Protein engineering in the big data era: harnessing near-redundant structural data

Samuel Coulbourn Flores[1,*], Athanasios Alexiou[2] and Anastasios Glaros[3]

[1]Department of Biochemistry and Biophysics, Stockholm University, SWEDEN., [2]Department of Computer Science and Biomedical Informatics, University of Thessaly, GREECE, [3]Eukaryotic Single Cell Genomics Facility, Science For Life Laboratory, Stockholm, SWEDEN

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Predicting the effect of mutations on protein-protein interactions is important for relating structure to function, as well as for *in silico* affinity maturation. The effect of mutations on protein-protein binding energy (ΔΔG) can be predicted by a variety of atomic simulation methods involving full or limited flexibility, and explicit or implicit solvent. Methods which consider only limited flexibility are naturally more economical, and many of them are quite accurate, however results are dependent on the atomic coordinate set used. In this work we perform a sequence and structure based search of the Protein Data Bank to find additional coordinate sets and repeat the calculation on each.

Results: . We improve increase precision and Positive Predictive Value, and decrease Root Mean Square Error and higher Positive Predictive Value, compared to using single structures. Given the ongoing growth of near-redundant structures in the Protein Data Bank, our method will only increase in applicability and accuracy.

Availability: Public web server at biodesign.scilifelab.se

Contact: Samuel.Flores@scilifelab.se

Supplementary information: Supplementary data are available online.

1 Introduction

In this work we are interested in predicting the change in protein-protein binding energy (ΔΔG) resulting from amino acid substitutions at the protein-protein interface. This quantity determines the change in protein-protein binding affinity and is thus important for understanding signaling, complex assembly, catalysis, host-pathogen interaction, and other functions. It is also industrially interesting, as it is key to designing proteins for therapy and diagnosis, as well as for purification and catalysis.

Methods of computing change in protein-protein binding energy (ΔΔG)

$\Delta\Delta G = \Delta G_{mutant} - \Delta G_{wild\text{-}type}$. $\Delta G_{wild\text{-}type}$ is the free energy change upon binding in the wild-type complex, while ΔG_{mutant} is the same quantity for the mutant complex. The equilibrium constant widely used in pharmacology is computed as Kd = exp(ΔG/RT), where R is the universal gas constant.

Many computational methods exist to calculate $\Delta\Delta G_{predicted}$, an estimate of the (known or unknown) experimental value, $\Delta\Delta G_{experimental}$. Some methods use reduced representations (Dehouck *et al.*, 2009, 2013), while others include all atoms. (Guerois *et al.*, 2002) such methods compute the protein-protein binding enthalpy (including electrostatic and van der Waals interactions) using physical formulae, but differ in the way they estimate the effect of solvent and side-chain entropy. The most successful and widely-used methods use implicit solvent to estimate these latter terms, namely they compute the solvent-accessible surface area on an atomic basis, then combine this quantity with the atom type and empirically-adjusted weight factors. (Guerois *et al.*, 2002; Cornell Cieplak, P., Bayley, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. & Kollman, P. A., 1995; Massova and Kollman, 2000; Li *et al.*, 2016) In recent years such approaches have made relatively small gains in accuracy.

Many limited-flexibility, implicit-solvent methods including FoldX offer good accuracy and economy. (Guerois *et al.*, 2002; Lowegard Anna U. AND Frenkel, 2020; Pires *et al.*, 2014) Multiple workers have found it is actually counterproductive to minimize the structure globally (e.g. by Molecular Dynamics or MD); (Petukh *et al.*, 2015a; Beard *et al.*, 2013) rather it is better to model the substitution and limit further modifications to those required for annealing the resulting steric clashes (a mostly-local minimization). All force fields are inevitably biased by the data they are fitted to, and most are classically formulated, considering quantum mechanical effects only indirectly. Thus modeling can only reduce the accuracy of 3D atomic coordinates, compared to X-ray crystallography or other high-resolution experimental methods. Also, potentials which successfully predict changes in protein-protein binding energy upon mutation ($\Delta\Delta G$) are typically trained on experimental structures (Guerois *et al.*, 2002). All of this argues in favor of limited, local minimization. The downside of *local* minimization is that it makes the results dependent on the idiosyncrasies of the experimental coordinates which may reflect crystallization conditions, which would change upon mutation, or which represent only one of many thermodynamically accessible configurations. This work addresses this limitation of local minimization, by identifying and using additional structural data.

The Protein Data Bank (PDB) is the main repository of protein structural data. Growth of data in the PDB is accelerating, and this includes growth in near-redundant structures – which reveal multiple observations of the same or closely related complexes under different experimental conditions. We can repeat the $\Delta\Delta G$ calculation over all such structures, and average over the results to increase precision.

Near-redundant structures in the Protein Data Bank (PDB)

Although the era of fold discovery is over, the growth of structural data is still accelerating. Many of the new structures differ only slightly from existing entries, having been obtained to e.g. seek higher resolution, determine the effect of a mutation, or add a ligand or subunit. In this work "near-redundant structures" refer to those which have the same (or nearly the same) composition, at least in the protein-protein interface of interest, but which were resolved in separate experiments or crystallographic units. The experimental effort is dramatically reduced when following proven protocols as opposed to solving proteins of previously unknown structure. Thus it is economical to solve the same complex to probe structural variations, improve resolution, etc. As an illustrative example, the complex of the human Growth Hormone (hGH) bound to two copies of its Receptor (hGHR) was resolved in (Receptor *et al.*, 1996). There exists a mutant of hGH which binds one one copy of hGHR, and the same work also reports the 1:1 complex structure (Receptor *et al.*, 1996). The same lab then remodeled part of the interface by point and phage display mutagenesis and reported the structure (Atwell *et al.*, 1997). Lastly, they solved the 1:1 complex again at improved resolution (Clackson *et al.*, 1998). Even minor differences in biopolymer sequence, number of additional subunits, experimental conditions, and fitting procedure can be expected to introduce differences in atomic positions on the order of tenths of Ångströms. While for some purposes these differences may be insignificant, they lead to variations in predicted $\Delta\Delta G$ when using FoldX and other limited-flexibility methods.

Why does averaging improve precision?

Accuracy is the closeness of the prediction or measurement to the correct value, whereas precision refers to the closeness of independent predictions or measurements to each other. The principle underlying homologyScanner is that ZEMu and other methods which perform only a local minimization are subject not only to limitations in accuracy (due to biases in the $\Delta\Delta G$ prediction potential, and to errors in the $\Delta\Delta G_{experimental}$ used as a gold standard) but also in precision (due to particularities of the coordinate sets used).

According to the central limit theorem, for N independent random variables distributed with mean μ and standard deviation σ, the sample mean approaches a normal distribution with mean μ and standard deviation $\sigma/\sqrt{N}$. Here the $\Delta\Delta G_{predicted}$ computed using a single structure would be a random variable sampled from an underlying normal distribution of unknown mean μ and standard deviation σ. A $\Delta\Delta G_{predicted}$ averaged from *multiple* calculations would have a smaller standard deviation, $\sigma/\sqrt{N}$, about the same mean μ.

Thus averaging multiple calculations obtained using independent structures provides a more-accurate estimate of the underlying mean μ. Note that even in the hypothetical case of very large N, μ would still be the mean of many individual values of $\Delta\Delta G_{predicted}$, subject to biases in the FoldX potential, and may not converge to an accurately measured $\Delta\Delta G_{experimental}$. Also the latter number depends on experimental conditions. And so though precision is increased by our method, one must consider limitations in the force fields and experimental measurements.

2 Methods

In our method, the user proposes a mutation, the PDB identifier of a single "query" structure, and two lists of chains, one for each subunit in the interaction of interest. Lastly, the user specifies the chain ID, residue position, and substituted residue type, for one or multiple simultaneous substitutions. The user typically needs not perform any further actions until the calculation completes, the remaining steps are automated, per the flowchart (Figure 1). The procedure is illustrated graphically in Figure 2. homologyScanner is an extension of MacroMoleculeBuilder (MMB).

MacroMoleculeBuilder (MMB)

MMB is a general-purpose, multiscale modeling code. Its internal-coordinate framework (Flores *et al.*, 2011) gives us full control over the flexibility of our molecular system, thus one can have chains which are fully rigid, fully flexible, or which are rigid in some parts and flexible in others. In past work we have used MMB for applications as diverse as morphing (Tek *et al.*, 2016), local minimization (Dourado and Flores, 2014), and fitting to low-resolution electron density maps (Flores, 2014). MMB can also do homology modeling, which may be considered as alignment of a flexible chain (of presumed unknown structure) to a rigid chain of known structure (Flores *et al.*, 2010). In a related technique, both chains can be partially-flexible or fully-rigid, and here the alignment can be called template docking (Dourado and Flores, 2016) or simply rigid alignment, depending on how constraints are applied. MMB can also use the Kabsch algorithm to compute the minimum Root Mean Square Deviation (RMSD) with which two complexes can be aligned – this is a very fast operation.

Sequence search and alignment

homologyScanner starts by searching the PDB (Lopez *et al.*, 2014) for chains which match the sequence of the user-specified chains with very high statistical significance (e-value <= 10^{-11}). Only structures which contain all the user-specified chains, each satisfying the e-value requirement, are kept and the rest are discarded. homologyScanner then uses the SeqAn-based (Döring *et al.*, 2008) alignment tools in MacroMoleculeBuilder (MMB) to compute the sequence identity for each corresponding chain, viz:

Matching residues / minimum(query chain length, subject chain length)

Structures which do not satisfy the minimum sequence identity (> 90%) are discarded. We thus know that the remaining structures have the relevant chains, but do not yet know whether they have the correct tertiary and quaternary structure. For that we perform the final structural check using MMB as follows.

Structural alignment

The Kabsch structural alignment is based on residue-residue (and ultimately atom-atom) correspondence between the query and subject structures, which we obtain from the mentioned sequence alignment. MMB can robustly deal with missing or non-canonical atoms (often encountered in PDB structures), usually without user intervention. (Tek *et al.*, 2016) The Kabsch alignment gives us the RMSD of the query vs. subject complex. If the RMSD meets a cutoff threshold (<= 6Å), the homolog complex is then passed on to the ΔΔG calculation. Note that this differs from the procedure of (Sippl, 1993), in which the PDB is searched on structure but not sequence; we wished to use only structures of very similar sequence to maintain accuracy.

ΔΔG calculation

The user specifies a mutation(s) with chain ID(s) and residue number(s) in the context of the query structure. However different subject structures may employ different residue numbering conventions. We translate the user-specified mutation into the numbering system of the subject structure on the basis of the sequence alignment. We the compute $\Delta\Delta G_{predicted}$ on the query and all subject structures, using FoldX.

3 Results

We benchmarked homologyScanner on the dataset used in (Dourado and Flores, 2014), comprising 1243 mutations (see Table 1, dataset A). This is a very diverse dataset of 1243 mutants, including some mutants with single-substitutions and some with multiple simultaneous substitutions. We first tried using only single structures, as done in (Dourado and Flores, 2014), and then repeated using multiple structures and quantified the improvement in correlation and Root Mean Square Error (RMSE) (Dourado and Flores, 2014):

$$RMSE = \frac{1}{N}\sum_{i=1}^{N}(\Delta\Delta G_{i,predicted} - \Delta\Delta G_{i,experimental})^2$$

We also tested a subset of the above, comprising single- and multiple-substitution mutants for which multiple structures were available (Table 1, dataset B). We further subdivided into single- and multiple-substitution mutants available (Table 1, datasets C and D). The RMSE decreases as number of available homologs increases from 1 to 4 (Supplementary Figure SF 1). However from 5 homologs onwards RMSE increases as number of data points becomes small and begins to consist of $\Delta\Delta G_{experimental}$ measurements from a single lab.

A scatterplot (Figure 3) of $\Delta\Delta G_{predicted}$ vs. $\Delta\Delta G_{experimental}$ for dataset C, shows the effect of averaging on outliers.

We also computed a Receiver Operating Characteristic (ROC) curve. This plots the True Positive Rate (TPR) vs. True Negative Rate (TNR) as the $\Delta\Delta G_{predicted}$ *threshold* is moved from $\Delta\Delta G_{predicted} = +\infty$ (loosest) to $-\infty$ (strictest). Mutations with $\Delta\Delta G_{predicted} <$ *threshold* are taken to be test positives, mutations with $\Delta\Delta G_{experimental} < 0$ are taken to be Gold Standard positives, and so e.g. mutations in the intersect set (i.e. those that meet both of these criteria) are True Positives (TP). Accordingly, FP (False Positives) are those mutations for which with $\Delta\Delta G_{predicted} <$ *threshold* but $\Delta\Delta G_{experimental} \geq 0$. The rest of the quantities (TN: True

Negatives, FN: False Negatives, etc.) are computed accordingly. The ROC was also computed based on single structures, and on multiple structures for comparison.

Another useful quantity is the Positive Predictive Value (PPV) = TP/(TP + FP). This answers the question: for a given threshold, what fraction of test positives will be TPs? If the goal is to get $\Delta\Delta G > 0$, then PPV tells us which fraction would have achieved this, for a given test threshold. Again we compute for single as well as multiple structures for the full range of test thresholds (Figure 5).

4 Discussion

As noted previously many ΔΔG prediction methods that limit structural rearrangements, particularly in regions distant from the mutation site, can yield good results, compared to no minimization or minimization of entire structures. (Guerois *et al.*, 2002) However a local minimization leaves us vulnerable to structural idiosyncrasies of the structure employed. In the present work, we diversify the coordinate data by identifying additional complexes including the relevant chains, in the relevant quaternary arrangement. Significantly, we introduce no adjustable parameters and so our results should apply to other potentials and localized minimization methods. The part of the error due to experimental uncertainty and limitations of the DDG potential, remains unchanged in our work but is the topic of ongoing research in the field. (Brender *et al.*, 2015; Petukh *et al.*, 2015b; Guerois *et al.*, 2002)

Our benchmarking was done using the dataset of (Dourado and Flores, 2014), itself compiled from SKEMPI (Moal and Fernández-Recio, 2012). The full dataset **A** contains mutants with single and multiple simultaneous substitutions. More to the point, for some mutants in dataset A only one structure is available, whereas for others there are several similar structures available (in one case 18 were found, see Supplementary Table S1). When only one structure is available of course homologyScanner makes no improvement, but since for many there were multiple structures, homologyScanner reduced RMSE by about 0.1 kcal/mol.

The more relevant comparison is the case for which multiple structures are available (datasets B, C, and D). Dataset C comprises only single-substitution, D comprises multiple-substitution mutations, while B contains both. homologyScanner presented the largest RMSE improvement for D, but error was high overall, so in general we do not recommend using homologyScanner for multiple-substitutions. For the single-substitutions (dataset C), the best RMSE of all, 1.04 kcal/mol, was obtained, better than reported in related work. (Guerois et al., 2002; Dourado and Flores, 2014)

Aggregate results however are only part of the story. Figure 3 highlights another important feature of homologyScanner:F whereas several outliers are evident when using single structures, there are arguably *zero* outliers when using multiple structures. For a given $\Delta\Delta G_{experimental}$, $\Delta\Delta G_{predicted}$ is consistently closer to the trendline for multiple than for single structures. One may note that the slope $\Delta\Delta G_{predicted}/\Delta\Delta G_{experimental}$ is not unity, this is a characteristic of FoldX which we do not reparameterize here; in any case the slope itself is not as important as the statistical measures of accuracy.

ROC curves are commonly used to evaluate binary classifiers with an adjustable threshold – in this case, we can classify mutations into those predicted to decrease ΔΔG (improve affinity), vs. those that should increase ΔΔG or leave it neutral. In ROC curves, Area Under the Curve (AUC) and slope at the point TNR = 1, TPR = 0 are two important measures. Larger AUC's correspond to more significant classifiers, whereas steeper slope indicates better performance for the highest-confidence cases, here those with lowest computed ΔΔG. Both quantities are larger when using multiple structures (Figure 4).

But perhaps the most important statistic, again for the purposes of design, is PPV, plotted in Figure 5. For single structures, PPV fluctuates around 0.5 for the highest-confidence mutants, that is to say in the range of $\Delta\Delta G_{predicted} < -1$ kcal/mol. For multiple structures, in contrast, PPV is a solid 1.0 in the same range of $\Delta\Delta G_{predicted}$. To reiterate, in an experiment *all* such mutants would have been found to improve affinity. While we believe this result is important and impressive, we also urge the reader to be cautious. This dataset is compiled from *published* data, which we strongly suspect contains more affinity-improving mutations than would be obtained by random mutagenesis. We reason that may investigators are looking to improve affinity, and will use tools at their disposal – published and unpublished data, structural calculations, bioinformatics, etc., prior to attempting a new mutation – and if they do not succeed they may decide not to publish. So while the case is strong for using multiple structures, we believe PPV will be less than unity in new applications.

In conclusion, we have presented a protocol for taking advantage of the growing accumulation of near-redundant structures in the PDB to improve prediction of ΔΔG. Though the approach is simple, it provides an improvement which is remarkable since clearly demonstrable improvements in ΔΔG accuracy have been slow in recent years. The method should be compatible with perturbative ΔΔG potentials other than FoldX. As multiple programs are required to implement homologyScanner efficiently, and since there is considerable incentive to reuse calculations, we make the method publicly available on an easy to use web server.

4 Distribution

HomologyScanner is available on a public server at biodesign.scilifelab.se. The setup is shown in Figure 6. To request a ΔΔG calculation, the user goes to the *Submit* tab and provides the PDB ID of one suitable structure, and specifies the relevant chains in subunit 1 and subunit 2 of the interaction of interest (chains not in the interface can be left out). The user then specifies the mutation to be computed (one to four simultaneous substitutions). homologyScanner is then invoked, meaning the PDB is searched for structurally similar complexes, the FoldX ΔΔG calculation is performed for all such complexes found, and the user is notified by email (usually within a few hours, depending on queue status and job characteristics) when the job is done. The server saves all results, so the structure search needs be done only once per PDB ID and definition of subunits 1 and 2, and each FoldX calculation is only done once in total.

There is also a *View* tab where the public can browse all results by selecting a PDB ID, subunits, and mutation (all from drop-down lists). They will then see the ΔΔG for all structural homologs, as well as the average ΔΔG. A Jsmol window displays the protein structure in *Cartoon* render style, with the mutation highlighted in *Sticks* style.

The web server itself comprises a computer with at least two CPU cores, one of which is responsible for running Apache, MySQL, and other web services. The remaining cores are managed by the SLURM queueing system. The web server submits homologyScanner jobs to this queue upon user request. These jobs call homologyScanner itself, which interacts with the PDB to perform the sequence and structure search. homologyScanner spawns one *breeder* job for each suitable homologous structure found. Breeder is a program introduced in (Dourado and Flores, 2014) which manages FoldX and stores ΔΔG and related results in the MySQL database.

Privacy may be important for some users, for example academics with unpublished data, or product developers in the pharmaceutical industry. For such users we have prepared a low-cost Udoo X86 Ultra single board computer implementation. The compact (120x85 mm) and light (under 200 g, excluding power adapter) format means it can easily be posted to the an academic or industry user. The X86 architecture ensures ease of compilation and update, compared to ARM architectures used by other single-board computers. The computer has a 2.56 GHz Intel Pentium quad-core processor. One core is used for running the web server, including MySQL database, and 1-3 cores are managed by SLURM for running homologyScanner. The computer can be fitted with an M.2-format Solid State Drive with up to 1TB capacity (we used 512 GB) as well as external HDs. We used a StarTech 300Mbps mini-wireless network adapter. An HDMI and three USB-3.0 type A ports mean it can be connected to a display, keyboard, mouse, and wireless network adapter. Alternatively such users can provide their own hardware (with mysql, docker, slurm) and use the dockerhub image (samuelflores/mmb-ubuntu:homologyscanner) which contains MMB, Breeder, and homologyScanner.

Acknowledgements

We thank Javier Delgado Blanco, Luis Serrano Pubul for discussions and help, including custom Rapsberry Pi compilations of FoldX. Andrei Rajkovic compiled homologyScanner, breeder, and MMB for the Raspberry Pi. We thank Chengxin Zhang for useful advice. The authors have no known conflicts of interest.

Funding

We gratefully acknowledge funding from the Swedish Research Council, the Swedish Foundation for International Cooperation in Research and Higher Education and Lars Hierta Memorial Foundation, and supercomputer time from the Swedish National Infrastructure for Computing

Conflict of Interest: none declared.

References

Atwell,S. *et al.* (1997) Structural Plasticity in a Remodeled Protein-Protein Interface. *Science (80-.).*, **278**.

Beard,H. *et al.* (2013) Applying physics-based scoring to calculate free energies of binding for single amino acid mutations in protein-protein complexes. *PLoS One*, **8**, e82849.

Brender,J.R. *et al.* (2015) Predicting the Effect of Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles. *PLOS Comput. Biol.*, **11**, e1004494.

Clackson,T. *et al.* (1998) Structural and functional analysis of the 1:1 growth hormone:receptor complex reveals the molecular basis for receptor affinity. *J. Mol. Biol.*, **277**, 1111–28.

Cornell Cieplak, P., Bayley, C. I., Gould, I. R., Merz, K. M., Ferguson, D. M., Spellmeyer, D. C., Fox, T., Caldwell, J. W. & Kollman, P. A.,W.D. (1995) A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.*, **117**, 5179–5197.

Dehouck,Y. *et al.* (2013) BeAtMuSiC: Prediction of changes in protein-protein binding affinity on

mutations. *Nucleic Acids Res.*, **41**, W333-9.

Dehouck,Y. *et al.* (2009) Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics*, **25**, 2537–2543.

Döring,A. *et al.* (2008) SeqAn An efficient, generic C++ library for sequence analysis. *BMC Bioinformatics*, **9**, 11.

Dourado,D.F.A.R. and Flores,S.C. (2014) A multiscale approach to predicting affinity changes in protein-protein interfaces. *Proteins*, **82**, 2681–90.

Dourado,D.F.A.R. and Flores,S.C. (2016) Modeling and fitting protein-protein complexes to predict change of binding energy. *Nat. Sci. Reports*, **6**, 25406.

Flores,S.C. (2014) Fast fitting to low resolution density maps: elucidating large-scale motions of the ribosome. *Nucleic Acids Res.*, **42**, 1–10.

Flores,S.C. *et al.* (2011) Fast flexible modeling of RNA structure using internal coordinates. *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, **8**.

Flores,S.C. *et al.* (2010) Predicting RNA structure by multiple template homology modeling. *Pac. Symp. Biocomput.*, 216–27.

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**, 369–87.

Li,M. *et al.* (2016) MutaBind estimates and interprets the effects of sequence variants on protein-protein interactions. *Nucleic Acids Res.*, **44**, W494–W501.

Lopez,R. *et al.* (2014) Using EMBL-EBI Services via Web Interface and Programmatically via Web Services. *Curr. Protoc. Bioinforma.*, **2014**, 3.12.1-3.12.50.

Lowegard Anna U. AND Frenkel,M.S.A.N.D.H.G.T.A.N.D.J.J.D.A.N.D.O.A.A.A.N.D.D.B.R. (2020) Novel, provable algorithms for efficient ensemble-based computational protein design and their application to the redesign of the c-Raf-RBD:KRas protein-protein interface. *PLOS Comput. Biol.*, **16**, 1–27.

Massova,I. and Kollman,P.A. (2000) Combined molecular mechanical and continuum solvent approach (MM-PBSA/GBSA) to predict ligand binding. *Perspect. Drug Discov. Des.*, **18**, 113–135.

Moal,I.H. and Fernández-Recio,J. (2012) SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics*, **28**, 2600–7.

Petukh,M. *et al.* (2015a) Predicting Binding Free Energy Change Caused by Point Mutations with Knowledge-Modified MM/PBSA Method. *PLoS Comput. Biol.*, **11**, e1004276.

Petukh,M. *et al.* (2015b) Predicting Binding Free Energy Change Caused by Point Mutations with Knowledge-Modified MM/PBSA Method. *PLoS Comput. Biol.*, **11**, e1004276.

Pires,D.E. V *et al.* (2014) mCSM: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, **30**, 335–42.

Receptor,I. *et al.* (1996) Protein Chemistry and Structure : Crystal Structure of an Antagonist Mutant of Human Growth Hormone , G120R , in Resolution Crystal Structure of an Antagonist Mutant of Human Growth Hormone , G120R , in Complex with Its Receptor at 2 . 9 Å Resolution *.

Sippl,M.J. (1993) Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Comput. Aided. Mol. Des.*, **7**, 473–501.

Tek,A. *et al.* (2016) MMB-GUI: a fast morphing method demonstrates a possible ribosomal tRNA translocation trajectory. *Nucleic Acids*

Res., **44**, 95–105.

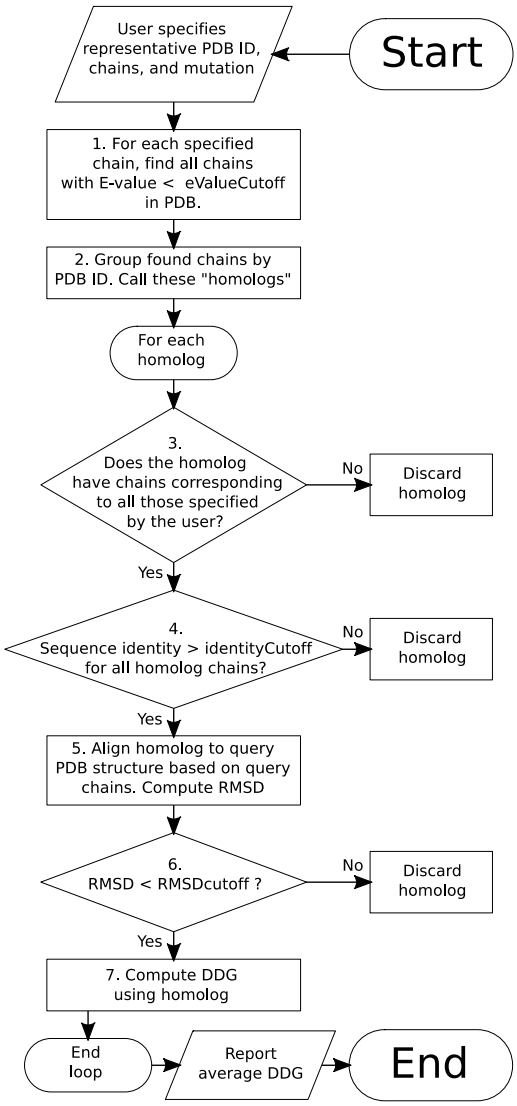

Figure 1 . Program flow. The user must provide an initial Protein Data Bank (PDB) ID, specify which relevant chains are in which of two interacting complexes (irrelevant chains may be left out). 1. The fasta_lwp program searches the PDB for structures containing chains homologous (E-value below eValueCutoff, here 10^{-11}) to those specified by the user. 2. We group the thus-discovered homolog chains by PDB ID, each such PDB ID is referred to as a "homolog." We loop over the homologs, performing three checks on each. 3. As a first check, we determine whether the thus-discovered homologs contain chains corresponding to all those specified by the user; those not having all such chains are discarded. 4. Homologs in which all chains do not have at least 90% sequence identity vs. the corresponding user-specified chain are discarded. 5. We perform a rigid alignment of the entire homolog against the user-specified structure, based only on the user-specified chains. Non-corresponding (extraneous) chains are moved along with the rest of the complex. This is the most computationally-expensive process, but only needs to be done once for homolog that makes it to this step; results of all three checks are saved persistently. 6. If RMSD > 6.0 Å (again based on corresponding chains), we discard the homolog. Most homologs which are rejected at this step contain the correct chains but in a different configuration. 7. We then compute the ΔΔG for the user-requested mutation, using the homolog structure and FoldX4. Steps 3-7 are repeated for each homolog. 8. We average ΔΔG over all homologs that reached and completed step 7 and report the result.

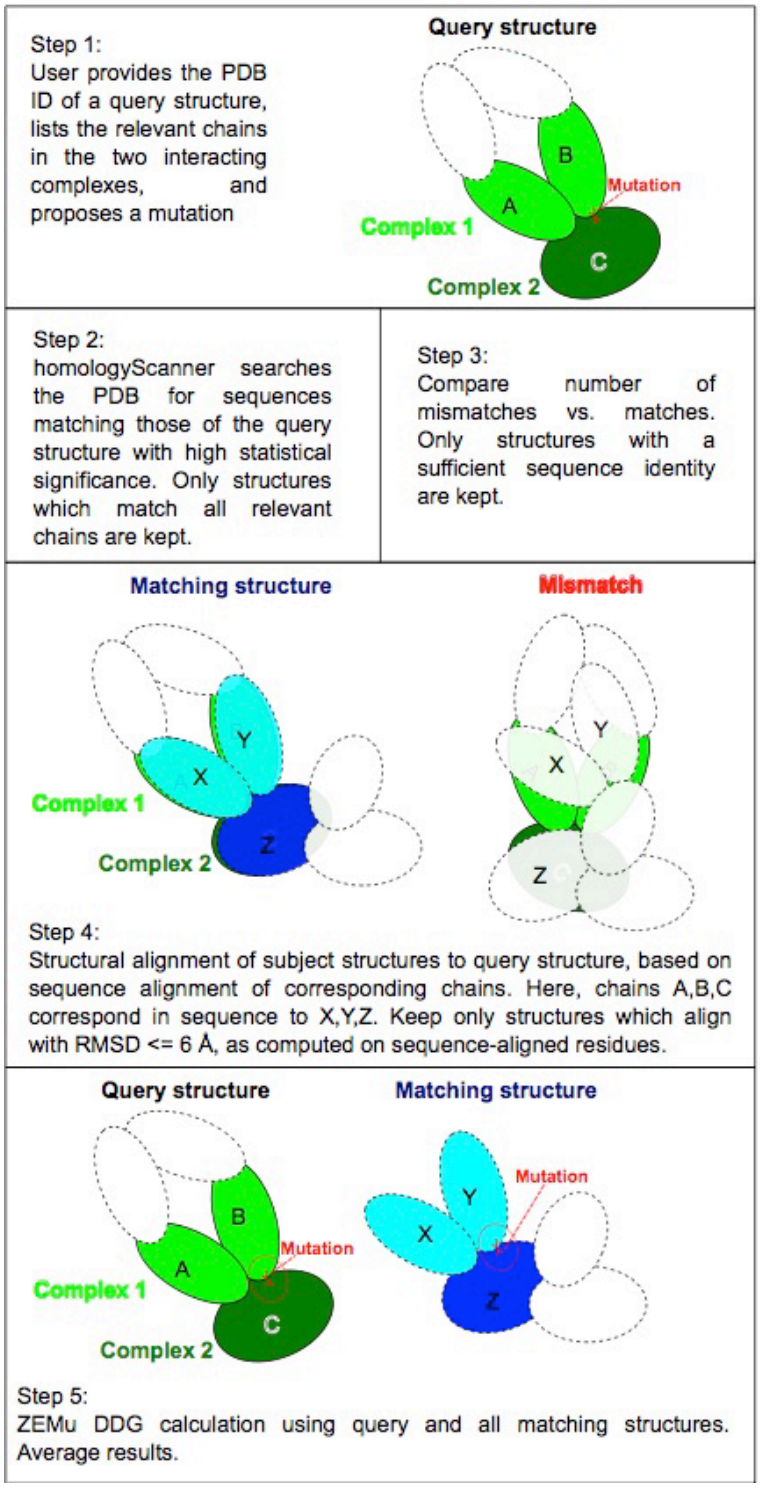

Figure 2. Illustration of the sequence and structure matching procedure.

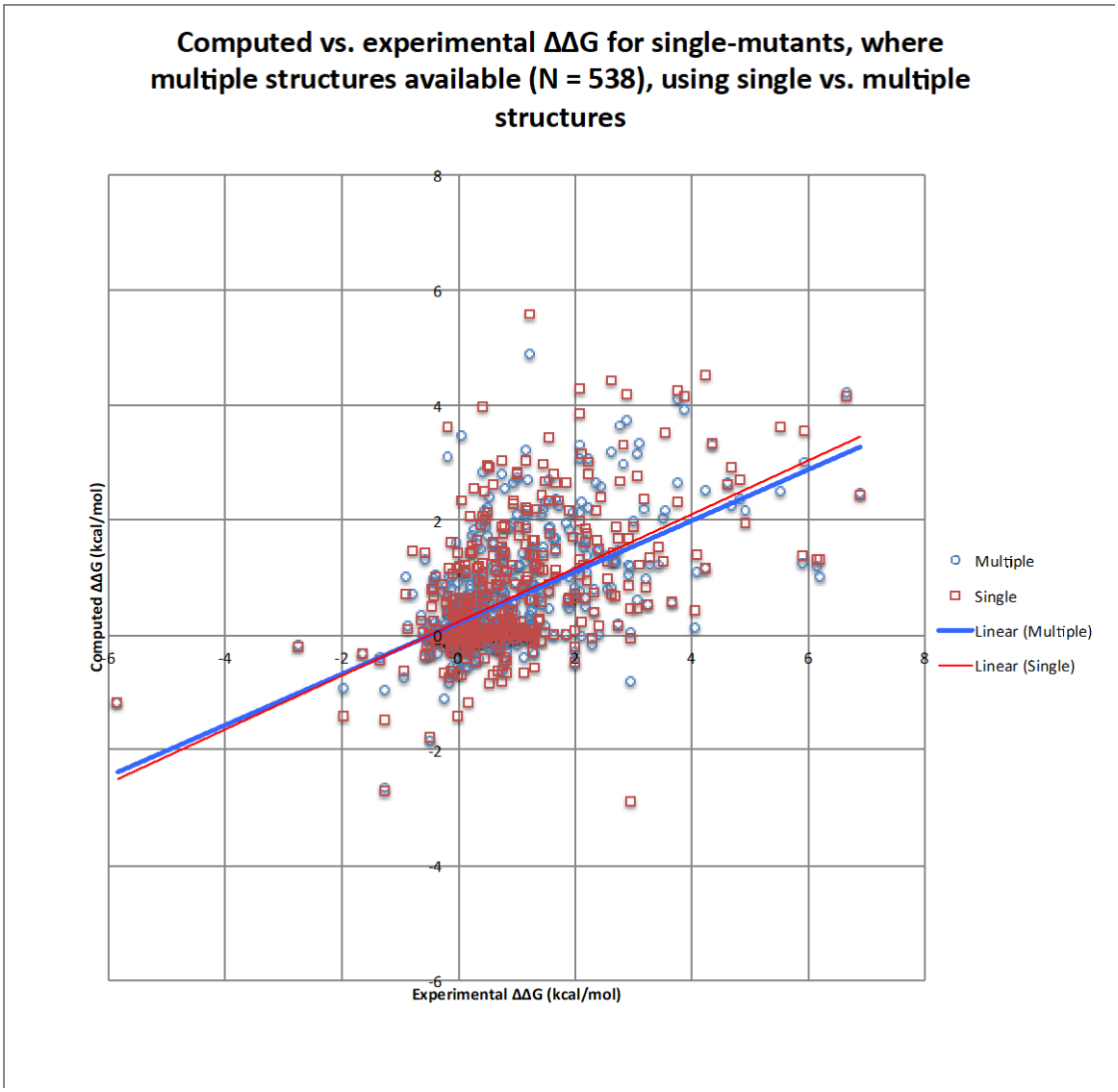

Figure 3. Scatterplot of $\Delta\Delta G_{predicted}$ vs. $\Delta\Delta G_{experimental}$. Shown for single calculations vs.averages over multiple structures.

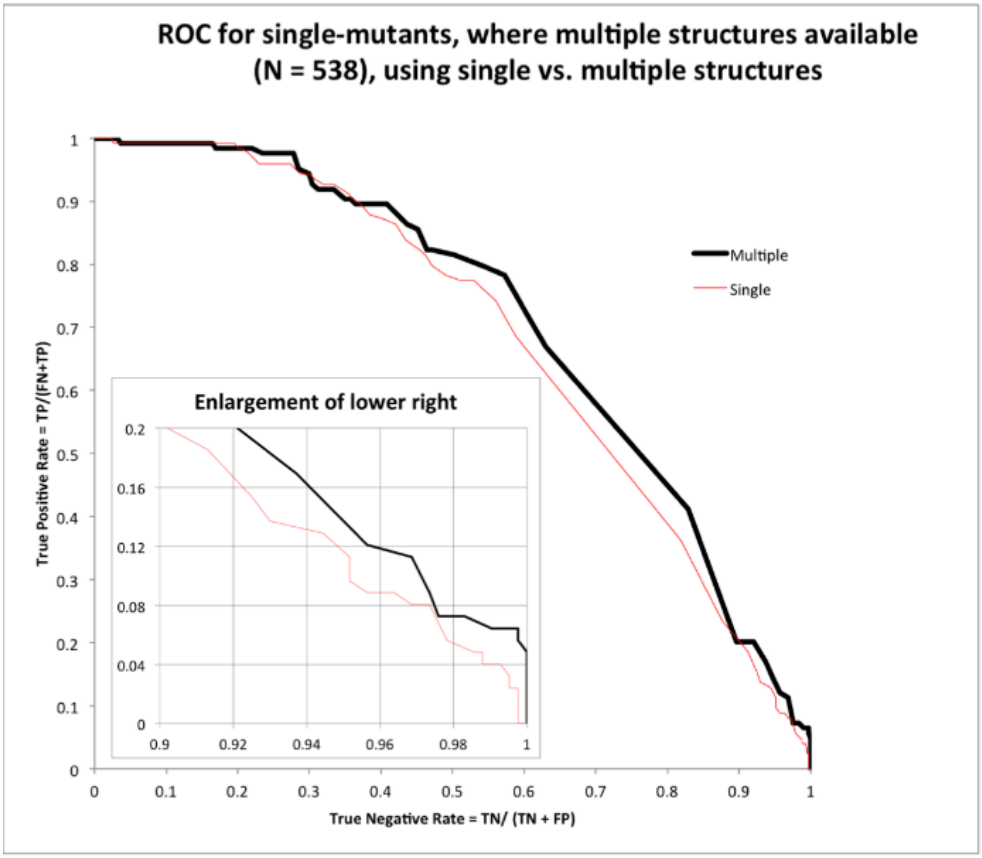

Figure 4. Receiver Operating Characteristic, comparing homologyScanner vs. calculation on single structures.

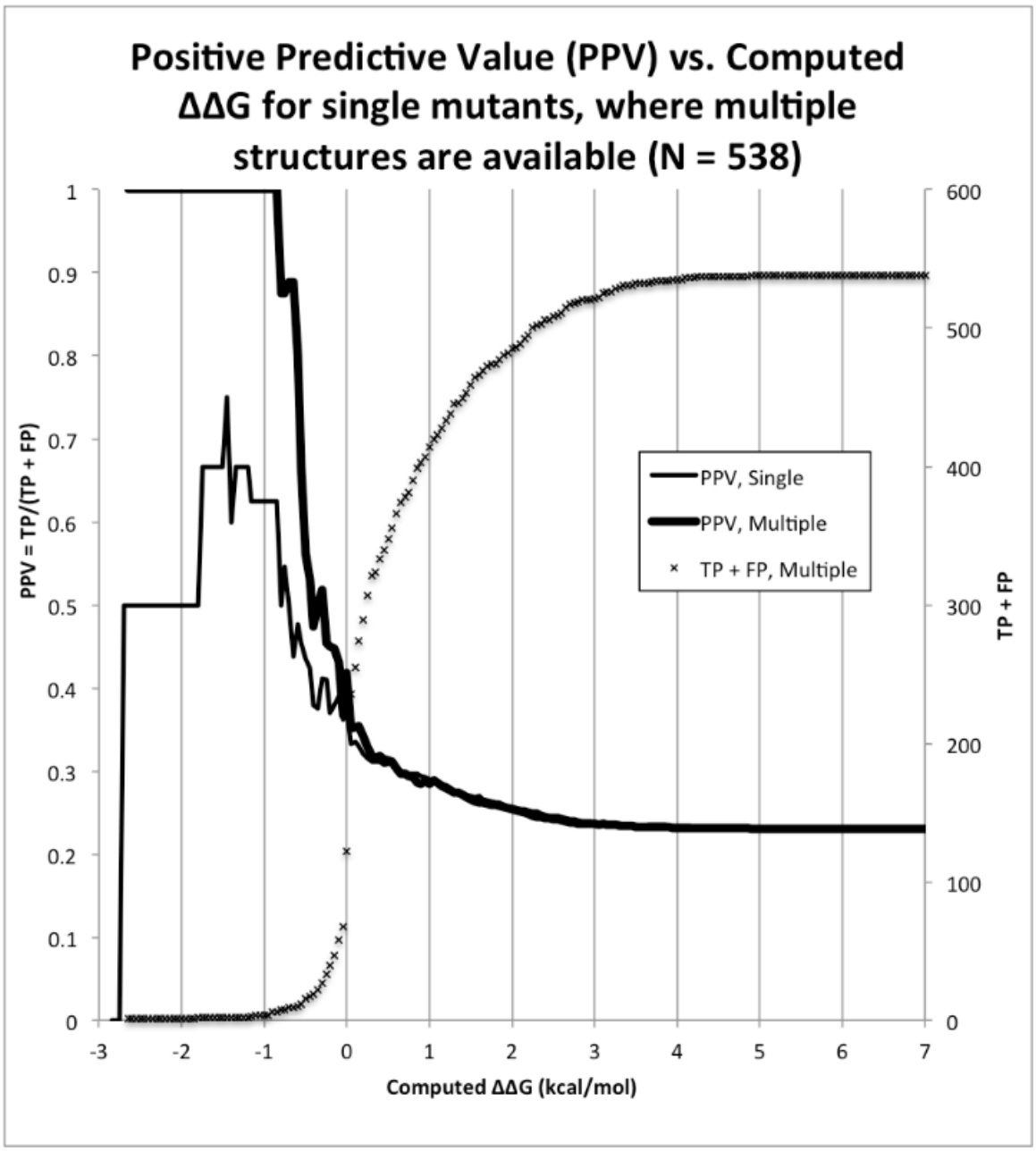

Figure 5. Positive Predictive Value (PPV) for single vs. multiple structures. TP + FP is the denominator of PPV, so we emphasize that this quantity becomes small for $\Delta\Delta G_{predicted}$ < -1 kcal/mol (crosses). This is why the PPV becomes erratic, at least for single structures.

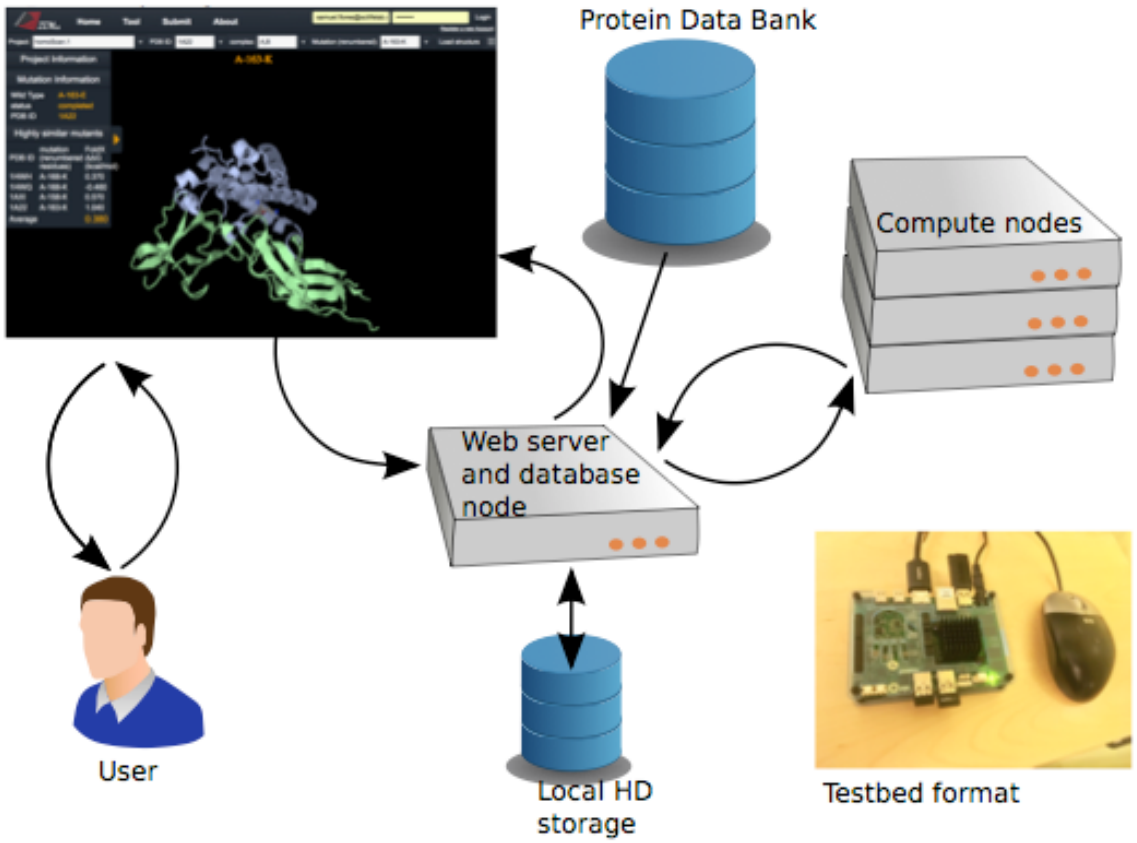

Figure 6. The homologyScanner public web server. Users can provide PDB ID, chose chains in each of two subunits, and specify a mutation to be computed. FoldX ΔΔG is computed for the query and all matching complexes and reported to the user. The results are available for browsing by others. The software components are available on github, simtk.org, and dockerhub. A server has also been set up on a single-board computer for private deployment.

Table 1

Comparison of computing ΔΔG using single vs. multiple structures, for several subsets of our benchmark set. Row label *Simultaneous substitutions*: how many simultaneous substitutions in each
mutant? Can be single-substitutions, multiple-substitutions, or a mixed set. Row label *Structures available*: how many structures are available for a given mutation in the dataset? For dataset in column A, used in (Dourado and Flores, 2014), multiple structures are available for
some but not all mutations, the remaining three datasets comprise only mutants for which multiple structures are available. Row label *Structures used*: How many of the available structures were used? For each dataset we compare use of all available structures (homologyScanner) vs. use of only one structure. Bottom: N, number of mutants in the dataset; RMSE, Root Mean Square Error; Correlation. Note that in all cases use of multiple structures yields lower RMSE and higher (or equal) correlation than use of single structures. Best results are with single-substitution mutants (as found in (Dourado and Flores, 2014)), for which multiple structures are available (dataset **A**).

	Dataset A		**Dataset B**		**Dataset C**		**Dataset D**	
Simultaneous substitutions	Single or multiple		Single or multiple		Single		Multiple	
Structures available	Single or multiple		Multiple		Multiple		Multiple	
Structures used	Single	Single or multiple	Single	Multiple	Single	Multiple	Single	Multiple
N	*1243*	*1243*	*736*	*736*	*538*	*538*	*198*	*198*
RMSE (kcal/mol)	*1.50*	*1.40*	*1.54*	*1.37*	*1.11*	*1.04*	*2.33*	*2.00*
Correlation	*0.69*	*0.64*	*0.64*	*0.65*	*0.56*	*0.60*	*0.56*	*0.56*