PROTEINS: Structure, Function, and Bioinformatics 60:325–332 (2005)

REVIEW

Comparing Protein–Ligand Docking Programs Is Difficult

Jason C. Cole,[1] Christopher W. Murray,[2] J. Willem M. Nissink,[1] Richard D. Taylor,[2] and Robin Taylor[1]*
[1]*Cambridge Crystallographic Data Centre, Cambridge, United Kingdom*
[2]*Astex Technology Ltd., Cambridge, United Kingdom*

ABSTRACT **There is currently great interest in comparing protein–ligand docking programs. A review of recent comparisons shows that it is difficult to draw conclusions of general applicability. Statistical hypothesis testing is required to ensure that differences in pose-prediction success rates and enrichment rates are significant. Numerical measures such as root-mean-square deviation need careful interpretation and may profitably be supplemented by interaction-based measures and visual inspection of dockings. Test sets must be of appropriate diversity and of good experimental reliability. The effects of crystal-packing interactions may be important. The method used for generating starting ligand geometries and positions may have an appreciable effect on docking results. For fair comparison, programs must be given search problems of equal complexity (e.g. binding-site regions of the same size) and approximately equal time in which to solve them. Comparisons based on rescoring require local optimization of the ligand in the space of the new objective function. Re-implementations of published scoring functions may give significantly different results from the originals. Ostensibly minor details in methodology may have a profound influence on headline success rates. Proteins 2005; 60:325–332.** © **2005 Wiley-Liss, Inc.**

Key words: affinity prediction; drug design; enrichment rates; protein–ligand binding; virtual screening

INTRODUCTION

Protein–ligand docking[1] is an increasingly important method of lead generation in the pharmaceutical and biotechnology industries. Several docking algorithms are in common use, for example, AutoDock,[2] DOCK,[3] FlexX,[4] FRED,[5] Glide,[6] GOLD,[7] ICM,[8] and QXP.[9] There have been many studies comparing such programs (a good list is given in Kellenberger et al.[10]). We believe that such comparisons are difficult to do and that it is easy to produce results that are misleading or not generally applicable. We highlight the major problems below and, where possible, suggest remedies. We confine ourselves to: (1) comparisons based on re-docking of ligands into their experimentally-observed protein binding sites, normally using protein–ligand X-ray structures from the Protein Data Bank[11] (PDB); and (2) enrichment studies, typically involving docking of data sets of random compounds spiked with known actives.

COMPARISONS OF DOCKING PROGRAMS

Measuring the Success of Pose Prediction

The success of a program in predicting a ligand binding pose is usually measured by the root-mean-square deviation (RMSD) between the experimentally-observed heavy-atom positions of the ligand and those predicted by the program (more specifically, the top-ranked solution from the program). Given docking results for a set of test complexes, the RMSD values can then be summarized in different ways. The most frequently used statistic is the percentage of test complexes for which the docking solution has RMSD < 2Å (termed here the success rate). The 2 Å threshold is arbitrary but rather commonly used.[12–14] Some authors count dockings whose RMSDs lie between 2 and 3 Å as partial successes.[15,16] However, only rarely are success rates quoted with error estimates. Such estimates are obtainable by bootstrapping and can be quite large.[17] (Bootstrapping involves taking random samples of test complexes and determining the docking success rate for each sample. This gives insight into how much the success rate varies when the test set is altered.) In consequence, the difference in the success rates of two programs may easily not be statistically significant. We are unaware of any published docking program comparison that addresses this possibility.

Friesner et al.[6] used the average RMSD over all their test complexes as an overall success measure. However, if two programs are run on a test complex and produce solutions with RMSDs of 4 Å and 8 Å, it may be more reasonable to conclude that both are wrong (i.e., of no practical use to a drug designer) than to say one is twice as accurate as the other. Conversely, if they give solutions

*Correspondence to: Dr. R. Taylor, Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK. E-mail: taylor@ccdc.cam.ac.uk

Received 1 September 2004; Revised 14 December 2004; Accepted 4 January 2005

Published online 3 June 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20497

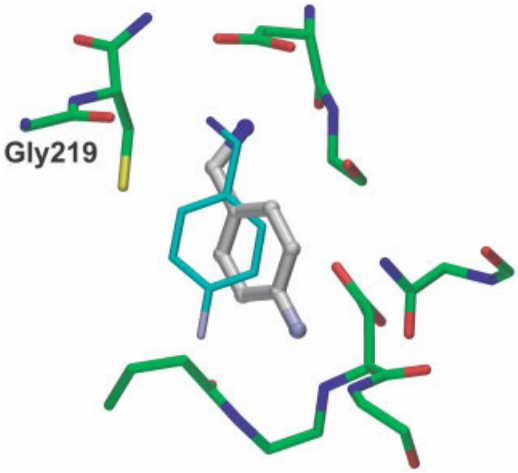

Fig. 1. Trypsin complexed with the inhibitor 4-fluorobenzylamine from PDB entry 1TNH (blue cylinders) with the top ranked docking solution from GOLD (grey ball and stick). The docking has an RMSD of 1.8 Å with the crystal structure, but the primary amine does not form the correct hydrogen bond to Gly219 and the fluorine is shifted by 3.1 Å from the experimental position. (All molecular plots constructed with AstexViewer2[21].)

with RMSDs of 0.3 Å and 0.6 Å, it might justifiably be argued that both are correct (i.e., of equal value to a drug designer), especially given the imprecision of the experimental results (see Goto et al.[18] for an interesting discussion of this point). Also, since the distribution of RMSDs is positively skewed, the average is a poor estimate of central location.

Another issue with RMSD is random expectation: small ligands can easily give low RMSDs even when randomly placed. For example, we have found that if arabinose is rotated randomly about its center of gravity, and the RMSD calculated between each random placement and the original orientation, some 10–15% of placements have RMSD < 2Å. This factor might become important for test sets containing significant numbers of small ligands, particularly if the size of the active site is also small.

Some authors[19,20] argue that RMSD is flawed as a success measure since it is possible to get solutions that have good RMSDs but form different interactions with the protein than are observed experimentally. For example, we have found that GOLD usually docks the ligand in the carbonic anhydrase complex 1CIL with an RMSD of below 1.5 Å, but the ligand sulphonamide group is invariably coordinated incorrectly to the active-site zinc atom (via oxygen rather than nitrogen). Another example is shown in Figure 1: the docking solution has a respectable RMSD (1.8 Å), but the primary amine does not form the correct hydrogen bond to Gly219. Conversely, solutions may have poor RMSDs but nevertheless be correct in essential features, such as an almost symmetric molecule or a lipophilic group that is docked the wrong way round (Figure 2). Another possibility is that a terminal group on the ligand is completely solvent exposed, forming no contacts at all to the protein (e.g. Figure 3). Misplacement of this group by a docking program would be of little or no importance but would increase the RMSD value.

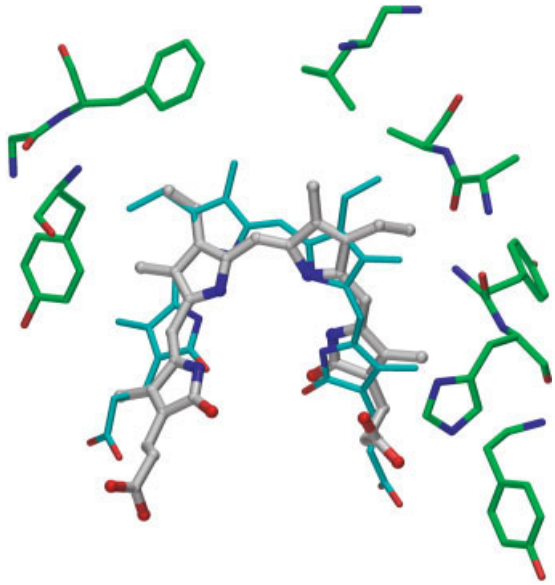

Fig. 2. Biliverdin IX gamma chromophore from PDB entry 1BBP (blue cylinders) with the top-ranked docking solution from GOLD (grey ball and stick). The ligand is almost symmetric and the essential features from the docking are correct. However, the RMSD with the crystal structure is 8.2 Å.

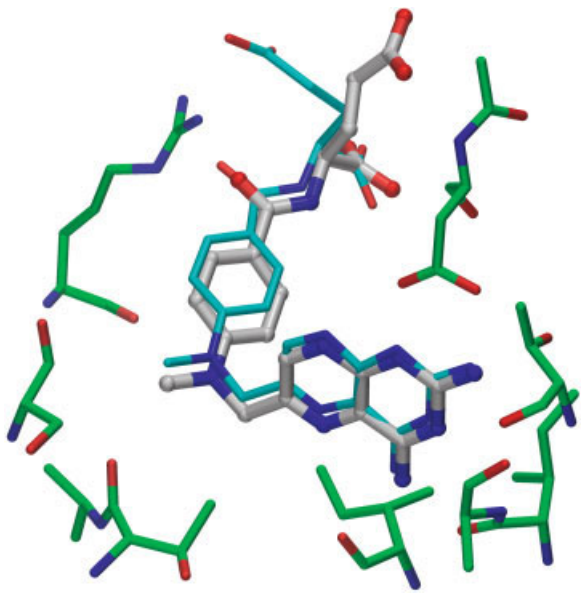

Fig. 3. Methotrexate bound to dihyrdofolate reductase from PDB entry 4DFR (blue cylinders) with the top-ranked docking solution from GOLD (grey ball and stick). The terminal $CH_2CH_2CO_2^-$ ligand group (top center of plot) is not in contact with any protein residue and as such its misplacement by a docking program is relatively unimportant.

In the light of these difficulties, an alternative measure ("IBAC") based on the number of correctly reproduced protein–ligand interactions has been proposed.[19] It was compared with RMSD on several test docking poses and its authors concluded that, although more subjective, it was "a more meaningful measure of docking accuracy in all these cases." Other investigators (e.g., in the CAPRI project[22]) have compared RMSD and interaction-based

measures and failed to come to a clear conclusion regarding their relative merits.

A concern with interaction-based measures is their possible sensitivity to fine details of the experimental protein X-ray structure, which may not always be accurate. For example, our experience is that GOLD often misplaces partially solvent-exposed ligand groups. In this situation, the observed and predicted interactions may be different but it is reasonable to question whether the fitting of the experimental electron density is unambiguous, given the likely high thermal mobility of the ligand group.

A further type of success measure is the "relative displacement error" (RDE),[23] given by:

$$\mathrm{RDE} = 100 \times \{1 - (\mathrm{L}/\mathrm{N}) \times (\Sigma\ 1/[\mathrm{L} + \mathrm{D_i}])\}$$

where the summation is over the N ligand atoms, D_i is the deviation between the observed and predicted position of the ith ligand atom, and L is a user-defined parameter, typically between 1.5–3 Å. RDE is 0% if all deviations are zero, 50% if all deviations are equal to L, 67% if they all equal 2L, 75% if they all equal 3L, etc. As can be seen, the influence of large discrepancies is down-weighted compared to the RMSD statistic, so an average RDE for a set of test complexes is less likely to be dominated by a small number of very bad docking solutions than the average RMSD. It still suffers, however, from the problems illustrated in Figures 1, 2, and to a lesser extent Figure 3. Also, there is a subjective element in the choice of L.

Measuring the Success of Affinity Prediction and Virtual Screening

Enrichment rates are usually quoted as the key success measure when the focus is on affinity prediction (i.e., molecules in the test database are ranked by docking score and the number of actives among the top x% counted). Halgren et al.[24] point out that the enrichment rate after x% of the database has been "screened" takes no account of the ranks of the active ligands; for example, the rate at x = 5 is the same whether all actives are in the top 1% or all lie between 4–5%. They propose an alternative formula that takes ranks into account; using this, or quoting enrichment rates at several values of x (or showing the enrichment curve plot) is advisable. Triballeau et al.[25] advocate the use of "receiver operating characteristic" (ROC) curves instead of conventional enrichment curves, since they are independent of the proportion of actives in the test set, and give information about false-positive and false-negative rates in the same plot.

Establishing statistical significance is again important: differences in enrichment rates may easily be due to chance. Suppose, for example, that two programs are used to dock a database of 990 inactives spiked with 10 actives (typical numbers[24,26]). If one program finds three actives among its 20 top-scoring molecules and the other finds seven (typical results), the difference is not statistically significant. This can be established by casting the data into a 2-by-2 contingency table and performing a χ^2 test (a Fisher test would be preferable if any expected cell total is < 5).[27] A final point is that apparent enrichments may be trivial anyway; for example, they may merely reflect the fact that scoring functions are usually correlated with molecular size.[28] The key question is whether the observed enrichment could have been obtained without docking. For example, if all the active molecules are charged, but few of the inactives, then a good enrichment could be obtained merely by focusing on this property, as demonstrated by Morley and Afshar.[29] Some workers recommend that the active molecules chosen for enrichment studies should have no better than micromolar binding affinities, since this is representative of real-world situations.[24] Others argue that use of nanomolar ligands is better since it increases the chance of establishing differences between scoring functions.[30] We tend to the former point of view since, even if it can be established that one scoring function is better than another at identifying nanomolar ligands, the result cannot necessarily be extrapolated to the micromolar compounds that are more likely to be relevant in practice.

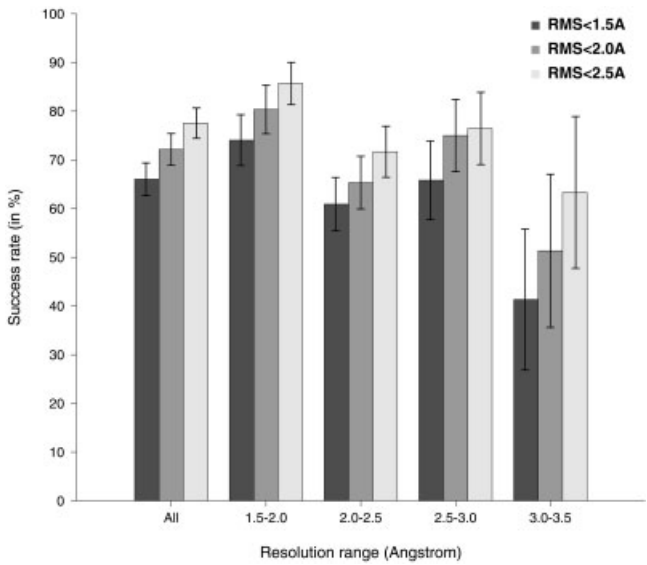

Fig. 4. Chart of GOLD success rates at different RMSD cutoff criteria as a function of the resolution of the PDB entries in the test set. Error bars indicate bootstrapped standard deviations.

Test Sets

Typically, docking programs work better on some targets than others[12,16,26,31] so a balanced assessment requires the test set to be large enough to cover diverse families, unless the point of the study is specifically to compare docking performance on a particular protein class (see for example, Hu et al.[32]). Perola et al. have recently investigated the performance of docking algorithms focusing on targets that have pharmaceutical importance;[33] such a data set will be less diverse but may be more relevant to contemporary drug discovery.

During the initial development of docking methods it is commonplace and legitimate to work with small numbers of complexes (e.g., Jones et al., 5 complexes[34]; Bursulaya et

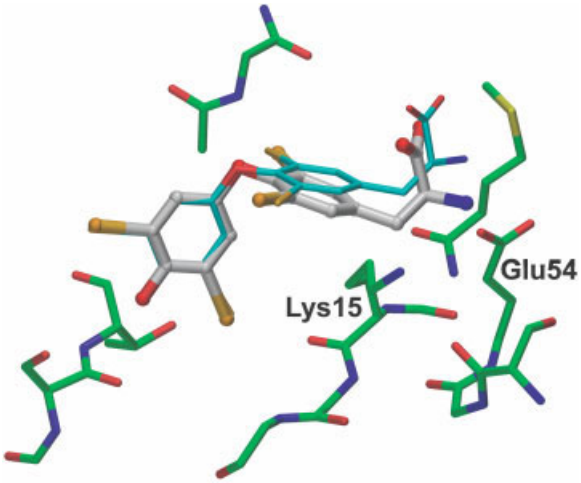

Fig. 5. Transthyretin (prealbumin) complexed with thyroxine from PDB entry 1ETA (blue cylinders). The manual binding mode (grey ball and stick) highlights the leverage effect of changing a C(ar)–O–C(ar) angle from the typical value in a small molecule crystal structure (120°) to a value taken from a previous version of Corina (106.8°), which causes the ligand to clash with Lys15 and Glu54.

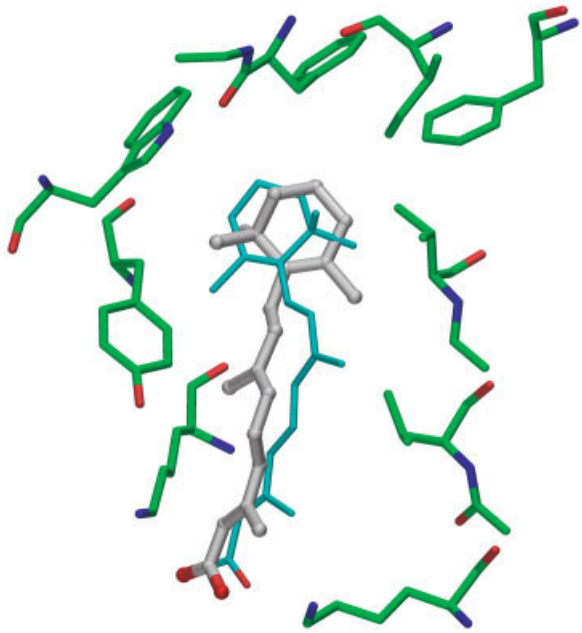

Fig. 6. Epididymal retinoic acid-binding protein (androgen dependent secretory protein) complexed with retinoic acid from PDB entry 1EPB (blue cylinders) with the top-ranked docking solution from GOLD (grey ball and stick). The head group, polar-end group, and lipophilic chain are all positioned moderately well but the RMSD is 2.4 Å. The RMSD is sensitive to the precise position of the long lipophilic chain and the orientation of the nearly-symmetrical head group, but the nonspecific nature of lipophilic interactions makes the precise orientation less important for such moieties.

al., 37 complexes[16]; Muryshev et al., 19 complexes[35]; Wang et al., 12 complexes[36]). However, comparison of the new program with others ultimately requires a large test set in order to establish that differences between program success rates are statistically significant.

Docking results tend to improve with the resolution of test-set complexes (Fig. 4), implying that some "failures" reflect experimental inaccuracies.[17] Unless the point of the study is specifically to test docking performance on poor-quality structures, it therefore makes sense to omit low resolution complexes so that failures can more definitely be ascribed to the docking algorithm itself. (Resolution is not the only indicator of experimental precision—for example, information in PDB header records is often informative. Goto et al.[18] recommend exclusion of structures whose ligands have high atomic displacement factors, viz. > 50 Å^2, or occupancies of less than 1.) For the same reason, structures with obviously incorrect features (e.g., close contacts or other clear errors) should be removed or annealed. If the latter option is chosen, the annealing needs to be done with care because it can introduce biases. For example, the original version of Chemscore[37] was parameterized on a set of complexes minimized with a force field that changed metal-acceptor bond lengths significantly from their experimental values. In consequence, the function performed better on complexes that had been annealed with that same force field than on raw X-ray structures and reparameterization was necessary for the function to perform optimally on the latter.[38]

It is desirable to recompute and visually inspect the ligand electron density to minimize the risk of using structures in which the ligand is misplaced, disordered, partially occupied or otherwise ill-defined. Ideally, the test set should not include complexes on which any of the docking programs were trained ("training" means any change made to the program to improve success rate). However, this is not always practicable since it involves significant extra work in finding and setting up additional test complexes. It is important to exclude structures in which neighboring chains or ligands in the crystal-packing environment are close enough to the binding site to influence the ligand-binding geometry. This is quite common[17] but rarely taken into account in test-set construction (identifying such problem cases is time consuming and prone to error). Alternatively, if the objective is to measure accuracy of pose prediction, such structures can be kept, but care must be taken to include in the binding-site description all moieties from the crystal-packing environment that interact with the ligand. If the aim is to measure affinity prediction, it is necessary to assess whether the contacts seen in the crystal structure are biologically relevant.

Randomization of the starting geometry and position of the ligand are important. Examples have been seen where the use of experimentally-observed ligand geometries as starting points has led to increased success rates, an observation rationalized by the investigators as indicating a lack of coverage of conformational space in the docking program's internal conformer generator.[31] Similar problems can arise if the initial coordinates are modified so that the ligand center of geometry coincides with that observed in the crystal structure.[16] Also, most docking programs keep ligand valence angles fixed, which can cause problems. For example, in the original GOLD validation, ligands were prepared by minimizing the experimental ligand geometry and then randomizing the conforma-

tion.[7] Optimization of valence angles in this way may artificially lower the strain energy of the experimental conformation (or something close to it) and therefore introduce an inadvertent bias. Starting geometry should not usually be an issue for programs that use an incremental ligand construction strategy, though Kellenberger et al. have pointed out that problems are still possible.[10]

The fact that most docking programs do not allow valence angles to vary may lead to a dependence of success rates on the program used to generate ligand 3D structures. For example, previous versions of Corina[39] assigned valence angles of 106.8° to aromatic ether oxygens, whereas small-molecule crystal structures in the Cambridge Structural Database[40] indicate that values at least 10° higher are more usual. Such a discrepancy can significantly alter the overall shape of a molecule, especially if the angle is near the center of the molecule where it has large leverage (Fig. 5). The consequences on success rates are likely to be more severe for docking programs that use steep potential functions. Possible solutions are to optimize the ligand geometry by using a state-of-the-art force field (although this, of course, will introduce its own errors) or explicitly fix known errors in the valence-angles delivered by the 3D structure-generation program. Alternatively, generating ligand geometries in two or more ways and docking with each is satisfactory,[6,41] though it produces more results and thereby complicates the analysis.

For completeness, we note that re-docking ligands back into the protein structures from which they were taken introduces a bias by hiding the problem of ligand-induced changes in protein conformations. This is highlighted by cross-docking studies (taking the observed binding site from one protein–ligand complex and docking into it a different ligand) which can give lower—sometimes much lower—success rates.[42]

Comparability of Search Spaces

Part of the docking problem involves seeking the global minimum of the objective function. The difficulty of this task depends on the size of the search space. In comparing docking programs, it is important to ensure that one algorithm is not gratuitously given a harder search problem than another. For example, if one program allows ligand ring conformations to vary and another does not, the former has a higher dimensional space to explore and the latter has an advantage if the input ring geometry is taken from the experimentally-observed ligand structure. Various authors point out the importance of the binding-site definition.[19,31] Provided it includes enough residues for the correct protein–ligand interactions to be found, the smaller the binding-site region that has to be explored, the easier the search problem. Since different programs require the binding site to be defined in different ways (e.g., a rectangular box, or a sphere centered on a point within the cavity) it is sometimes impossible to equalize this factor. In their comparison of FRED, FlexX, and Glide, Schulz-Gasch and Stahl observe that much of the difference in success rates can be ascribed to the way in which the binding site was defined for each program.[31]

Equally, it is important to present the same amount of information to each program. For example, if one program is evaluated on a set of binding sites in which crystallographically-observed water molecules are included, the results cannot be compared with those from another program tested on the same set, but without the water molecules. As a referee of this paper has said, "[in receptor preparation,] what is wrong is to mix different assumptions between different docking programs."

It is necessary to ensure that each program has approximately the same amount of time to do the search, since the real issue is docking accuracy per unit time. Comparisons in which one algorithm is allowed much more time than another are not uncommon,[41] although some workers have been careful to choose program settings that give very nearly the same time per ligand to each program.[26,43] Reports of docking-program comparisons frequently include comments about the relative speeds of different algorithms. Occasionally, we feel these are unduly damning. For example, programs such as Glide, ICM, and GOLD have been described as being of low suitability for high-throughput virtual screening[44] although such programs are clearly being used for this purpose[28] and the opinion is directly contradicted by other workers.[10] Some programs have different speed settings, so may be classed as either slow or fast, depending on the settings chosen. The invariable approach in comparisons is to select a single setting for each program. This makes it difficult to compare the speed/accuracy profiles of the programs, it being entirely possible that one program might be better than another when fast settings are used but worse when slow settings are employed. Finally, Muryshev et al. note that most programs take significant time to initialize the protein.[35] In a virtual screening run, this normally only has to be done once. For that reason, they subtracted initialization time when tabulating the cpu times taken by their Algodock program. Unfortunately, they were unable to make corresponding corrections for the other programs discussed in their study, which made speed comparisons between these programs and Algodock difficult.

Rescoring

A common methodology is to dock with one scoring function, rescore the solution with several other functions, compare results and draw conclusions about the relative accuracy of the functions.[45,46] In this situation, it is essential to perform local optimization of the docking solution with respect to each of the scoring functions. We have experimented by docking ligands into their parent proteins using the Chemscore fitness function, and then computing their GoldScore fitness values with and without local optimization. The effect of local optimization is often to alter the GoldScore value by 50 or more units, which is a huge shift for this scoring function. This may occur even if the RMSD between optimized and unoptimized ligand positions is 0.5 Å or less. It is clear that any attempt to draw conclusions from the unoptimized values is problematic. Moreover, minimizing this particular function involves optimizing some protein atom positions (rotatable

polar hydrogens such as serine OH groups) as well as ligand atom positions. The problems caused by failing to optimize before rescoring will be worse for functions that give rise to rugged response surfaces. Published works are sometimes unclear about whether local optimizations have been performed prior to rescoring, in which case it is difficult to know whether a problem exists.

Other Factors

Many of the papers referred to here compare programs developed by those papers' authors with other docking programs. It is difficult to get the best out of a docking program without experience in its use, as noted by several authors,[19,31] and since workers will be more expert in their own software than other programs, this may bias comparisons. Kontoyianni et al.[12] make this point, commenting that they, as independent workers in an industrial computational-chemistry laboratory, are unable to match the success rates achieved and published by vendors of docking-programs. Also, authors will be working with the very latest versions of their software but may be unable to use contemporary versions of other programs. It may not be possible to remedy the situation but is obviously sensible to make it as explicit as possible, for example, by stating the date of release of the program versions used and the version numbers of the latest releases of the programs.

Confusion can arise from the divergence of scoring-function implementations. It is very difficult to re-implement a scoring function with complete accuracy given only the information in a published paper. Authors would often need to make source code accessible to fully document an implementation and this is normally impossible for commercial reasons. Also, changes are sometimes deliberately made when functions are re-implemented.[38,45] Re-implementations of scoring functions often perform differently from the original. For example, Wang et al[46] achieved 42% and 35% success rates with third-party re-implementations of GoldScore and Chemscore, respectively, while the original GoldScore implementation and GOLD's re-implementation of Chemscore gave 75% and 73% respectively (Taylor, R.D. and Verdonk M.L., personal communication) (and a similarly improved success rate for Chemscore was also reported by Ferrara et al.[47]). While authors may be meticulous in documenting the exact variants of the scoring functions they use, we fear that the subtleties are often lost when their original work is cited.

SUMMARY AND RECOMMENDATIONS

Characterizing and comparing the performance of docking programs is a surprisingly complex problem. We have pointed out many of the difficulties and now make recommendations for the design of docking-program comparisons. Some of these recommendations are strong; that is, we believe that current, common practice is flawed and clearly open to improvement. Notable among these is the necessity for establishing the statistical significance of differences between the success rates of docking algorithms. Other recommendations are less definite, either because there is no unique "right way" or because the optimum protocol will depend on circumstances.

Measure Statistical Significance

In most branches of science, a comparison of two or more treatments will routinely be supported by an analysis determining whether differences between the treatments are statistically significant. There is no reason why docking-program comparisons should not meet the same standard. The method employed for estimating significance levels must depend on the type of success measure being used. For pose prediction, when results have been categorized in some way as successes or failures, bootstrapping is particularly to be recommended. This is because it takes some account of sampling uncertainties in the test set. If a standard two-sample hypothesis test is performed for estimating whether RMSD or RDE values from one program differ significantly from those from another, a paired test (e.g., the Wilcoxon test or a paired t-test) will normally be more powerful than the corresponding unpaired test, given that both programs have been run on the same set of complexes. Two-tailed rather than one-tailed tests will generally be appropriate. For enrichment studies, the use of a Kolmogorov-Smirnov test allows the whole enrichment curve to be taken into account. If many programs are being compared in the same paper, an analysis of variance can be used to establish that there is an overall, statistically-significant variation between the programs, prior to performing two-sample hypothesis tests on individual pairs of programs. In summary, a formal statistical analysis should be done, but, depending on circumstances, a variety of methods may legitimately be chosen.

Use an Information-Rich Success Measure

Presentation of enrichment or ROC curves appears to be a reasonably satisfactory way of presenting results on the identification of active ligands. The situation with pose prediction is far less clear. Here, it is essential that the success measure is not influenced by how wrong the wrong answers are. Use of RMSD is problematic if the ligand is nearly symmetrical. In this situation, an RMSD based on the largest symmetrical substructure is an alternative, but is not necessarily safe if the remaining groups (i.e., those that break the symmetry of the molecule, for example, the methyl and vinyl groups of the molecule in Fig. 2) are not chemically innocent. A pragmatic solution is simply to reject nearly-symmetrical ligands when creating test sets. Additionally, RMSD can be misleading when there are ligand groups (normally terminal) that are misplaced but do not interact with the protein. A reasonable solution is to identify such groups, either visually or algorithmically, and remove them from the RMSD calculation, although this involves some subjectivity. Once again, an alternative pragmatic approach is to avoid such ligands.

Even with the above refinements, measures such as RMSD or RDE are not adequate on their own because of the possibility that solutions may, e.g., have good RMSDs, yet form different interactions with the protein from those observed experimentally. Since this may affect scoring-function values, it must count as an important error. Unfortunately, interaction-based measures such as IBAC are sensitive to what is defined as an interaction; this is

particularly a problem for weak contacts with little or no directional preference.

A referee of this paper has pointed out the advantage of using a continuous measure such as RDE rather than categorizing docking solutions as successes or failures based on an arbitrary RMSD or RDE cutoff. One advantage of categorization, however, is that it becomes trivial to include both RMSD (or RDE) and interaction-based measures in the success criteria.

We are therefore of the opinion that there is no unique "best" success measure. A recommended (but by no means the only) compromise is to begin by counting as a success any complex whose RMSD is < 2 Å (assuming that nearly-symmetrical ligands have been avoided). A 1.5-Å cutoff might be preferable for small ligands. This initial categorization should then be refined by use of an interaction-based measure and/or by visual inspection, and with the main aim of ensuring that solutions are only counted as successes if all key interactions with the protein (viz. unequivocal hydrogen bonds) are correctly reproduced. We favor at least some visual inspection. It could be argued that this is lacking in objectivity, prone to mistakes and slow. While acknowledging these weaknesses, it is also difficult and time consuming to write a computer algorithm capable of detecting all the factors that might be relevant for judging the quality of docking poses. Figure 6, for example, illustrates a mitigating factor that is easy to spot by eye, and might reasonably result in re-categorizing a docking pose as a success, despite it having an RMSD > 2 Å. Comprehensive algorithmic detection of situations such as this is difficult.

Avoid Suspect Structures in Test Sets

Testing docking programs on suspect experimental data adds unwanted noise, making it more difficult to identify differences in program performance. For testing pose prediction, it is best if protein–ligand X-ray structures are only used if the structure factors are available, so that the difference electron density associated with the ligand can be calculated and examined. If there is doubt regarding how the electron density is fitted, the complex is best avoided. We are currently setting up a new test set consisting entirely of complexes validated in this way and will make this test set publicly available. In addition, the checks listed by us and others previously[17,18] are recommended: in particular, testing for nearby chemical species in the crystal-packing environment is important. High-resolution complexes should be chosen if at all possible. Annealing of protein structures may cause unexpected biases and should be avoided (and will be unnecessary if structures of good quality have been selected). The size of the test set will determine the sensitivity of the docking-program comparison: a test set of less than 100 complexes is unlikely to be large enough to produce statistically-significant results. A test set is publicly available that meets most of these criteria.[17]

A key issue when comparing enrichment rates is the number and nature of the active compounds that are included in the test set. First, if there are few of them (10 or less), it is highly unlikely that differences in enrichment rates will be statistically significant. Secondly, it cannot be assumed that an ability to detect high-affinity ligands indicates a similar ability to identify the weak actives that are much more likely in practice. Consequently, we recommend that test sets should contain micromolar rather than nanomolar ligands, and as many of these as possible. Finally, it must be demonstrated that there are no obvious, systematic difference between the structures of the active and inactive ligands in the test set. If there are, enrichment can be achieved without docking and the test set is therefore inappropriate for comparing docking programs.

Randomize Starting Ligand Geometries

Both the ligand position and conformation should be randomized before input to the docking program (ensuring there are no unrealistic intramolecular clashes), even if the docking-program documentation suggests this is unnecessary. If ligand bond lengths and angles are optimized, this should be done starting from a random conformation.

Ensure Equivalence of Search Spaces

The volume of the binding-site cavity in which the ligand can be placed should be similar, preferably identical, for all the programs being compared. The number of conformational degrees of freedom in the ligand should be similar, preferably identical, for all programs, that is, all programs should allow the same geometrical features to vary. Programs should be allowed similar amounts of time on any given test complex, and, in judging this, time spent on protein initialization should be ignored. Binding sites should be set up as similarly as possible, for example, with the same policy regarding the retention or exclusion of water molecules.

Optimize when Rescoring

If docking solutions are rescored with another scoring function, it is essential to perform a local optimization of the ligand pose in the space of the new function. This optimization should be over all the degrees of freedom that are relevant to the function.

Report Caveats

The conclusions from any docking-program comparison are likely to be subject to several caveats, which should obviously be stated explicitly. Among these might be: findings relate to only a certain type of protein; programs offer different speed/reliability settings but were only compared at one setting; search space (size of binding site, number of ligand degrees of freedom) could not be made identical or near-identical for all programs; old versions of programs were used; ligand geometries were generated with a particular 2D-to-3D conversion program.

Concluding Remarks

We have reviewed recent comparisons of protein–ligand docking programs and have highlighted the difficulties associated with these comparisons. We have proposed a list of recommendations that we hope will be helpful to

researchers in subsequent assessments of docking performance.

ACKNOWLEDGMENTS

The authors thank Mike Hartshorn and Marcel Verdonk of Astex Technology Ltd. for many helpful discussions and comments on the manuscript. We thank the referees for helpful comments.

REFERENCES

1. Taylor RD, Jewsbury PJ, Essex JW. A review of protein-small molecule docking methods. J Comput Aided Mol Des 2002;16:151–166.
2. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, Belew RK, Olson AJ. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comp Chem 1998;19:1639–1662.
3. Makino S, Kuntz ID. Automated flexible ligand docking method and its application for database search. J Comput Chem 1997;18: 1812–1825.
4. Rarey M, Kramer B, Lengauer T, Klebe G. A fast flexible docking method using an incremental construction algorithm. J Mol Biol 1996;261:470–489.
5. OpenEye Scientific Software, Sante Fe, New Mexico, USA.
6. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, Repasky MP, Knoll EH, Shelley M, Perry JK, and others. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 2004;47: 1739–1749.
7. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol 1997;267:727–748.
8. Abagyan RA, Totrov MM, Kuznetsov DA. ICM: a new method for structure modeling and design: applications to docking and structure prediction from the distorted native conformation. J Comp Chem 1994;15:488–506.
9. McMartin C, Bohacek RS. QXP: powerful, rapid computer algorithms for structure-based drug design. J Comput Aided Mol Des 1997;11:333–344.
10. Kellenberger E, Rodrigo J, Muller P, Rognan D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. Proteins 2004;57:225–242.
11. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
12. Kontoyianni M, McClellan LM, Sokol GS. Evaluation of docking performance: comparative data on docking algorithms. J Med Chem 2003;47:558–565.
13. Kramer B, Rarey M, Lengauer T. Evaluation of the FlexX incremental construction algorithm for protein-ligand docking. Proteins 1999;37:228–241.
14. Gohlke G, Hendlich M, Klebe G. Knowledge-based scoring functions to predict protein-ligand interactions. J Mol Biol 2000;295: 337–356.
15. Vieth M, Hirst JD, Kolinski A, Brooks III CL. Assessing energy functions for flexible docking. J Comp Chem 1998;19:1612–1622.
16. Bursulaya BD, Totrov M, Abagyan R, Brooks III CL. Comparative study of several algorithms for flexible ligand docking. J Comput Aided Mol Des 2003;17:755–763.
17. Nissink JWM, Murray C, Hartshorn M, Verdonk ML, Cole JC, Taylor R. A new test set for validating predictions of protein-ligand interaction. Proteins 2002;49:457–471.
18. Goto J, Kataoka R, Hirayama N. Ph4Dock: pharmacophore-based protein-ligand docking. J Med Chem 2004;47:6804–6811.
19. Kroemer RT, Vulpetti A, McDonald JJ, Rohrer DC, Trosset J-Y, Giordanetto F, Cotesta S, McMartin C, Kihlen M, Stouten PFW. Assessment of docking poses: interaction-based accuracy classification (IBAC) versus crystal-structure deviations. J Chem Inf Comp Sci 2004;44:871–881.
20. Pang Y-P, Perola E, Xu K, Prendergast FG. EUDOC: a computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases. J Comp Chem 2001;22:1750–1771.
21. Hartshorn MJ. AstexViewer: a visualization aid for structure-based drug design. J Comput Aided Mol Des 2002;16:871–881.
22. CAPRI: Critical assessment of prediction of interactions. http://capri.ebi.ac.uk/
23. Abagyan RA, Totrov MM. Contact area difference (CAD): a robust measure to evaluate accuracy of protein models. J Mol Biol 1997;268:678–685.
24. Halgren TA, Murphy RB, Friesner RA, Beard HS, Frye LL, Pollard WT, Banks JL. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. J Med Chem 2004;47:1750–1759.
25. Triballeau N, Acher F, Bertrand H-O. A plea for a wider application of ROC curve analysis in drug design. Evaluation of virtual screening workflows applied to metabotropic glutamate receptor sub-type4. XVIIIth Int Symp Med Chem 2004, Copenhagen.
26. Bissantz C, Folkers G, Rognan D. Protein-based virtual screening of chemical databases. 1. Evaluation of different docking/scoring combinations. J Med Chem 2000;43:4759–4767.
27. Siegel S. Nonparametric statistics for the behavioral sciences. London: McGraw-Hill; 1956.
28. Verdonk ML, Berdini V, Hartshorn MJ, Mooij WTM, Murray CW, Taylor RD, Watson P. Virtual screening using protein-ligand docking: avoiding artificial enrichment. J Chem Inf Comput Sci 2004;44:793–806.
29. Morley SD, Afshar M. Validation of an empirical RNA-ligand scoring function for fast flexible docking using RiboDock. J Comput Aided Mol Des 2004;18:189–208.
30. Krovat EM, Langer T. Impact of scoring functions on enrichment in docking-based virtual screening: an application study on renin inhibitors. J Chem Inf Comput Sci 2004;44:1123–1129.
31. Schulz-Gasch T, Stahl M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. J Mol Model 2003;9:47–57.
32. Hu X, Balaz S, Shelver WH. A practical approach to docking of zinc metalloproteinase inhibitors. J Mol Graph Model 2004;22:293–307.
33. Perola E, Walters WP, Charifson PS. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. Proteins 2004;56:235–249.
34. Jones G, Willett P, Glen RC. Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. J Mol Biol 1995;245:43–53.
35. Muryshev AE, Tarasov DN, Butygin AV, Butygina OYu, Aleksandrov AB, Nikitin SM. A novel scoring function for molecular docking. J Comput Aided Mol Des 2003;17:597–605.
36. Wang J, Kollman PA, Kuntz ID. Flexible ligand docking: a multistep strategy approach. Proteins 1999;36:1–19.
37. Eldridge MD, Murray CW, Auton TR, Paolini GV, Mee RP. Empirical scoring functions: 1. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. J Comput Aided Mol Des 1997;11:425–445.
38. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD. Improved protein-ligand docking using GOLD. Proteins 2003;52: 609–623.
39. Molecular Networks GmbH, Erlangen, Germany.
40. Allen FH. The Cambridge Structural Database: a quarter of a million crystal structures and rising. Acta Crystallogr B Biol Crystallogr 2002;58:380–388.
41. Erickson JA, Jalaie M, Robertson DH, Lewis RA, Vieth M. Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. J Med Chem 2004;47:45–55.
42. Murray CW, Baxter CA, Frenkel AD. The sensitivity of the results of molecular docking to induced fit effects: application to thrombin, thermolysin and neuraminidase. J Comput Aided Mol Des 1999;13:547–562.
43. Jenkins JL, Kao RYT, Shapiro R. Virtual screening to enrich hit lists from high-throughput screening: a case study on small-molecule inhibitors of angiogenin. Proteins 2003;50:81–93.
44. Lyne PD. Structure-based virtual screening: an overview. Drug Discov Today 2002;7:1047–1055.
45. Clark RD, Strizhev A, Leonard JM, Blake JF, Matthew JB. Consensus scoring for ligand/protein interactions. J Mol Graph Model 2002;20:281–295.
46. Wang R, Lu Y, Wang S. Comparative evaluation of 11 scoring functions for molecular docking. J Med Chem 2003;46:2287–2303.
47 Ferrara P, Gohlke H, Price DJ, Klebe G, Brooks CL III. Assessing scoring functions for protein-ligand interactions. J Med Chem 2004;47:3032–3047.