HHS Public Access

Author manuscript

J Comput Chem. Author manuscript; available in PMC 2015 July 22.

Published in final edited form as:

J Comput Chem. 2011 March ; 32(4): 568–581. doi:10.1002/jcc.21642.

VoteDock: Consensus Docking Method for Prediction of Protein–Ligand Interactions

Dariusz Plewczynski[1,*], **Michałłaźniewski**[1,2], **Marcin Von Grotthuss**[3], **Leszek Rychlewski**[4], and **Krzysztof Ginalski**[1]

[1]Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw, Pawinskiego 5a Street, 02-106 Warsaw, Poland [2]Department of Physical Chemistry, Faculty of Pharmacy, Medical University of Warsaw, Banacha 1 Street, Warsaw, Poland [3]Department of Chemistry and Chemical Biology, Harvard University, Cambridge, Massachusetts 02138 [4]BioInfoBank Institute, Poznan, Poland

Abstract

Molecular recognition plays a fundamental role in all biological processes, and that is why great efforts have been made to understand and predict protein–ligand interactions. Finding a molecule that can potentially bind to a target protein is particularly essential in drug discovery and still remains an expensive and time-consuming task. *In silico*, tools are frequently used to screen molecular libraries to identify new lead compounds, and if protein structure is known, various protein–ligand docking programs can be used. The aim of docking procedure is to predict correct poses of ligand in the binding site of the protein as well as to score them according to the strength of interaction in a reasonable time frame. The purpose of our studies was to present the novel consensus approach to predict both protein–ligand complex structure and its corresponding binding affinity. Our method used as the input the results from seven docking programs (Surflex, LigandFit, Glide, GOLD, FlexX, eHiTS, and AutoDock) that are widely used for docking of ligands. We evaluated it on the extensive benchmark dataset of 1300 protein–ligands pairs from refined PDBbind database for which the structural and affinity data was available. We compared independently its ability of proper scoring and posing to the previously proposed methods. In most cases, our method is able to dock properly approximately 20% of pairs more than docking methods on average, and over 10% of pairs more than the best single program. The RMSD value of the predicted complex conformation versus its native one is reduced by a factor of 0.5 Å. Finally, we were able to increase the Pearson correlation of the predicted binding affinity in comparison with the experimental value up to 0.5.

Keywords

drug discovery; PDBbind database; docking; consensus; molecular recognition

Correspondence to: D. Plewczynski; darman@icm.edu.pl.

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

Introduction

A typical drug design campaign requires substantial costs and is time consuming due to the fact that for thousands of chemical compounds biochemical screening has to be performed before proceeding to a more refined analysis. The *in silico* methods promise to shorten the time and decrease the amount of work needed when searching for a new inhibitor. One of the most important methods used here is the molecular docking that predicts a preferable conformation of a ligand when bound to a receptor molecule. Docking is used frequently in a high-throughput virtual screening where large libraries of commercially available compounds are searched to find the most active compound for a selected protein target. The aim of docking procedure is to predict the correct pose of a ligand in the binding site of the protein as well as to score it according to the strength of interaction in a reasonable time frame. As all programs exploit empirically based scoring functions and algorithms, docking results are sometimes far from reality.

Among the most frequently reported is the docking accuracy of small organic compounds to a given protein,[1–6] yet the nucleic acids can also be considered as a target for ligand molecules.[7,8] In the protein–protein docking,[8–10] the interactions between two identical or different proteins are studied. In the case of protein– ligand docking, various algorithms address different representations of a ligand and a receptor, their intrinsic chemical properties, and detailed characteristics of intramolecular interactions between their atoms. As in recent years, the crystallography and multidimensional NMR provided a wealth of structural information about various biological targets, several protein–ligand docking programs have been proposed.[11,12] Usually, the receptor is treated as a rigid molecule because of high computational costs, whereas conformational flexibility of ligands is taken into account leading to different placement algorithms.[13] The scoring procedure of such docked conformers is still regarded as one of the most difficult tasks in molecular docking because of their empirical nature. In our work, we used only software that considers flexibility of ligands, not proteins, and thus structure of protein before docking was not changed in comparison with original pdb file, assuring that protein is already in bounded state.

There are three major goals of docking simulations: (1) the native conformation of ligand in the active site should be predicted; (2) the binding energy should be estimated allowing for arrangement of the tested set of ligands according to their affinity toward the protein target; (3) in addition, it should be fast enough to screen large collections of small chemical molecules. The typical docking procedure is performed in two steps. The first step is focused on placing a small molecule into the binding site of the protein using mostly geometrical features and searching for its best three-dimensional (3D) conformation inside the cavity. The second step is performed using different scoring functions and it leads to the estimation of the binding affinity between the protein and the ligand.

During the last two decades, a set of different docking programs has become available both for commercial and academic use, such as DOCK,[14] AutoDock,[15] FlexX,[16] Surflex,[17] GOLD,[18] ICM,[19] Glide,[20] CDocker,[21] LigandFit,[22] MCDock,[23] and many others. They are based on different algorithms and can be grouped into four general categories: stochastic

Monte Carlo, fragment-based, genetic algorithms, and, finally, shape complementary methods. None of those programs uses systematic search to fully explore all degrees of freedom in both receptor and ligand molecules because of enormous computational cost of such a procedure.[2] That is why docking programs avoid systematic search and perform only guided search in conformational space. Our consensus algorithm attempts to combine those independent docking approaches into a single and powerful prediction method. We select a set of representative conformations from each docking algorithm to efficiently inspect different guided search algorithms for correct conformation of a protein–ligand complex.

The binding affinity of generated output protein–ligand conformations is calculated here by using different scoring functions. More than 30 different scoring functions were published until 2009[2,24–40] and they can be classified into three major categories. The first group applies force fields functions to calculate the energy of a complex as the sum of the ligand and the receptor internal interaction energies and also the energy of intermolecular interactions between those two molecules. Typically, the force fields such as Assisted Model Building With Energy Refinement (AMBER)[41] or Tripos[42] are employed, considering two energy terms, i.e., van der Waals and electrostatic interactions between molecules. Additionally, to improve the accuracy of those functions, sometimes the solvation energy term is also included, usually using a distance-dependent dielectric function.[43] Most of the docking programs do not support ligand binding to protein via covalent bond. However, when applied to protein–ligand complexes, the force fields are often found to overestimate the binding affinity,[2] even when using very precise and time-consuming procedures. Therefore, the scaling coefficients multiplying both terms are used to resolve this problem. The second group, i.e., the empirical scoring functions, describes interactions between a protein and a ligand as scalable parameters. Almost all of the proposed parameters exploit hydrogen bonds, hydrophobic interaction, metal bonds energy, typical force fields energies, and finally, the solvation energy term. The scaling parameters together with the empirical functions are trained on the selected dataset of complexes with known binding affinity for which scaling factors for each energy term can be optimized. Empirical scoring functions are often able to recreate binding affinities of original training dataset with very high accuracy,[24] yet the results on previously unconsidered protein– ligand complexes are not always successful. The third group, namely knowledge-based scoring functions, is developed from the statistical analysis of X-ray and NMR structures of protein– ligand complexes. The distribution of different pairs of atom types is gathered using a set of pairs of atoms, one from a protein and the other from a ligand, and then converted into pair-wise atom–atom statistical potentials. The final interaction energy is calculated as the sum of all pairwise interactions between atoms from a ligand and a protein lying within the sphere of the given cutoff (usually from 6 Å up to 12 Å).

The consensus is a novel technique recently used in various applications, mostly in bioinformatics. The main rationale behind is that although individual approaches can generate some misleading results, yet the distribution of those errors is random.[32] That is why even a simple majority voting of a set of programs providing different results can be in principle much closer to the correct answer, than even the best single program. In the context of docking problem, several attempts to transfer that approach have been made. However, as the general opinion is that posing is not the main drawback of docking

programs, typically consensus approach is applied in prediction of ligands activities. Nevertheless, some cases where authors applied this technique to poses selection were also reported. For example, Wolf et al.[44] merged two docking algorithms, namely genetic-and fragment-based method into single AutoxX protocol. The software used FlexX and AutoDock algorithms for choosing optimal ligand conformation, and it was able to decrease the mean root mean square distance (RMSD) of top score conformations by 0.3 Å in comparison with best individual program from those two. This approach allowed to predict correct conformation of ligand for 126 pairs of the 206 tested (RMSD below 2 Å), more than six for AutoDock alone. However, no consensus scoring was proposed there, thus, scoring functions were omitted and reported separately from those two programs.

Up to now, the research community focused mostly on improving scoring predictions, because in common opinion, calculating a ligand *in vitro* activity is very difficult task. Therefore, typical strategy is to gather data from diverse set of scoring functions representing different approaches to create new function using simple linear regression technique. Typically, this procedure allows for development of the function working for specific protein families; therefore, it cannot be transferred from one family to another. Similar approach was used by Teramoto et al.[28] where authors used the support vector regression performed on three protein families, acetylcholine esterase, thrombin, phosphodiesterrase 5, and proliferator-activated receptor gamma. New functions were used as an input scoring results obtained from *F*-score, *D*-score, Potential of Mean Force (PMF), *G*-score, and ChemScore. Those authors in 2007 used “rank-by-vote” approach, where instead of the absolute scores values, each ligand was given the rank based on its position in ligand list ordered by particular scoring function. Ligand with lowest average rank from the set of scoring functions was then chosen in this method as the most active one. Similar approach is also used in Sybyl’s Consensus Score (CS) model. The successful modification of “rank-by-vote” approach was implemented in SeleX-CS algorithm developed by Bar-Haim.[32] Here, the Monte Carlo simulated annealing is used to choose functions that can vote for a particular ligand. Two types of votes are allowed: “primary” rank-by-vote value, and “secondary” rank-by-number value. Authors reported three times increase in enrichment factor value obtained for studied small set of proteins. Summarizing, according to our knowledge, no single workflow that combines consensus both in pose prediction and score prediction has been introduced up to now.

Here we propose the consensus docking protocol that allows for massive docking of ligands into their corresponding protein targets using several independent docking algorithms and scoring functions running in parallel. Our approach combines the results from various programs into a single consensus prediction of the 3D protein–ligand complex structure. The clustering of results from those several docking algorithms is performed to select the poses that are close to the corresponding native conformation, and then the consensus scoring is performed using the multivariate linear regression to select the strongly binding conformations. The consensus docking method is evaluated here in terms of both posing and scoring abilities on the large dataset of protein–ligand complexes with known 3D structures and binding affinities.

Materials and Methods

Here, we present a novel docking method for selecting potent inhibitors using the results of docking performed by several programs. Seven docking software packages were used to perform the consensus procedure (AutoDock 4.2.1, Glide 4.5, GOLD 3.2, Surflex 2.2, FlexX 2.2.1, eHiTS 9.0, and LigandFit2.3). This selection covers a variety of types of docking algorithms, thus representing a rich data source for optimizing the consensus between the most popular docking programs. The method is optimized on a large benchmarking dataset of 1300 protein– ligand pairs to provide the accurate posing (RMSD value for each predicted conformation versus the corresponding native one and the percentage of successfully docked pairs in the whole collection of inhibitors) and the scoring ability (correlation of the obtained score with experimental pK_d or pK_i value). The predicted consensus pose for a given ligand on the protein target has on an average lower RMSD value in comparison with that obtained by any individual program. Furthermore, compounds predicted by us as the active ones, on an average have higher correlation coefficient calculated using the experimental binding value than scores predicted by any particular scoring functions.

Benchmark Dataset

To benchmark our method, we selected the PDBbind 2007 data-base[45–47] containing 3124 protein–ligand complexes with known 3D structure and the corresponding ligand-binding affinity. The dataset is selected as the richest and diverse dataset used in various evaluation studies.[30,48] From it, authors of PDBbind extracted a subset of 1300 protein–ligand pairs creating its “refined” set that was used in this work (see Supporting Information Table S1). The requirements for a given complex from the Protein Data Bank (PDB) database [45,46] to be included in refined, and in consequence, in our benchmark dataset were as follows.

The experimental resolution of a crystal structure has to be lower than 2.5 Å. Other studies performed previously on the GOLD original benchmark set by Jones et al.[18] confirm that selecting structures with poor resolution may produce false predicted conformations. Both a ligand and a protein structure have to be complete without any chain breaks or unsolved regions. No NMR-solved structures are involved in creating the refined data-set. In addition, only complexes with known binding affinity are considered for the refined set. The activity should be given as either pK_i (an inhibition constant), or pK_d (a dissociation constant). The results given as IC50 are rejected as such values depend on the design of a binding assay. However, recently, the accuracy of this data was questioned[49] as for 36% of the complexes the calculations of binding affinities were affected by crystal artifacts such as water molecules. We excluded complexes with ligands containing other than standard atom types (like Be or Si). Ligands that are not covalently bound with protein were discarded. Ligand mass should not be larger than 1000 amu. A complex is rejected if the distance between its ligand and the protein heavy atoms is closer than 2 Å. Finally, only complexes with a single ligand in the active site were chosen for docking simulations.

The native conformation of each ligand was extracted from the protein–ligand complex; similarly, the corresponding protein target’s 3D structure was prepared. Each ligand was then converted into a two-dimensional representation; later the Simplified Molecular Input

Line Entry Specification (SMILES)[50] chemical name of each inhibitor was created by using OpenBabel (http://openbabel.sourceforge.net/) and Marvin software (http://www.chemaxon.com). Then, the 3D input ligands structures were generated *ab initio* from their SMILES names using two typically used programs, namely the Corina[51] and Omega2[52]. Therefore, four different datasets were created. First, as a starting point, a single low-energy conformation for each ligand was generated using the Corina program. The second dataset consists of 10 conformers for each ligand generated by Corina. The third dataset contains only a single low-energy conformation for each ligand generated using Omega2 software. Finally, the fourth input dataset for docking was created using 10 of the conformers generated by Omega2 for each ligand of the PDBbind 2007 database. In addition to performing the optimization and evaluation of our docking procedure, we tested whether the docking results depend on the ligand size, type, and structure, such as the number of rotatable bonds in a molecule or its chemical properties such as hydrophobic and hydrophilic features given by the ligand partition coefficient between water and octanol:

$$\log P_{\mathrm{oct/wat}} = \log\left(\frac{C_{\mathrm{oct}}}{C_{\mathrm{wat}}}\right). \quad (1)$$

Proteins extracted from crystal structures undergo the following preparation steps. First, hydrogen atoms are added with the protonation state simulated to pH = 7. Therefore, aspartate and glutamate amino acids were negatively charged, histidine was neutral, and arginine and lysine amino acids were positively charged. The terminal carboxyl groups were deprotonated, whereas amine groups were protonated. Atoms and bonds types were inspected by using Sybyl software, yet no geometry optimization was performed. In addition, we decided to remove all water molecules and metal atoms from the protein pdb files, as no significant changes in docking accuracy of our method were observed (for more details see Supporting Information Table S2).

Docking Software

The prepared dataset of 1300 proteins and ligands was then redocked by seven independent docking programs. Here the protein active site was defined as a collection of residues in the vicinity of the bound ligand (in most cases, within the surrounding box of 12 Å size). This strategy, although seems to be different than real-life drug discovery process, was employed by us because of two reasons. The first goal of our work was to evaluate the ability to place ligand in an active site, when the active site is known, not the quality of searching for an active site by each docking program. Moreover, the most of the programs, when ligand's initial position is not known, can propose the large number of possible active sites, yet user has to make final decision and select by hand the most promising one. Our work was performed in the context of virtual high-throughput screening (vHTS), where all steps have to be automated without human intervention. Therefore, the selection of the active site was not evaluated here. The second reason is that presently the knowledge about protein structures is increasing rapidly (currently more than fifty thousands different structures is deposited in the PDB database); therefore, the comparative homology techniques could be employed to precisely locate an active site for a selected protein target. BLAST/PSI-BLAST tools, or more advanced homology search engines, such as MetaServer,[53,54] allow to find

typically a set of close homologues that are likely to share both the same protein fold, as well as the active site position. The impact of staring conformation on the final docking result was also analyzed. Four different ensembles of starting conformations for a given ligand were used. Either two single low-energy conformations from Corina or Omega2 or two sets of 10 low-energy conformers from Corina and Omega2 programs were docked separately. Each docking program predicts 10 highest scoring poses for each of those input ligands structures. We used the following docking algorithms:

- Fragment-based incremental methods: Surflex (Jain, A.N. et al.), eHiTS (SimBioSys Inc.) and FlexX (BioSolveIt) that splits ligand into pieces which are docked in an incremental way;
- Evolutionary methods: GOLD (CCDC) and AutoDock (The Scripps Research Institute) that use genetic algorithms;
- Force field-based method: Glide (Schrodinger Inc.), which has implemented Monte Carlo based engine;
- Shape complementary-based method: LigandFit (Accelrys Software Inc.), which exploits grids to fit the shape of the ligand to the target.

Therefore, in the case of the single input structure, in total, 70 different poses were generated, and in the case of 10 different input conformations of a given ligand, 700 poses were prepared; as from one input structure, 10 output conformations are usually generated. Those poses predicted by seven docking programs were compared with the 3D native structure of each ligand extracted from the corresponding protein–ligand complex. In addition, the docking scores of the top conformation returned by each docking program were compared with the experimental value of the binding affinity.

All docking programs were tested using their default parameters (see Supporting Information Table S3 for more details), although we are aware that proper selection of the scoring function and docking algorithm parameters can impact the obtained results. However, in the case of our diverse and highly populated benchmark dataset, the additional computational time used for such testing and further optimization of those parameters would make it impossible to perform using our limited hardware resources.

Docking Accuracy

The ability of docking software to predict the correct ligand poses close to the native one (found in X-ray complexes) is crucial in achieving success in docking experiment. Typically, two approaches are used for evaluating the success. The first one describes how many specific contacts between the ligand molecule and the protein are recreated (for example, hydrogen bonds or hydrophobic interactions) rather than focusing on exact placement of all ligands atoms. In our studies, millions of poses are generated; therefore, it is impossible to follow such a detailed protocol. Instead, we decided to use the second approach, i.e., calculation of the RMSD value for heavy atoms between the predicted pose and the corresponding native conformation of the cocrystallized ligand. Such a measure is well known and accepted as a reliable structural quality parameter by the whole docking community. The RMSD value between two poses is given by the equation:

$$\mathrm{RMSD}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}d_i^2}, \quad (2)$$

where N is the number of atoms and d_i is the distance between the corresponding atoms.

Of all poses that are predicted by a given docking program, one can select two poses of crucial importance. The first one is the conformation returned with the highest score by the corresponding scoring function. Here, such a pose is called the top score pose and its RMSD value to the native ligand conformation is calculated for each docking program and all analyzed protein–ligand complexes. In our approach, each program was run to predict 10 different poses for each input molecule. Therefore, for each program, we select the best pose conformation of all predicted ones, which has the lowest RMSD value to the native conformation. Those two RMSD values are very useful for benchmarking purposes, yet the best pose cannot be used in real-life experiments, where the native structure of the protein–ligand complex is not known. Usually, the best pose conformation is not returned with the highest score by a program, so it is not the top score pose.

The second useful quality parameter is the percentage of protein–ligand complexes for which the RMSD value of the top score pose is lower than 2 Å, a threshold commonly assumed as the acceptable accuracy by the docking community. In this work, we use two values of this measure, one calculated using the top score poses, and the other using the best pose conformation. Contrary to previously described RMSD values, the percentage of successfully docked complexes does not depend much on the results obtained for wrongly predicted complexes, which can significantly change the mean RMSD value averaged over the whole benchmarking dataset.

Initial Ligand Conformations Ensembles

The number of selected initial conformers for docking evaluation study may impact the docking results, and, in fact, many previous benchmarks reported that docking the native structure of a ligand extracted from X-ray crystallographic structure provides better results. Therefore, we decided to find out what is the optimal number of the prepared conformers that should be used in docking procedure and further in the consensus approach. Typically, two programs are used in research community: Corina and Omega2 to recreate crystallographic structure of molecule. Therefore, we have used them both to generate the predefined number of conformers for each input 2D ligand chemical representation. We tested the quality of those predicted 3D structures by comparing them to the known 3D structure. We have compared three ensembles of initial ligand conformations: a single most optimal structure, 10 low-energy conformers, and, finally, the 100 of predicted conformers. We have noticed that 10 conformers ensemble seems to be the optimal number, because for almost all ligands, we were able to find at least one conformer within those predicted that was almost identical to the native 3D ligand structure (RMSD below 1 Å). This observation was confirmed for both Corina and Omega2 programs. Therefore, we decided to perform four different docking experiments for each protein–ligand pair using four different ensembles of ligands' initial conformations. In two cases, we have selected only single low-

energy conformations (comprising corina one and omega one datasets). In next two cases we docked two ensembles of 10 conformations generated by either Corina or Omega2 (comprising corina ten and omega ten datasets). Therefore, we collected the docking results in those four independent experimental setups; each performed docking using different ensembles of the input ligand 3D conformations. The evaluation parameters (such as the RMSD values or the percentage of correctly docked complexes) were then averaged over those datasets, as no significant changes between those were observed. In addition, we carefully checked the docking results performed on the subset of complexes using hundreds of input conformers. The results were quite similar (in terms of both the minimal RMSD value and the percentage of accepted pairs below 2 Å). However, it takes on an average almost 10 times longer computational time in comparison with the docking of 10 conformers; therefore, the experiment was not repeated on the whole benchmarking dataset.

The binding affinity

Another important issue in prediction of protein–ligand complexes is the ability of docking programs to correctly predict the strength of ligand binding to its protein target, i.e., the binding affinity. It describes the strength of intermolecular binding between the ligand and its receptor, and it can be described by the number of parameters, such as dissociations constant K_i, concentration of ligands that decrease the activity of particular enzymes by 50% IC_{50}, or by the Gibbs free energy ΔG. The calculation of binding affinity was done here by each docking program internal scoring function. The PDBbind database reports the experimental values of the activity for all evaluated protein–ligand complexes. Therefore, we are able to compare the docking score with the corresponding experimental value of binding affinity, and to calculate the Pearson correlation coefficient between those two values. Scoring functions should order the list of poses in accordance with their binding strength to select those that are close to the native structure (hopefully the strongest bound conformation).

In our benchmark, we used the following scoring functions: GoldScore, LigScore1, GlideScore SP, Total score (the combination of several scoring functions used by Surflex: Chem-Score, *F*-Score, PMF-Score, and others), FlexX score, eHiTS score, and AutoDock scoring function. For each scoring function, we calculated Pearson and Spearman correlations for four different sets of conformers (generated using Corina and Omega2 software). The Spearman correlation is much less sensitive to a few outsiders, and on the contrary, a few wrong protein–ligand complexes can significantly affect the Pearson correlation coefficient.

Selection of the optimal pose

Our benchmarking results show that no single docking software is more reliable than others. Therefore, the consensus approach seems to be the right tool to boost the overall docking accuracy. Several consensus approaches were successfully applied in the context of bioinformatics,[54] chemoinformatics,[55] and general computer sciences. For example, the 3D-Jury[53] algorithm predicts a 3D protein structure using several autonomous methods. Here, we present our novel method MetaPose that is able to improve the selection of the best pose using a set of conformations obtained from various docking programs. The *n* predicted poses

from seven tested programs are used, neglecting the scores obtained using their scoring functions (n = 70 for Corina and Omega2one, and n = 700 for Corina and Omega2 ten input datasets). Those 3D conformations are compared with each other by calculating the RMSD value between them. Even if a subset of the predicted poses is obviously wrong, or contaminated (for example, as a result of weak docking program), yet this does not affect our method because the majority of molecules would be placed correctly in a protein active site. In this way, the similarity matrix is created, where the ith row represents the similarity of the pose i to all other poses. The pose score 3DScore for this conformation is calculated as the arithmetical mean of RMSD values from the selected row:

$$\mathrm{3DScore}_i = \frac{1}{n}\sum_{j=1}^{n} w_{ij}, \quad (3)$$

where n is the number of conformations and w_{ji} is the RMSD value calculated between i and j poses.

The pose with the lowest value of the score is selected by our method as the final result. We search here for the conformation that represents the entire set of possible poses; therefore, it is most similar to others. The conformational search of the MetaPose method is based only on the geometrical similarity between poses, without taking into account the scores given by docking programs.

The strongly binding poses

The second step of our analysis is designed to improve the correlation between the docking score and its experimental values of the binding affinity. The MetaScore algorithm builds the consensus scoring function using multivariate linear regression optimization guided by the experimental results. We assume that the consensus scoring function is described by a linear combination of scores from seven docking programs with the weights describing the influence of each scoring function on the final result.

The optimization method has to be constructed and tested separately using different and not overlapping datasets of protein–ligand complexes. Therefore, we divided our benchmark dataset into two parts. The first dataset contains 400 pairs and was used for optimization of the new metascoring function. The second dataset (the rest of database) was used as the independent testing dataset. Each protein–ligand complex can be represented as an ensemble of seven hundreds or tens of poses (depending on the size of input conformers used in docking); therefore, each pose can be described by seven different docking scores of top score poses. The multivariate linear regression was then used to obtain the single MetaScore function for our four benchmarking subsets, namely Corina one, Corina ten, Omega2 best, and Omega2 ten generated conformers. However, surprisingly, the coefficients for each consensus scoring function were tested to be almost equal for each of those four optimization datasets, therefore the single MetaScore function can be proposed, and it is given by the equation:

$$\begin{aligned}\text{MetaScore} = &-0.378*\text{eHiTSScore}+0.015*\text{FlexXScore}\\ &-0.358*\text{GildeSPScore}\\ &+0.014*\text{GoldScore}+0.004*\text{LigScore}+0.15*\text{SurflexScore},\end{aligned} \tag{4}$$

where each scoring function is described by the name of the corresponding docking program. Result of this function is value that represents ($-\log K_d$) value for a given protein–ligand complex.

The MetaScore subset of poses, i.e., all top score conformations from seven docking programs, are used for consensus procedure. We created the 7 × 7 similarity matrix calculating the RMSD values between those seven top score conformations. Then, for each of them, the 3DScore was calculated and the conformation with the lowest value was chosen as the predicted pose. The results of this procedure are summarized below and are presented in Tables 1 and 2 (the MetaScore row).

In addition, having all conformations ordered by their corresponding 3DScore value, we chose the top representative for all seven docking program, each with the lowest 3Dscore value of all predicted poses by this program. MetaScore eq. (4) can be calculated using docking scored of those six conformations. Similarly, the Pearson correlation between those values and experimentally determined binding affinities can be calculated. The results are presented in Table 1.

The consensus docking method

In the third step of our analysis, we designed the consensus method that is able both to predict the correct conformation of a protein-ligand complex, and its binding affinity value. Previously described methods, namely MetaPose and MetaScore, focus on different goals for virtual screening. The first algorithm selects the pose inside the protein active site that is close to the native one, and the second one focuses on the calculation of correct binding affinity for the analyzed ligand. Here, we introduce VoteDock algorithm that predicts both the correct pose and the strength of binding between the ligand and the receptor. The VoteDock uses modified MetaPose algorithm to select the correct conformation. The MetaPose neglects the information about the source of analyzed conformations (namely the docking programs from which they are taken). Therefore, although it is proven to be more effective than any individual program, yet the influence of each docking program is similar for all used algorithms, even if there is a single one among them that has a very weak ability to pose the ligand inside the active site, or to score it.

For each protein-ligand complex, we create subsets of poses, each containing the poses selected independently by a certain number of docking programs. Therefore, we can assign to each predicted conformation the number from one to six depending on the number of programs that confirmed that pose as the correct one. The similarity matrix for each subset of poses is calculated separately for all poses that were confirmed by the predefined number of docking programs, and called vote2, vote3, up to vote7. For example, if a conformation from GOLD was also predicted by eHiTS, i.e., if there exists at least one conformation from eHiTS's predicted poses that has the RMSD value between it and the GOLD's conformation

lower than the threshold value of 2 Å, such a conformation is qualified to be included in vote2 subset. If the next docking program (for example FlexX) has another predicted conformation closer than 2 Å from the original pose, then the pose is also included in vote3 subset. In the case of vote7 dataset, the highest quality predicted conformations; all seven docking programs predicted each of them as the possible ones in our test. The voting procedure not only narrows down the number of predicted conformations but also eliminates incorrect poses that influence the center of the solutions domain. In the case of some protein–ligand complexes, the vote7 dataset is empty within the given RMSD threshold, therefore, the hybrid approach is proposed. For every protein–ligand pair, we select the subset of predicted conformations, which have the highest available vote order, and subsequently, we use MetaPose approach on such highest vote dataset. The similarity matrix is constructed and the 3DScore [see eq. (3)] is assigned to those poses. The conformation with the lowest value of 3DScore is chosen as the best one. If no vote subset is present (even the vote2 subset), as the result for such a protein–ligand complex, the pose selected by original MetaPose algorithm is returned.

The conformations selected by the first step of VoteDock procedure are then scored using the MetaScore scoring function, as each pose is now described by more than one docking score. A detailed analysis of vote subsets allows one to select a single program that is eliminated from each vote. In the case of vote7, the scores from all seven programs describe each predicted conformation. In the case of vote6, the AutoDock predictions are excluded for most of the complexes, vote5 eliminates Glide, vote4 FlexX docking program, vote3 removes LigandFit, and finally, in vote2, in most cases, eHiTS is lacking, leaving only two programs, namely Surflex and GOLD. The MetaScore procedure is here modified by optimizing weights of each docking program for each vote dataset separately, using multivariate linear regression (details on scoring function used here are listed in Supporting Information Table S4). Therefore, six new Meta-Score functions are calculated, each for the particular subset of docking programs. To compare correlation values from the VoteDock with individual docking programs, here we use the hybrid approach similarly to the previously described optimal conformation selection prediction procedure. We select the highest vote subset for each protein–ligand complex and apply the MetaScore optimized function suitable for this particular vote order. If no vote is present or a conformation is described by a different combination of programs, we apply the original Meta-Score procedure. The highest scores from the available program are collected, and the eq. (4) is used to calculate the predicted value of the ligand binding affinity. The workflow of data is presented in Figure 1, together with the schematic diagram showing how consensus methods work.

Results

In this section, we summarize the quality of the three proposed consensus methods, namely MetaPose, MetaScore, and Vote-Dock and compare them with the results of seven diverse docking programs: AutoDock, eHiTS, FlexX, Glide, GOLD, LigandFit, and Surflex. First, we describe the ability of all used algorithms to correctly predict the ligand binding poses. In addition, we test whether the consensus docking results depend on the ligand size, type, and structure, such as the number of rotatable bonds in a molecule or its chemical properties like

hydrophobic and hydrophilic features given by the ligand partition coefficient between water and octanol. Then, we report the ability of all programs' scoring functions to accurately predict the experimental binding affinities (pK_i and pK_d). The proposed novel methods that are based on the consensus between various docking algorithms are proved below to considerably increase docking accuracy and proper sorting of predicted poses.

In Tables 1 and 2, we present the evaluation of our consensus algorithms in both, the correct poses and the binding affinities prediction. Each consensus method is compared with the results obtained by the best and the second best docking program on the whole benchmarking dataset. In addition, we provide the mean docking result calculated by averaging the results of all seven docking programs on the same dataset.

We also explored several physicochemical features of the ligand, which are often used in various docking programs evaluations. First, we divided our datasets using the number of rotatable bonds in a ligand molecule. We created subsets of small compounds, which have five bonds or less, and large compounds, which have more than five rotatable bonds. It is obvious that for smaller molecules, the results will be better, yet our main goal was to identify, which program depends less on ligand size. Next, the hydrophobic/hydrophilic properties of ligands were analyzed. As previously, two datasets are created using the log*p* values. This parameter also covers the number of possible hydrogen bonds, which a ligand can create with a protein, as for more hydrophobic ligands, fewer contacts are usually built. Another ligands subset that we explored contains proteinlike molecules. This subset is interesting because of the growing number of protein-like drugs that are introduced to the market. We wanted to evaluate the quality of prediction for those types of molecules. Finally, the benchmarking dataset was divided based on the strength of the binding between the ligand and its corresponding receptor. Here, our goal was to determine if there is a preferable compound type that docking programs could handle more precisely, for example, small and strongly binding molecules. The results of those evaluations are presented for two preselected conformations: top score conformation and best pose conformation. The top score pose for each docking program, or consensus method, is the conformation that achieved the highest docking score of all generated by the program, whereas best pose is the one with the lowest RMSD value to the native conformation. In the case of MetaScore, top score conformation has the lowest 3DScore value of all conformations with the highest docking score from individual programs, whereas best pose has the lowest RMSD value of those conformations. MetaPose and VoteDock algorithms similarly have top score pose as the pose with the lowest value of 3DScore function among all poses generated by those consensus algorithms (MetaPose), or among those from the highest order vote subset of conformations, preferably vote7. In the case of MetaPose and VoteDock, the best pose was selected as the pose with the lowest RMSD value of the first 10 (corina one and omega one), or 100 (corina ten and omega ten) poses ordered by the 3DScore. Those limits in the number of analyzed poses simulates the use of our consensus algorithms as the stand-alone programs; therefore, they would generate only tens or one hundreds of poses.

In Table 1, we present the Pearson and Spearman correlations between the experimentally determined binding affinities and the scores from all scoring function. In our work, those correlations are calculated for both top score conformation and the best pose. To calculate

the correlation between the score of best pose for MetaScore and MetaPose algorithms, the scores of best poses for individual docking programs are used, and the total score is calculated using eq. (4). However, MetaPose uses only the first ten or one hundred conformations, and sometimes some docking programs are not represented in this dataset, therefore, not always all six scores were used in eq. (4), and zero value was used for such missing docking scores. In the case of VoteDock, instead of the highest 3DScore conformations, we used individual docking scores for conformations with the lowest RMSD values of the first ten or one hundred conformations as ordered by their 3DScore values. Our evaluation was done separately on four different datasets (corina one, corina ten, omega one, and finally omega ten), therefore, the values in Tables 1 and 2 are averaged over those subsets.

Quality of pose generation

The best docking program is the GOLD software, which in top score prediction outperforms other programs. However, the GOLD program uses the slowest, time-consuming algorithm that takes more than 5 min to dock a ligand to the receptor. In Table 1, GOLD was not chosen as the best program only twice, for the hydrophobic dataset of ligand eHiTS is the first one, and GOLD is the second one. The weakness of GOLD program in the case of hydrophobic ligands was already pointed out in other benchmarks.[1] The GOLD scoring function was also not the first one in terms of the correlation with the experimental binding affinity (see Table 1). We report the eHiTS and Surflex scoring functions as the best and the second best program. In Table 2, we present the results of the pose generation that are analyzed in terms of ligand binding strength, where we report GOLD as the first program; however, in the case of strongly binding and small ligands, eHiTS achieves better results. In most cases, eHiTS is the second best docking program and sometimes switches its position with GOLD. In the case of protein-like ligands, AutoDock is chosen as the second program.

In the case of best pose presented in Tables 1 and 2, the best docking programs are usually the same as in the case of top score conformations. More diversity is observed for the second best program, namely Surflex (large ligand and hydrophobic ligands datasets) and LigandFit (protein dataset) are better than others. For the entire benchmark dataset of 1300 complexes, MetaPose algorithm is nearly 18% more accurate than averaged docking results, and its mean RMSD value drops by more than 1 Å. A smaller but still significant change can be observed when comparing MetaPose to the best docking program, where the increase is almost 5%, and the RMSD value is improved by almost 0.2 Å. Even more accurate is the VoteDock consensus docking method where the average docking accuracy increases by almost 23% in comparison with the average docking programs accuracy, and by more than 10%, when compared with the best result obtained by GOLD. The mean RMSD value is also increased by 0.5 Å in comparison with GOLD. In the case of MetaScore, structural results are above average docking, yet less than a 10% increase in successfully docked pairs can be observed. However, when compared with the best and the second best program, the obtained results are not so good as before. Therefore, the results on top score subset prove that MetaScore is very efficient in the prediction of binding affinities, yet this does not correspond to the good overall structural prediction.

The difference between MetaPose and the best docking program for all analyzed types of physicochemical features is typically around 5% increase in docking accuracy and 0.3 Å in the mean RMSD value. The VoteDock is usually even more accurate with a 10% increase in comparison with the best docking program, and the mean RMSD value usually drops by more than 0.6 Å. A similar difference can be seen when comparing the dataset composed of small molecules with one composed of large molecules. The number of ligand rotatable bonds describes its flexibility. The worst docking results are obtained for large ligands (the high number of rotatable bonds), which is due to the fact that the size of explored conformation space increases dramatically. MetaPose is on average 3 and 4% more accurate than the best docking program for, respectively, the small and large molecules subsets, whereas for VoteDock, 9 and 11% increase is observed. The gap in docking accuracy between small and large dataset of ligands is smaller when a consensus-docking algorithm is used. In the case of the best docking algorithm, there is an almost 20% drop in accuracy between small and large datasets. In the case of VoteDock, the drop of accuracy is close to 15%, MetaPose achieves intermediate results of around 17%.

The same conclusions can be observed when dividing the entire benchmarking dataset using hydrophobic and hydrophilic characteristics of ligands. Those features result from some important aspects of ligand behavior, mostly the ability to create hydrogen bonding between a ligand and a protein, as well as forming interactions with the hydrophobic pocket in the protein active site. As expected, hydrophilic ligands are predicted with a much higher rate of success than hydrophobic ones. However, it should be remembered that for the hydrophilic dataset, GOLD was chosen as the best program, and for the hydrophobic data-set, eHiTS was most successful. MetaPose and VoteDock are more accurate than any individual docking software. For hydrophilic ligands, there is, respectively, a 2 and 7% increase between single docking and the consensus result. Those two metaalgorithms are even more accurate for hydrophobic ligands with a 9 and 14% increase in docking accuracy. Those results clearly suggest that the consensus approach can be very effective in avoiding the weaknesses of individual docking programs. Similarly, as in the case of the large and small molecules subsets, there is a much smaller gap when comparing hydrophilic ligands with hydrophobic ones. For the consensus approach, the difference is close to 10% in the number of successfully docked pairs, and 0.4 Å for the mean RMSD value, whereas for the single docking programs, those differences are significantly higher, reaching almost 20% and 0.7 Å.

The increasing role of short peptides as potential drugs such as antibiotics, antihistamine, or antitumor agents create a unique opportunity to benchmark available docking software on known protein–peptide complexes. We have created a small benchmark dataset containing proteins with cocrystallized peptides, or other protein-like molecules. The complexes were extracted from PDBbind 2007 dataset, assuming that a selected ligand contains at least one amino acid-like substructure, and therefore, it is not always identical to the naturally occurring peptides. In the case of such polymers, we have checked the presence of protein bonding between molecule substructures, and structures with nontypical atoms (all types except oxygen, hydrogen, carbon, and sulfur) were discarded. Fourteen complexes contain phosphate atoms, and for six complexes, some fluorine atoms were found. Following this procedure, we have created the peptide benchmark dataset that contains in total 143

complexes for which ligand size may vary from single amino acids up to longer peptide chain created from tens of mers. The best docking software was able to find the correct top score pose only for 46% of those complexes within 2 Å cut off from the native ligand structure. The mean RMSD value for the top score pose was typically higher than 4 Å. eHiTS as the second best docking program achieved a very similar result. Our consensus approaches are able to increase the accuracy up to 50% for MetaPose and 57% for VoteDock. The mean RMSD value decreased to almost 3.5 Å and 3.3 Å, respectively. Therefore, we prove that our method is better in predicting the correct conformations for small proteins inside the receptor active site. Finally, we divided the ligands from the whole benchmarking dataset into three groups, according to their experimentally measured binding affinities to the corresponding protein receptors. The first group (strong) contains ligands for which their concentration necessary to inhibit the enzyme is lower than 45 nM; the second (medium) has their pK_i or pK_d between 45 nM and 3.6 μM, and, finally, the third group of inhibitors (weak) with the concentration of a compound greater than 3.6 μM. For all those groups, we calculated how many small and large molecules fall into each category, to check if the dependence of the benchmarking results is based only on ligand binding strength and not on its size. In the case of strong dataset, there are 271 large ligands and 159 small ligands; for medium dataset, there are 213 and 222; and finally for weak dataset, there are 165 and 270, respectively. In Table 2, we summarize the results for each subset, additionally divided into small and large molecules. In the case of individual programs, there is a small difference between particular datasets, however, small&weak and small&-medium molecules are usually better predicted than small&-strong ones. A similar trend is observed when looking for large molecules where large&weak molecules are predicted to be 10% more accurate than large&strong ligands, in case of GOLD, best docking program. This result is very unfortunate as in typical drug design strong-binding molecules are searched for. Our consensus method follows individual docking programs trend, however, both MetaPose and VoteDock seem to be less affected when comparing strong to medium, or medium to weak subsets. In the case of VoteDock, the drop of accuracy between for example small&weak and small&strong bound molecules is only 3%, whereas the results for large&weak and large&strong are almost identical. The mean RMSD value seems to change between those datasets only marginally. Similar behavior can be observed for MetaPose algorithm.

Summarizing, the consensus structural methods (MetaPose and VoteDock) seem to be more effective than any individual docking program. The percentage of successfully docked pairs is 5 and 10% higher than the best individual docking program, and the observed result is even higher when comparing the consensus with the averaged docking results. The significant improvement in the mean RMSD value is observed with VoteDock close to the cut-off value of 2 Å. Similar results are observed when dividing our dataset based on physicochemical properties of a ligand. Interestingly, the consensus methods prove to be successful even when individual docking programs fail. The hydrophobic dataset can be given as an example here, where there is more than 10% increase between the best program and VoteDock algorithm.

Evaluation of binding affinity prediction

The second important issue in the prediction of protein–ligand complexes is the ability of docking programs to predict the strength of a ligand binding to its protein target. The best docking program with the highest correlations is eHiTS for which the correlations are 0.39 and 0.47 for Pearson and Spearman correlations, respectively. Surflex follows the eHiTS scoring function closely, with the correlation equal to 0.3 and 0.34, respectively. The averaged results for all seven scoring functions are around 0.2 for both Pearson and Spearman correlations, proving that programs have significant problems when binding affinities have to be predicted. The consensus docking procedure, namely Meta-Score [see eq. (4)] scoring function significantly improves the Pearson correlation between the final docking metascore and the experimental value of the binding affinity (see Table 1). In the case of training dataset (randomly selected subset of 400 protein–ligand complexes), the MetaScore reaches the value of 0.46 for the Pearson correlation. The optimized MetaScore was later tested on the rest of the protein–ligand complexes from the whole benchmarking dataset that were not used in the training. The Pearson correlation dropped slightly to 0.44, yet still it is much higher than any single docking program. If the MetaScore was trained on the whole PDBbind dataset, the Pearson correlation for the whole benchmarking dataset is equal to 0.48. Similar values were observed for Spearman correlation. What is more, the values of weights multiplying each used docking programs scores were found to be equal for all analyzed subsets used in optimization of the MetaScore equation, for example, for smaller and larger ligands if used separately for optimization procedure. Therefore, the MetaScore scoring function is the universal one, and it is applicable to various types or classes of inhibitors.

VoteDock is able to increase the Pearson correlation up to 0.49 and Spearman up to 0.5. However, in the case when protein–ligand complexes have at least vote2 or higher order subset and only the optimal combination of single docking programs is found (such conditions are fulfilled for half of the entire benchmarking dataset), the correlation is even higher and is close to 0.6. Summarizing, the drop in the values of Pearson and Spearman correlation for the whole dataset is explained by the fact that only half of the whole subset meets our selection criteria.

The third algorithm MetaPose achieved the worst results of all consensus approaches close to 0.4 for both Pearson and Spearman correlations, yet it is still higher than the correlation achieved by any single docking program. Nevertheless, MetaPose is designed to be the pose-prediction algorithm and should not be used for binding affinities prediction, as more accurate consensus docking algorithms are optimized and presented in this article.

Consensus docking strategy

In Table 3, we present the results for each step of VoteDock procedure. VoteDock is the hybrid approach that uses for an individual protein–ligand pair, the highest possible vote order dataset created by our consensus procedure. In Table 3, we present how many pairs can be classified as each vote order in the hybrid VoteDock procedure based on the selected threshold. The less strict threshold is selected, the more pairs pass to higher vote order dataset. For example, for the threshold of 1 μ for only 43 pairs, the vote7 subset of poses

exists, whereas for the less restricted threshold of 3 Å, more than 220 pairs have vote7 as the final VoteDock set of solutions.

On the other hand, when the threshold increases the docking accuracy decrease is observed. For the threshold of 1 Å, the accuracy is 97%; in the case of the threshold of 1.5 Å, almost 95% is achieved; the threshold of 2 Å has 87%, the threshold of 2.5 Å 81.5%; and finally for the threshold of 3 Å ,only 78% of pairs is docked successfully. A similar trend can be observed for the mean RMSD value, which increases from 0.6 Å for the threshold of 1 Å up to 1.42 Å for the threshold of 3 Å. Therefore, we selected for VoteDock algorithm the optimal threshold equal to 2 Å, as it maximizes both the mean RMSD and the percentage of successfully docked pairs.

Conclusions

Very low values of correlation coefficients between docking score and pK_d (pK_i) indicate that all molecular docking programs that have been tested are unable to predict binding affinities correctly. Our results clearly show that still there is the lack of scoring function that would be universal for all kinds of ligands and protein families. Although reports of increasing accuracy of scoring functions have already been published,[30] yet none of the functions we studied could be classified as reliable, and so further analysis is necessary. On the other hand, protein– ligand docking programs can predict with high accuracy poses of ligands in the binding site of protein. As they are usually very fast, this capability is extremely valuable.

Our novel methods, namely MetaPose, MetaScore, and Vote-Dock, use as the source data the results of individual programs, and by consensus approach, they attempt to overcome the weaknesses of each docking program to better predict both a ligand conformation in the active site and the strength of that interaction. The VoteDock, which is a combination of MetaPose and MetaScore algorithms, outperforms each individual docking program, both as concerns the correct pose predictions and the scoring. Those observations lead to the conclusion that applying metaapproach is a very successful procedure and worth exploring in the near future when even more docking programs will be available. Although individual programs in some particular cases can be close to our consensus methods, none of them can reach the quality of VoteDock on the large dataset of more than a thousand ligand–protein complexes. The future improvements of the existing software will strongly and positively affect the accuracy of VoteDock algorithm. The consensus will benefit from those advances, and its quality will be further increased.

Consensus algorithms increase the number of successfully docked pairs up to 70% for VoteDock and 63% for MetaPose, whereas the best docking software in our evaluation reached less than 60% docking accuracy. The mean RMSD of top score pose for VoteDock is equal to 2 Å, and for MetaPose is nearly 2.5 Å, which confirms that the consensus approach is a powerful tool in predicting ligand conformation inside a protein active site. In 93% of the protein–ligand complexes, at least one program was able to predict a single conformation with the RMSD value to the ligand native structure less than 2 Å. We were unable to exceed that value because a consensus method does not create ligands or poses *de*

novo but only allows for selection of poses out of those that were previously generated. Moderate success was achieved in terms of the binding affinity prediction; the correlations for VoteDock and MetaScore are close to 0.5. Although the correlation for the best scoring function in our evaluation is substantially lower and equal to 0.38, we are aware that our results are still about halfway to achieve the perfect 1.0 value. Correlation values for docking programs show that there is still a lot to be done to increase the accuracy of scoring functions. Further improvements of docking software will significantly improve VoteDock accuracy. In addition, if we could generate with VoteDock procedure at least to the level of vote2 for all benchmarked ligands, the overall correlation for them would improve substantially up to 0.6. This value is twice as good as the best docking program correlation (eHiTS score).

Furthermore, our method could be an important contribution to the vHTS. vHTS is a computational method, which is widely applied to *in silico* screening of commercial collections of compounds to select the most potent inhibitors for a selected protein receptor. Typically, because of its speed and prediction accuracy several ligand-based methods make use of the information provided by already known inhibitors, such as pharmacophore matching, 3D shapes matching methods,[56] or clustering and machine-learning techniques.[55,57,58] However, when the target structure is known and there is no prior knowledge about inhibitors, typically docking techniques have to be used.

In addition, libraries of compounds for vHTS, such as Ligand.Info,[59] can contain millions of molecules. The time required for docking large datasets of compounds with the programs we used in this article on typical 100 nodes cluster would be close to 3 months (AutoDock 192 s/molecule, eHiTS 168 s/ molecule, FlexX 34 s/molecule, LigandFit 110 s/molecule, Surflex 100 s/melocule, GOLD 305 s/molecule, and Glide 660 s/molecule). The time of running our consensus docking procedure to analyze all predicted poses is marginal in comparison with the time of individual docking. We need only a few seconds per molecule to process all poses for a selected compound, therefore, the overall increase in time is less than a week. The extra time needed to process docking results will improve the accuracy of predictions by almost 15%, and eliminates many false positives from vHTS experiment saving months of experimental work on biochemical analysis. Recently, our consensus method was successfully applied for screening possible drugs against H1N1 influenza virus; the details will be provided in a forthcoming paper.

Summarizing, our VoteDock consensus docking algorithm is able to predict a structure of a protein–ligand complex within 1400 s/molecule on the 2-GHz single-core processor. When a set of input ligands or multiple protein targets are used, the time needed to perform the prediction is scaling linearly with the size of the input dataset. However, because of the parallelization on the linux cluster, those calculations can be submitted at once, the overall time is similar to that for a single submission. In the case of large datasets of proteins or ligands (for example, whole proteomes and metalobomes), we suggest performing the vHTS experiment using MLdock service (machine learning–based algorithm[55,57,60] combined with ICM-Pro docking[58]) instead of slower VoteDock algorithm.

The above VoteDock consensus method will be used in our new internet server that is now under development. Although most of docking programs used in our work cannot be distributed under academic license agreement, we have decided to combine VoteDock approach for pose prediction (using conformations obtained using AutoDock and DOCK software) with scoring prediction using the statistical function SMOG (Small Molecule Growth). Moreover, the VoteDock pipeline allows user with access to local versions of software described above to perform pose selection by combining several files with decoys prepared using those different docking programs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The calculations were performed in the Interdisciplinary Centre for Mathematical and Computational Modelling (ICM) at Warsaw University under the computational grants G14-6 and G30-2. The authors thank Prof. Iwona Wawer for fruitful discussions and Kamil Steczkiewicz for his help during the preparation of the article.

Contract/grant sponsor: OxyGreen (6FP project); contract/grant number: KBBE-2007-21228

Contract/grant sponsor: EMBO Installation

Contract/grant sponsor: National Institute of Health; contract/grant number: 1R01GM081680-0

Contract/grant sponsor: Foundation for Polish Science (Focus)

Contract/grant sponsor: Polish Ministry of Science and Higher Education; contract/grant numbers: N301 159735, N301 159435, N301 24643

Contract/grant sponsor: PepLaser; contract/grant number: HEALTH-2007-22324

References

1. Perola E, Walters WP, Charifson PS. Proteins. 2004; 56:235–249. [PubMed: 15211508]
2. Moitessier N, Englebienne P, Lee D, Lawandi J, Corbeil CR. Br J Pharmacol. 2008; 153(Suppl 1):S7–S26. [PubMed: 18037925]
3. Chen H, Lyne PD, Giordanetto F, Lovell T, Li J. J Chem Inf Model. 2006; 46:401–415. [PubMed: 16426074]
4. Bissantz C, Folkers G, Rognan D. J Med Chem. 2000; 43:4759–4767. [PubMed: 11123984]
5. Perola E, Walters WP, Charifson P. J Chem Inf Model. 2007; 47:251–253. [PubMed: 17260981]
6. Kellenberger E, Rodrigo J, Muller P, Rognan D. Proteins. 2004; 57:225–242. [PubMed: 15340911]
7. Holt PA, Chaires JB, Trent JO. J Chem Inf Model. 2008; 48:1602–1615. [PubMed: 18642866]
8. Vajda S, Kozakov D. Curr Opin Struct Biol. 2009; 19:164–170. [PubMed: 19327983]
9. Andrusier N, Mashiach E, Nussinov R, Wolfson HJ. Proteins. 2008; 73:271–289. [PubMed: 18655061]
10. Ritchie DW. Curr Protein Pept Sci. 2008; 9:1–15. [PubMed: 18336319]
11. Krovat EM, Fruhwirth KH, Langer T. J Chem Inf Model. 2005; 45:146–159. [PubMed: 15667140]
12. Krovat EM, Steindl T, Langer T. Curr Comput Aided Drug Des. 2005; 1:93–102.
13. Rester U. QSAR Comb Sci. 2006; 25:605–615.
14. Ewing TJ, Makino S, Skillman AG, Kuntz ID. J Comput Aided Mol Des. 2001; 15:411–428. [PubMed: 11394736]
15. Morris M. J Comput Chem. 1998; 19:1639–1662.

16. Rarey M, Kramer B, Lengauer T, Klebe G. J Mol Biol. 1996; 261:470–489. [PubMed: 8780787]
17. Jain AN. J Med Chem. 2003; 46:499–511. [PubMed: 12570372]
18. Jones G, Willett P, Glen RC, Leach AR, Taylor R. J Mol Biol. 1997; 267:727–748. [PubMed: 9126849]
19. Abagyan RA, Totrov MM, Kuznetsov DA. J Comput Chem. 1994; 15:488–506.
20. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, Shaw DE, Francis P, Shenkin PS. J Med Chem. 2004; 47:1739–1749. [PubMed: 15027865]
21. Wu G, Robertson DH, Brooks CL III, Vieth M. J Comput Chem. 2003; 24:1549–1562. [PubMed: 12925999]
22. Venkatachalam CM, Jiang X, Oldfield T, Waldman M. J Mol Graph Model. 2003; 21:289–307. [PubMed: 12479928]
23. Liu M, Wang S. J Comput Aided Mol Des. 1999; 13:435–451. [PubMed: 10483527]
24. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. J Comput Aided Mol Des. 1997; 11:425–445. [PubMed: 9385547]
25. Zsoldos Z, Reid D, Simon A, Sadjad BS, Johnson AP. Curr Protein Pept Sci. 2006; 7:421–435. [PubMed: 17073694]
26. Meng EC, Shoichet BK, Kuntz ID. J Comput Chem. 1992; 13:505–524.
27. Wang R, Lu Y, Wang S. J Med Chem. 2003; 46:2287–2303. [PubMed: 12773034]
28. Teramoto R, Fukunishi H. J Chem Inf Model. 2008; 48:288–295. [PubMed: 18229906]
29. Perez C, Ortiz AR. J Medl Chem. 2001; 44:3768–3785.
30. Cheng T, Li X, Li Y, Liu Z, Wang R. J Chem Inf Model. 2009; 49:1079–1093. [PubMed: 19358517]
31. Paulsen JL, Anderson AC. J Chem Inf Model. 2009; 49:2813–2819. [PubMed: 19950979]
32. Bar-Haim S, Aharon A, Ben-Moshe T, Marantz Y, Senderowitz H. J Chem Inf Model. 2009; 49:623–633. [PubMed: 19231809]
33. Zavodszky MI, Stumpff-Kane AW, Lee DJ, Feig M. J Com-put Aided Mol Des. 2009; 23:289–299.
34. Korb O, Stutzle T, Exner TE. J Chem Inf Model. 2009; 49:84–96. [PubMed: 19125657]
35. Kortagere S, Chekmarev D, Welsh WJ, Ekins S. Pharm Res. 2009; 26:1001–1011. [PubMed: 19115096]
36. de Azevedo WF Jr, Dias R. Bioorg Med Chem. 2008; 16:9378–9382. [PubMed: 18829335]
37. Yin S, Biedermannova L, Vondrasek J, Dokholyan NV. J Chem Inf Model. 2008; 48:1656–1662. [PubMed: 18672869]
38. Fukunishi H, Teramoto R, Takada T, Shimada J. J Chem Inf Model. 2008; 48:988–996. [PubMed: 18426197]
39. Teramoto R, Fukunishi H. J Chem Inf Model. 2008; 48:747–754. [PubMed: 18318474]
40. Dias R, Timmers LF, Caceres RA, de Azevedo WF Jr. Curr Drug Targets. 2008; 9:1062–1070. [PubMed: 19128216]
41. Weiner SJK, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta S Jr, Weiner P. J Am Chem Soc. 1984; 106:765–784.
42. Matthew C, Cramer RD III, van Opdenbosch N. J Comput Chem. 1989; 10:982–1012.
43. Cheng TM, Blundell TL, Fernandez-Recio J. Proteins. 2007; 68:503–515. [PubMed: 17444519]
44. Wolf A, Zimmermann M, Hofmann-Apitius M. J Chem Inf Model. 2007; 47:1036–1044. [PubMed: 17492829]
45. Wang R, Fang X, Lu Y, Wang S. J Med Chem. 2004; 47:2977–2980. [PubMed: 15163179]
46. Wang R, Fang X, Lu Y, Yang CY, Wang S. J Med Chem. 2005; 48:4111–4119. [PubMed: 15943484]
47. Wang R, Lu Y, Fang X, Wang S. J Chem Inf Comput Sci. 2004; 44:2114–2125. [PubMed: 15554682]
48. Andersson CD, Thysell E, Lindstrom A, Bylesjo M, Raubacher F, Linusson A. J Chem Inf Model. 2007; 47:1673–1687. [PubMed: 17559207]

49. Sondergaard CR, Garrett AE, Carstensen T, Pollastri G, Nielsen JE. J Medi Chem. 2009; 52:5673–5684.
50. Weininger D. J Chem Inf Comput Sci. 1988; 28:31–36.
51. Sadowki, J.; Schwab, C.; Gasteiger, J. CORINA version 3.4 available at http://www.mol-net.de
52. Bostrom J, Greenwood JR, Gottfries J. J Mol Graph Model. 2003; 21:449–462. [PubMed: 12543140]
53. Ginalski K, Elofsson A, Fischer D, Rychlewski L. Bioinformatics. 2003; 19:1015–1018. [PubMed: 12761065]
54. Ginalski K, von Grotthuss M, Grishin NV, Rychlewski L. Nucleic Acids Res. 2004; 32:W576–W581. Web server issue. [PubMed: 15215454]
55. Plewczynski D, Spieser SA, Koch U. J Chem Inf Model. 2006; 46:1098–1106. [PubMed: 16711730]
56. Wolber G, Seidel T, Bendix F, Langer T. Drug Discov Today. 2008; 13:23–29. [PubMed: 18190860]
57. Plewczynski D, Spieser AH, Koch U. Comb Chem High Throughput Screen. 2009
58. Plewczynski D, von Grotthuss M, Rychlewski L, Ginalski K. Comb Chem High Throughput Screen. 2009
59. von Grotthuss M, Koczyk G, Pas J, Wyrwicz LS, Rychlewski L. Comb Chem High Throughput Screen. 2004; 7:757–761. [PubMed: 15578937]
60. Plewczynski D, von Grotthuss M, Spieser SA, Rychlewski L, Wyrwicz LS, Ginalski K, Koch U. Comb Chem High Throughput Screen. 2007; 10:189–196. [PubMed: 17346118]

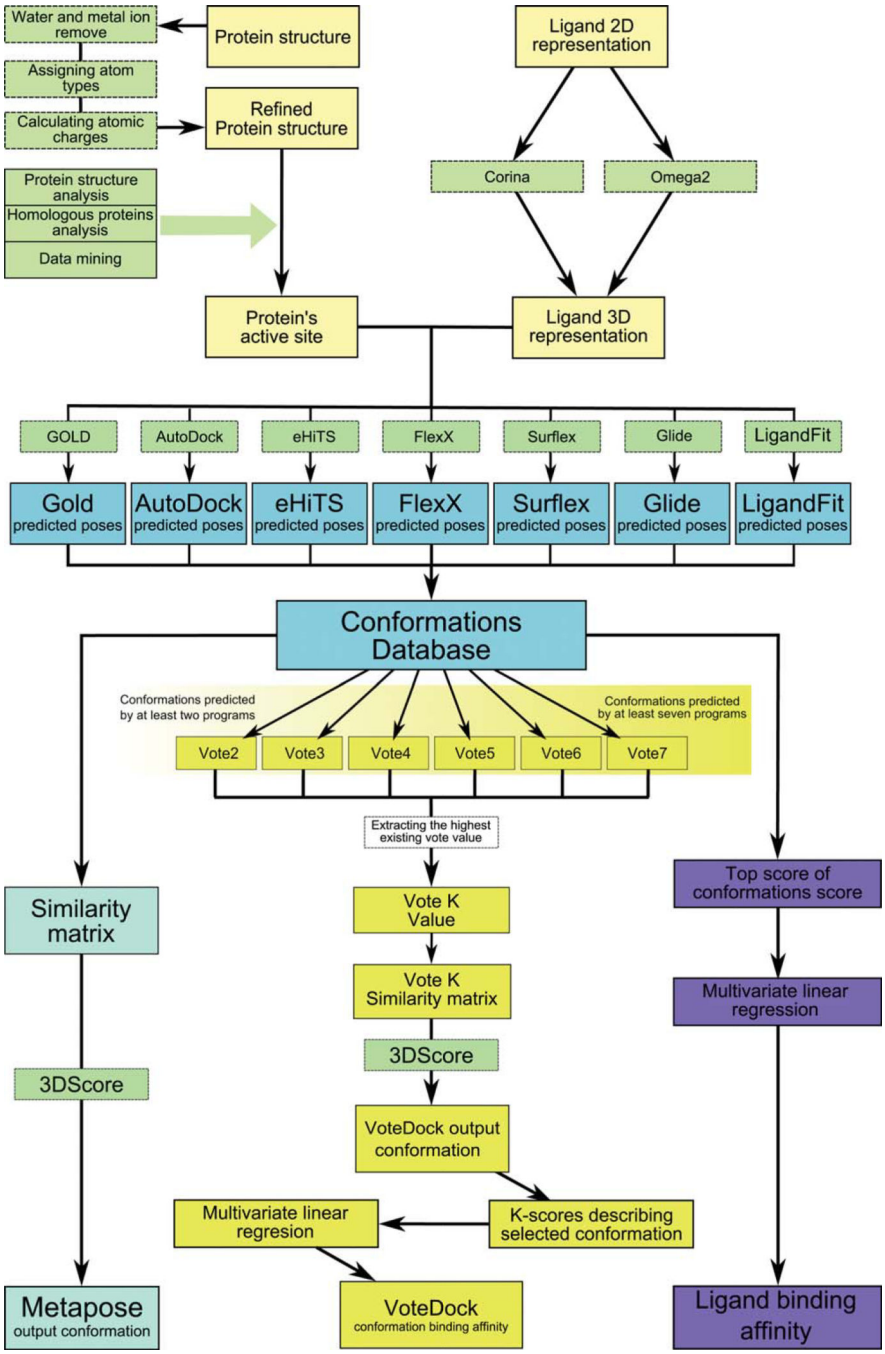

Figure 1.

The VoteDock protein–ligand docking algorithm. The main goal of the VoteDock is to provide fast and accurate prediction method for 3D structure of a protein–ligand complex. It facilitates data exchange between various prediction docking methods, publicly available software, evaluation programs and visualization modules. The general model of the information flow and components of the algorithm are presented in the following diagram.

Table 1

Evaluation of Three Consensus Docking Protocols, Namely MetaPose, MetaScore, and VoteDock, Built Using the Results of the Diverse Set of Docking Programs.

		Correlation		Entire set		Small		Large		Hydrophilic		Hydrophobic		Proteins	
		Pearson	**Spearman**	**RMSD**	**% Pairs**	**RMSD**	**% Pairs**	**RMSD**	**% Pairs**	**RMSD**	**% Pairs**	**RMSD**	**% Pairs**	**RMSD**	**% Pairs**
Top score															
Best docking	Program	eHits	eHits	GOLD	GOLD	GOLD	GOLD	GOLD	GOLD	GOLD	GOLD	eHits	eHits	GOLD	GOLD
program	Results	0.38	0.29	2.68	58.45	1.96	67.11	3.50	48.45	2.30	65.67	2.91	48.33	4.03	46.50
Second best	Program	Surflex	Surflex	eHits	eHits	eHits	eHits	eHits	eHits	eHits	eHits	GOLD	GOLD	AutoDock	AutoDock
program	Results	0.33	0.22	2.76	54.02	1.96	64.80	3.59	37.71	2.62	50.96	3.06	50.28	4.06	46.86
Averaged results*	Results	0.23	0.16	3.63	45.13	2.50	56.69	4.53	34.84	3.26	49.19	4.02	42.10	5.45	36.80
MetaPose	Results	0.42	0.43	2.57	62.67	1.75	70.60	3.11	52.25	2.08	67.31	2.63	57.90	3.59	50.00
MetaScore	Results	0.48	0.47	3.43	52.70	2.17	62.47	3.70	42.81	2.67	56.65	3.19	48.46	4.36	28.78
VoteDock	Results	0.49	0.50	2.20	68.67	1.58	76.18	2.82	59.04	1.90	72.49	2.28	62.70	3.26	57.00
Best pose															
Best docking	Program	eHits	eHits	GOLD	GOLD	eHits	eHits	GOLD	GOLD	GOLD	GOLD	eHits	eHits	GOLD	GOLD
program	Results	0.29	0.38	1.66	73.83	1.20	83.30	2.01	66.38	1.43	79.05	1.74	72.42	2.63	61.77
Second best	Program	Surflex	Surflex	eHits	eHits	GOLD	GOLD	Surflex	Surflex	eHits	eHits	Surflex	Surflex	LigandFit	LigandFit
program	Results	0.22	0.30	1.69	73.80	1.31	81.07	2.42	64.82	1.64	72.30	2.04	69.04	2.71	49.87
Averaged results*	Results	0.16	0.19	2.22	65.63	1.54	76.92	2.99	53.54	2.02	68.46	2.52	62.44	3.64	48.60
MetaPose	Results	0.30	0.34	1.54	82.53	0.97	87.83	1.86	74.62	1.29	84.64	1.59	77.89	1.86	74.62
MetaScore	Results	0.39	0.44	1.29	82.12	0.93	89.91	1.65	72.35	1.17	84.52	1.41	77.73	1.94	65.84
VoteDock	Results	0.40	0.45	1.64	78.91	1.35	83.75	2.09	70.19	1.46	80.01	1.83	74.05	2.34	65.17

The posing and scoring abilities of three consensus docking protocols are compared with the best docking programs, the second best programs, and the averaged results over seven tested docking programs that were used in benchmarks. Typically, the individual docking programs have problems with proper scoring (the best correlation is 0.38 for eHiTS). The reported problem was also observed for best poses, which was quite a surprise as for those conformations contacts between ligand and protein are recreated similar to those present in crystallographic structure. It seems that today present scoring functions lack some crucial components responsible for in vitro activity of ligands. However, they are able to identify the correct pose of the ligand with reasonable accuracy of more than 50%. Our consensus methods that are using the results from those programs seem to work better on the entire benchmarking dataset. VoteDock designed for both the ligand pose and the binding affinity prediction achieves a relatively high correlation with the experimental value close to 0.5. The pose prediction reaches almost 70% on the entire set, which is almost 10% higher than the best docking program. Our method overcomes the drawbacks of individual docking software, and better predicts the results regardless of the type of ligand (small/large, hydrophobic/hydrophilic, and peptides), or the protein family. The presented results are averaged over the large benchmark dataset of 1300 protein-ligand complexes, and moreover, four types of preprocessing algorithms used to prepare the 3D structure of a ligand, namely Corina one, Corina ten, Omega one, and Omega ten. The results are presented for both top score conformation (the pose with the highest result from the scoring function) and best pose (the pose that is closest to the corresponding native ligand among all proposed by the program). Furthermore, we present a more detailed analysis of docking programs and consensus algorithms on data-sets sharing similar physicochemical properties of ligand molecule.

* Presented numbers are arithmetic mean value calculated based on the results of seven docking programs.

Table 2

Detailed Analysis of Docking Programs and The Consensus Algorithms for Subsets of the Entire Database, Selected Based on a Ligand Binding Strength to its Protein Receptor.

		Strong				Medium				Weak			
		Small		Large		Small		Large		Small		Large	
		RMSD	% Pairs	RMSD	% Pairs	RMSD	% Pairs	RMSD	% Pairs	RMSD	% Pairs	RMSD	% Pairs
Top score													
Best docking program	Name	eHits	eHits	GOLD	GOLD	GOLD	GOLD	GOLD	GOLD	eHits	eHits	GOLD	GOLD
	Results	2.00	66.40	3.80	45.00	1.70	73.60	3.60	48.30	1.90	68.60	2.80	56.60
Second best program	Name	GOLD	GOLD	eHits	eHits	eHits	eHits	eHits	eHits	GOLD	GOLD	eHits	eHits
	Results	2.20	61.90	3.60	44.60	2.00	62.90	3.70	41.70	2.00	66.40	3.50	40.70
Averaged results*	Results	2.34	62.46	5.11	31.51	2.23	60.80	4.80	32.81	2.44	58.61	4.29	35.50
MetaPose	Results	1.84	68.30	3.54	46.39	1.74	73.84	3.00	50.54	1.72	71.11	2.81	51.50
MetaScore	Results	2.32	62.44	3.81	42.47	2.14	59.86	3.74	43.78	2.10	64.47	3.47	41.42
VoteDock	Results	1.57	74.03	3.08	57.85	1.53	76.42	2.66	61.92	1.57	77.13	2.61	57.03
Best pose													
Best docking program	Program	eHits	eHits	GOLD	GOLD	GOLD	GOLD	GOLD	GOLD	GOLD	GOLD	GOLD	GOLD
	Results	1.21	88.32	3.80	45.00	1.10	85.35	2.11	65.95	1.38	80.28	1.71	72.53
Second best program	Program	GOLD	GOLD	eHits	eHits	eHits	eHits	eHits	eHits	eHits	eHits	eHits	eHits
	Results	1.30	85.00	2.12	64.23	1.20	85.60	2.33	59.98	1.38	80.28	2.13	63.30
Averaged results*	Results	1.44	81.78	3.22	50.15	1.33	80.93	3.07	51.03	1.55	78.13	2.53	56.24
MetaPose	Results	1.36	85.39	1.91	72.27	0.82	90.37	2.13	73.44	0.99	87.18	1.85	70.14
MetaScore	Results	0.96	90.48	1.65	71.44	0.88	90.29	1.75	70.31	0.95	89.26	1.47	76.50
VoteDock	Results	1.21	85.86	2.26	70.01	1.18	83.07	1.94	73.44	1.21	82.26	2.01	66.32

Strong ligands are defined as those for which the concentration to inhibit the receptor is lower than 45 nM, medium ligands are between 45 nM and 36 M, and weak ones have more than 36 M. We divided each of those three datasets into two subsets based on ligand size, as the number of small and large ligands varies between individual subsets. Both top score conformation and best pose are evaluated. The consensus algorithms VoteDock and MetaPose achieve a higher rate of success than any individual docking program. Those metaalgorithms are also less sensitive to transition from the set of strong to medium and weak ligands, and they predict correct poses of strongly binding ligands with almost the same accuracy as the weakly binding ones. For example, the drop in docking accuracy between strong&small, medium&small, and weak&small ligands for VoteDock is less than 3%, whereas the best docking program has the difference of about 7%, and for MetaPose it is 5%. As can be seen below, docking programs have a problem in predicting poses for large molecules regardless of the binding strength of the ligand. A similar discrepancy is particularly well observed for strong and medium sets where the difference in the successfully docked pairs is more than 20% for the strong set of ligands and 25% for the medium set of ligands, whereas for weakly bound ligands, there is only a 12% difference between large and small molecules. In the case of consensus algorithms that gap still exists, but it is smaller than for any individual docking program. VoteDock has 17, 15, and 20%, respectively, for strong, medium, and weak sets of ligands. MetaPose has respectively 22, 23, and 20%.

* Presented numbers are arithmetic mean value calculated based on the results of seven docking programs.

Table 3

Quality of Predictions for VoteDock Consensus Docking Algorithm When Different Threshold Values are Chosen.

	Threshold 1 Å			Threshold 1.5 Å			Threshold 2 Å			Threshold 2.5 Å			Threshold 3 Å		
	RMSD	**% Pairs**	**Quantity**	**RMSD**	**% Pairs**	**Quantity**	**RMSD**	**% Pairs**	**Quantity**	**RMSD**	**% Pairs**	**Quantity**	**RMSD**	**% Pairs**	**Quantity**
VoteDock	2.48	61.49	1300	2.37	64.05	1300	2.20	69.67	1300	2.47	60.65	1300	2.48	59.35	1300
Vote2	1.74	73.00	979	2.04	69.08	1142	2.17	66.57	1204	2.31	63.92	1234	2.37	63.00	1249
Vote3	1.42	80.36	728	1.66	75.25	976	1.95	70.44	1090	1.98	71.80	1149	2.03	69.20	1188
Vote4	1.15	86.92	491	1.46	81.08	709	1.70	75.84	883	1.85	71.29	974	1.99	68.29	1037
Vote5	0.95	91.43	268	1.22	85.49	458	1.48	80.27	583	1.67	74.43	676	1.81	71.15	743
Vote6	0.81	95.23	101	1.10	89.79	191	1.34	83.74	248	1.42	79.10	314	1.56	74.19	363
Vote7	0.60	97.60	43	0.92	95.28	106	1.12	87.30	148	1.25	81.5	184	1.42	76.00	222

The threshold is used to determine, which poses from different docking programs are considered to be similar. The threshold value equal to 2 Å proves to be most efficient. In addition, we evaluate each vote order result, such as the docking accuracy and the number of pairs that pass each vote order when a different threshold value is applied.