebrate lysozyme	HEW Lysozyme	5	0	0	1	0	0	20	0	80	40

(continued)

Table 1. Continued

Protein 1	Protein 2	#muts	#kin	#ther	#PDBs	INT	SUR	COR	SUP	RIM	Ala
Urokinase-type plasminogen activator	Urokinase plasminogen activator receptor	5	0	0	1	40	20	0	0	40	100
CAR D1 domain	AD12 knob protein	4	0	0	3	0	0	75	0	25	0
Mutant trypsinogen	BPTI	4	0	0	3	100	0	0	0	0	0
Bovine trypsin	Mung bean inhibitor peptide	4	0	0	1	0	0	75	0	25	100
Subtype N9 neuraminidase	Antibody NC41 scFv	4	0	0	1	0	0	100	0	0	0
Human Angiotensin-converting enzyme 2	SARS spike protein receptor binding domain	4	4	0	1	0	50	50	0	0	0
Bovine beta-actin	Bovine profilin	4	0	0	1	0	0	50	25	25	50
Cbl-b UBA	Ubiquitin	4	0	0	1	50	16.7	16.7	0	16.7	0
CAR D1 domain	CAV-2	3	0	0	1	0	33.3	33.3	0	33.3	66.7
Bovine alpha-chymotrypsin	Ecotin	3	0	0	1	0	0	100	0	0	0
TRP region of PEX5	Sterol carrier protein 2	3	0	3	1	100	0	0	0	0	33.3
Arc1p	GluRS	3	0	0	1	0	0	66.7	0	33.3	33.3
IgG1 lambda FAB	Flu virus hemagglutinin	3	2	0	2	0	0	100	0	0	0
Mlc transcription regulator	PTS glucose-specific enzyme EIICB	3	3	0	1	0	0	50	50	0	50
HIV-1 Nef	Fyn SH3 domain R96I mutant	3	0	0	1	0	25	75	0	0	25
B. subtilis endoxylanase	TAXI-I	3	3	0	1	0	0	100	0	0	33.3
Staphylococcal enterotoxin B	Beta-chain of 14.3.d	3	0	0	1	0	0	33.3	0	66.7	0
Bovine alpha-chymotrypsin	PMP-D2v insect inhibitor	2	1	0	1	0	0	100	0	0	0
Bovine alpha-chymotrypsin	PMP-C insect inhibitor	2	1	0	1	0	0	100	0	0	0
Ephrin-B2	Ephrin type-B receptor 4	2	0	2	1	0	0	50	50	0	0
Type II IgNAR	HEW Lysozyme	2	2	0	1	0	50	50	0	0	0
Subtilisin BPN	Eglin c	2	0	0	2	0	0	50	50	0	0
CAR D1 domain	AD37 knob protein	1	0	0	1	0	100	0	0	0	100
Cdc42-GAP	Cdc42	1	0	0	1	0	100	0	0	0	0
Leech metallocarboxypeptidase inhibitor	Carboxypeptidase A	1	0	0	1	100	0	0	0	0	0
Mono-ADP-ribosyltransferase C3	Ras-related protein Ral-A	1	0	1	1	0	0	100	0	0	0
Complement C3d	Ehp	1	0	1	1	0	0	100	0	0	100
	ALL COMPLEXES	3047	713	127	158	6.3	17.4	43.3	10.9	22.1	34.9

Reported for each system is the number of $\Delta\Delta G$ values (#muts), the number of entries with kinetic rate constants (#kin), the number of entries with enthalpy and entropy changes (#ther), the number of relevant PDB structures (#PDBs), the proportion of mutations at the interior (INT), surface (SUR), core (COR), support (SUP) and rim (RIM), as well as the proportion of alanine mutations (Ala).

the specific residues or complexes in the dataset. Thus, instead of using leave-one-out cross-validation, leave-residue-out or better still leave-complex-out cross-validation should be used to estimate the efficacy of the model when applied to residues or proteins other than those in the database. However, even doing this neglects biases towards certain protein families or interfaces. This can sometimes be desirable; the large number of antibody/antigen complexes would be favourable for those looking to apply models to antibody affinity optimization. We may wish to estimate the performance of a model specifically on this class of mutations and leave-complex-out cross-validation would suffice. However, to avoid overestimating the general predictive power of a model on any future unspecified protein complex, this must be accounted for in the validation process, for instance by removing entire classes of complexes together when performing cross-validation. However, in order to do this, one must carefully specify how classes are defined, as we have attempted to do below. Care should be made not to aggregate complexes that need not be put together, as this will reduce the number of cases in the cross-validation training sets, and can lead to overestimation of the generalisation error.

When defining classes to be simultaneously held out during cross-validation, we wanted to hold out pairs of interactions where at least one protein in each pair was using the same, or a homologous, binding site. We did not wish to simultaneously hold out interactions at sites distal to each other, such as the factor VIIa/tissue factor site, where binding is antipodal to the inhibitor binding site. Furthermore, we wanted to capture common binding modes that may not be reflected by sequence homology. In serine protease inhibitor interactions, which are highly represented in the benchmark, inhibition is achieved by a loop that adopts a canonical conformation. The convergent evolution of this local loop structure in unrelated protein families of different global structure and sequence prevents some complexes involving this loop from being classified together using structural or sequence alignment. However, the change in binding energy of mutations within this loop is often independent of the inhibitor family, as shown by interscaffolding additivity cycles (Krowarsch *et al.*, 1999; Qasim *et al.*, 1997). In light of this, we propose a scheme based on the grouping together of interactions based on cross-reactive and homologous binding sites. Supplementary Figure S1 shows the homologies and interactions between the proteins in the database. For all pairs of proteins with a shared binding partner or homologous binding partners, we checked whether the interactions took place at the same binding site (see Section 2). Figure 1 shows pairs of proteins that bind to the same binding site on a third protein (red), or that bind to the same binding site on two homologous proteins (blue). For more than half of the proteins, binding is to a unique site. For many of the rest, the site is shared in one or two other interactions (for instance, the binding site on the β-lactamase inhibitory protein BLIP is shared by both TEM-1 and SHV-1 β-lactamase), and thus these interactions should be simultaneously held out during cross-validation. The remainder fall into three groups. The first is the protease inhibitors, including the inhibitory antibodies E2 and S4, which share protease binding sites. The second is the antigens, which bind to the same region of their respective antibodies. The third group is the proteases, which share a common binding site on the inhibitors. Thus, all protease-inhibitor interactions should also be simultaneously held out, as should all the antibody–antigen interactions. For the membrane-type serine protease, the active site is shared by both protease inhibitors and the S4 and E2 inhibitory antibodies. Thus, interactions involving this protein should be either ignored during evaluation or both classes of interaction should be held out. The classes, as well as all the complexes which should be omitted, are outlined in the SKEMPI database. This scheme allows the evaluation of the generalization error with the minimal amount of grouping required to account for, to the authors' knowledge, all biases arising from overrepresentation of specific types of mutation in the SKEMPI database.

5 CONCLUSION

Presented here is SKEMPI, a large database of experimental binding free energy changes upon mutation, alongside the first large collection of Δk_{on}, Δk_{off}, $\Delta\Delta H$ and $\Delta\Delta S$ values. These data will allow the evaluation of hypotheses regarding the relationship between sequence, structure and binding properties, as well as the evaluation of techniques and the parameterization of empirical functions, which can then be applied to large-scale investigations. While previous benchmarks contained up to 699 structurally cross-referenced $\Delta\Delta G$ values, and very few stabilizing mutations, SKEMPI contains 3047 entries including many that are energetically favourable. Thus, this will allow the training of models that span the whole range of affinities, from greatly

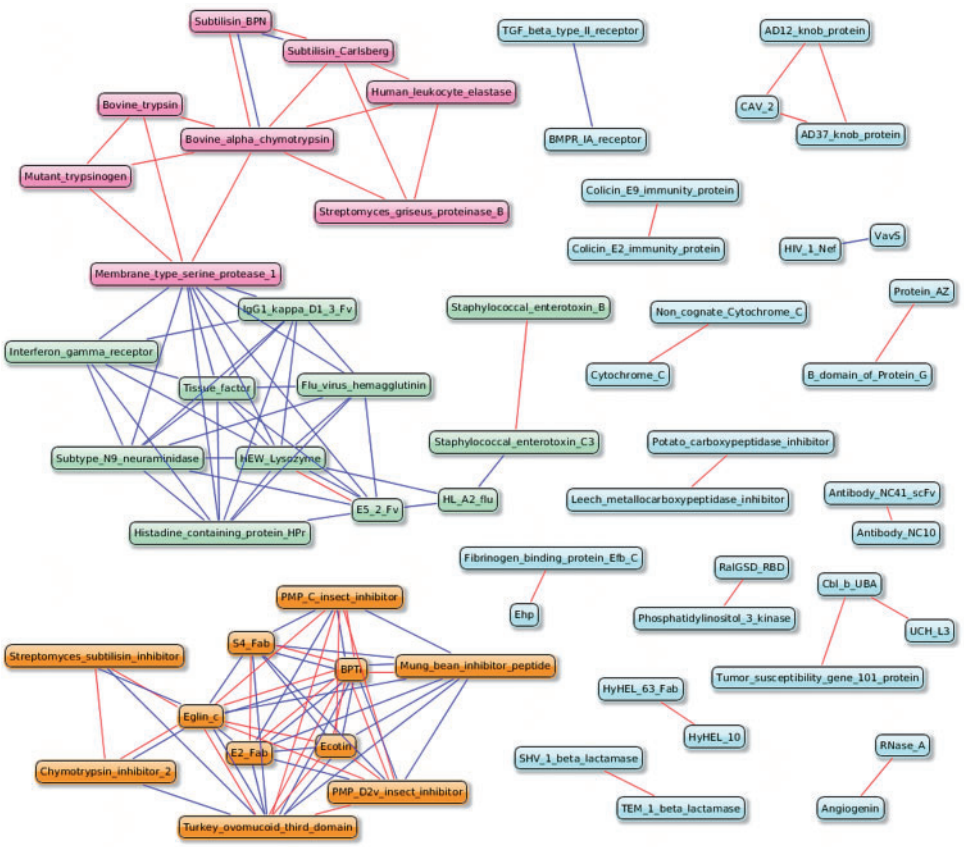

Fig. 1. The proteins which bind to share, or homologous, binding sites. Those connected in red share a binding site on the same protein, while those connected in blue share a binding site on homologous proteins. For instance, the colicin E9 and E2 immunity protein both bind colicin E9 Dnase at the same binding site, and are connected in red. The TGF-β type-II receptor binds to TGF-β3, whose binding site is homologous to the BMPR IA receptor binding site on BMP2. Thus, the TGF-β type-II receptor and the BMPR IA receptor are connected in blue. Pink nodes correspond to proteases, orange nodes correspond to protease inhibitors, the green correspond to the antigens, and cyan to the remaining proteins. Proteins with no shared or homologous binding sites are not shown

stabilizing to highly destabilizing. The applicability of empirical energy functions to the investigation of mutations in proteins outside of the training set depends upon two things. First, the magnitude of energy difference between wild-type and mutant, and second, the generalization error of the model. A model may, for instance, be able to detect mutations that strongly enhance binding, or destroy the interaction, while not being sufficiently accurate to detect subtle changes in ligand specificity. However, in order to accurately estimate the generalization error, the model must be validated in such a way that accounts for the biases in the training data; biases that reflect the interests of the experimentalists who have measured the data. We have investigated these biases in the SKEMPI database, towards protein families, proteins, binding sites, interfaces, positions within the interface, specific residues and alanine mutations. Using these data, we present a leave-complex-out cross-validation scheme that has been adjusted so as to simultaneously hold out interactions at the same, or homologous, binding sites. This scheme prevents binding-site-specific information in the training sets from leaking into the validation sets, allowing the estimation of generalization error, and thus the magnitude of the energy gaps amenable to large-scale investigation.

ACKNOWLEDGEMENTS

The authors thank Brian Jiménez-García for help in constructing the website.

Funding: This work was supported by grant BIO2010-22324 from MICINN, Spain.

Conflict of Interest: none declared.

REFERENCES

Altshuler,D. *et al.* (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Anderson,D.E. *et al.* (1990) pH-induced denaturation of proteins: a single salt bridge contributes 3–5 kcal/mol to the free energy of folding of T4 lysozyme. *Biochemistry*, **29**, 2403–2408.

Araya,C.L. and Fowler,D.M. (2011) Deep mutational scanning: assessing protein function on a massive scale. *Trends Biotechnol.*, **29**, 435–442.

Ashkenazi,A. *et al.* (1990) Mapping the CD4 binding site for human immunodeficiency virus by alanine-scanning mutagenesis. *Proc. Natl. Acad. Sci. USA.*, **87**, 7150–7154.

Bass,S.H. *et al.* (1991) A systematic mutational analysis of hormone-binding determinants in the human growth hormone receptor. *Proc. Natl. Acad. Sci. USA.*, **88**, 4498–4502.

Bateman,K.S. *et al.* (2000) Deleterious effects of beta-branched residues in the S1 specificity pocket of Streptomyces griseus proteinase B (SGPB): crystal structures of the turkey ovomucoid third domain variants Ile18I, Val18I, Thr18I, and Ser18I in complex with SGPB. *Protein Sci.*, **9**, 83–94.

Bateman,K.S. *et al.* (2001) Contribution of peptide bonds to inhibitor-protease binding: crystal structures of the turkey ovomucoid third domain backbone variants OMTKY3-Pro18I and OMTKY3-psi(COO)-Leu18I in complex with Streptomyces griseus proteinase B (SGPB) and the structure of the free inhibitor, OMTKY-3-psi(CH2NH2+)-Asp19I. *J. Mol. Biol.*, **305**, 839–849.

Benedix,A. *et al.* (2009) Predicting free energy changes using structural ensembles. *Nat. Methods*, **6**, 3–4.

Bogan,A.A. and Thorn,K.S. (1998) Anatomy of hot spots in protein interfaces. *J. Mol. Biol.*, **280**, 1–9.

Bonfield,J.K. *et al.* (1995) A new DNA sequence assembly program. *Nucleic Acids Res.*, **23**, 4992–4999.

Draghi,J.A. *et al.* (2010) Mutational robustness can facilitate adaptation. *Nature*, **463**, 353–355.

Empie,M.W. and Laskowski,M. (1982) Thermodynamics and kinetics of single residue replacements in avian ovomucoid third domains: effect on inhibitor interactions with serine proteinases. *Biochemistry*, **21**, 2274–2284.

Ernst,A. *et al.* (2010) Coevolution of PDZ domain–ligand interactions analyzed by high-throughput phage display and deep sequencing. *Mol. Biosyst.*, **6**, 1782–1790.

Farady,C.J. *et al.* (2007) The mechanism of inhibition of antibody-based inhibitors of membrane-type serine protease 1 (MT-SP1). *J. Mol. Biol.*, **369**, 1041–1051.

Fields,B.A. *et al.* (1996) Hydrogen bonding and solvent structure in an antigen-antibody interface. Crystal structures and thermodynamic characterization of three Fv mutants complexed with lysozyme. *Biochemistry*, **35**, 15494–15503.

Fleishman,S.J. and Baker,D. (2012) Role of the biomolecular energy gap in protein design, structure, and evolution. *Cell*, **149**, 262–273.

Fleishman,S.J. *et al.* (2011) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. *Science*, **332**, 816–821.

Fleury,D. *et al.* (1998) Antigen distortion allows influenza virus to escape neutralization. *Nat. Struct. Biol.*, **5**, 119–123.

Fowler,D.M. *et al.* (2010) High-resolution mapping of protein sequence–function relationships. *Nat. Methods*, **7**, 741–746.

Grosdidier,S. and Fernandez-Recio,J. (2012) Protein–protein docking and hot-spot prediction for drug discovery. *Curr. Pharm. Des*, In Press.

Grosdidier,S. and Fernandez-Recio,J. (2008) Identification of hot-spot residues in protein-protein interactions by computational docking. *BMC Bioinformatics*, **9**, 447.

Guerois,R. *et al.* (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.*, **320**, 369–387.

Harel,M. *et al.* (2007) On the dynamic nature of the transition state for protein–protein association as determined by double-mutant cycle analysis and simulation. *J. Mol. Biol.*, **371**, 180–196.

Haspel,N. *et al.* (2008) Electrostatic contributions drive the interaction between Staphylococcus aureus protein Efb-C and its complement target C3d. *Protein Sci.*, **17**, 1894–1906.

Helland,R. *et al.* (1999) The crystal structures of the complexes between bovine beta-trypsin and ten P1 variants of BPTI. *J. Mol. Biol.*, **287**, 923–942.

Hou,T. *et al.* (2011) Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations. *J. Chem. Inf. Model*, **51**, 69–82.

Howard,B.R. *et al.* (2003) Structural insights into the catalytic mechanism of cyclophilin A. *Nat. Struct. Biol.*, **10**, 475–481.

Huang,M. *et al.* (1998) The mechanism of an inhibitory antibody on TF-initiated blood coagulation revealed by the crystal structures of human tissue factor, Fab 5G9 and TF.G9 complex. *J. Mol. Biol.*, **275**, 873–894.

Kamisetty,H. *et al.* (2011) Accounting for conformational entropy in predicting binding free energies of protein–protein interactions. *Proteins*, **79**, 444–462.

Kang,S.A. and Crane,B.R. (2005) Effects of interface mutations on association modes and electron-transfer rates between proteins. *Proc. Natl. Acad. Sci. USA.*, **102**, 15465–15470.

Kastritis,P.L. *et al.* (2011) A structure-based benchmark for protein–protein binding affinity. *Protein Sci.*, **20**, 482–491.

Keeble,A.H. *et al.* (2008) Experimental and computational analyses of the energetic basis for dual recognition of immunity proteins by colicin endonucleases. *J. Mol. Biol.*, **379**, 745–759.

Kelley,R.F. *et al.* (1995) Analysis of the factor VIIa binding site on human tissue factor: effects of tissue factor mutations on the kinetics and thermodynamics of binding. *Biochemistry*, **34**, 10383–10392.

Kiel,C. *et al.* (2004) Electrostatically optimized Ras-binding Ral guanine dissociation stimulator mutants increase the rate of association by stabilizing the encounter complex. *Proc. Natl. Acad. Sci. USA.*, **101**, 9223–9228.

Kortemme,T. and Baker,D. (2002) A simple physical model for binding energy hot spots in protein-protein complexes. *Proc. Natl. Acad. Sci. USA.*, **99**, 14116–14121.

Kortemme,T. and Baker,D. (2004) Computational design of protein–protein interactions. *Curr. Opin. Chem. Biol.*, **8**, 91–97.

Kotzsch,A. *et al.* (2008) Structure analysis of bone morphogenetic protein-2 type I receptor complexes reveals a mechanism of receptor inactivation in juvenile polyposis syndrome. *J. Biol. Chem.*, **283**, 5876–5887.

Krowarsch,D. *et al.* (1999) Interscaffolding additivity: binding of P1 variants of bovine pancreatic trypsin inhibitor to four serine proteases. *J. Mol. Biol.*, **289**, 175–186.

Kumar,M.D. and Gromiha,M.M. (2006) PINT: protein–protein interactions thermodynamic database. *Nucleic Acids Res.*, **34** (Database issue), D195–198.

Kuroda,D. *et al.* (2012) Computer-aided antibody design. *Protein Eng. Des. Sel*, In Press.

Lang,S. *et al.* (2000) Analysis of antibody A6 binding to the extracellular interferon gamma receptor alpha-chain by alanine-scanning mutagenesis and random mutagenesis with phage display. *Biochemistry*, **39**, 15674–15685.

Levy,E.D. (2010) A simple definition of structural regions in proteins and its use in analyzing interface evolution. *J. Mol. Biol.*, **403**, 660–670.

Li,W. *et al.* (1997) Protein–protein interaction specificity of Im9 for the endonuclease toxin colicin E9 defined by homologue-scanning mutagenesis. *J. Biol. Chem.*, **272**, 22253–22258.

Lu,W. *et al.* (1997) Binding of amino acid side-chains to S1 cavities of serine proteinases. *J. Mol. Biol.*, **266**, 441–461.

Lu,S.M. *et al.* (2001) Predicting the reactivity of proteins from their sequence alone: Kazal family of protein inhibitors of serine proteinases. *Proc. Natl. Acad. Sci. USA.*, **98**, 1410–1415.

Mandell,D.J. and Kortemme,T. (2009) Computer-aided design of functional protein interactions. *Nat. Chem. Biol.*, **5**, 797–807.

Moal,I.H. and Bates,P.A. (2012) Kinetic rate constant prediction supports the conformational selection mechanism of protein binding. *PLoS Comput. Biol.*, **8**, e1002351.

Moal,I.H. *et al.* (2011) Protein–protein binding affinity prediction on a diverse set of structures. *Bioinformatics*, **27**, 3002–3009.

Moreira,I.S. *et al.* (2007) Hot spots–a review of the protein–protein interface determinant amino-acid residues. *Proteins*, **68**, 803–812.

Pal,G. *et al.* (2005) Alternative views of functional protein binding epitopes obtained by combinatorial shotgun scanning mutagenesis. *Protein Sci.*, **14**, 2405–2413.

Pasternak,A. *et al.* (2001) The energetic cost of induced fit catalysis: crystal structures of trypsinogen mutants with enhanced activity and inhibitor affinity. *Protein Sci.*, **10**, 1331–1342.

Qasim,M.A. *et al.* (1997) Interscaffolding additivity. Association of P1 variants of eglin c and of turkey ovomucoid third domain with serine proteinases. *Biochemistry*, **36**, 1598–1607.

Radisky,E.S. *et al.* (2004) Binding, proteolytic, and crystallographic analyses of mutations at the protease-inhibitor interface of the subtilisin BPN'/chymotrypsin inhibitor 2 complex. *Biochemistry*, **43**, 13648–13656.

Radisky,E.S. *et al.* (2005) Role of the intramolecular hydrogen bond network in the inhibitory power of chymotrypsin inhibitor 2. *Biochemistry*, **44**, 6823–6830.

Reichmann,D. *et al.* (2005) The modular architecture of protein–protein binding interfaces. *Proc. Natl. Acad. Sci. USA.*, **102**, 57–62.

Reichmann,D. *et al.* (2007) Binding hot spots in the TEM1-BLIP interface in light of its modular architecture. *J. Mol. Biol.*, **365**, 663–679.

Reynolds,K.A. *et al.* (2008) Computational redesign of the SHV-1 beta-lactamase/beta-lactamase inhibitor protein interface. *J. Mol. Biol.*, **382**, 1265–1275.

Schreiber,G. and Fersht,A.R. (1995) Energetics of protein-protein interactions: analysis of the barnase-barstar interface by single mutations and double mutant cycles. *J. Mol. Biol.*, **248**, 478–486.

Selzer,T. *et al.* (2000) Rational design of faster associating and tighter binding protein complexes. *Nat. Struct. Biol.*, **7**, 537–541.

Thorn,K.S. and Bogan,A.A. (2001) ASEdb: a database of alanine mutations and their effects on the free energy of binding in protein interactions. *Bioinformatics*, **17**, 284–285.

Tischkowitz,M. *et al.* (2008) Pathogenicity of the BRCA1 missense variant M1775K is determined by the disruption of the BRCT phosphopeptide-binding pocket: a multi-modal approach. *Eur. J. Hum. Genet.*, **16**, 820–832.

Tong,W. *et al.* (2004) Computational prediction of binding hotspots. *Conf. Proc. IEEE Eng. Med. Biol. Soc.*, **4**, 2980–2983.

Tuncbag,N. *et al.* (2009) Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics*, **25**, 1513–1520.

Vaughan,C.K. *et al.* (1999) Structural response to mutation at a protein–protein interface. *J. Mol. Biol.*, **286**, 1487–1506.

Weikl,T.R. and von Deuster,C. (2009) Selected-fit versus induced-fit protein binding: kinetic differences and mutational analysis. *Proteins*, **75**, 104–110.

Weiss,G.A. *et al.* (2000) Rapid mapping of protein functional epitopes by combinatorial alanine scanning. *Proc. Natl. Acad. Sci. USA.*, **97**, 8950–8954.

Whitehead,T.A. *et al.* (2012) Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nat. Biotechnol.*, **30**, 543–548.

Wodak,S.J. (2012) Next-generation protein engineering targets influenza. *Nat. Biotechnol.*, **30**, 502–504.

Wu,X. *et al.* (2011) Focused evolution of HIV-1 neutralizing antibodies revealed by structures and deep sequencing. *Science*, **333**, 1593–1602.

Xia,J.F. *et al.* (2010) APIS: accurate prediction of hot spots in protein interfaces by combining protrusion index with solvent accessibility. *BMC Bioinformatics*, **11**, 174.