J Mol Model (2009) 15:329–341
DOI 10.1007/s00894-008-0390-0

ORIGINAL PAPER

How sophisticated should a scoring function be to ensure successful docking, scoring and virtual screening?

Dmitry Tarasov · Dmitry Tovbin

Received: 28 August 2008 / Accepted: 2 October 2008 / Published online: 9 December 2008

Abstract To estimate how sophisticated should an empirical scoring function be to ensure successful *docking, scoring* and *virtual screening* a new *scoring function NScore* (naive score) has been developed and tested. NScore is an extremely simple function and has the minimum possible number of parameters; nevertheless, it allows all the main effects determining the *ligand–protein interaction* to be taken into account. The fundamental difference of NScore from the currently used empirical functions is that all its parameters are selected on the basis of general physical considerations, without any adjustment or training with the use of experimental data on ligand–protein interaction. The results of docking and scoring with the use of NScore in an independent test sets of proteins and ligands have proved to be as good as those yielded by the ICM, GOLD, and Glide software packages, which use sophisticated empirical scoring functions. With respect to some parameters, the results of docking with the use of NScore are even better than those obtained using other functions. Since no training set is used in the development of NScore, this scoring function is indeed versatile in that it does not depend on the specific goal or target. We have performed virtual screening for ten targets and obtained results almost as good as those yielded by the Glide and better than GOLD and DOCK software.

Keywords Docking · Interaction · Ligand-protein · New scoring function NScore · Scoring · Virtual screening

D. Tarasov (✉) · D. Tovbin
DIDIALL, INC.,
382 Central Park W, #5K,
New York, NY 10025, USA
e-mail: tarasov@didiall.com
URL: www.didiall.com

Introduction

The docking, scoring, and virtual screening of large libraries of chemical compounds are a widely used approach to lead discovery in the pharmaceutical industry when a high-resolution structure of the biological target of interest is available [1–4]. These numerical methods are based on the prediction of the binding affinity of ligand–protein interaction. For practical uses, the methods for predicting the binding affinity should be not only accurate, but also sufficiently quick; therefore, programs that use fast scoring functions (GOLD [5], FlexX [6], Glide [7, 8], ICM [9], Fred [10], AutoDock [11], DOCK [12]), are the most widely employed for the prediction of ligand-protein interaction.

Practically all quick scoring functions, including knowledge based ones, included in currently used software are empirical. A common approach to the development of empirical scoring functions is the use of a physical model, some parameters of which are adjusted with the use of different methods and different training sets of experimental data.

Although currently used empirical scoring functions achieve increasing reliability in affinity prediction for some biological targets; for many others, predictions remain rather unsatisfactory. Numerous independent studies have demonstrated that the results of docking, scoring, and virtual screening by means of empirical functions in sets different from their training sets may be considerably worse than the results obtained in the course of training [13–17]. The main causes of errors occurring when empirical scoring functions are used are not always obvious. They may be related to the physical models, the experimental data used for training, or the training technique.

It is always possible to begin the development of an empirical scoring function with a function whose parameters have been selected on the basis of general

considerations and have not been trained. Such a function is often used as a starting scoring function for training. In the course of training, the performance of the scoring function on the training data set is improved compared to the untrained scoring function as estimated by the trained parameters. However, if in the training there are problems (e.g., overfitting, noise in the training set and so on), the performance of the scoring function on the test set may be worse than that of the original, untrained scoring function. In addition, the performance of the trained scoring function with respect to untrained but still important parameters may become worse than the performance of the untrained function with respect to these parameters.

In our opinion, it is of utmost importance to know how much better the performance of currently used scoring functions is than that of entirely untrained scoring functions. An answer to this question will allow us

(1) to estimate the dependence of empirical scoring functions on the physical models, training sets, and the training itself;
(2) to determine the main cause of errors in currently used empirical scoring functions;
(3) to assess how versatile empirical scoring functions are; and
(4) to understand how, and to what degree, the performance of empirical scoring functions can be improved.

For this purpose, we developed NScore (naive score), a very simple scoring function with as few parameters as possible that nevertheless takes into account all the main parameters determining the ligand–protein interaction. The fundamental difference of NScore from the currently used empirical scoring functions is that all its parameters are selected on the basis of general physical considerations, without any adjustment or training with the use of experimental data on ligand–protein interaction.

The development of the scoring function NScore is described in detail below. We have tested this function for docking, scoring, and virtual screening in independent sets of protein–ligand complexes. We have compared the results of NScore working with the results obtained by scoring functions in software Glide, GOLD, ICM, DOCK.

Methods

Scoring function

Our goal was to develop the simplest possible scoring function that would nevertheless be suitable for docking, scoring, and virtual screening.

A scoring function in which the ligand score is proportional to the number of heavy atoms in the ligand is the simplest. Surprisingly, such a scoring function is often closely correlated with the free energy of the ligand–protein interaction for some test sets of active ligands. However, such a function is inapplicable to docking or virtual screening, because its value is independent of the ligand pose in the active site.

To develop NScore, we used the form that is the simplest and most convenient for calculations, namely, the so-called atom–atom approximation, where the scoring function is represented as

$$S = \sum_{i,j} S_{PL}(r_{i,j}) + S_{int} + S_S, \tag{1}$$

where i and j are the ordinal numbers of atoms in the protein and ligand, $r_{i,j}$ is the distance between the protein and ligand atoms, $S_{PL}(r_{i,j})$ is a function depending on the types of atoms in the protein and ligand and the distance between them, S_{int} is internal energy of the ligand which depends only on internal ligand state, and S_S is entropy loss because of the restriction of ligand movement upon binding.

The main effects that we took into account when developing NScore were the hydrophobic effect $\sum_{i,j} S_{lipo}(r_{i,j})$, formation of hydrogen bonds $\sum_{i,j} S_{HB}(r_{i,j})$, interaction between the ligand and metal ions in the active site $\sum_{i,j} S_{ME}(r_{i,j})$, repulsion between the ligand and protein atoms $\sum_{i,j} S_{rep}(r_{i,j})$, internal energy of the ligand S_{int}, and entropy loss because of the restriction of ligand movement upon binding S_S.

$$S = \sum_{i,j} S_{lipo}(r_{i,j}) + \sum_{i,j} S_{HB}(r_{i,j}) + \sum_{i,j} S_{ME}(r_{i,j}) + \sum_{i,j} S_{rep}(r_{i,j}) + S_{int} + S_S \tag{2}$$

The interaction of the ligand and protein with surrounding water molecules is one of the main factors determining the binding of the ligand with the protein. To take this interaction into account in an explicit form is difficult and entails too much calculation. The interaction with water molecules can be taken into account implicitly in the form of the so-called hydrophobic effect, which is sufficiently accurate in many cases. The hydrophobic effect is determined by the change in the area of contact between the interacting molecules that is accessible for water molecules and is approximately –25 cal/(mol·Å^2), [18]. For simplicity, we used the atom–atom approach to take into account the hydrophobic effect when developing NScore. According to this approach, the change in the free surface was taken to be proportional to the number of contacts between atoms of the molecules approximately recalculated to the score under the assumption

that the energy of the formation of one optimal hydrophobic contact was –0.1 kcal/mol (the first term in Eq. (1)).

Another important factor determining the ligand–protein interaction is the formation of a hydrogen bond between ligand and protein atoms. In NScore, all hydrogen bonds except the bond between the atoms contained in charged groups (e.g., amino and carboxyl groups) were assumed to be equivalent. A hydrogen bond was treated as an attraction between the hydrogen and the acceptor. We also ignored the direction of hydrogen bonds in an explicit form versus common practice, e.g., as it made in program FlexX [6]. A hydrogen bond has a more negative energy than a hydrophobic bond; therefore, on the basis of estimations reported in [19], we estimated the score for the formation of one optimal hydrogen bond at –1.5 kcal/mol.

When describing hydrophobic and hydrophilic interactions, one should take special care to ensure that the ratio between the scores for hydrophobic interaction and formation of hydrogen bonds be correct. Since we used the atom–atom approximation (form (1)) for the scoring function, we could calculate the score for each ligand atom that was in a certain pose in the protein binding site if we neglected the ligand internal energy and entropy loss related to the limited mobility of the ligand, whose contributions to the score are usually smaller than those of other effects. Having calculated the score with the use of the NScore scoring function for each atom in the native pose of the ligands for the set of some protein-ligand complexes, we found that average scores per hydrophobic atom and per atom capable of forming a hydrogen bond were –1.0 and –0.7 kcal/mol, respectively. The ratio between these values is physically unreasonable, because a hydrogen bond has a more negative energy than a hydrophobic contact. Therefore, we corrected the starting NScore scoring function by decreasing the score per contact between hydrophobic atoms at the optimal pose by a factor of 2, i.e., to –0.05 kcal/mol.

If a protein contains a metal ion, a bond about as strong as a hydrogen bond is formed between the metal and ligand atoms. This bond is characterized by shorter interaction distances compared to a hydrogen bond. We assumed the score for the formation of one optimal bond between the ligand and the metal ion in the active center to be –1.5 kcal/mol (i.e., equal to that for an optimal hydrogen bond); as in the case of a hydrogen bond, we ignored the direction of the bond with the metal ion in an explicit form.

Hydrogen bonds are known to be strictly oriented; however, for simplicity's sake, we did not take their orientation into account in an explicit form when constructing the model. Surprisingly, hydrogen bonds formed between the ligand and the protein as a result of docking with the use of the NScore scoring function still proved to be rather strictly oriented, mainly because the atoms adjacent to those forming the bonds were involved in the interaction.

At short distances, there is noticeable repulsion between atoms of the ligand and protein molecules. To describe this repulsion, we used the Lennard–Jones 6–12 potential for $r<r_1$:

$$V_{LJ}(r) = \varepsilon\left(\frac{r_1}{r}^{12} - 2\frac{r_1}{r}^{6}\right). \quad (3)$$

We assumed the parameters ε and r_1, to be $\varepsilon_{l\text{-}l}$=0.06 kcal/mol and $r_{l\text{-}l}$=4.1 Å for describing the repulsion of hydrophobic atoms and $\varepsilon_{H\text{-}H}$=0.6 kcal/mol and $r_{H\text{-}H}$=1.8 Å for describing the repulsion of atoms capable of forming hydrogen bonds. The values of ε and r_1 were chosen almost arbitrarily, except that we bore in mind that the distance between hydrophobic atoms of carbon in an aqueous solution is ~4.0–4.2 Å and the distance between heavy atoms forming a hydrogen bond is ~2.8 Å. The values of $\varepsilon_{l\text{-}l}$ for the interaction of hydrophobic atoms were assumed to be half as high as those corresponding to the repulsing component of the energy of interaction between the aliphatic carbon atoms in the force fields AMBER [20], $\varepsilon_{l\text{-}l}$= 0.12 kcal/mol. The $\varepsilon_{l\text{-}l}$ values were twofold decreased so that the contacts between the protein and ligand atoms would be tighter, which allowed us to implicitly take into account the local mobility of protein atoms induced by the protein–ligand interaction.

To avoid steric clashes of the ligand when searching for its optimal pose, we introduced into NScore the internal energy of the ligand in the form

$$S_{\text{int}} = \sum_{i,j} f_{\text{int}}\left(r_{i,j}\right), \quad (4)$$

where i and j are the ordinal numbers of atoms in the ligand, $r_{i,j}$ is the distance between the atoms, and

$$f_{\text{int}}(r) = \begin{cases} k(r - r_{\text{int}})^2, r < r_{\text{int}} \\ 0, r \geq r_{\text{int}} \end{cases}. \quad (5)$$

with the repulsion being calculated only for nonhydrogen atoms separated by three or more covalent bonds. The values k=20 kcal/(mol·Å^2) and r_{int}=2.5 Å were taken arbitrarily on the basis of general physical considerations, with the r_{int} being somewhat underestimated intentionally lest the estimate of internal overlap clashes interfere with the search for the optimal ligand pose in the site.

When a ligand molecule is bound with the protein, both the mobility of the ligand molecule and its internal degrees of rotation are restricted, which leads to entropy loss. We took this loss into account in NScore in the form $\boldsymbol{S}_S$, where

$$S_S = k_{rot} \cdot N_{rot} + S_0, \quad (6)$$

where N_{rot} is the number of covalent bonds in the ligand that are capable of rotation; k_{rot}=0.33 kcal/mol is the entropy loss resulting from the restriction of rotation of one bond, which is equal to 0.5 kT (the energy per degree of freedom; a more detailed theoretical estimation yields practically the same

entropy loss [21, 22]), and S_0 is the entropy loss accounted for by the restriction of movement of the ligand as a whole. This entropy loss only slightly depends on the ligand size. Since only the relative score is important for selecting the best ligands or the best pose of one ligand, scoring functions are usually determined to an accuracy of a constant; therefore, we assumed S_0=0 kcal/mol.

Electrostatic interactions are an important part of the interactions between ligands and proteins. Electrostatic interactions are not always clearly distinguishable from other interactions; e.g., hydrogen bonds are sometimes considered entirely in electrostatic terms. We have not explicitly included electrostatic interactions into NScore for the following reasons: (1) electrostatic interactions are already partly included in an implicit form as hydrogen bonds and the bonds between ligand and metal ions in the active site; (2) if electrostatic interaction is explicitly taken into account, the predicted distribution of charges among ligand and protein atoms becomes complicated and ambiguous; and (3) attempts at taking into account the dielectric permeability and the environment effect on it also result in complexities and ambiguity. We additionally took into account electrostatic interaction in the form of a stronger hydrogen bond between atoms belonging to oppositely charged groups with an additional contribution to the score of –1.5 kcal/mol.

When developing NScore, we did not take into consideration many physical processes, such as polarization and stacking interaction. We believe that, in general, they do not affect substantially ligand–protein interactions, although they may even entirely determine the interaction in some special cases.

In accord with the physical models forming the basis of the new scoring function that were described above, we suggested the following classification of the types of ligand and protein atoms:

- hydrophobic atoms (carbon atoms),
- hydrogen involved in hydrogen bonds (e.g., the hydrogen atoms of the OH and NH_2 groups),
- hydrogen acceptors (e.g., the oxygen atoms of the COOH and C=O groups),
- nitrogen atoms covalently bound with the hydrogen atoms that are involved in hydrogen bonds (e.g., the nitrogen atom in the NH_2 group), and
- metal ions in the active site.

The scoring function describing repulsion and attraction between atoms of the ligand and protein of different types had the following general form:

$$G(r)=\begin{cases} e+\varepsilon\left(\left(\frac{r_1}{r}\right)^{12}-2\left(\frac{r_1}{r}\right)^{6}+1\right), r<r_1 \\ f(r), r_2\geq r\geq r_1 \\ 0, r>r_2 \end{cases}. \quad (7)$$

where $f(r)=ar^3+br^2+cr+d$ and $f(r_1)=e$, $f(r_2)=0$, $f'(r_1)=0$, $f'(r_2)=0$. The parameters e, r_1, and r_2 for each pair of types A and B were selected according to the physical models described above, for example, e may be e_{lipo}, e_{HB}, e_{ME} and r_1 may be $r_{l\text{-}l}$, $r_{l\text{-}h}$, $r_{h\text{-}h}$, r_{ME} and r_2 may be $r_{2\_l\text{-}l}$, $r_{2\_l\text{-}h}$, $r_{2\_h\text{-}h}$, r_{2_ME}. In this form, the scoring function is continuous and continuously differentiable for any $r>0$. Table 1 shows the main parameters of the NScore scoring function.

NScore developed in this study is functionally similar to many scoring functions that are currently used, e.g., ChemScore [23], which is not unexpected, because the main physical effects determining ligand–protein interaction are well known and should be taken into consideration in any scoring function. The fundamental difference of NScore from all the empirical and knowledge based scoring functions is that all its parameters are selected on the basis of general physical considerations alone, without adjustment to any experimental data on ligand–protein interaction. We did not try to select these parameters too accurately. They were chosen almost arbitrarily, the only requirement being that they should be of the same order of magnitude as actual physical effects.

The algorithm of docking

We developed an algorithm of scaling search for the ligand pose with the lowest score and used it for docking.

1. We determined several thousands of active points in a protein binding site, where, in principle, an atom of the ligand may be located.
2. A ligand was placed into the binding site randomly except that one of the atoms of the ligand was in an active point. Local score minimization of the ligand was performed, and the score at the minimized pose was calculated. This procedure was repeated several thousands of times.
3. The ligand minimized poses that were close to one another were grouped into clusters, and one pose with the best score was selected in each cluster.
4. For each ligand pose selected in the clusters at the previous step, the pose in the binding site was randomly changed on the distance less than the distance a. For each new pose, we performed local minimization and calculated the score at the minimized pose. This procedure was repeated several tens of times.
5. Steps 3 and 4 were repeated several times, the parameter a varying according to the power law $a_{n+1}=a_n^{0.5}$.

To accelerate the calculations, we calculated the score by means of a grid (the potentials for an atom of the ligand of every type in the protein binding site calculated previously).

Table 1 The main parameters of the NScore scoring function

Parameter	Value	Comment
e_{lipo}	–0.05 kcal/mol	formation of a hydrophobic contact
e_{HB}	–1.5 kcal/mol	hydrogen bond
e_{ME}	–1.5 kcal/mol	bond with a metal ion
$r_{l\text{-}l}$	4.1 Å	optimal distance for a hydrophobic contact
$r_{l\text{-}h}$	3.6 Å	optimal distance for a contact between hydrophobic and hydrophilic atoms
$r_{h\text{-}h}$	1.8 Å	optimal distance for the formation of a hydrogen bond
r_{ME}	2.0 Å	optimal distance for the formation of a bond between a ligand atom and a metal ion in the active site
$r_{2\_l\text{-}l}$	6.5 Å	maximum distance for a hydrophobic interaction
$r_{2\_l\text{-}h}$	5.5 Å	maximum distance for a hydrophobic and hydrophilic interaction
$r_{2\_h\text{-}h}$	3.5 Å	maximum distance for a hydrogen bond interaction
r_{2_ME}	3.5 Å	maximum distance for a metal ion and a ligand atom interaction
$\varepsilon_{l\text{-}l}$	0.06 kcal/mol	coefficient in the repulsing part of the Lennard–Jones potential for a contact between hydrophobic atoms
$\varepsilon_{l\text{-}h}$	0.3 kcal/mol	coefficient in the repulsing part of the Lennard–Jones potential for a contact between hydrophobic and hydrophilic atoms
$\varepsilon_{h\text{-}h}$	0.6 kcal/mol	coefficient in the repulsing part of the Lennard–Jones potential for a contact between atoms forming a hydrogen bond
r_{int}	2.5 Å	distance between atoms within the ligand in Eq. (5)
k_{int}	20 kcal/(mol·Å^2)	coefficient of internal repulsion in Eq. (5)
k_{rot}	0.33 kcal/mol	entropy loss accounted for by the rotating bond in Eq. (6)

Docking was performed into the protein whose structure did not change during the docking procedure.

The docking algorithm was tested as follows. We used known three-dimensional structures of ligands in protein binding sites. The ligand was removed, and docking into the binding site of the removed ligand was performed for each complex. The docking algorithm was considered to be correct if the ligand pose determined by docking had the minimum score, and all other poses of this ligand had higher scores. Indeed, the score of the best pose of the ligand determined by docking was no higher than at the native pose in 98% of test dockings, which is indirect evidence that the algorithm for the search of the best pose of the ligand was correct in most cases, the errors being mainly accounted for by problems with docking target functions.

In the course of docking, all atoms of the protein, including all hydrogens, were assumed to be rigid, and rotation within the ligand molecules occurred about single bonds not involved in cyclic structures. To take into consideration different conformations of cyclic structures, we used the CORINA software [24] (docking of several conformations was performed, and the results were merged).

The test set for scoring and docking

For testing the NScore scoring function in the course of docking and scoring, we used a set of 100 proteins and ligands with known native poses taken from the Protein Data Bank [25] in the study [15] – Vertex test set and an over diverse, high-quality test set of 85 proteins and ligands from Protein Data Bank – Astex test set [26]. We used these sets of data because (1) they are independent, (2) the selection of complexes for these sets had been substantiated and explained, and (3) the well-known ICM, GOLD, and Glide software packages had been independently tested using the Vertex test set of data and GOLD had been tested using the Astex test set.

Complex preparation

Each of the complexes in the test set was prepared for docking and scoring as follows. The ligand was removed, and bonds were manually set in the correct order; protonated states of the ligand were generated automatically; initial three-dimensional ligand structures for docking were taken from native structure or automatically generated by means of the CORINA software, with the stereoisomeric form of the native ligand taken into account. All water molecules and cofactors were removed from the protein that was left after the removal of the ligand; if metal ions were involved in the ligand–protein binding, they remained in the binding site and were taken into consideration when docking was performed. We ignored water molecules in the binding site when performing docking and scoring. Hydrogen atoms were automatically added to the protein by means of the REDUCE software [27]. We did not use any visual corrections of protonation, tautomeric forms, or local minimization of hydrogens or heavy atoms for either ligands or proteins, because this correction is not always self-evident and makes the results of docking and scoring somewhat subjective.

For a more correct comparison of the results of docking with the use of NScore and other programs, we adjusted the conditions of each docking so that they were as close as possible to the docking conditions in the studies [15] and [26]. Therefore, if not indicated otherwise, the active site for the Vertex test set was determined by the native pose of the ligand in the form of a 18×18×18 Å box with a center coinciding with the center of the ligand; for the Astex test set, the active site was determined by the native pose of the ligand in the form of a box with the distance 6 Å from the ligand to a side of the box.

Preparation of targets and ligands for virtual screening

For correct comparison of the results of virtual screening with the use of NScore and other scoring functions, we selected as targets 10 proteins from the study [28], where they were used for comparing the results of virtual screening by means of the GOLD, Glide, and DOCK programs. These 10 targets were among the 15 targets used for training in the Glide software. For the remaining five targets, we could not reliably determine the structures of active ligands used for virtual screening in [28] and for training scoring functions in the Glide software [7, 8]. Inactive ligands were the same as in [7, 8, 28]. The original three-dimensional structures of active ligands and decoys for virtual screening were automatically generated using the CORINA software. The same pdb files as in [7, 8, 28] were used as 3D protein structures for all targets. We prepared the protein targets for virtual screening in the same way as for docking and scoring, without any visual correction or minimization of binding sites.

Results and discussion

Scoring

Figure 1 shows the correlation between the scores calculated using NScore for native, locally minimized poses of ligands from the Vertex test set and the experimentally estimated binding affinities. The coefficients of correlation between the experimental data and the binding affinities calculated using NScore and other scoring functions for various test sets of proteins and ligands are shown in Table 2. As evident from the table, NScore showed about the same bad correlation in the independent test set as other, sometimes much more detailed, scoring functions whose parameters were selected on the basis of experimental data on proteins and ligands. For some scoring functions, Table 2 also shows the correlation coefficients that were obtained in the course of training [29]. According to these data, the correlation was drastically decreased when going from training sets to independent ones; it becomes about the same as in the case of NScore, which was not trained at all.

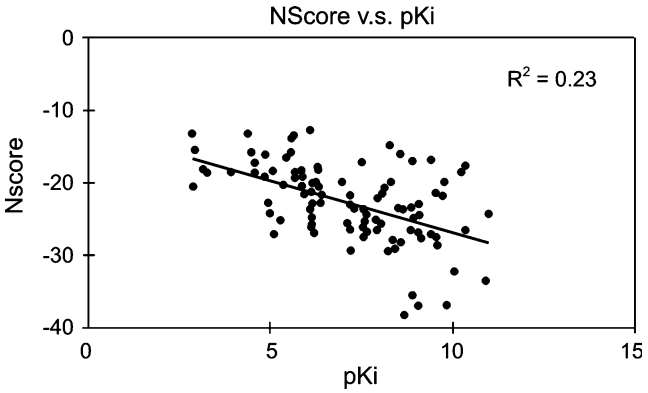

Fig. 1 NScore vs pKi

We varied all the main parameters in NScore in a wide range and performed scoring with the use of the modified functions. Energy parameters, such as the score for a hydrogen bond at the optimal pose, increased or decreased within a factor of two, and the optimal distances changed by 0.1 Å as the scoring function was modified. This modification only slightly affected the results of scoring.

Scoring results obtained by NScore and other scoring function on Vertex test set indicate that scoring functions trained for scoring predict the binding affinity with the same quality that has been obtained during training only for the protein–ligand complexes that are sufficiently similar to training complexes. If the complexes are not similar to those from the training set, the prediction quality will be considerably worse, in fact, as bad as that for an untrained scoring function.

Docking

Figure 2 shows the numbers of complexes for which the top-ranked solution obtained by docking with the use of the NScore scoring function on Vertex test set differs from the native one with respect to the root mean square deviation (RMSD) for heavy atoms less than by certain values. Figure 3 shows the numbers of complexes for which the solution that is the closest to the native pose among the top 20 solutions obtained by docking with the use of the NScore scoring function differs from the native one with respect to the RMSD less than by certain values. For all dockings, we selected the native geometry from the PDB structure as the initial ligand geometry. These figures also show the data obtained in [15] using the ICM, GOLD, and Glide software packages. As evident from these data, the results of docking by means of NScore for the first solution were almost the same as those yielded by other programs.

Table 2 Correlation coefficients between calculated and experimentally determined binding affinities for different scoring functions and different test sets of proteins and ligands

Score	R2	Test set
NScore	0.23	100 complexes; source: [15]
ChemScore	0.26	150 complexes; source: [15]
ChemScore	0.71	original test set [30]
GlideScore	0.31	150 complexes; source: [15]
PLP	0.31	150 complexes; source: [15]
PMF	0.11	150 complexes; source: [15]
PMF	0.61	original test set [31]
PMF612	0.26	150 complexes; source: [15]

The ICM, GOLD, and Glide software packages use rather detailed empirical scoring functions. The scoring functions used in GOLD were specially optimized for docking, and the authors of this software obtained successful docking in ~75% of cases when testing it. The obtained results is additional evidence that the results of docking in test sets are usually worse than the results obtained by the authors of the software when testing the scoring functions in training sets if the complexes in independent test sets are not similar to those from the training set. Moreover, the results of docking for independent sets of proteins and ligands with the use of scoring functions with parameters obtained without any adjustment or training, on the basis of general physical considerations alone, are almost the same as those yielded by more sophisticated, trained empirical scoring functions.

Many details of docking with the use of different programs may eventually have a substantial effect on the results. For example, programs differ from one another in models of active sites (this is a sphere in the GOLD software and a box in ICM and Glide) and arrangement of electrical charges. It is reasonable to consider the results

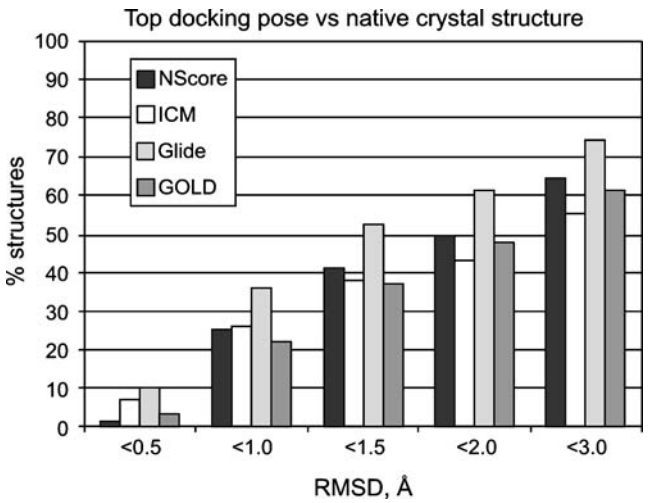

Fig. 2 Top docking pose vs native crystal structure

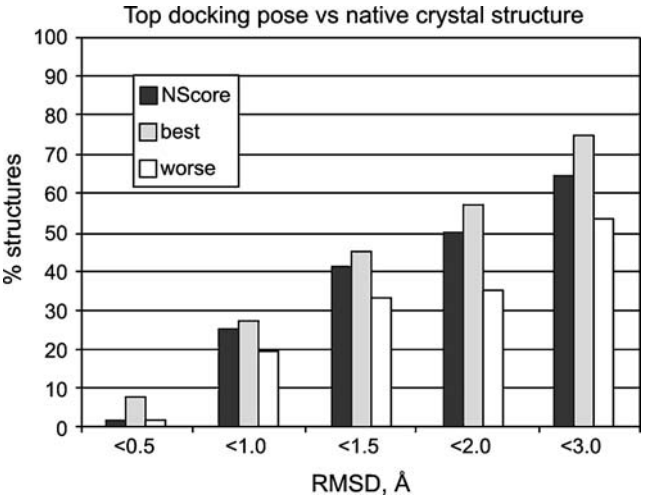

Fig. 3 Top docking pose vs native crystal structure

within a certain accuracy when comparing different programs; this also applies to the results of docking and scoring reported here. The better results in the case of the Glide software may be explained by the fact that the size of the active site, 12×12×12Å (V=1728 Å^3), used for this software in paper [15] was smaller than that used in the GOLD software, R=10 Å (V=4187 Å^3). In addition, docking in the study [15], from which the results of docking by means of GOLD, ICM, and Glide were taken, was performed in 150 complexes; 100 of them are available in the Protein Data Bank, and docking by means of NScore has been performed for these 100 complexes.

It was unlikely that we guessed the optimal docking parameters when selecting the parameters of NScore; therefore, we varied all the main parameters in a wide range and performed docking with the use of the modified functions. Energy parameters, such as the score for a hydrogen bond at the optimal pose, increased or decreased within a factor of two, and the optimal distances changed by 0.1 Å as the scoring function was modified. This only slightly affected the results of docking and scoring; Figs. 4 and 5 show the results of docking for the best and the worst cases. As can be seen in the figures, the initial parameters of NScore were not optimal for docking in the Vertex test set used, which is not surprising, because the parameters in NScore were selected without adjustment to experimental data. The resistance of the results of docking to varying the parameters within a reasonably wide range is a useful property of the NScore scoring function, because this allows NScore to be used for various classes of proteins and ligands with about the same expected quality of docking for all targets.

Table 3 shows the results of docking in the Astex test set with the use of the NScore scoring function and docking in the same test set obtained in the study [26] using the GOLD software. Dockings using both NScore and GOLD were

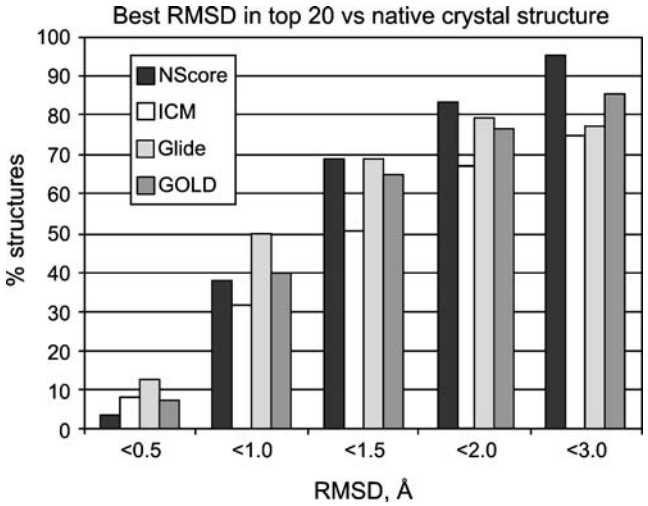

Fig. 4 Best RMDS in top 20 vs native crystal structure

performed under different conditions: the size of the active site was varied, either the native geometry or the geometry obtained by means of the CORINA software was used as the initial geometry of the ligand, and water molecules contained in the active site were taken into account in an explicit form. As can be seen in the table, the results obtained using NScore are only slightly (on average, less than 10%) worse than those obtained using GOLD.

Table 3 also shows the best and the worst results of docking obtained upon varying the parameters of the NScore scoring function, the energy parameters being decreased or increased within a factor of two and the optimal distance being changed by 0.1Å. We performed docking with modified parameters into an active site with the sizes exceeding those of the native ligand by 6 Å in every direction. The initial ligand geometry was generated using the CORINA software; the docking did not take water molecules into account in an explicit form. After

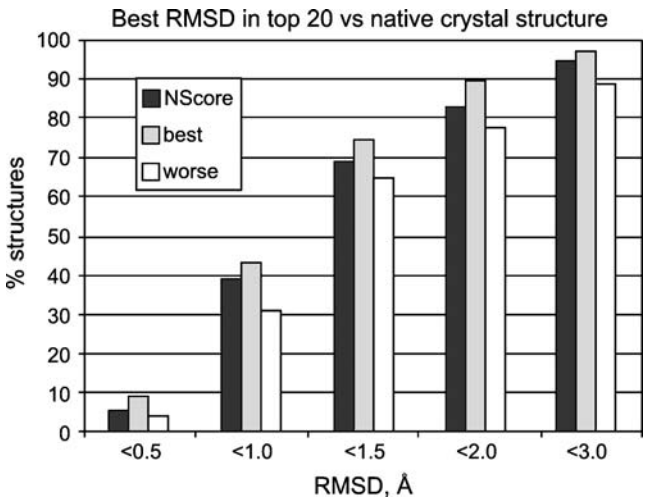

Fig. 5 Best RMDS in top 20 vs native crystal structure

Table 3 Docking performance on astex test set

	NScore	GOLD
4 Å frame in binding site	75	86.5
6 Å frame in binding site	73	80.5
10 Å frame in binding site	73	80.4
X-ray waters present	96	98.6
Corina ligand geometry, 6 Å frame in binding site	72	75.2
Best results for modified NScore	79	
Worse results for modified NScore	55	

certain modifications of the parameters (doubling the score of hydrophobic interaction), NScore yielded even better results than the GOLD software did.

We visually examined the cases of docking, in both the Vetex and Astex test sets, where the solution with the best score was not close to the native one. As a result, we found that docking errors may have occurred for one of the following reasons: (1) the ligand formed one or several hydrogen bond(s) with water molecules, which, in turn, formed one or several hydrogen bond(s) with the protein; (2) the orientation of hydrogens or their spatial arrangement in the protein did not correspond to the case where the protein bound the ligand; (3) the ligand–protein interaction was accompanied by protonation of some groups, e.g., the carboxyl group, which were not taken into account correctly during docking; (4) the ligand conformation at which the binding occurred was predicted incorrectly; (5) errors occurred in scoring function performance; or (6) there were problems with the protein structure used for docking.

We classified an error resulting from the formation of hydrogen bonds between the ligand and water molecules only if one of the solutions found in the course of docking was close to the native one, this solution becoming the best if −0.5 kcal/mol per hydrogen bond formed with a water molecule was added to the score. It was also assumed that a water molecule could form a hydrogen bond with the ligand only if this water molecule had formed one or several hydrogen bond(s) with the protein. The causes of errors in protein structures for docking are described in detail elsewhere

Table 4 Causes of docking errors in percent of the total number of errors

	Vertex test set	Astex test set
Score failures	44	30
Water interaction failures	34	30
Hydrogen direction failures	6	14
Hydrogen protonation failures	8	14
Ligand conformation failures	0	12
Protein structure problems	8	0

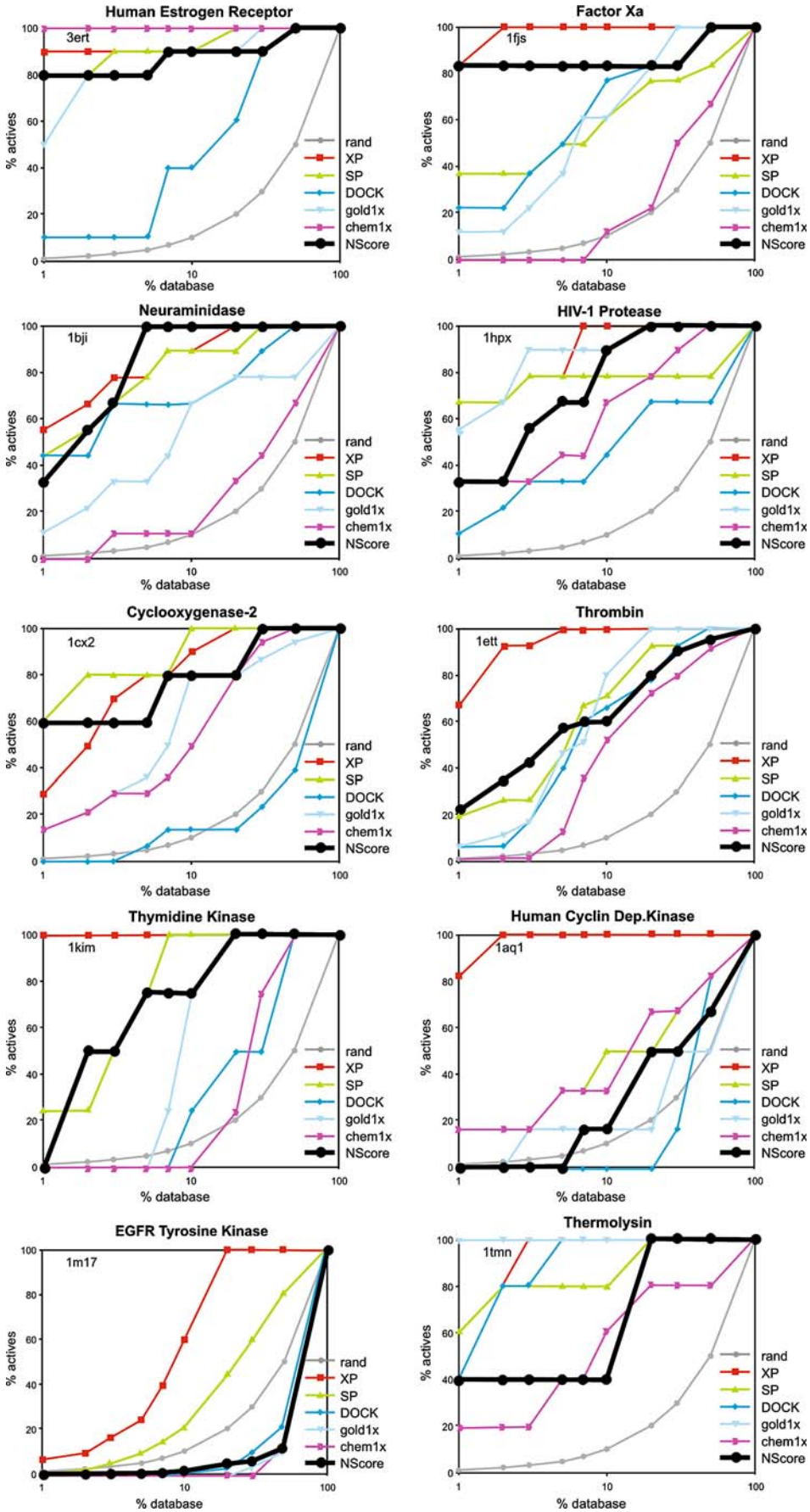

Fig. 6 Results of virtual screening

[26]. We classified errors caused by incorrect performance of the scoring function in all cases when the errors were not attributed to other causes.

Table 4 shows the distribution of errors in docking. As evident from these data, there was no single main cause of docking errors in the case of either Vertex or Astex test set. Most interestingly, problems with scoring, in the case of the Astex test set, were not the main cause of docking errors, although scoring errors were the most frequent in docking, along with the errors occurring because water was incorrectly taken into account in an explicit form for ligand–protein interaction. There were no errors in predicting the ligand conformation in the case of docking using the Vetex test set because, in this case, the native conformation taken from the pdb file was used as the initial conformation of the ligand in the course of docking. The absence of errors related to protein structure problems in the case of the Astex test set was explained by a more careful selection of complexes.

Virtual screening

Virtual screening is the docking of numerous ligands into an active site, their scoring, and the selection of ligands with the best score for further analysis. Although docking and scoring are used in virtual screening, its success determined by the scoring function used is not directly related to the quality of this scoring function in terms of docking or scoring, because a scoring function is mainly used in virtual screening to differentiate between active and inactive ligands. To estimate how applicable not training simple scoring function NScore is to virtual screening, we performed virtual screening of ten target proteins. All of these targets were the same that were used for developing and testing the GlideScore scoring functions for the Glide software. These targets cover a wide spectrum of protein classes.

Figure 6 shows the results of virtual screening. Averaged results of virtual screening are shown in Fig. 7. As can be seen in the figures, the results of virtual screening with the use of an entirely untrained, simple scoring function NScore proved to be better than those obtained by means of the GOLD and DOCK software using detailed empirical scoring functions but worse than those obtained by means of the Glide software, especially if the GlideScore XP scoring function was used. The GOLD software uses an empirical scoring function trained for both docking and scoring, and DOCK uses an extremely detailed scoring function. A characteristic feature of the Glide software is that the scoring functions used in it, GlideScore XP and GlideScore SP, were trained specially for virtual screening, the training procedure using the same complexes that we used for virtual screening.

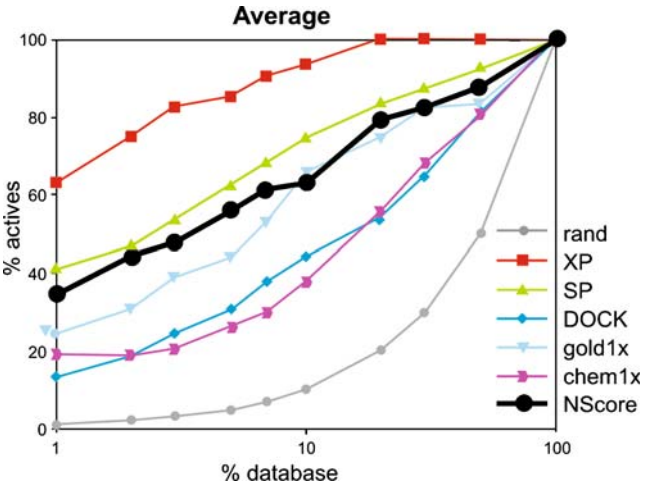

Fig. 7 Averaged results of virtual screening

Virtual screening with NScore yielded the worst results for three targets: Human Cyclin Dep. Kinase, EGFR Tyrosine Kinase, and Thermolysin. Analysis of the results showed that, in the case of Thermolysin, problems occurred because the interaction of the ligands with the Zn ion in the protein active site was incorrectly taken into account with the use of the NScore scoring function; regarding the other two targets, errors were caused by common problems of scoring using NScore.

It is known that, in the cases of Thymidine Kinase and EGFR Tyrosine Kinase, the active site contains water molecules through which the proteins may bind active ligands. When the interaction of the ligands with these water molecules was explicitly taken into account, the results were substantially improved only for Thymidine Kinase; in the case of EGFR Tyrosine Kinase, the results of virtual screening remained unsatisfactory.

Successful docking is a necessary condition for good results of virtual screening. If the most probable pose of the ligand is not found correctly, it is impossible to perform correct scoring and, hence, differentiate between active and inactive ligands. For all ten proteins, the most probable native poses of active ligands could be found, irrespective of whether it was determined from the results of X-ray analysis, or the native pose of a very similar active ligand was known. Detailed analysis of the results showed that docking for all the ten targets usually yielded ligand poses that were close to the native one, the proportion of successful dockings varying from 40 to 95%, depending on the protein structure. This permitted subsequent correct scoring and selection of these ligands among random ones.

To estimate the tolerance of the results of virtual screening to variation of parameters, we varied all the main parameters of the NScore scoring function in a wide range and performed virtual screenings again using the modified scoring functions. Figure 8 shows the results. The results of

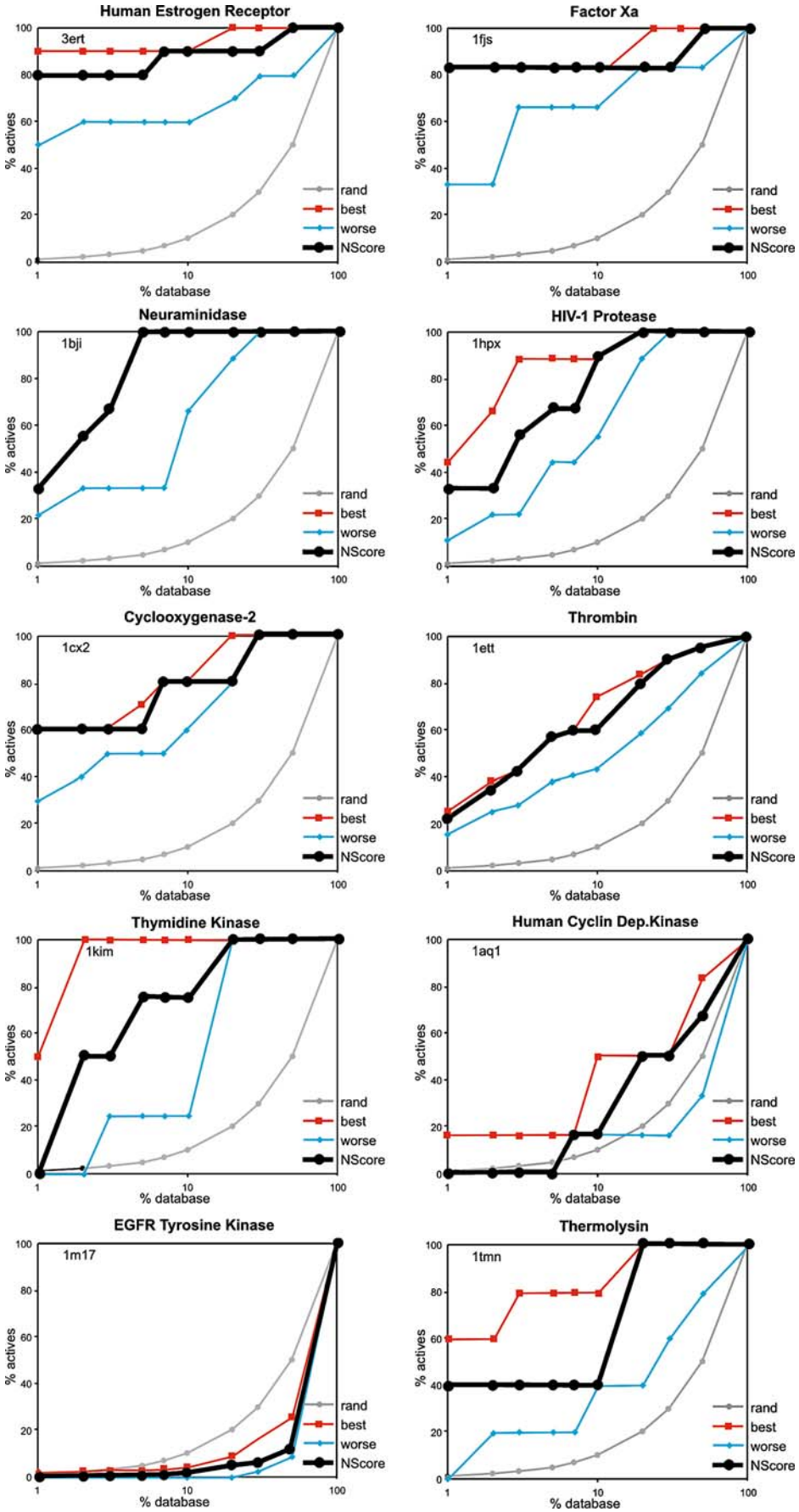

Fig. 8 Results of variation of parameters

virtual screening proved to be more sensitive to the variation of the parameters than the results of docking and scoring, which was not unexpected, because virtual screening is a complicated process including the docking and scoring of both active and inactive ligands. However, virtual screening remained stable to a certain degree; e.g., the results do not change dramatically if the energy parameters of the score for the hydrophobic effect and hydrogen bond were changed by 25%, although the results of docking and scoring were unchanged if these parameters were changed even by 50%.

Conclusions

We developed and tested a very simple scoring function NScore, all parameters of which were chosen almost arbitrarily, i.e., on the basis of general physical considerations alone, without any training and any adjustment to any experimental data. The results of docking, scoring, and virtual screening with the use of NScore in independent test sets proved to be almost as good as those obtained by means of the ICM, GOLD, DOCK and Glide software packages whose rather detailed empirical scoring functions were trained using protein–ligand complexes. We believe that this, somewhat unexpected result was accounted for by the following problems in the empirical scoring functions development:

Training empirical scoring functions always faces problems with the selection of the training set of complexes: this set may be either insufficiently complete or insufficiently balanced. The results of docking, scoring, and virtual screening with the use of empirical scoring functions for targets differing from those used for training the scoring functions may be substantially worse than the results obtained during training.

Whereas a scoring function has been trained for docking in the GOLD software, and a scoring function has been trained for virtual screening in the Glide software, both functions are used for three purposes: docking, scoring, and virtual screening. The objective of docking is to predict the most probable pose of the ligand, the objective of scoring is to predict the binding affinity, and the objective of virtual screening is to perform docking and scoring in order to differentiate between active and inactive ligands on the basis of their scores. Although these objectives have something in common, they are still not the same. Indeed, virtual screening by means of Glide yields better results than other programs, because the scoring function in Glide has been mainly trained for virtual screening, the improvement being considerably more pronounced in training sets than in test ones.

Satisfactory results obtained by means of the NScore scoring function are largely determined by the low sensitivity of docking, scoring, and virtual screening to the variation of the parameters of the NScore scoring function. The results only slightly change if the energy parameters are changed by 50% (in the cases of docking and scoring) or 25% (in the case of virtual screening). This stability is mainly accounted for by the simplicity and clear physical meaning of the NScore scoring function and is an extremely useful property of this function. The low sensitivity of docking, scoring, and virtual screening to the parameters of scoring functions may account for the fact that scoring functions obtained after incorrect training on incorrect training sets are fairly acceptable for docking, scoring, and virtual screening if the parameters of these functions are physically reasonable.

Analysis of errors in docking by means of the NScore scoring function has shown that, although many docking errors result from incorrect performance of scoring functions, no fewer errors are caused by other factors. The main of them is that the contribution of the water molecules simultaneously interacting with both the protein and the ligand into the ligand–protein interaction is estimated incorrectly. Most errors of virtual screening by means of the NScore scoring function are explained by incorrect performance of scoring functions. This usually occurs in tests using the targets for which other programs, such as GOLD and DOCK, also yield unsatisfactory results.

Comparison of the results of virtual screening by means of an entirely untrained, extremely simple function NScore with the results of virtual screening by means of sophisticated, trained empirical functions used in the GOLD, Glide, and DOCK software has shown that, for a better performance, the empirical scoring functions should be trained specially for virtual screening, with the use of the same targets for which these scoring functions are intended to be used, as was the case with the Glide software.

References

1. Böhm H-J, Schneider G (2003) Protein-ligand interactions. From molecular recognition to drug design. Methods and principles in medicinal chemistry. Wiley, Weinheim
2. Alvarez J, Shoichet B (2005) Virtual screening in drug discovery. CRC Press, Boca Raton
3. Kubinyi H (2006) Success stories of computer-aided design. In: Ekins S (ed) Computer applications in pharmaceutical research and development. (Wiley Series in Drug Discovery and Development). Wiley-Interscience, New York, pp 377–424
4. Klebe G (2006) Virtual ligand screening: strategies, perspectives and limitations. Drug Discov Today 11:580–594
5. Jones G, Willett P, Glen RC (1995) Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. J Mol Biol 245:43–53

6. Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261:470–489
7. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shaw DE, Shelley M, Perry JK, Francis P, Shenkin PS (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 47:1739–1749
8. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL (2004) Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J Med Chem 47:1750–1759
9. Totrov M, Abagyan R (1997) Flexible protein–ligand docking by global energy optimization in internal coordinates. Proteins Suppl 1:215–220
10. McGann M, Almond H, Nicholls A, Grant JA, Brown F (2003) Gaussian docking functions. Biopolymers 68:76–90
11. Goodsell DS, Olson AJ (1990) Automated docking of substrates to proteins by simulated annealing. Proteins 8:195–202
12. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) A geometric approach to macromolecule-ligand interactions. J Mol Biol 161:269–288
13. Bissantz C, Folkers G, Rognan D (2000) Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem 43:4759–4767
14. Wang R, Lu Y, Wang S (2003) Comparative evaluation of 11 scoring functions for molecular docking. J Med Chem 46:2287–2303
15. Perola E, Walters WP, Charifson PS (2004) A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. Proteins 56(2):235–249
16. Stoermer J et al (2006) Current status of virtual screening as analysed by target class. Med Chem 2:89–112
17. Warren GL et al (2006) A critical assessment of docking programs and scoring functions. J Med Chem 49:5912–5931
18. Richards FM (1977) Areas, volumes, packing, and protein structure. Ann Rev Biophys Bioeng 6:151–176
19. Fersht AR, Shi JP, Knill-Jones J, Lowe DM, Wilkinson AJ, Blow DM, Brick P, Carter P, Waye MM, Winter G (1985) Hydrogen bonding and biological specificity analysed by protein engineering. Nature 314:235–238
20. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Weiner P (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. J Am Chem Soc 106:765–784
21. Christopher W, Verdonk ML, Verdonk ML (2002) The consequences of translational and rotational entropy lost by small molecules on binding to proteins. J Comput-Aided Mol Des 16:741–753
22. Finkelstein AV, Janin (1989) J Protein Engineering 3:1–3
23. Eldridge M, Murray C, Auton T, Paolini G, Mee R (1997) Empirical functions: I. the development of a fast empirical scoring function to estimate the affinity of ligands in receptor complexes. J Comput-Aided Mol Des 11:425–445
24. Sadowski J, Gasteiger J (1993) From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. Chem Reviews 93:2567–2581
25. http://www.pdb.org
26. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WTM, Mortenson PN, Murray CW (2007) diverse, high-quality test set for the validation of protein-ligand docking performance. J Med Chem 50:726–741
27. Word JM, Lovell SC, Richardson JS, Richardson DC (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J Mol Biol 285:1735–1747
28. Zhou Z, Felts AK, Friesner RA, Levy RM (2007) Comparative performance of several flexible docking programs and scoring functions: enrichment studies for a diverse set of pharmaceutically relevant targets. J Chem Inf Model 47:1599–1608
29. Gohlke H, Klebe G (2002) Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. Angew Chem Int Ed 41:2644–2676
30. Eldridge M, Murray C, Auton T, Paolini G, Mee R (1997) Empirical functions: I. the development of a fast empirical scoring function to estimate the affinity of ligands in receptor complexes. J Comput-Aided Mol Des 11:425–445
31. Muegge I, Martin Y (1999) A general and fast scoring function for protein-ligand. interactions: a simplified potential approach. J Med Chem 42:791–804