Evaluating docking programs: keeping the playing field level

John W. Liebeschuetz

Received: 31 October 2007 / Accepted: 28 December 2007 / Published online: 15 January 2008

Abstract Over recent years many enrichment studies have been published which purport to rigorously compare the performance of two or more docking protocols. It has become clear however that such studies often have flaws within their methodologies, which cast doubt on the rigour of the conclusions. Setting up such comparisons is fraught with difficulties and no best mode of practice is available to guide the experimenter. Careful choice of structural models and ligands appropriate to those models is important. The protein structure should be representative for the target. In addition the set of active ligands selected should be appropriate to the structure in cases where different forms of the protein bind different classes of ligand. Binding site definition is also an area in which errors arise. Particular care is needed in deciding which crystallographic waters to retain and again this may be predicated by knowledge of the likely binding modes of the ligands making up the active ligand list. Geometric integrity of the ligand structures used is clearly important yet it is apparent that published sets of actives + decoys may contain sometimes high proportions of incorrect structures. Choice of protocol for docking and analysis needs careful consideration as many programs can be tweaked for optimum performance. Should studies be run using 'black box' protocols supplied by the software provider? Lastly, the correct method of analysis of enrichment studies is a much discussed topic at the moment. However currently promoted approaches do not consider a crucial aspect of a successful virtual screen, namely that a good structural diversity of hits be returned. Overall there is much to consider in the experimental design of enrichment studies. Hopefully this study will be of benefit in helping others plan such experiments.

Keywords Docking · Enrichment · Factor Xa · GOLD · Protein–ligand interaction · Thrombin · Virtual screening

Abbreviations

CCDC	Cambridge Crystallographic Data Centre
RCSB	Research collaboratory for structural bioinformatics
PDB	Protein data bank
COX2	Cyclooxygenase 2
ER	Oestrogen receptor
sPLA2	Secretory phospholipase 2
RMSD	Root mean square deviation
VS	Virtual screening
ADME	Absorption, distribution, metabolism, excretion
ROC	Received operating characteristic
AUC	Area under curve

Introduction

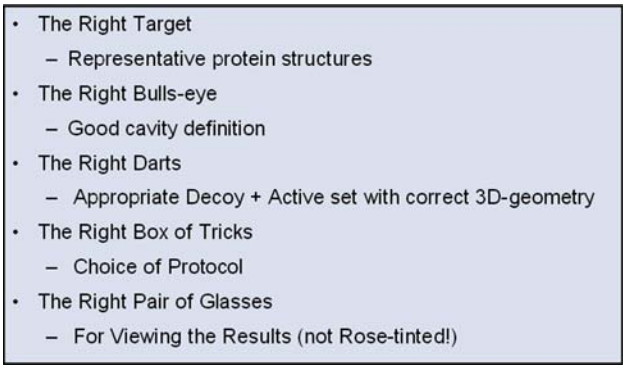
One thing that most experienced practitioners of molecular docking might agree upon is that comparing different docking packages for performance in virtual screening (VS), is fraught with pitfalls [1]. Many studies have been published [2] yet there is a question of how many of these can be considered free of bias, unintentional or not, or flawed in their set up in some way, or deficient in the analysis tools used. Some studies are either carried out or sponsored by commercial companies with a vested interest in a particular program. Such studies need to be scrutinized

J. W. Liebeschuetz (✉)
Cambridge Crystallographic Data Centre, 12 Union Rd,
Cambridge CB2 1EZ, UK
e-mail: john@ccdc.cam.ac.uk

with particular vigour. In some cases it has not been easy to obtain sufficient information to repeat the work. In cases where a re-examination of a published comparison has in fact been possible, it was found that at least some results were not reproducible [3].

Many studies have been carried out by non-affiliated organisations most notably by modelling groups in the major pharmaceutical companies [4–7]. The likelihood of bias is less in these cases, but not completely excluded as certain programs may have strong advocates within the organisation and there may also be internally developed programs included in the comparison. Another point is that molecular modellers in the pharmaceutical industry have limited time to carry out these studies. They earn their bread and butter after all by working on highly directed drug discovery programs. Time pressure may lead to some aspects of the study not being designed and audited with sufficient rigour. This is not to criticize the authors of these studies too harshly as there are many errors and oversights it is possible to make and there exist as yet no accepted set of standards to which a study design can be held up to. Setting up these studies to exclude flaws is difficult to do and in practice all published studies are likely on close examination to be flawed in some respect or other.

A recent published study is that by Chen et al. a substantial part of which was available to be repeated in entirety [6]. This study compared the docking programs FlexX, GOLD, GLIDE and ICM in the context of both crystallographic pose retrieval and enrichment within a virtual screen. This study reported poor results for GOLD as a tool for VS. This concerned us sufficiently to wish to repeat the work. As will be shown we were able to get better results than those published, partly because we were able to improve on certain aspects of the original protocol and also use VS search settings not available to the authors of the Chen study. We thank the authors for supplying protein and ligand files, and GOLD protocols, and providing useful assistance. Whilst carrying out this work we became motivated to identify and analyse issues that need careful consideration whilst setting up enrichment studies. This analysis makes no attempt to be comprehensive. However some of the areas in which mistakes can be made will be looked at. Figure 1 lays out five different aspects of a typical enrichment study, each of which requires care and attention to get right. We will examine these issues in turn.



- The Right Target
  - Representative protein structures
- The Right Bulls-eye
  - Good cavity definition
- The Right Darts
  - Appropriate Decoy + Active set with correct 3D-geometry
- The Right Box of Tricks
  - Choice of Protocol
- The Right Pair of Glasses
  - For Viewing the Results (not Rose-tinted!)

Fig. 1 The make-up of a good enrichment study

Enrichment study using the GOLD docking program

Chen et al. used 12 targets in their original study [6]. Four of these target protein structures were supplied to us in mol2 format. Each was derived from a deposited structure in the research collaboratory for structural bioinformatics (RCSB) protein data bank (PDB) [8]. The targets and corresponding PDB accession codes, are thrombin (1dwd), cyclooxygenase 2 (COX2) (1cx2), oestrogen receptor (ER) (1err) and secretory phospholipase 2 (sPLA2) (1db4). Files containing the 3D structures for the active molecules for these targets were also supplied, embedded within the list of decoy molecules used in the original study. The active lists for the remaining eight targets contained hits proprietary to the authors' parent company, therefore the enrichment experiments could not be repeated against these targets. GOLD configuration files were provided for the original docking experiments. Therefore a near replication of the original work was in principle possible for four targets.

The total number of actives numbered 344 (125 thrombin, 125 ER, 75 Cox2 and 19 sPLA2) and decoys, 20,000. However the number of structures docked was over 50% greater as alternative reasonable protonation states, tautomers and ring conformers had been included for both actives and decoy molecules.

Results and discussion

The results from the original experiments are given in Table 1. Enrichment factors based on actives retrieved are quoted for a cut of the top 10% of the set. Enrichment is poor for thrombin and worse than random (i.e. <1) for ER and sPLA2. Modest enrichment was reported for COX2, and in fact this enrichment was the best reported against this particular target over all docking packages.

Results obtained when these experiments were repeated are also tabulated in Table 1. GOLD version 2.2, as used in the original study, and the GOLD protocols supplied by the original authors were employed. These protocols include the active site definition so this should also be replicated. Enrichment factors are quoted at a 10% cutpoint.

There appears to be a significant difference between the two sets of data for three of the four targets. Discussions

Table 1 Comparison of VS results obtained in the Chen study [6] using GOLD 2.2 with those generated in house using an identical protocol, and those using GOLD 3.0.1 with a more advanced VS protocol

Docking protocol	Thrombin	COX2	ER	sPLA2
Chen. GOLD 2.2	1.44	3.55	0.75	0
In house GOLD 2.2	2.8	4.1	2.2	3.5
GOLD 3.0.1	5.8	4.7	3.3	5.9

The top 10% of the database as ranked by GoldScore was examined in each case. Figures quoted (the enrichment factor) are the actual number of actives retrieved over the number expected from a random selection (Max = 10 for a 10% cut)

with the authors of the original study have not been able to satisfactorily resolve why this difference exists. If the latest results are correct, then GOLD is performing rather better than was originally reported.

Chen et al. used a fast docking protocol that was entirely consistent with guidelines provided within the GOLD documentation for the version of GOLD used at the time. Current recommendations for a default VS protocol differ, partly because of the implementation of new features that improve docking efficiency. Results are presented in Table 1 for GOLD 3.0.1 using this newer default protocol. Only the search protocol has been updated. The protein model and active site definition remain unchanged. These results are significantly better and now reasonable enrichment is achieved for three of the four targets, ER being the exception. Development of docking methodology is a fast moving field and new versions of docking programs come out frequently so it is inevitable that docking program comparisons rapidly get out of date.

We will now attempt to analyse the make up of a VS experiment as given in Fig. 1. In some instances the Chen study is used for reference. It is not the intention to criticize this study in particular. It happened to be the one with which we were most familiar and if any shortcomings are implied, many of these are likely also to be found in a number of other published studies.

Attributes of high quality enrichment studies

The right target

Choosing good protein structures to include in a study is obviously important. Choosing a representative set of target types is a topic that has been addressed by a number of publications [7, 9–13]. Here we will consider another aspect, namely the choice of structural model. Choice will not always be available in a design program on a new target. However enrichment studies often use well established targets, and the modern success of structural biology groups in generating protein crystal structures often leaves the modeller with a plenitude of structures to choose from. The quality of each structure needs to be assessed and, although good resolution is important, other factors also need to be considered such as fit of protein and ligand structure to the electron density in the active site region [12]. No less important is the typicality of the structure with respect to its peers. It is important that a superposition and visual comparison of all available structures should be made to ensure that one with no unusual features is selected.

Figure 2 shows all thrombin structures with resolution better than 1.9 Å superimposed on the region of the active site. This superposition was created using the program Relibase+ [14] by first searching for all structures with 100% homology to 1dwd and resolution better than 1.9 Å, and then minimum root mean square deviation (RMSD) superimposition of all Cα carbons within 7 Å of the 1dwd ligand. 1dwd, the thrombin model used in the Chen study, is highlighted including its ligand. Mobility of some active site residues can be observed. These side-chains are in unexceptional geometry in 1dwd and, on the whole 1dwd can be considered a reasonable choice of model. Figure 3 shows all sPLA2 structures superimposed on the region of the binding site. Again there is side-chain mobility. 1db4, the model used in the Chen study is highlighted including the ligand. Again this appears a reasonable model to use. Four structures have a histidine residue that impinges on the region occupied by the 1db4 ligand. These would appear to be poor choices of model if we wished to retrieve ligands similar to that of 1db4. It turns out that these are

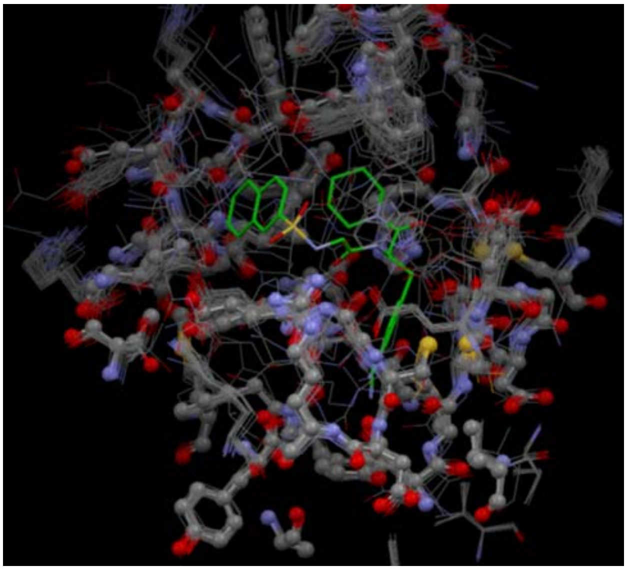


Fig. 2 Superposition of all thrombin structures deposited with resolution better than 1.9 Å (53 at time of writing) using 1dwd (ball and stick) as reference

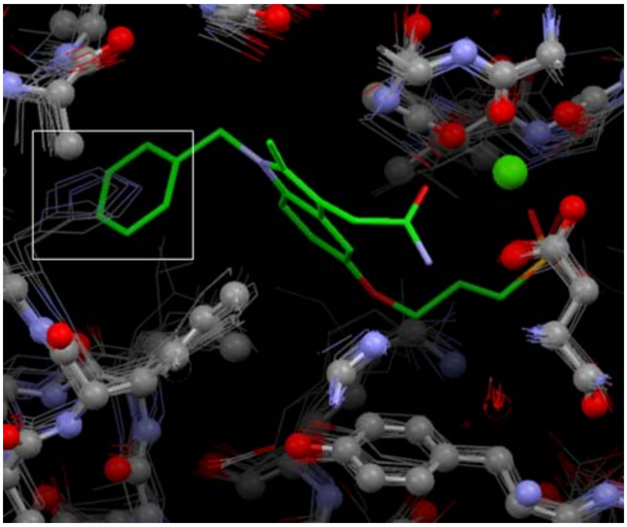


Fig. 3 Superposition of all sPLA2 structures deposited (23 at time of writing) using 1db4 (ball and stick) as reference. The inset box highlights a histidine residue folded in to the active site in four *apo* structures

apo structures. Whilst some *apo* structures may make suitable protein models for VS, side-chain collapse into the binding region needs always to be considered as a drawback.

The two examples given so far are cases where the active site is on the whole conserved in size and shape between 100% homologous structures. However some proteins can take up two distinct structural forms. Figure 4 shows all ER structures superimposed with 1err, the model used in the Chen study, highlighted. The ligand in 1err clashes badly with protein structure present in about half the superimposed models. ER is a nuclear hormone receptor and its mode of transcriptional regulation involves switching between an agonist and an antagonist form [15]. Which form is present is predicated by the bound ligand. 1err is in the antagonist form, notable for having a portion of the structure, helix12, displaced away from the binding site, exposing it to solvent. The antagonist form has a considerably larger binding site and a slightly different disposition of hydrogen bonding functionality to the agonist form. Ligands which selectively stabilise one or other form are known as agonists and antagonists. It turns out that the ER actives used in the Chen study are a roughly equal mix of both types.

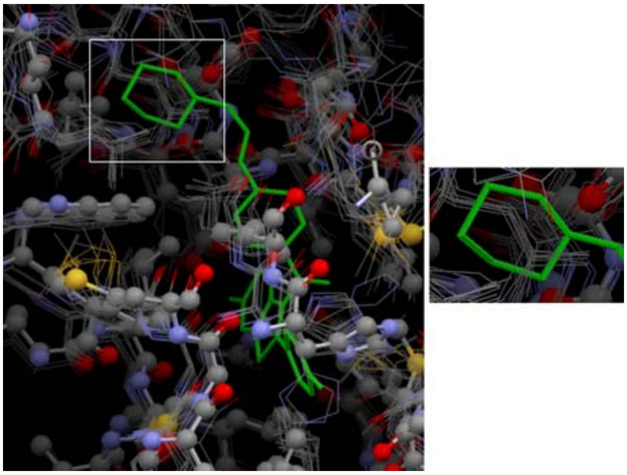


Fig. 4 Superposition of all estrogen receptor structures deposited (32 at time of writing) using 1err (ball and stick) as reference. The inset box highlights where part of the ligand of 1err coincides with helix 12 in the agonist forms of ER

Table 2 compares enrichment calculated over all ligands, using the more thorough Gold 3.0.1 protocol described earlier but with either ChemScore or GoldScore as the scoring function. One would conclude correctly that the ChemScore protocol was more useful than the GoldScore protocol at retrieving ER ligands *in toto*. What is revealed however, if enrichments are also calculated considering antagonist ligands only (also in Table 2), is that the GoldScore protocol is actually a reasonably efficient and highly selective protocol for picking out antagonist ligands using an antagonist protein model. This is a desirable characteristic. It is true that the ChemScore protocol may have greater utility in a VS campaign against ER. The key point however is that the goal of enrichment studies is to compare and characterize various docking methodologies as being able to retrieve actives cognate to a particular active site. If so the structures classified as actives must be true ligands for the form of protein model used.

The right Bulls-eye

Most docking programs require the user to define the volume constituting the desired binding site. Decisions also need to be made regarding which crystallographic waters to

Table 2 Enrichment experiment against the antagonist form of the ER illustrating the dependence of enrichment on the type of ligand analysed

Scoring function	Ligand type	Enrichment
GoldScore	Agonists + Antagonists	3.3
GoldScore	Antagonists	6.0
GoldScore	Agonists	1.1
ChemScore	Agonists + Antagonists	9.1
ChemScore	Antagonists	10
ChemScore	Agonists	8.5

GOLD 3.0.1 was used with the ‘fast’ auto-settings protocol. The top 10% of the database as ranked by the scoring function used, was examined in each case

retain and what protonation state certain active site residues should take up. When comparing docking programs, it is clearly essential to keep active site definitions as consistent as possible. Ensuring similar extent of binding site is itself not trivial as some programs force users to define binding sites in a certain way (e.g. as a box) which may be inconsistent with other programs. Other matters also merit consideration. The optimum density of grid points for some grid-based docking programs may be linked to binding site size, for instance. In a commentary on one comparison study Schulz-Gasch and Stahl stated that "... much of the differences in performance of FRED, Glide and FlexX can be attributed to [differences] in binding site definition" [16].

Water molecules in the active site present their own special problems. Again careful study of all homologous structures is advisable to identify those waters which are invariably displaced, and those which are retained and sometimes form a mediated hydrogen bond between the protein and ligand. It is well known, however, that, for certain targets, there are waters which can take up either of these roles, depending on the ligand bound. Docking programs are beginning to treat such waters dynamically [17]. However consideration of this extra layer of complexity within a general enrichment comparison, is probably not desirable.

Chen et al. in setting up the thrombin model 1dwd, chose to retain one water that sits in the S1 pocket, a binding pocket that avidly accepts protonated planar functionality such as amidines or guanidines. This water is known to provide a useful extra hydrogen bonding anchor for such moieties. However non-polar S1 binders are also now well known and these generally displace this water and gain entropically from doing so [18]. Therefore the argument previously put, that the active ligand list must be appropriate to the protein model, applies here as well. The decisions made in regard to which waters are retained also may impact which ligands should be chosen as 'active' for that target and vice versa. Detailed knowledge of the structure-binding relationships that pertain to the target is required for this assessment.

Inspection reveals that, by and large, only a small proportion of ligands selected as thrombin actives in the Chen study potentially suffer from this issue. However examples can be found. Figure 5a shows the GOLD binding mode and score for one such ligand which almost certainly displaces the S1 water in reality. Most highly active thrombin inhibitors fill the S1 (Asp 189, Tyr 228), S2 and S3 pockets. However in this case a high quality binding mode cannot be achieved as the S2 pocket is not filled; and the score attained by this structure is not high enough to promote it into the top 10%. Figure 5b illustrates the considerably improved binding mode and score achieved using a GOLD protocol which can make the choice to exclude this water molecule or treat it as part of the binding site [19]. GOLD excludes the water and finds a more realistic binding pose.

The right Darts

An enrichment study requires, for each target protein, a set of actives embedded within a larger set of decoy molecules. There is much debate as to how closely the decoys should resemble the actives and what the correct ratio of actives to decoys should be. It is not intended to add to that

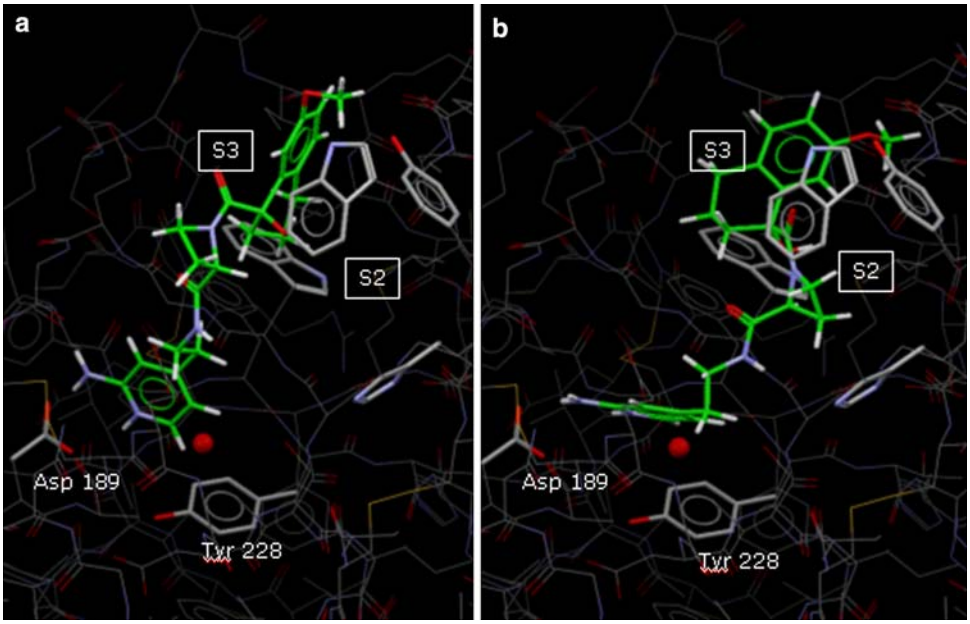


Fig. 5 (**a**) Docked pose for a ligand that in practice is likely to displace the displayed water molecule in the S1 pocket. This pose has a GoldScore of 46.2, obtained with the water kept in place. (**b**) Docked pose for the same ligand, this time allowing the water to be ignored if in so doing a better score can be attained. This pose has a GoldScore of 54.8. The ligand and water overlap the same space

debate here. Instead we will look at the difficulties inherent in generating a set of homogenous and essentially correct 3D structures for a set of ligands. One principle to abide by rigorously is to treat the actives and decoys in an identical fashion. This is obvious but nevertheless isn't always adhered to.

The file of 3D structures will normally be created from a line notation such as SMILES, or from a list of 2D structures. State of the art software is commercially available for this purpose. Some differences will be found in the exact geometries generated by these programs, so it is important to stick to one methodology throughout. Figure 6 illustrates the sort of problem that is sometimes encountered. Corina 3.1 generates aryloxyaryl bond angles of 106.8° whereas an analysis of the Cambridge Structural Database suggests a range of 117–124°. Comparison of GOLD results using crystallographic ligand geometry and Corina generated geometries indicate a degradation of performance at regenerating the binding mode in cases where the Corina geometries are in error [R.Taylor (Cambridge Crystallographic Data Centre, CCDC), P.Mortensen (Astex), unpublished results]. A recent VS comparison between three docking programs suggested that Corina generated geometries lead to lower hit rates than the geometries of the same molecules taken from the PDB [20].

Protonation state, tautomeric form and ring conformation are three additional factors that need to be considered. Getting these right can be crucial not least because some docking protocols may be less or more sensitive to accuracy in this regard. The simplest option is to assume one most likely form for active and decoy and to stick with that. However this presupposes that the correct form is known for each active, in context of the binding conformation (getting the decoys wrong is not such an issue). Such an assessment requires a modeller highly knowledgeable in the structure-binding mode relationship and, even so, grey areas may be encountered where the correct state is not known with absolute certainty. So it is a difficult problem. The accepted and commonly used way of avoiding this issue is to calculate all plausible protonation, tautomer and ring conformation states for both actives and decoys [21]. This was the approach taken by Chen et al. The decoy and active sets comprise over 20,000 molecules but over 33,000 structures. One drawback to this approach is that it complicates the analysis of the results. A second, and far from insignificant problem, is that unless actives and decoys are very similar in complexity, actives may sometimes be represented by significantly more structures than decoys. The reason for this is that if one molecule in an active set is characterised by possible multiple forms then there is likely to be a correlation in the number of forms for the other molecules in the set. This is likely to be especially true if there is a lot of 2D similarity in the set of active structures. This may give those programs that are less discriminatory an unfair advantage, since the actives are effectively being allowed several docking runs to find an acceptable binding mode rather than a single docking run. If short VS type protocols are used, this becomes important, as finding the highest possible scoring pose cannot necessarily be guaranteed for complex molecules in a single run, and it is a legitimate aim of an enrichment study to take into account the speed of the docking program in finding a good high scoring pose rapidly. One check that could be done is to see if the ratio of active ligand structures to active ligands is similar to the ratio of decoy ligand structures to decoy ligands. The ratio of ligand structures to ligands for the four targets of the Chen study, are respectively, thrombin: 5.5, ER: 1.5, COX2: 1.3 and sPLA2: 1.0. The ratio of inactive structures to inactives is 1.65. Therefore, for one target, thrombin, it appears that the ratio of structural forms is highly biased towards the actives. Consequently there is a possibility that this might artificially improve enrichment rates for some programs.

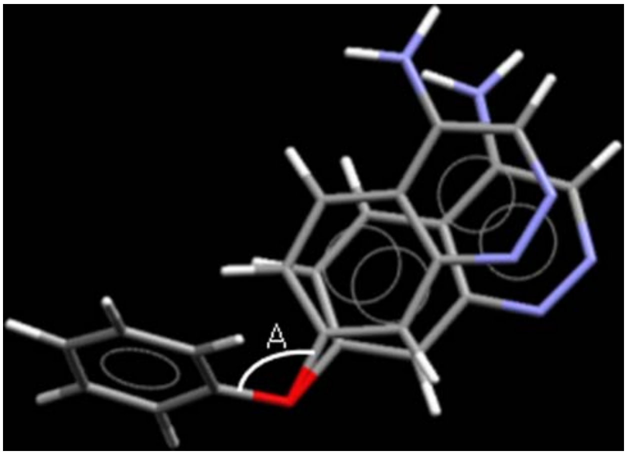


Fig. 6 Comparison of Corina 3.1 generated structure with similar structure from the Cambridge Structural Database. Angle A is 106.8° in the Corina generated structure but 117–124° in CSD structures

Structure generation of large VS sets is necessarily a process that has no manual checking component. However things can go wrong which are hard to pick up. Figure 7 illustrates a thrombin ligand from the Chen set. This thrombin ligand should contain an amidine functionality, an important binding motif for thrombin and many other serine proteases. However two extra hydrogens have been added and this portion of the molecule is incorrectly non-planar. The wire-frame diagram gives an indication of the correct 3D representation of benzamidine. Forty percentage of the thrombin ligands in this set are similarly incorrect. Clearly even if these ligands are usually docked in approximately the correct pose, as is found for GOLD, the fitness score obtained for them is unlikely to truly represent their binding ability relative to other structures in the set.

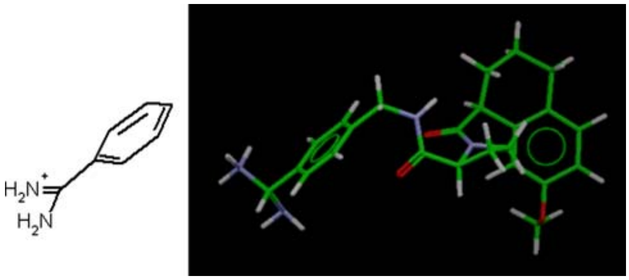


Fig. 7 Benzamidine with two extra H's and non-planar geometry found in 40% of thrombin actives of a published dataset [6]. The correct structure and geometry are illustrated in the 2D diagram

Structural misrepresentation of actives appears to be a common problem. For instance it was found in a study using the DUD [22] factor Xa active set that three of the amidine actives had precisely the same structural problem as those in the Chen set and a further ten other structures also needed correction (9% of the total number of structures). In real VS exercises it may not be time efficient to ensure absolute 3D integrity of all structures. However, when the aim is to compare different docking methodologies the care required to generate structure files of high integrity cannot be overstated.

Some would advocate incorporating 3D structure generation and protonation state, tautomer and ring conformation assignment as part of the docking procedure. This has the advantage that it allows no impediment to starting with the simplest form of molecular representation, with benefit to the user. It also would remove another source of bias within enrichment experiments. However in counter, it surely makes sense to use the most advanced purpose built software if realistic geometries, protonation states etc., are to be achieved reliably. This may not be possible with less sophisticated methods incorporated within the docking program. An alternative approach would be to have a way of automatically and accurately checking geometries in a large file and excluding those of poor geometry prior to docking. This appears an attractive idea and could be applied as part of a standard set-up prior to docking. Such software is not currently openly available as far as this author knows.

The right Box of Tricks

Choice of appropriate docking and analysis protocol is a potential minefield. Docking packages are becoming ever more advanced and it is normal that new options become available on each release. Therefore the best practice could easily change for a given program over time. Choosing protocols therefore requires care, attention and experiment. To do this fairly in the sometimes limited time available for the work, may be hard. Protocol choice must be fair across all different packages assessed, a difficult task unless a number of different program experts are at hand. Additionally it is well known that protocol choice is target specific. Different scoring functions may perform better on one target than another and longer docking times may be appropriate for some targets over others. Again some experimentation may be required to establish a good general protocol.

Going further, modellers experienced in working with some targets may wish to apply knowledge based tweaks that they know will give good results for that target (e.g. constraining that a certain hydrogen bond be present, or that a particular pocket be filled). Such tweaks may be invaluable when using a program in a drug discovery situation. However they are questionable when applied in a docking comparison unless they can be applied consistently for all targets and programs. This will be very hard to achieve. One study that comes close is that by Warren et al. [5] where a key goal was to use both program experts and target experts to ensure as much comparability as possible. Even though this was the laudable aim it was apparently not possible to always achieve equivalent expertise across all modellers who took part in the study [23].

The results in Table 1 illustrate how protocol design may affect results. The GOLD 2.2 protocol using library settings could be considered a best practice VS screening protocol for GOLD up to the release of GOLD 3.0.1. The overall improvement in results obtained using the GOLD 3.0.1 auto settings protocol is due to implementation of a feature that allows GOLD to spend more time docking large flexible molecules than small inflexible ones. This effectively allows the VS protocol to be kept short, whilst giving big structures a good chance. The actives for sPLA2 and thrombin are predominantly large and flexible, hence 'improvement' in performance is most noted for these. The GOLD 3.0.1 protocol developed from work in house and can now be considered a current candidate for best practice VS protocol for GOLD. However, without doing their own in depth protocol optimisation this wouldn't have been obvious to an evaluator until such time as communicated to by the software provider.

This highlights the point that sometimes software companies are not as proactive as they should be in providing guidelines for best VS practice. However it is also behoves the evaluators to check with software providers that their protocols are suitable for the job in hand. If this is done then the software provider can have little complaint as to the outcome of the evaluation.

Overall then, protocol choice is possibly one of the most difficult aspects of a docking comparison to get right. Getting as much assistance from the software suppliers, perhaps to the extent of asking for 'black box' protocols standardised over all targets, would at least remove some of this burden from the evaluator. This is not to say that comparison studies using highly tweaked protocols don't have a significant role to play. However published results using 'black box' protocols will be of significant benefit as a starting point to those new to the field or those who have little time and effort available to fully optimise their own protocols. The comparison of these results with the tweaked protocols then gives a measure of what advantage target specific protocols may generate.

The right Glasses

The correct tools to use in the analysis of enrichment experiments form another area about which debate rages. It is not the intent here to add greatly to that debate, though one or two comments will be made. Another issue entirely is whether any single simple numerical method can truly quantify the effectiveness of a particular docking protocol in what is a simulation of a true VS situation.

The 'standard' tool for measuring enrichment is enrichment factor, as used in this study. The ratio is calculated of the number of actives retrieved in a specified top *n*% of the database to the number of actives expected at random. This methodology has two advantages and rather more disadvantages. One advantage is simplicity, the other is that it is easily relatable to actual VS practice. The principle disadvantages are four-fold. First, considerable variation may be experienced in numbers calculated for each run, especially if the number of actives included is low. So repetition of the experiments is required. Second, enrichment factors are highly sensitive to the ratio of actives and decoys, and this means it is hard to compare results obtained using different ligand sets. Third a decision needs to be made as to where to the cut-point in the database. This is not always obvious. In practice it might be necessary to select several cut-points (e.g. top 1% and top 10%) to attempt to gain a fuller picture of the enrichment profile. Fourth, no measure is made of the diversity of active ligands captured. However a good list of retrieved actives should display as wide diversity as possible. In fact arguably diversity of hits is more important practically speaking than number of hits retrieved from a screen. The importance of avoiding existing patent art, the desirability to bypass difficult synthetic chemistry and the possibility of unwelcome adverse absorption, distribution, metabolism, excretion (ADME) issues arising in any given series, all demand that as many unique starting points for lead optimisation programs be found. In addition the modeller may well search in the dataset for similar ligands to those found by VS, thereby uncovering actives not picked out immediately by the docking program. A docking methodology which captures some active types well but rates others poorly is therefore not as good as one exhibiting similar or perhaps even poorer enrichment factors but capturing a wider diversity of actives.

Area under the received operating characteristic (ROC) curve is a metric used in other fields of study such as engineering and medicine, and is now being adopted for enrichment studies. This metric has advantages over enrichment factor because it is holistic, independent of active/decoy ratio and generally exhibits less variation. It can be argued that a disadvantage is that it doesn't say much about the sharp end of the curve (i.e. the top 5%) which is where most drug designers want to be. This has lead to the development of metrics that specifically address the 'Early recognition problem' [24, 25]. However an interesting point arises. An ROC curve may for example be very steep at the beginning but plateau later on, before getting steeper later (Fig. 8, curve A). It might be argued that such a curve is more desirable than one which maintains a steady rise but has the same area under curve (AUC) (Fig. 8, curve B), as better enrichment factors are found at the sharp part of the curve. This may be so. However if one asks why the ROC curve should have a shape such as A, a doubt arises. The most likely reason a curve such as A would be generated is that the docking protocol is good at picking out a certain class of active but deficient at picking out a second class. If so, it is not at all clear that the protocol generating curve A is any better than that generating curve B. If the second protocol is successfully finding both classes of hits at low % cut off then arguably this is the best protocol.

The arguments presented above suggest that a metric that looks at the diversity of actives retrieved is also worth calculating. In the Warren study a measure was made of the percent of score-ordered structures that had to be counted before at least one of all active types had been retrieved [5]. This appears to be a laudable attempt to assess the value of a screening protocol in way other than a figure measuring

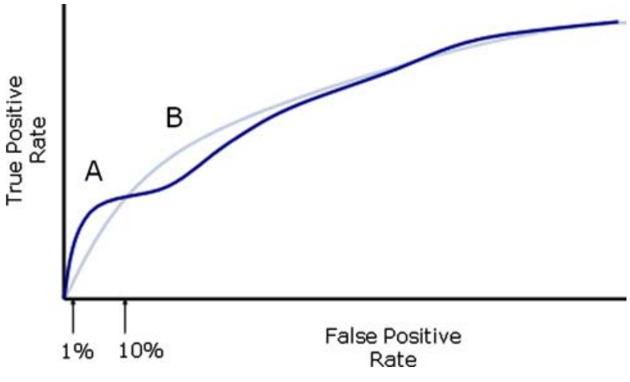


Fig. 8 Comparison of two hypothetical receiver operator characteristic curves with similar area under the curve

whole enrichment. Other measures of diversity could be used also. When diversity of structure is discussed in the context of two of the three areas described above, patent cover, chemistry (and even to a certain extent the third, ADME), it is usually 2D diversity that is being considered and, more particularly, it is often diversity of scaffold construction that is important. Therefore a case could be made for using as a suitable metric for diversity of hit retrieval, the reduced graph set clustering method presented by Barker et al. [26] and used by Good [27] to evaluate diversity in the DUD [22] and WOMBAT [28] datasets.

Enrichment factors and AUC's say something about the retrieval of active molecules. However they do not say anything about whether the poses obtained for the retrieved actives are reasonable or not. This seems to be an important issue that has not been well addressed so far. One reason for this of course is that, without crystal structures available for each active, no binding modes can be defined with certainty. However it is certainly true that some high scoring but clearly wrong binding poses can be normally be identified for any enrichment experiment. For example an analysis has been carried out on the poses of actives found in the top 10% of the dataset in the thrombin enrichment studies quoted in Table 1. Only those thrombin ligands that did not have incorrect amidine or guanidine representations were looked at. Poses were classed by eye as incorrect if they didn't make good H-bonding interactions in the S1 pocket or significantly failed to fill S2 or S4 pockets in the manner expected from crystallography of related protein/ligand complexes. About 26% of the ligands selected as 'active' were not represented by any 'correct' pose for the results obtained using GOLD 2.2. The figure was 19% for those poses selected using the more efficient protocol within GOLD 3.0.1. Should such ligands be classed as 'successes' in an enrichment study? Arguably not.

Conclusions

Comparative evaluations of docking methodologies and protocols should usefully inform the experimenter who wishes to employ structure-based VS to find novel bioactive leads. However many precautions need to be taken to ensure that these evaluations are meaningful. Care needs to taken in choice and set up of the protein structures employed in such a study. Where a number of protein structures are available for a given target, it is important to do a rigorous structural comparison. Then a structure should be chosen that is not atypical in its conformation. *Apo* structures in particular need to be considered with care. Another case is that of a protein with two functional forms. It is important to clearly delineate which one is being used. In such a case it is important that the active ligands selected for the study are compatible with the protein structural form used (for example use only antagonists with an antagonist protein model). A more subtle issue is inclusion of active site waters in the protein model. Some waters are known to be displaced by certain ligands and not by others, and it is advisable to use such knowledge if available, to either leave out such waters, or alternatively, to tailor the ligand list to exclude those that displace water.

Obtaining good initial geometries for the structures making up the ligand + decoy set is crucial and it should be noted that some published sets of structures contain examples of poor ligand structures which no docking package can be expected to dock correctly. It is common to generate multiple structures to represent different tautomers, ring conformers and protonation states. However it is important to ensure that some ligand sets don't become over- (or even under-) represented by such an approach.

Choice of docking protocol is often a problem area. It is suggested that standard 'black box' protocols deemed acceptable by software providers should be used to provide a baseline benchmark of a program's performance. The effects of additional tweaks and biases introduced by an experienced user for a particular target, are worth knowing about, but have greatest value when published alongside those obtained with a standard protocol.

Lastly, most measures of success in enrichment studies examine only how well actives are concentrated in the higher ranking docking poses of the study. Whilst this seems an obvious parameter to choose, it ignores the assertion made here, namely that diversity of hits retrieved is more important practically speaking, than number of hits retrieved. Therefore, it may be very useful to assess the diversity of hits obtained by each docking protocol, in addition to measure success of enrichment. The consequence of this may be to alter the viewpoint as to what methodology or protocol is, for a given case, the best for the task of finding new bioactive leads.

Acknowledgements The author thanks Willem Nissink and Noel O'Boyle for assistance in setting up the docking studies presented here, Hongming Chen of AstraZeneca for supplying structure files and protocols for these studies and Richard Sykes for the GoldMine program used to carry out the analysis. Robin Taylor and Jason Cole are thanked for making perceptive and useful comments on the manuscript, and Simon Bowden, Marcel Verdonk, Chris Murray and Paul Mortensen are thanked for useful discussions.

References

1. Cole JC, Murray CW, Nissink JWM, Taylor RD, Taylor R (2005) Proteins 60:325
2. Kellenberger E, Rodrigo J, Muller P, Rognan D (2004) Proteins 57:225
3. Perola E, Walters WP, Charifson P (2007) J Chem Inf Model 47:251

4. Kontoyianni M, McClellan LM, Sokol GS (2004) J Med Chem 47:558
5. Warren GL, Andrews CW, Capelli A-M, Clarke B, LaLonde J, Lamber MH, Lindvall M, Nevins N, Semus S, Senger S, Tedesco G, Wall ID, Woolven JM, Peishoff CE, Head MS (2006) J Med Chem 49:5912
6. Chen H, Lyne PD, Giordanetto F, Lovell T, Li J (2006) J Chem Inf Model 46:401
7. Perola E, Walters SP, Charifson PS (2004) Proteins 56:235
8. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig I, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28:235
9. Jones G, Willett P, Glen R, Leach AR, Taylor R (1997) J Mol Biol 267:727
10. Kramer B, Rarey M, Lengauer T (1999) Proteins 37:22
11. Nissink JWM, Murray CW, Hartshorn MJ, Verdonk ML, Cole JC, Taylor R (2002) Proteins 49:457
12. Hartshorn MJ, Verdonk ML, Chessari G, Brewerton SC, Mooij WTM, Mortensen PN, Murray CW (2007) J Med Chem 50:726
13. Friesner RA, Murphy RB, Repasky MP, Frye LL, Greenwood JR, Halgren TA, Sanschagrin PC, Mainz DT (2006) J Med Chem 49:6177
14. Hendlich M, Bergner A, Günther J, Klebe K (2003) J Mol Biol 326:607
15. Pike ACW, Brzozowski AM, Walton J, Hubbard RE, Thorsell A-G, Li Y-L, Gustafsson J-A, Carlqusit M (2001) Structure 9(2):145
16. Schulz-Gasch TA, Stahl M (2003) J Mol Model 9:47
17. Verdonk ML, Chessari G, Cole JC, Hartshorn M, Murray CW, Nissink JWM, Taylor RD, Taylor R (2007) J Med Chem 50:726
18. Tucker TJ, Brady SF, Lumma WC, Lewis SD, Gardell SJ, Naylor-Olsen AM, Yan Y, Sisko JT, Stauffer KJ, Lucas BJ, Lynch JJ, Cook JJ, Stranieri MT, Holahan MA, Lyle EA, Baskin EP, Chen I-W, Dancheck KB, Krueger JA, Cooper CM, Vacca JP (1998) J Med Chem 41:3210
19. Verdonk ML, Chessari G, Cole JC, Hartshorn MC, Murray CW, Nissink JWM, Taylor RD, Taylor R (2005) J Med Chem 48:6504
20. Onodera K, Satou K, Hirota H. (2007) J Chem Inf Model 47:1609
21. Taken from I. Dramburg et al from http://cisrg.shef.ac.uk/shef2004/abstracts.htm
22. Huang N, Shoichet BK, Irwin JJ (2006) J Med Chem 49(23):6789
23. Communicated in a seminar by M. Nevins, 234th ACS National Meeting, COMP 150
24. Truchon J-F, Bayly CI (2007) J Chem Inf Model 47:488
25. Sheridan RP, Singh SB, Fluder EM, Kearsley SJ (2001) J Chem Inf Comput Sci 41:1395
26. Barker EJ, Gardiner EJ, Gillet VJ, Kitts P, Morris J (2003) J Chem Inf Comput Sci 43:346
27. Good AC (2007) 234th ACS meeting 2007, Abs 266
28. Olah M, Mracec M, Ostopovici L, Rad R, Bora A, Hadaruga N, Olah I, Banda M, Simon Z, Mracec M, Oprea TI (2004) In: Oprea TI (eds) Chemoinformatics in drug discovery. Wiley-VCH, New York, p 223