GOLD scoring function performance against the Astex Diverse set of protein-ligand complexes: side-by-side comparison with other docking programs

Aim

To identify which GOLD scoring function performs best for pose prediction and to compare GOLD's performance against other state-of-the art docking packages.

Introduction

This work arose out of the comparative docking experiment that was carried out by many suppliers of docking programs and presented at the Docking and Scoring Symposium at the 241st ACS Meeting in Anaheim. We summarise here the part of the experiment devoted to analysing pose prediction. The test set provided was broadly based on the Astex Diverse Set (ADS)[1]. Protein and ligand files were prepared by the organisers with the intention that all participants would use these files without further modification. This experiment provided an opportunity to assess all the GOLD scoring functions against each other and against other commonly used docking packages.

Method

Protein and ligand structures were supplied for all 85 Astex Diverse targets. In some 38 cases there are multiple sites of binding for the ligand. In such cases docking was carried out to all such binding sites even where such a site would not be considered the 'normal' binding site for the ligand. 25 protein structures were found to have either a) alternate conformations for the ligand or for side chains near the ligand (11), b) crystal packing interaction with the ligand c) incomplete density for the ligand (4). Structures where these problems occurred were put on a 'rejected list'. The remaining 60 structures were designated the 'white list'.

Docking was carried out using GOLD 5.0.1. Binding sites were defined as being 6Å (standard default) from the ligand in the supplied protein structure. No water was present in any binding site. The default docking protocol was applied (1.0x auto Settings, 10 GA) and the best pose saved. Each experiment was then repeated 25X. This protocol was repeated for all four scoring functions GoldScore, ChemScore, ASP and ChemPLP.

.

Results

Table 1 shows the relative pose prediction performance of GOLD for the 4 scoring function over the entire set of proteins. These figures are calculated using only the best scored pose out of the 25 saved for each binding site. We present two sets of results; namely, those that treat each binding site in each protein as a separate observation (All Sites) and results which give the best possible result we could have obtained by selecting one binding site for docking. The latter results are broadly comparable to the results obtained by Verdonk et al [1]. In that study, the authors selected a binding site by visual inspection: if there were no obvious reasons for selecting one site over another, then the first site in the structure was used, but in cases where there were obvious reasons to choose one site over another an intelligent choice was made. The recent scoring function, ChemPLP [2], appears to be the most successful scoring function with a top-ranked success rate over all sites of 81%. Also notable are the results obtained at the far tighter threshold of 1.0A. ChemPLP achieves a success rate of 59% at this threshold for all sites and 68% for the best site. This is significantly better than the other scoring functions provided with GOLD, and as such further analysis in this work focuses on results achieved with this scoring function.

	All Sites				Best Sites Only			
	Top Ranked		Closest		Top Ranked		Closest	
	2.0 Å	1.0 Å	2.0 Å	1.0 Å	2.0 Å	1.0 Å	2.0 Å	1.0 Å
ChemPLP	81%	59%	91%	76%	87%	68%	93%	80%
GoldScore	69%	50%	82%	68%	78%	58%	88%	74%
ChemScore	76%	48%	87%	66%	82%	55%	91%	74%
ASP	72%	44%	86%	61%	79%	53%	89%	71%

Table 1 - Success rates for GOLD with the four scoring functions.

Pose prediction performance for ChemPLP is shown in table 2 for the 'white list' structures versus 'black list' structures and for the best site. Numerically, it appears that the 'rejected' list out performs the 'white list' at the 2.0Å threshold, but it is questionable whether this is significant within the bounds of uncertainty, as the rejected list contains just 25 structures (43 sites): the difference is just 5%, which equates to only 2 sites.

	All Sites				Best Sites Only			
	Top Ranked		Closest		Top Ranked		Closest	
	2.0	1.0	2.0	1.0	2.0	1.0	2.0	1.0
White list	79	66	92	77	85	69	92	78
Rejected list	84	42	86	72	92	64	96	84

Table 2 - Comparison of pose prediction success with ChemPLP for Black List and White list structures.

At the 1Å cut off the success rate for top ranked poses is better for the 'white list' structures. This is particularly true when all binding sites are considered. This is a more expected result: we would expect the difference between 'white list' and 'rejected' to become clearer at the lower cut-off. Nevertheless not too much should be read into this results.

The large number of dockings performed using different scoring functions provides us with a window into the overall sampling performance of GOLD using the 25 repeats for the different scoring functions.

The percentage of poses generated that were within 2.0A with each scoring function was recorded. The average retrieval rates for the 4 scoring functions in gold are given in Table 3. As is apparent, ChemPLP significantly out-performs the other 3 scoring functions for pose prediction, retrieving, on average, 76.6% of poses within 2.0A of the correct answer for the complete test set. For 'Challenging' structures a slightly larger difference is observed: Challenging structures are defined as those where at least one scoring function fails to achieve a 100% success rate for pose retrieval.

	% of poses within 2.0A RMSD	
	All Structures	Challenging Structures
ChemPLP	77%	59%
GoldScore	70%	48%
ChemScore	71%	50%
ASP	68%	44%

Table 3 - Sampling analysis for each scoring function.

ChemPLP shows a 68% success rate at retrieving structures within 1.0Å of the correct structure for the best site and is ranked the best scoring function in this regard. It might be asked whether a better scoring function might improve on this value. However there are limits to how much improvement is possible without resorting to deliberate bias. Firstly water has been deliberately excluded from all structures. This is because inclusion of many waters makes pose reproduction an almost trivial exercise as demonstrated by Verdonk et al 1. Conversely though, leaving all the waters out as has been done here, inevitably creates difficulties in retrieving precisely the right pose. Docking into 1XM6 with ChemPLP, results in a best pose with RMSD 1.2 Å. The difference between the docked pose and the correct pose is shown in figure 1. A water mediated interaction causes the imidazoline ring to rotate to the right in the correct structure whereas it is found rotated left in the docked structure. Solvent mediated contacts are likely to cause similar issues in 3 other complexes where the 1.0Å threshold is not achieved (5% in total).

Docking failures due to near-symmetry between two possible binding modes can also occur (Figure 2). In these cases the incorrect binding mode should probably be considered reasonable binding modes. This occurs in two structures 1jje and 1q41 (2% of total).

Lastly three other failures for ChemPLP (4%) are on the 'rejected' list of protein structures and there is therefore some doubt as to the 'correct' ligand binding mode.

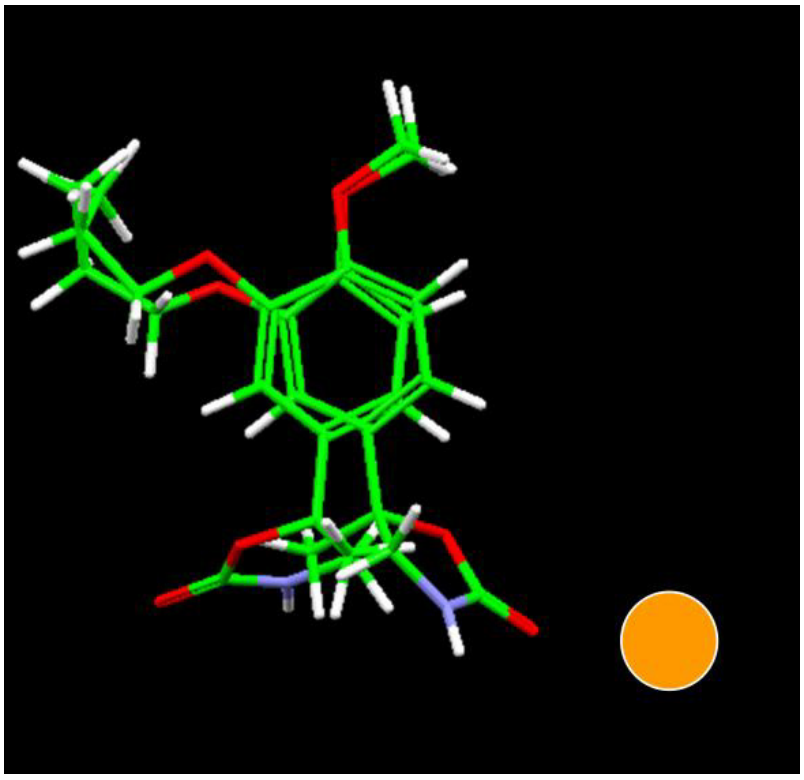

Figure 1 - Docked and experimental binding modes in 1xm6. The orange circle represent position of a water which mediates an interaction with the ligand in the experimental structure.

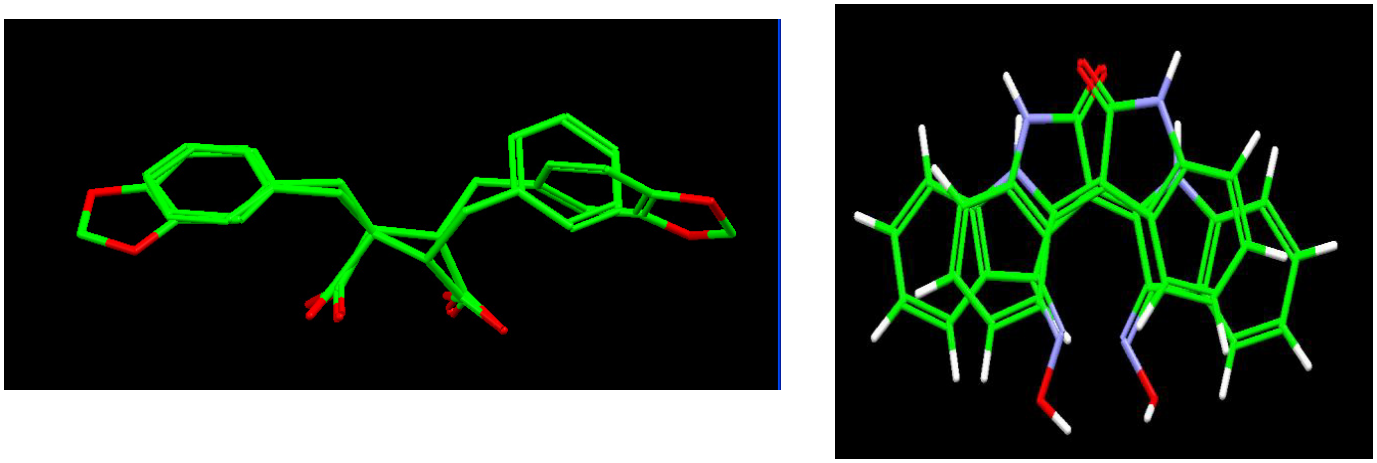

Figure 2 - Docked and experimental binding modes for 1jje and 1q41. These symmetry related binding modes are reasonable and are difficult to separate.

Figure 3 compares the performance of GOLD+ChemPLP with other docking programs in this comparative experiment (Raw data supplied by G. Warren, Open Eye). Data is shown for top-ranking poses only. GOLD performance is excellent compared to other programs. This is encouraging especially as binding site sizes were defined as no less than default size and standard length docking protocols were employed.

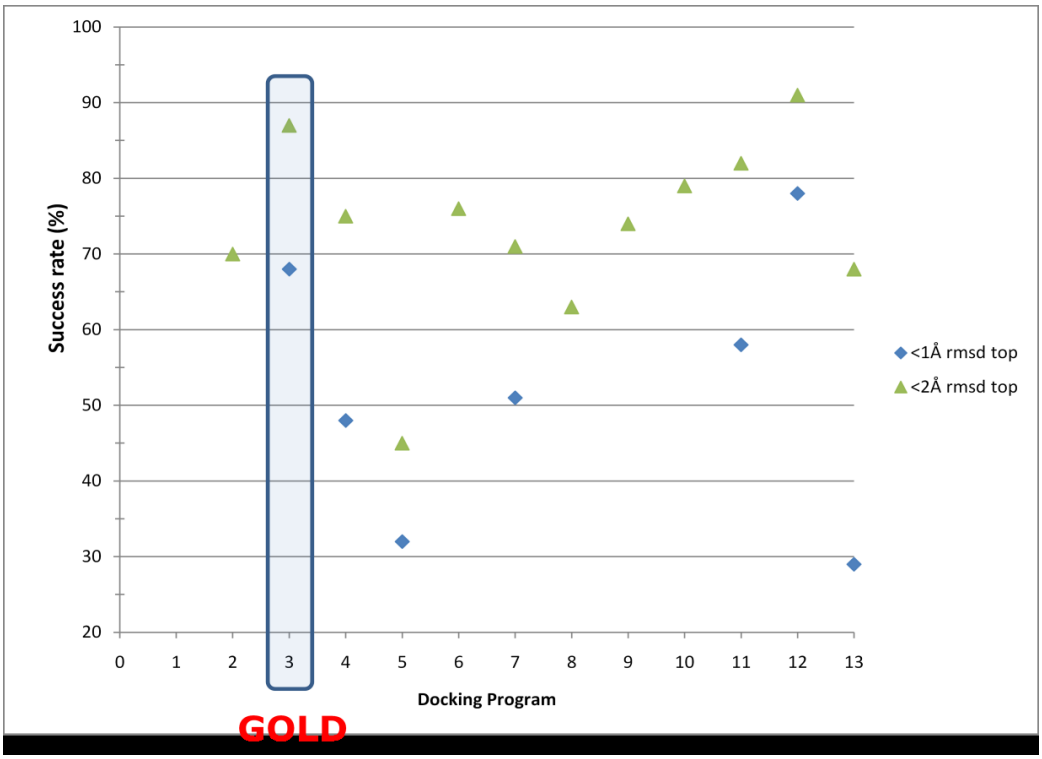

Figure 3 - Pose prediction performance of GOLD alongside other programs. Note: Docking program 12 used a binding site definition of 4Å around the ligand.

Conclusions

This study has demonstrated that our most recently introduced scoring function, ChemPLP2 is the most effective scoring function for pose prediction in cognate protein-ligand complexes, out of the four available in GOLD. Although we should be cautious in suggesting that ChemPLP will give the best results no matter what the protein target is, we recommend that ChemPLP should be the scoring function of choice for pose prediction given no further information is available.

These results also confirm that GOLD remains one of the most accurate docking programs available for prediction of binding poses in protein/ligand complexes.

A separate use case illustrates the performance of the four GOLD scoring functions in the virtual screening part of the same comparative experiment. ChemPLP is again shown to be the best all round scoring function of the four available.

References

1. Diverse, High-Quality Test Set for the Validation of Protein-Ligand Docking Performance, M. J. Hartshorn, M. L. Verdonk, G. Chessari, S. C. Brewerton, W. T. M. Mooij, P. N. Mortenson, C. W. Murray, J. Med. Chem., 50, 726-741, 2007.

2. Empirical Scoring Functions for advanced Protein-Ligand Docking with PLANTS, O. Korb, T Stützle, T. E. Exner, J. Chem, Inf. Mod., 49, 84-96, 2009.

Products

CSD – the world's only comprehensive, fully curated database of crystal structures, containing over 1,000,000 entries

Hermes – CCDC's life science visualiser, used by GOLD, GoldMine, Relibase+ and SuperStar

GOLD – an accurate and reliable protein-ligand docking program

For further information please contact:
Cambridge Crystallographic Data Centre, 12 Union Road,
Cambridge CB2 1EZ, UK. Tel: +44 1223 336408, Fax: +44 1223 336033,

Email:admin@ccdc.cam.ac.uk.