JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

CSAR Scoring Challenge Reveals the Need for New Concepts in Estimating Protein−Ligand Binding Affinity

Fedor N. Novikov,[†] Alexey A. Zeifman,[‡] Oleg V. Stroganov,[†,‡] Viktor S. Stroylov,[†] Val Kulkov,[§] and Ghermes G. Chilov[*,†,‡]

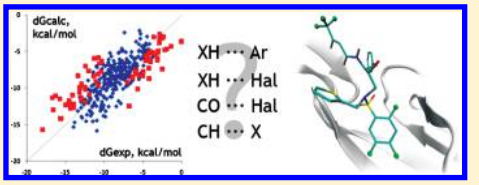
[†]MolTech Ltd., Russian Federation, 119992 Moscow, Leninskie gory, 1/75A

[‡]N.D.Zelinsky Institute of Organic Chemistry, Russian Federation, 119991 Moscow, Leninsky pr-t, 47

[§]BioMolTech Corp., 226 York Mills Road, Toronto, Ontario M2L 1L1, Canada

Supporting Information

ABSTRACT: The dG prediction accuracy by the Lead Finder docking software on the CSAR test set was characterized by R^2=0.62 and rmsd=1.93 kcal/mol, and the method of preparation of the full-atom structures of the test set did not significantly affect the resulting accuracy of predictions. The primary factors determining the correlation between the predicted and experimental values were the van der Waals interactions and solvation effects. Those two factors alone accounted for R^2=0.50. The other factors that affected the accuracy of predictions, listed in the order of decreasing importance, were the change of ligand's internal energy upon binding with protein, the electrostatic interactions, and the hydrogen bonds. It appears that those latter factors contributed to the independence of the prediction results from the method of full-atom structure preparation. Then, we turned our attention to the other factors that could potentially improve the scoring function in order to raise the accuracy of the dG prediction. It turned out that the ligand-centric factors, including Mw, cLogP, PSA, etc. or protein-centric factors, such as the functional class of protein, did not improve the prediction accuracy. Following that, we explored if the weak molecular interactions such as X-H...Ar, X-H...Hal, CO...Hal, C-H...X, stacking and π-cationic interactions (where X is N or O), that are generally of interest to the medicinal chemists despite their lack of proper molecular mechanical parametrization, could improve dG prediction. Our analysis revealed that out of these new interactions only CO...Hal is statistically significant for dG predictions using Lead FInder scoring function. Accounting for the CO...Hal interaction resulted in the reduction of the rmsd from 2.19 to 0.69 kcal/mol for the corresponding structures. The other weak interaction factors were not statistically significant and therefore irrelevant to the accuracy of dG prediction. On the basis of our findings from our participation in the CSAR scoring challenge we conclude that a significant increase of accuracy predictions necessitates breakthrough scoring approaches. We anticipate that the explicit accounting for water molecules, protein flexibility, and a more thermodynamically accurate method of dG calculation rather than single point energy calculation may lead to such breakthroughs.



INTRODUCTION

The dG prediction accuracy has remained practically unchanged in the past 5−10 years at the level of 2−3 kcal/mol, as evidenced by the various studies employing different computational methods and different test sets.[1−6] Furthermore, the addition of nonbinding molecules to the current test sets typically reduces the quality of predictions since the existing scoring functions in general have not been able to perform well in the differentiation of active molecules from inactive ones.[7] That has been a major problem of the structure-based virtual screening for years. It appears that the current methods of dG estimation have stalled, presumably due to the fact that all of these methods are simply variations of the same general approach of combining the well-known energy factors in different ways.

It seems evident that a critical evaluation of the current state of affairs is necessary in order to find a way to overcome the present limitations. The existing methods need to be evaluated for their strengths and weaknesses in order to define possible ways forward. In this respect, the CSAR competition created a fertile ground for research in the docking and scoring community. The ability to curate the CSAR test set by the community members and the ability to use the test set without a prior preparation, such as the addition of protons or hydrogen atoms that can introduce a significant uncertainty in the prediction results, seem to be the two obvious advantages of the CSAR test set. We felt that our participation in the CSAR competition would present us a good chance to evaluate what energy terms carry the greatest significance in the dG prediction and which ones are simply correlated and degenerate and to attempt to identify what new ligand-centric or protein-centric interactions are unaccounted for that, when considered, might increase the accuracy of the dG predictions. Also, we felt it was important to explore the scope of the applicability of the existing scoring functions and to pinpoint

Special Issue: CSAR 2010 Scoring Exercise

Received: January 24, 2011
Published: May 25, 2011

areas where we have little or no understanding of the mechanisms at play and to suggest possible approaches to extend the applicability and increase accuracy of the dG predictions.

METHODS

Preparation of the Full-Atom Models. The full-atom structures of the CSAR test set were prepared by three distinct protocols. The first version comprised the CSAR NRC-HiQ test-set, which had been downloaded from the CSAR Web site[8] without modifications. The second version comprised the full-atom models which had been automatically prepared using our own Model Builder software, which is available as a part of Lead Finder package, or as an individual application (upon request from the authors). Model Builder calculates pK_a of protein side-chains according to our new graph-theoretical algorithm called TSAR (Thermodynamic Sampling of Amino acid Residues) and selects the most probable ionization state of a system at a given pH.[9,10] Positions of the polar hydrogens are globally optimized, while heavy atoms are not moved. All water molecules were deleted upon model preparation. In addition, Model Builder adds symmetric protein subunits if they are in contact with the ligand's binding site. Finally, the third protocol implied the default protonation states for all side-chains (the residue is deprotonated at pH > pK_a and protonated at pH < pK_a, where pK_a corresponds to the ionization of side-radical of the free amino acid in aqueous solution) and random placement of functional hydrogen atoms (along its terminal rotational degree of freedom). The latter protocol, referred to here as the null model, had been used to probe the influence of full-atom model quality on the accuracy of dG prediction.

Energy Scoring. The estimations of the protein−ligand binding affinity were performed with the Lead Finder v 1.1.15 docking software.[11] Since Lead Finder performs dG estimations for the locally optimized ligand position (rather than for the crude input coordinates) and due to the stochastic nature of the Lead Finder local optimization algorithm, all dG estimations were averaged over 5 independent runs of the program.

Lead Finder scoring function is a semiempiric molecular mechanical functional, which energy terms are represented in Table 1. Among standard terms common for many scoring functions of such kind, van der Waals and volume-based solvation contributions can be mentioned. The expressions of other energy terms have been considerably revised by us in order to gain more accurate description of protein−ligand interactions. In particular, the nonpolar volume-based solvation has been supplemented with four additional terms accounting for the polarity of ligand atoms (polar/nonpolar) interacting with either protein or solvent and the areas of corresponding contacts. These additional terms were required to introduce more accurate balance into ligand desolvation and hydrophobic interactions with protein.

The screened Coulomb potential has been used in our scoring function to account for electrostatic interactions.[12] This model is favorable due to its physics-based description of the reaction field, and it is not computationally expensive compared to Poisson−Boltzmann models. In addition, the screened Coulomb potential can be easily implemented in the grid-based calculations, and thus it is suitable for docking.

Three extra energy terms were added to the direct energy of H-bonds to account for the whole contribution of H-bonds gains and losses upon ligand binding. Two of these new terms describe energetic penalties on ligand and protein atoms that do not form

Table 1. Construction of the Lead Finder Scoring Function

energy term		equation[a] $i \in$ ligand, $j \in$ protein	description
van der Waals		$k_{VdW}\Sigma LJ_{ij}(r_{ij})$	$LJ_{ij}(r_{ij})$ — is a smoothed Lennard-Jones potential
nonpolar solvation	volume-based	$k_{sol}\Sigma S_i V_j e^{-r_{ij}^2/3.6}$	S_i and V_i — are atomic solvation parameters r_{ij} — interatomic distance
	surface-based	$E_{L(pol)-P} + E_{L(pol)-S} + E_{L(a-pol)-P} + E_{L(a-pol)-S}$	surface-based solvation components of ligand's (L) polar (pol) or nonpolar (n-pol) atoms with protein (P) and solvent (S)
electrostatic interactions		$\Sigma k_{elec,ij} E_{elec,ij}$	$k_{elec,ij}$— is a scaling coefficient, which depends on the hydrophilicity of a microenvironment of atoms i,j and their buried fraction $E_{elec,ij}$ — is the energy of electrostatic interaction, which depends on the microenvironment-specific dielectric screening functions
hydrogen-bonding energy		$K_{HB}E_{HB} + K_{L(pen)}\Delta E_{L(pen)} + K_{P(pen)}\Delta E_{P(pen)} + \Delta E_{corr}$	E_{HB} — energies of H-bonds between protein and ligand $E_{L(pen)}$, $E_{P(pen)}$ — energetic penalties for nonforming H-bond by protein and ligand in the complex ΔE_{corr} — reward for the correlated H-bonds
internal energy	nonbonded	$E_{ES} - E_W$	E_{ES} — scaled nonbonded energy of ligand in protein−ligand complex E_W — scaled nonbonded energy of ligand in water
	entropic losses	$k_{tors}n_{tors}$	n_{tors}— the number of freely rotatable ligand bonds
	dihedral	$k_{dihedral}\Sigma E(\alpha_i)$	$E(a_i)$ — energy of the ligand's dihedral angle i
interactions with metals		$k_M\Sigma\alpha_{ij}LJ_{ij}(r_{ij})$	a_{ij}— coefficient which depends on the metal coordination state and relative orientation of ligand and metal orbitals

[a] The detailed description of formulas and corresponding coefficients and parameters can be found elsewhere.[9]

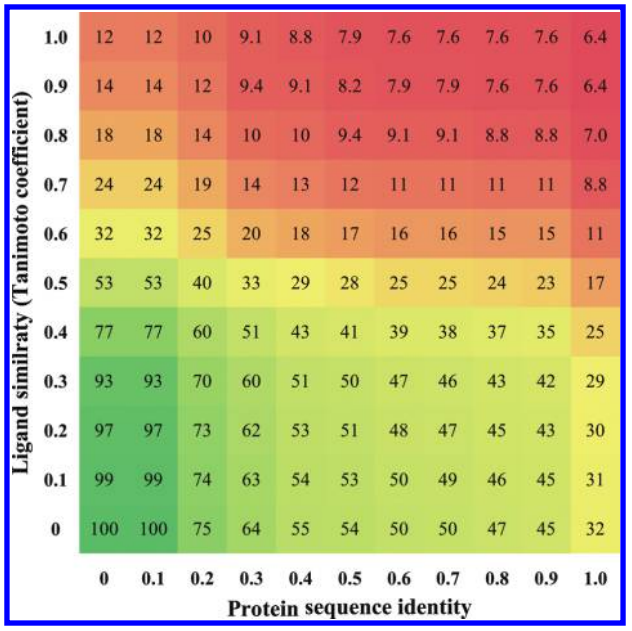


Figure 1. Similarity of the Lead Finder scoring function training set (330 protein−ligand complexes) and the CSAR test set (343 protein−ligand complexes). The numbers represent the percent of complexes in two sets sharing common sequence identity and ligand similarity above given thresholds.

H-bonds in the complex but are able to form them with water in the unbound state. One additional term accounts for the correlated H-bonds, which take place when the ligand's group is capable of forming more H-bonds with the protein than the water molecules in unbound state can form instead.

Finally, we have performed a thorough parametrization of ligand interactions with metal ions. The corresponding parameters were adjusted to reproduce the geometry of ligand binding in a set of ~100 structures from the PDB containing Fe^{2+}, Fe^{3+}, Zn^{2+}, Mg^{2+}, Ca^{2+}, and Mn^{2+} metal ions and coordinated with ligand's O, N, and S atoms.

The parametrization of our scoring function was performed on the training set of 330 protein−ligand complexes with high-resolution crystal structures and robust experimental data on the binding affinity available. This training set was compiled using the AffinityDB,[13] PDBbind,[14,15] BindingDB.[16] databases. The adequacy of the corresponding experimental data was manually annotated for each structure using primary literature sources. Finally, a set of 20 energy-scaling coefficients was fit, and the accuracy of dG prediction on the described training set was characterized by the rmsd=1.60 kcal/mol and R^2=0.66.

RESULTS AND DISCUSSION

Comparison of the CSAR Test Set and the Lead Finder Training Set. The comparison of our training set and the CSAR test set is presented in Figure 1. It can be seen that these sets share about 1/3 of common proteins, about 1/8 of common ligands and in total contain only 6.4% of identical protein−ligand complexes. Thus we believe that overlapping of these sets has not noticeably influenced the results of the CSAR competition. A number of additional conclusions can be drawn from Figure 1. For example, a well-known twilight zone can be seen, which is a steep borderline corresponding to the ~20% of sequence identity, meaning that almost any two arbitrary proteins would have such similarity with each other. Moreover, such borderline can be observed for ligands as well, and it is placed near the Tanimoto distance of 0.4.

Table 2. Accuracy of dG Prediction for the Three Protocols of Preparing the Full-Atom Models of Protein−Ligand Complexes

	rmsd, kcal/mol[a]	RMSE, kcal/mol[b]	R^2	τ^c	ρ^d	ddG, kcal/mol average	ddG, kcal/mol min	ddG, kcal/mol max
Null model	2.08	1.40	0.58	0.56	0.77	−0.10	−5.2	6.4
NRC-HiQ	2.05	1.34	0.59	0.56	0.76	−0.10	−5.5	4.9
MLT[e]	1.93	1.29	0.62	0.58	0.79	−0.20	−5.2	4.4

[a] Root Mean Squared deviation. [b] Root Median Squared Error. [c] Kendall tau. [d] Spearman rho. [e] Models had been automatically prepared using Model Builder software (see Methods section for details).

Table 3. Main Energy Contributions to the Accuracy of dG Prediction

energy terms	R^2
VdW+sol	0.50
VdW+solv+internal	0.55
VdW+solv+Elec+HB	0.51
VdW+solv+Elec+HB+internal	0.62
Complete scoring function	0.62

Accuracy of Binding Energy Calculation and How the Full-Atom Model Preparation Technique Influences It. The overall accuracy of dG prediction on the NRC-HiQ version of the CSAR test set is described by rmsd=2.05 kcal/mol and R^2=0.59. Thus, the accuracy of our scoring function on the CSAR test set differs from that on our training set. This means that our scoring function is overfitted on the consolidated set comprised by our training and CSAR test sets. According to the information theory[17] we could further reduce the number of parameters in our model to achieve equal accuracy on the training and test sets. Alternatively, we can assume that such consolidated data set is just a (small) subset of all experimental data on protein−ligand binding affinity and that the sizes of training and test sets can be further increased.

Another issue, which appeared to be crucial for many (if not for all) of the participants, was the quality of the full-atom models of protein−ligand complexes and its influence on the accuracy of dG prediction. Many of the participants, including ourselves, observed a number of disagreements in the protonation of the active sites of some structures. That raised a series of successive versions of the CSAR set, including the final NRC-HiQ version. To study this issue in more details we prepared two additional versions of the test set: the first one using our own Model Builder program (described in the Experimental Section), and the second one using the default ionization state of protein side-chains and a random rotation of the terminal functional hydrogen atom (the so-called null model). It appeared that the use of Model Builder software returned slightly better results compared to the NRC-HiQ models; however, the difference was not dramatic (Table 2). However it also appeared that the null model returned results which were not dramatically less accurate. This observation was quite surprising to us since the energy terms describing the H-bond energy in our scoring function are quite sensitive to the proton orientation.

Table 4. Correlation between the Ligand-Centric Properties and the Error of dG Prediction

parameter	R^2
cLogP (ACD)	4.5E-03
cLogP (ChemAxon)	6.3E-04
molecular weight	2.8E-02
number of freely rotatable bonds	3.7E-02
number of HB acceptors	1.1E-02
number of HB donors	2.5E-02
number of rings	9.0E-06
polar surface area	2.1E-02

The Main Contributions to the Protein−Ligand Binding Energy. Thus we have asked ourselves which energy terms of our scoring function were the most significant for the accurate dG prediction? To answer this question we studied the accuracy of dG prediction (on the CSAR test set) in which particular energy terms were switched off. It appeared that the most significant contribution to the accuracy was from the combination of van der Waals and solvation energy terms (Table 3). This means that the bigger the area of protein−ligand interaction and the more complementary it is in terms of polarity, the more favorable energetic estimate our scoring function will return. It is critically important that the solvation terms depend on the polarity of ligand atoms contacting with protein and solvent, as was described above. In the absence of such a balance, the energy of protein−ligand interactions would be consistently overestimated.

The internal ligand energy is a contribution which is ranked next by its impact on the accuracy of dG prediction. This contribution is comprised by the difference of ligand energies in two conformations (bound and unbound) and by the ligand entropic losses. At the same time, the consolidated account of electrostatic interactions and H-bonds yields a smaller impact (separately these contributions yield a negligible improvement). However, if the electrostatic and H-bond interactions are summed with the internal, van der Waals and solvation energies, the obtained function returns almost the same correlation as the complete model. The remaining part of the correlation is produced by the electrostatic (Born) component of ligand desolvation and the ligand interaction with metal ions.

We can notice that the relative importance of energy contributions, obtained for our scoring function, was quite expected for the model of implicit solvent account. The main role in such models is attributed to the complementarity of protein and ligand interface including its shape and polarity. The H-bonds have smaller contribution because of their compensative nature. In addition it seems that the complexes of relatively good binding ligands from the CSAR test set are well balanced with respect to H-bond losses and gains. Probably, the situation for the test set enriched with poor binders would be different. We can also suggest that due to the relatively small impact of H-bond energy the accuracy of dG prediction in our case was not very sensitive to the protocol of full-atom model preparation.

The Influence of Ligand-Centric and Protein-Centric Properties on the Accuracy of dG Prediction. Next we questioned whether we could improve the accuracy of our scoring function by introducing additional terms into it. The most obvious step was to add some correction for the ligand-centric properties, such as cLogP, Mw, PSA, etc. To test this hypothesis we analyzed the correlation of dG prediction error with a set of typical physicochemical ligand-based descriptors (Table 4). No correlation was observed. Thus we had to conclude that a primitive account of ligand-centric properties would not allow to improve the accuracy of dG prediction using our scoring function.

Table 5. Accuracy of dG Prediction for Different Functional Classes of Proteins

		dG prediction		ddG, kcal/mol		
protein group	N	rmsd, kcal/mol	R^2	average	min	max
oxidoreductases	12	1.37	0.83	−0.29	−2.8	1.9
transferases	58	1.73	0.53	−0.02	−5.1	4.3
hydrolases	129	2.04	0.69	0.33	−4.7	4.3
lyases	21	1.56	0.48	−0.04	−2.6	4.1
isomerases	6	1.17	0.71	0.22	−2.2	1.4
ligases	2	2.07	n/a	1.81	0.8	2.8
lectines	13	1.13	0.69	−0.48	−3.2	0.6
membrane proteins	13	2.12	0.03	−1.69	−5.2	−0.6
transport proteins	18	1.48	0.62	−0.94	−2.8	0.6
other proteins	71	2.26	0.56	−0.93	−5.1	4.4
all proteins	343	1.93	0.62	−0.20	−5.2	4.4

Then we checked whether an additional account of protein-centric properties could improve the accuracy. The proteins from the test set were split into groups by their functional activity, and the accuracy of dG prediction was analyzed separately for the obtained groups (Table 5). It appeared that for most of the groups the rmsd was quite close to the average, and the signed error was close to zero (Table 5). This means that the addition of protein class-specific correction would not improve average accuracy by more than 0.2−0.3 kcal/mol for such proteins. However a few classes displayed notably shifted dG estimates, namely ligases and membrane and transport proteins. While ligases can be ignored due to their low representation (2 structures), membrane and transport proteins deserve more close attention. In the case of the membrane proteins we not only have the large signed error of −1.69 kcal/mol but also the virtual absence of correlation (R^2=0.03), which is due to 3 outlier structures. This is why a simple shift of dG for these structures just slightly improves the rmsd (from 2.12 to 1.87 kcal/mol). Thus, the appropriateness of such protein-centric correction is questionable. We can also mention that all membrane proteins in the CSAR test set are glutamate receptors, and their ligand-binding domain is not a transmembrane one. Thus, to analyze specific features of the membrane-embedded ligand binding sites, the CSAR test set needs additional relevant structures.

In the case of the transport proteins a correction for the average signed error also does not improve the rmsd significantly (from 1.48 to 1.18 kcal/mol). Further analysis revealed that those proteins were quite structurally diverse despite their common (transport) function, and we could not gain additional understanding of that subset.

The Influence of Weak Interactions on the Accuracy of dG Prediction. Our previous analysis revealed that a primitive account of either ligand- or protein-centric properties was not fruitful. Thus we questioned whether an inclusion of some additional types of molecular interactions into our scoring function could improve the accuracy of dG prediction. The rationale was that the current molecular mechanical approaches poorly reproduce interactions for which rearrangement of electron densities is crucial and which can be robustly described only at the *ab initio* level. In addition to that we can directly observe such interactions by analyzing the

Table 6. Statistical Significance of Including Additional "Weak" Molecular Interactions into the Lead Finder Scoring Function

interaction type	N	ddG, kcal/mol[g]	rmsd, kcal/mol[g]	P[h]
X-H...Ar[a]	3	0.20	0.95	0.62
X-H...Hal[b]	7	1.64	2.38	0.046
CO...Hal[c]	7	2.19	2.30	0.005
π-cationic interaction[d]	18	−1.02	2.30	0.47
C-H...X[e]	248	−0.15	1.91	0.66
stacking[f]	62	−0.50	1.88	0.20

[a] X=N, O, d < 5 Å, 120° < α < 180°. [b] X=N, O, Hal=F, Cl, Br, I, d < 5 Å, 120° < α < 180°. [c] Hal=F, Cl, Br, I, d < 4 Å, 100° < ∠CHalO < 180°, 110° < ∠COHal < 180°. [d] d < 5 Å, 120° < α < 180°. [e] X=N, O, d < 3 Å, 120° < α < 180°. [f] d < 10 Å, 120° < α < 180°. [g] Averaged over the structures with a given potentially new type of interactions. [h] Student's t test (for π-cationic, C-H...X and stacking interactions) and bootstrapping (for X-H...Ar, X-H...Hal, and CO...Hal interactions) probability of a null hypothesis.

growing domain of structural data, for example the PDB database. One of the recent reviews brightly describes such "weak" interactions, which are hypothesized from the medicinal chemistry experience and supported by the analysis of existing crystallographic structures of protein−ligand complexes.[18]

To analyze this hypothesis we chosen a series of interactions from the review,[18] which lacked proper molecular mechanical parametrization but had reasonable geometrical definitions. The list of such interactions included the following: the interaction of polar hydrogen with the electrons of aromatic ring (X-H...Ar), polar hydrogen with halogen (X-H...Hal), halogen with C-O,[18,19] or C═O[18] (CO...Hal), nonpolar hydrogen with polar atom (C-H...X), stacking interaction, and π-cationic interaction. Using the geometric definitions of those interactions as described in ref 18, we selected structures satisfying the corresponding criteria. It appeared that stacking and C-H...X were quite widespread, while the rest of the interactions were present in less than 5% of structures (Table 6). Then we determined that the signed error for dG prediction was comparable with the rmsd only for structures with X-H...Hal and CO...Hal interactions, while for the remaining structures it was significantly smaller. That meant that from the point of improving accuracy of dG prediction the introduction of new energy terms into our scoring function was justified only for structures with X-H...Hal and CO...Hal interactions. To prove that point further, we used the Student's t test (for two-tailed distribution) in order to test the statistical significance of C-H...X, π-cationic, and stacking interactions, that is to probe whether the error of dG prediction on those structures was distributed differently as compared to the rest of the test set. Application of Student's criterion returns the probability that two hypotheses are indistinguishable, thus a low probability value suggests that the introduction of a new potential into the scoring function is statistically justified. It appeared that none of these interactions were statistically significant (P < 0.05) (Table 6). As for the X-H...Ar, CO...Hal, and X-H...Hal interactions, since the number of structures possessing these interactions was too small to use the Student's t test criterion we performed the so-called bootstrapping procedure in those cases. During this procedure an appropriate number of structures (by the number of complexes with corresponding interaction) was repeatedly selected at random from the test set structures containing halogen atoms in the ligand molecule (in order to diminish possible effects connected with errors in the halogen atoms force field parametrization). Then the probability of finding the subset with the average signed error (in dG prediction) above or equal to the given and the rmsd below or equal to the given was calculated. If thus obtained probability was small enough, we could state that the deviation of predicted dG for such subset of structures was not by chance. The bootstrapping procedure has the advantage over the analytical criteria in that it does not impose any assumptions on the nature of the distribution (in most cases the normality of distribution is implied). On the other hand the bootstrapping is practically feasible only for small samples, since the evaluation of a sample of size 20 already leads to an astronomically large number of variants (C_{343}^{20} or approximately 10^{51}). We found that the observation of similar error distribution as for CO...Hal or X-H...Hal by chance had probabilities of 0.005 and 0.05 correspondingly. Thus, the inclusion of a new interaction type CO...Hal into our scoring function was rigorously proved by bootstrapping statistical test, while the inclusion of X-H...Hal interaction was at the threshold level of justification. Accounting for CO...Hal interactions was also much more effective compared to X-H...Hal from the view of absolute contribution to the accuracy of dG prediction. In the former case addition of the signed dG error, averaged over the subset of complexes possessing such interaction, to the predicted dG led to the reduction in rmsd from 2.19 to 0.69 kcal/mol, while in the latter case - from 2.38 to only 1.72 kcal/mol. A more detailed analysis of complexes possessing these two new potential types of interactions (Tables S2 and S3, Supporting Information) confirmed that the error in dG prediction had the same sign and close absolute value for almost all of the 7 complexes with CO...Hal interaction, while in the case of X-H...Hal the distribution of dG error was quite uneven. However, we should warn against the immediate inclusion of CO...Hal interaction into scoring functions, since the current analysis is performed only on 7 structures, 6 of which represent F and one I, while other halogens are not represented at all (Table S2). Obviously, more structures with such interaction type should be considered in order to confirm its statistical significance in scoring and to elaborate proper parametrization. Our quick analysis of the PDB revealed 486 crystal structures which include totally 848 of such potential new interactions (661 - for F, 41 - for Cl, 18 - for Br, and 128 - for I), almost 179 of which had experimentally measured binding affinity according to the PDB annotation (from BindingDB[20] and BindingMOAD[21] databases). So, a proper analysis of these structures should be the subject of our future work inspired by the participation in the CSAR scoring challenge.

In addition to finding potentially new interactions contributing to our dG estimates, we have retrospectively studied (using the Student's t test) the significance of individual energy components of H-bond contribution, which we had previously introduced in the Lead Finder scoring function.[11] It appeared that the direct H-bond energy and the penalty for nonforming H-bonds by protein atoms and correlated H-bonds energy were statistically significant components of our scoring function, while the penalty for nonforming H-bonds by the ligand's atom was not. Surprisingly, that component seemed quite important to us from the general considerations at that time; however, now we can notice its redundancy. Thus, our participation in the CSAR scoring challenge allowed us to resolve new, yet unaccounted significant components in the scoring function as well as to notice redundancy of some old components. Obviously, our future versions of the Lead Finder scoring function must account for these findings.

The Need for New Concepts in Estimating Protein−Ligand Binding Affinity. Although the current analysis revealed new,

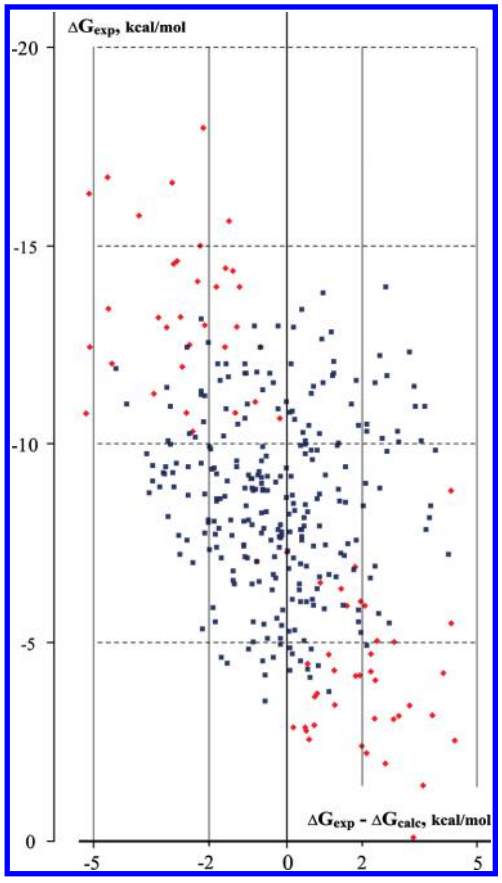


Figure 2. Binding energy prediction errors for the CSAR test set plotted against the experimentally measured binding energy. Red dots correspond to the structures, which received universally poor dG estimates among participants of the CSAR challenge.

previously unnoticed features of our scoring approach and highlighted its weak points, we have yet to answer the basic questions. When will the accuracy of dG prediction overcome the barrier of 2 kcal/mol? When can we overcome the errors of 3, 4, and 5 kcal/mol? As one can see from Figure 2, such cases are common. Can we at least surmise which direction in the development of scoring approaches we should follow in order to resolve these issues step by step?

In principle, mastering of the scoring approaches may follow two independent directions. The first one concerns the refinement of potentials of molecular interactions. Generally speaking, such refinement may be gained by learning from the growing domain of (structural and functional) experimental data or from the results of *ab initio* calculations. The second direction consists of expanding the explicitly treated part of the system that is in accounting for the protein flexibility and explicit solvent in a thermodynamically consistent manner to approach the free energies of protein−ligand binding. Our current experience suggests that the significant improvement in scoring cannot be achieved without advances in the latter direction. Moreover, the lack of convincing, robustly validated results attained by the current approaches to treat protein flexibility or explicit water, suggests that the corresponding layer of chemical theory, neighboring with statistical physics, needs a serious reconsideration.

In order to gain a qualitative estimate of the impact of protein flexibility and explicit water treatment we analyzed the complexes for which the error in dG prediction exceeded 2 kcal/mol. First of all we selected those structures in which the protein−ligand interaction was mediated by water molecules forming at least 2 H-bonds (in total with protein and ligand). Second, we selected those complexes where protein residues that formed H-bonds with ligands were quite flexible (that is they could lose H-bonds upon side-chain movements), or where such residues did not form H-bonds but they were able to do so due to the side-chain flexibility. The latter criterion was quite subjective of course, but we did not try to elaborate a more quantitative measure since we were primarily interested in the qualitative analysis. Thus we tried to understand in which cases the explicit account of (at least tightly bound) water or protein flexibility could improve description of protein−ligand binding. As a result, out of 104 poorly scored complexes 44 were marked as having tightly bound water molecules, 41 - as influenced by protein flexibility, and 14 - as having both features simultaneously. On one hand this means that we have at least a hint of how to improve the dG prediction for 70% of the poorly scored structures. On the other hand the signed error for those complexes appeared to be close to zero, suggesting no obvious correlations could be conjectured *per se*. Moreover, we do not have an even weaker hypothesis for the remaining 30% of the poorly scored structures, and these facts eloquently depict our current understanding of the scoring approaches. However, owing to the open competitions, like CSAR, we had a valuable opportunity to examine our weak points and to do our homework.

■ ASSOCIATED CONTENT

Supporting Information. Detailed results of binding free energy calculations for CSAR test-set, detailed analysis of complexes possessing CO...Hal and XH...Hal interactions. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: +7(495)135-53-13. Fax: +7(495)135-53-13. E-mail: ghermes@moltech.ru.

■ ACKNOWLEDGMENT

The work was supported by the Foundation for assistance to small enterprises in the scientific area (Contract 6332p/7168).

■ REFERENCES

(1) Ryangguk, K.; Skolnick, J. Assessment of Programs for Ligand Binding Affinity Prediction. *J. Comput. Chem.* **2008**, *29*, 1316–1331.

(2) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function. *J. Comput. Chem.* **1998**, *19*, 1639–1662.

(3) Ruth, H.; Garrett, M.; Morris, A.; Olson, D.; Goodsell, A. Semiempirical Free Energy Force Field with Charge-Based Desolvation. *J. Comput. Chem.* **2007**, *28*, 1145–1152.

(4) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. *J. Med. Chem.* **2006**, *49*, 6177–6196.

(5) Muegge, I. PMF Scoring Revisited. *J. Med. Chem.* **2006**, *49*, 5895–5902.

(6) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., III Assessing Scoring Functions for Protein-Ligand Interactions. *J. Med. Chem.* **2004**, *47*, 3032–3047.

(7) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49*, 5912–5931.

(8) Community Structure-Activity Resource. http://www.csardock.org/ (accessed May 20, 2011).

(9) Stroganov, O. V.; Novikov, F. N.; Zeifman, A. A.; Stroylov, V. S; Chilov, G. G. TSAR — a new graph-theoretical approach to computational modeling of protein side-chain flexibility. Modeling of ionization properties of proteins. In submission.

(10) Stroganov O.; Novikov F.; Zeifman A.; Stroylov V.; Kulkov V.; Chilov, G. . TSAR — a new graph-theoretical approach to computational modeling of ionization properties of proteins. 241st ACS National Meeting, Anaheim, CA, March 27−31, 2011; COMP 36.

(11) Stroganov, O. V.; Novikov, F. N.; Stroylov, V. S.; Kulkov, V.; Chilov, G. G. Lead finder: an approach to improve accuracy of protein-ligand docking, binding energy estimation, and virtual screening. *J. Chem. Inf. Model.* **2008**, *48*, 2371–2385.

(12) Mehler, E. L.; Guarnieri, F. A Self-Consistent, Microenvironment Modulated Screened Coulomb Potential Approximation to Calculate pH-Dependent Electrostatic Effects in Proteins. *Biophys. J.* **1999**, *75*, 3–22.

(13) Block, P.; Sotriffer, C. A.; Dramburg, I.; Klebe, G. AffinDB: a freely accessible database of affinities for protein−ligand complexes from the PDB. *Nucleic Acids Res.* **2006**, *34*, D522–D526.

(14) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database: Methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111–4119.

(15) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.* **2004**, *47*, 2977–2980.

(16) Liu, T.; Lin, Y.; Wen, X.; Jorrisen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2006**, *00*, D1–D4.

(17) MacKay D. J. C. Model Comparison and Occam's Razor. In *Information Theory, Inference, and Learning Algorithms*; Cambridge University Press: Cambridge, UK, 2003; pp 343-357.

(18) Bissantz, C.; Kuhn, B.; Stahl, M. A Medicinal Chemist's Guide to Molecular Interactions. *J. Med. Chem.* **2010**, *53*, 5061–5084.

(19) Auffinger, P.; Hays, F. A.; Westhof, E.; Ho, P. S. Halogen bonds in biological molecules. *Proc. Natl. Acad. Sci. U.S.A.* **2004**, *48* (101), 16789–16794.

(20) Liu, T.; Lin, Y.; Wen, X.; Jorrisen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2006**, *35*, D198–D201.

(21) Hu, L.; Benson, M. L.; Smith, R. D.; Lerner, M. G.; Carlson, H. A. Binding MOAD (Mother Of All Databases). *Proteins* **2005**, *60*, 333–40.