GOLD scoring function performance against the DUD decoy/active set

Aims

- To identify which scoring function performs best overall for Virtual Screening applications, and which may be best applied to certain target classes.

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, 2011. We summarise here the part of the experiment devoted to analysing virtual screening. The DUD Active/Decoy[1] set of forty diverse targets was chosen by the organisers of the experiment to be the test bench for evaluating virtual screening performance. 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

Binding sites were defined as being 6Å (standard default) from the ligand in the supplied protein structure unless it was deemed that a larger binding site needed to be defined. Binding site definition was set at 6Å for all proteins in an additional experiment for one scoring function (ChemPLP)[2]. Water molecules in the active site, if deemed they could be involved in ligand binding, were allowed to optimize their hydrogen positions. For the three targets dhfr, gpb and hsp90 it was deemed that some waters might be displaced on binding of certain ligands and these were allowed to toggle on or off during docking and, if 'on', were allowed to optimize hydrogen positions and move up to 1Å from starting positions.

50% search efficiency settings were used for docking, repeated ten times and the top pose in each case used in the enrichment calculations. This is a relatively slow protocol for virtual screening purposes so the experiment was repeated for one scoring function (ChemPLP) using 10% search efficiency.

Virtual screening experiments were carried out using all four GOLD scoring functions, GoldScore, ChemScore, Astex Statistical Potential (ASP) and ChemPLP, the most recent scoring function to be introduced.

The total area under the Receiver Operating Characteristic (ROC) curve was used as the primary measure of enrichment success. In addition, early enrichment was measured by calculating an enrichment factor over the top 0.1%, 1% and 2% of the database as ranked by score of the top-ranked pose in each case.

AUCs and enrichment metrics were averaged over all the DUD sets to assess which scoring functions had best overall performance. In addition the DUD sets were divided into the constituent target classes (*nuclear hormone receptors* (8 members), *kinases* (9 members), *serine proteases* (3 members), *metalloproteases* (4 members), *folate enzymes* (2 members), and *others* (14 members) and the enrichment metrics calculated for these families separately.

Results

Figure 1 shows the histogram of AUC values for all scoring functions over all targets. It is clear there is considerable variation in enrichment success over the different targets with extremely good enrichment achieved for some targets and no enrichment (AUC ≤ 0.5) for others. Although poor success for a given target may be because of a poor docking and scoring protocol it is also possible that a poor choice of protein structure where protein mobility is likely; and incorrectly prepared ligands, may also contribute to poor enrichment. Other docking programs that were tested in exactly the same test set were found to show a similar distribution of poor and good enrichment.

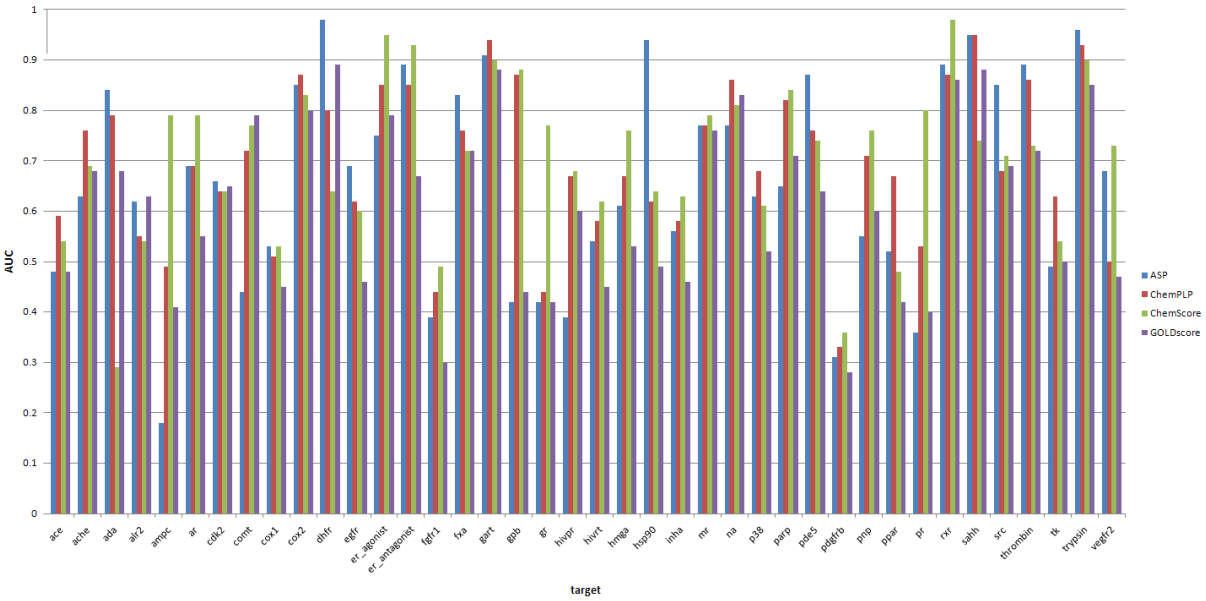

Figure 1 - ROC AUC's for all 40 DUD active/decoy sets, over the four GOLD scoring functions.

For many targets the performance is similar over all four scoring functions. However there are some cases where one scoring function performs better than the others.

Table 1 shows the average enrichment metrics for all four scoring functions and early enrichment AUC metrics for different fractions of the false positive rate. Figure 2 gives the histogram for the same figures including standard deviations. Table 2 shows average enrichment factors for the top 0.1, 1 and 2% of the ranked datasets.

% FPR	**0.1**	**1**	**2**	**100**
ASP	0.05	0.11	0.15	**0.66**
ChemPLP	*0.08*	*0.14*	*0.17*	**0.70**
ChemScore	0.03	0.08	0.12	**0.70**
GoldScore	0.06	0.11	0.14	**0.61**

Table 1 - Average AUC of the ROC curve for different fractions of False Positive Rate (An AUC at an FPR of 2% represents the portion of the ranked active-decoy list that contains the top 2% of the total number of decoys).

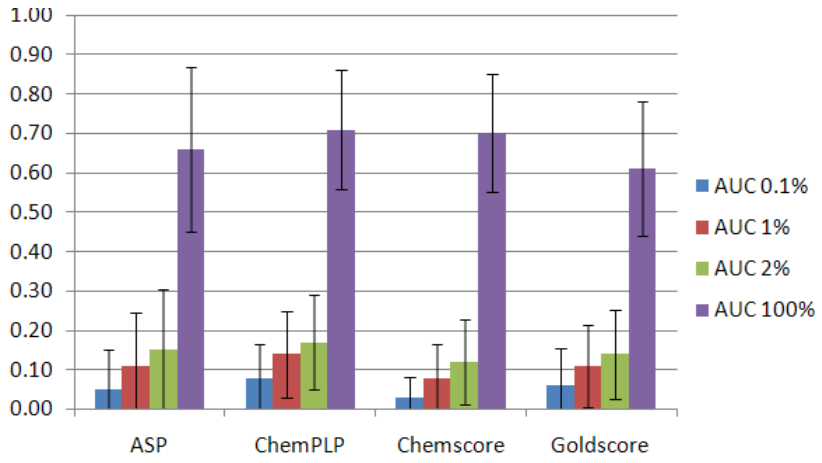

Figure 2 - Average AUC of the ROC curve for different fractions of False Positive Rate. Error bars are standard deviations.

Enrichment Factor	0.1	1	2
Max	31	31	31
ASP	13	9	8
ChemPLP	*20*	*12.5*	*9*
ChemScore	10.5	8.6	7.1
GoldScore	17	10.5	8

Table 2 - Enrichment factors calculated for the top 0.1, 1 and 2% of ranked active-decoy lists.

These results clearly suggest that, overall for this dataset, ChemPLP is the strongest performing scoring function for virtual screening. This is especially true if we examine the early enrichment metrics.

We can break these results down into the different target classes available within the DUD set of 40 proteins. However when doing so it is essential to bear in mind the sample sizes may be small in some cases. Figures 3-7 show the average early enrichment (at 2% FPR), and full AUCs for five different subclasses of protein. Figure 8 gives the same figures for the remaining proteins in the DUD set. The error bars on the graphs represent the expected error of the mean in each case, rather than the standard deviation.

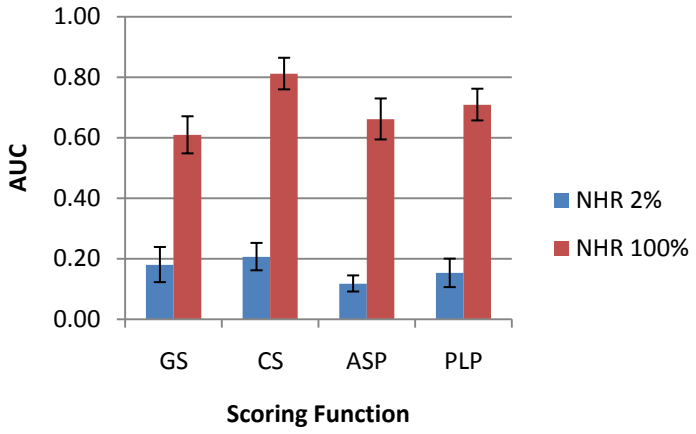

Figure 3 - Average Early enrichment (at 2% FPR) and Full ROC AUCs for the Nuclear Hormone Receptor subset of proteins (n= 8).

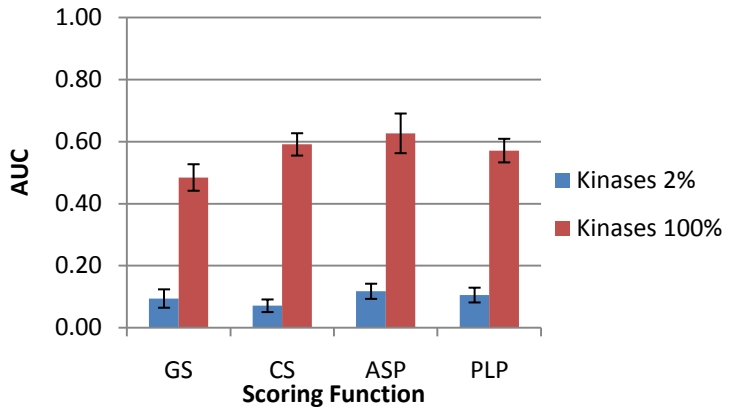

Figure 4 - Average Early enrichment (at 2% FPR) and Full ROC AUCs for the Kinase subset of proteins (n= 9).

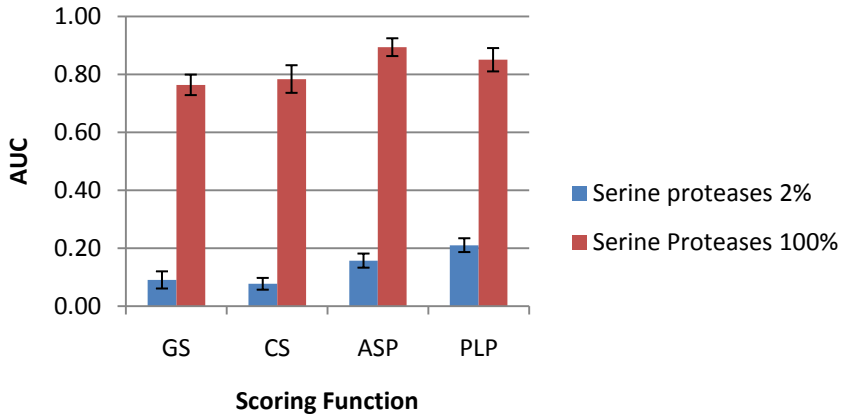

Figure 5 - Average Early enrichment (at 2% FPR) and Full ROC AUCs for the Serine protease subset of proteins (n= 3).

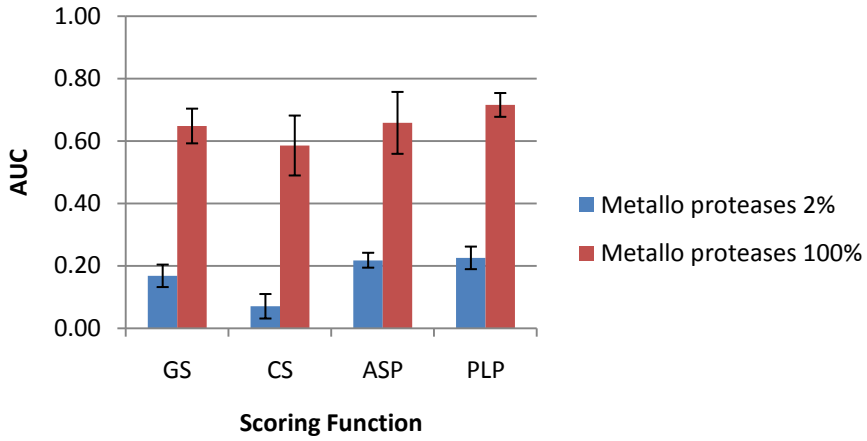

Figure 6 - Average Early enrichment (at 2% FPR) and Full ROC AUCs for the Metallo protease subset of proteins (n= 4).

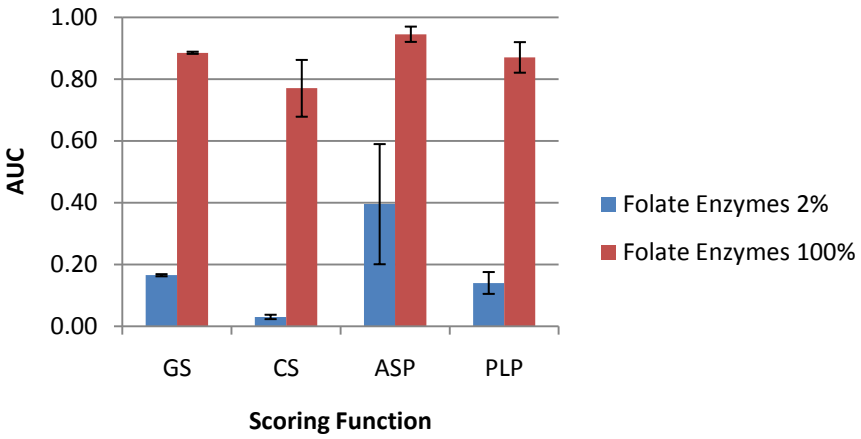

Figure 7 - Average Early enrichment (at 2% FPR) and Full ROC AUCs for the subset of proteins where the inhibitors mimic folate (n= 2).

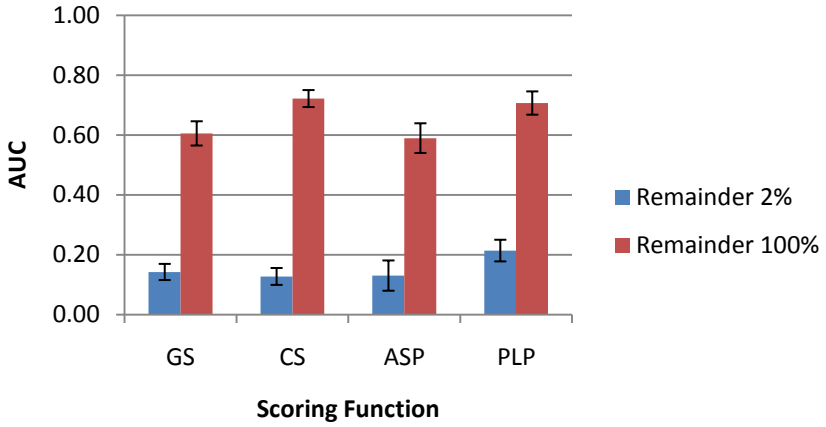

Figure 8 - Average Early enrichment (at 2% FPR) and Full ROC AUCs for the remaining proteins in the DUD set (n= 14).

When analysing these graphs early enrichment is used as the principle guide, although where this cannot provide a clear distinction between scoring functions, the full AUC is also taken into account.

These graphs suggest that different scoring functions may prove best for difference enzyme classes. ChemScore for instance, works best on the whole for nuclear hormone receptors, whereas ASP is the preferred scoring function overall for proteins where the inhibitors mimic folate.

For three classes of proteins, kinases, metalloproteases and serine proteases, either ChemPLP or ASP could be chosen as an effective scoring function. ChemPLP is the scoring function of choice for those proteins not belonging to any of the above classes.

The preference for a particular scoring function may reflect a particular characteristic of the enzyme class. Nuclear hormone receptors usually have lipophilic cavities and the surface contact term in ChemScore usually rewards that strongly. Folate mimics have to make strong and specific hydrogen bonds and these may be captured particularly well by a knowledge based scoring function such as ASP.

Enrichment for certain classes of protein is worse than for others. The average AUC for ChemPLP against the kinase set is particularly bad at about 62%. Kinases are known to usually have considerable flexibility in the active site and induced fit may often occur. Here we are docking to a single protein model in each case and this may be one reason why the overall enrichment statistics are poor.

Although the above graphs can be used for guidance when selecting the best scoring function to be used, it is possible that the best scoring function for the protein structure actually available may differ. Therefore these graphs should be used for indication only. If time permits and appropriate actives and decoys are available, some experimentation is recommended.

Conclusions

It has been demonstrated that overall, when both early enrichment and full AUC's are considered, ChemPLP, the most recently introduced GOLD scoring function[2], appears to be the best performing scoring function, at least for the DUD test set. This conclusion is consistent with that found in the pose-prediction part of this study and leads us to the recommendation that this scoring function be the one used preferentially for general purpose work.

Other scoring functions may work better for individual target classes. It is suggested ChemScore be used for nuclear hormone receptors and other enzymes or receptors with highly hydrophobic active sites. ASP appears to be a particularly good scoring function for folate enzymes and appears to work well for proteins where specific hydrogen bonds need to be made by the ligand, such as serine proteases.

GoldScore, the original default scoring does not perform as well in comparison and is also a slower scoring function to calculate. It is not recommended to use GoldScore for general purpose virtual screening, unless it is used for docking and then the poses rescored with a second scoring function which may significantly improve enrichment. For instance rescoring GoldScore generated poses with Asp was found to be successful in the SAMPL blind virtual screening challenge. Rescoring results on these DUD test sets will be presented separately.

References

1. Benchmark sets for molecular docking. N. Huang, B.K. Shoichet, and J.J. Irwin, *J. Med. Chem.*, **49**(23), 6789-6801, 2006.

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

GOLD – an accurate and reliable protein-ligand docking program

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

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