Customised Scoring Functions for Docking

The University Of Sheffield.

NOVARTIS

Iain Mott†, Peter Gedeck‡ and Val Gillet†

1. Introduction

Although great progress has been made, current docking methodologies often fail to correctly prioritise the crystallographic observed ligand pose. Table 1 presents the findings of a recent study[1]; whilst the correct binding mode is elucidated in many cases, the scoring functions employed *fail* to rank the correct poses above all others:

Method	Binding mode is found	**Binding mode is top ranked pose**
LigandFit	24	**7**
GOLD	47	**10**
DOCK	4	**0**
FlexX	39	**17**
Glide	26	**9**

Table 1 - Variation in correctly ranked binding modes for different docking methodologies against 69 protein targets.

This poster introduces a novel approach to scoring function development that applies a multiobjective methodology. Given the diversity and varied nature of binding sites, we argue that a single scoring function can never adequately approximate relative binding affinities in different target classes. This work explores the potential to develop class or protein specific scoring functions, initially investigating the degree of contention between different proteins.

2. Scoring Function and Descriptor Selection

A subset of the ChemScore[2] (CS) scoring function terms are used (table 2). The scoring function employed in CS is typical of those used in docking; it consists of a series of terms, each having a coefficient and a calculated ligand-protein interaction term. The sum of these terms is an approximation of $\Delta G_{binding}$. In the case of CS the weight coefficients for each term were determined by multiple linear regression against complexes with known binding affinities.

Term	Contribution
$P_{h\text{-}bond}$	Ligand-protein hydrogen bonding
P_{lipo}	Lipophilic potential
P_{clash}	Clash penalty
$P_{internal}$	Torsion penalty
P_{rot}	Rotational bond freezing term

Table 2 - ChemScore terms used in optimisation.

3. Dataset and descriptor generation

The newly available Astex diverse dataset[3] has been employed for this work. For each of the 85 protein co-crystals the ligand was rescored using the CS function. The ligand was then docked using GOLD[4], however the GOLD GA parameters were deliberately set such that the docking run produced a range of suboptimal poses, these are hence termed the *decoy* poses. The decoys were then inspected visually by overlay to the crystal pose to ensure sufficient deviance from the crystallographic ligand binding mode.

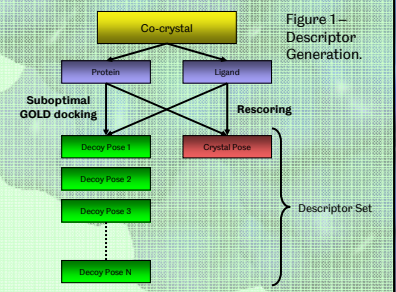

Figure 1– Descriptor Generation.

For each decoy pose and the rescored binding mode pose for each co-crystal a subset of the CS ligand-protein binding descriptor terms was extracted. A principal components analysis (PCA) was then performed to understand the distribution of descriptor values across the protein classes. The PCA indicated some localisation of the descriptors between classes.

4. Multiobjective Optimisation

In a previous study[5] an evolutionary algorithm (EA) was applied to optimise scoring functions for docking. This method met with limited success because it was not possible to evolve a scoring function that performed well across a wide range of protein targets. In this work we use a similar ranking methodology, however the EA has been replaced with a multiobjective evolutionary algorithm (MOEA). This class of machine learning algorithms is useful for real world problems where there exists no single response variable, or *objective*, and where each different objective conflicts.

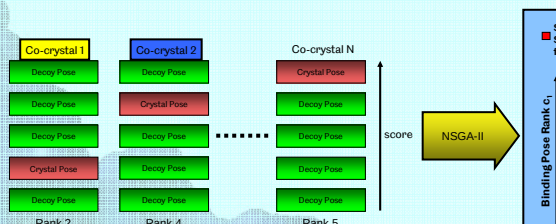

Figure 2 - Ranking Function.

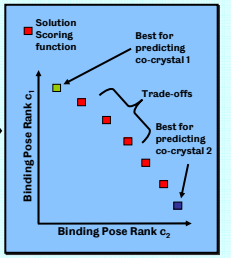

Figure 3 - Approximation of the Pareto-optimal set of scoring functions.

Here a custom implementation of the *nondominated sorting genetic algorithm-II* (NSGA-II) is employed[6]. The optimisation evolves a population of solutions by selection, mutation and recombination, each representing a set of scoring function coefficients. Each of the solutions are evaluated against the descriptors of every pose. All poses are then ordered by their relative score (figure 2). The multiobjective optimisation problem is then to maximise the rank position of the known ligand binding mode in every co-crystal. The output of the NSGA-II is the set of least worst solutions to this multiobjective ranking problem (figure 3).

5. Results: Four Co-crystal Optimisation

Four co-crystals were selected as representatives of their respective protein class (table 3). A population of 100 solutions was evolved in 40 generations of the NSGA-II algorithm to produce a non-dominated set as in figure 4. There were 401 poses (400 decoy + 1 binding) with a value of 400 being the best possible ranking for a scoring function, and 0 the poorest.

Co-crystal PDB id	Class
1xm6	Phosphdiesterases
1sqn	Nuclear receptors
1pmn	Kinases
1oyt	Serine Proteases

Table 3 – Class membership of the four co-crystals used.

Analysis: The curved **1xm6 and 1pmn** (figure 4) plot indicates a trade-off front behaviour, good solutions for one are mutually exclusive of the other - no single function can be found to satisfy both. This behaviour is also observed clearly for **1xm6 and 1oyt**. However, **1sqn and 1oyt** show complementary scoring functions, a cluster of good solutions for both is visible close to optimal. It is also observed that suboptimal solutions score equally as badly only for a limited range then show an independent, and more scattered, relationship. The **1pmn and 1sqn** plot is highly scattered suggesting that scoring functions are independent of each other.

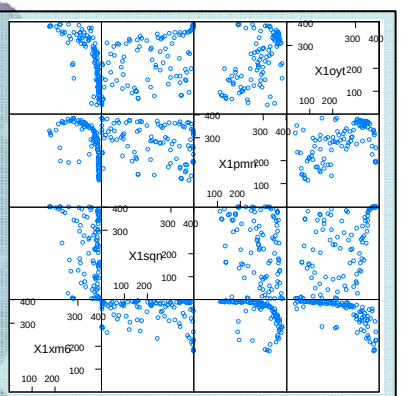

Figure 4 - Scatter plot of scoring functions for a four co-crystal optimisation.

6. Results: Two class optimisation – Kinases / Nuclear Receptors

Figure 5 shows the two objective - average rank - optimisation of eleven Kinase and nine Nuclear Receptor targets. Three repeat optimisations were performed, all of which approximated the same front. As expected no single scoring function (such as the hypothetical utopian solution - blue triangle) could be produced that is able to simultaneously correctly rank the known binding poses for both classes. Also shown is the result of a single objective optimisation for both classes, these scoring functions coincide with the best ranking functions for the individual classes derived from the NSGA-II optimisation.

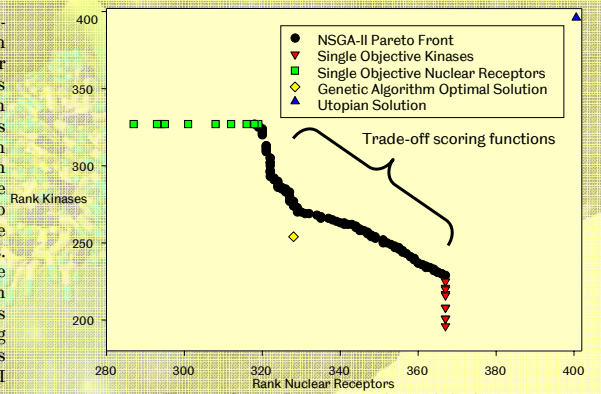

Figure 5 - Scoring Functions two class optimisation.

The result of a single objective genetic algorithm (using a different scoring function combining both objectives) is shown in yellow. Note this solution represents a poor scoring function for both protein classes. This demonstrates that a single scoring function is unlikely to adequately score both classes.

7. Conclusions

- A single scoring function for docking - derived from regression or other single objective techniques - cannot perform well against all targets.
- Initial findings indicate clear contentions do exist in optimisation against different co-crystals.
- By applying a MOEA, it is demonstrated that customised scoring functions can be derived that are appropriate for either a particular protein or protein class.
- Further refinement of the protocol is required including the cross-docking of a diverse set of ligands to each co-crystal – better approximating the actual difficulty inherent in the docking problem.

† Department of Information Studies, University of Sheffield, Sheffield S10 2TN, UK
‡ Novartis Institutes for Biomedical Research, Novartis Horsham Research Centre, West Sussex RH12 5AB, UK

This work is a product of Knowledge Transfer Partnership No. KTP001223 between Novartis and University of Sheffield.

References

1. Kontoyianni, M., L. M. McClellan and G. S. Sokol (2004). Journal of Medicinal Chemistry **47**(3): 558-565.
2. Eldridge, M. D., C. W. Murray, T. R. Auton, G. V. Paolini and R. P. Mee (1997). Journal of Computer-Aided Molecular Design **11**(5): 425-445.
3. Hartshorn, M. J., M. L. Verdonk, G. Chessari, S. C. Brewerton, W. T. M. Mooij, P. N. Mortenson and C. W. Murray (2007). **50**(4): 726-741.
4. Jones, G., P. Willett, R. C. Glen, A. R. Leach and R. Taylor (1997). Journal of Molecular Biology **267**(3): 727-748.
5. Smith, R., R. E. Hubbard, D. A. Gschwend, A. R. Leach and A. C. Good (2003). Journal of Molecular Graphics and Modelling **22**(1): 41-53.
6. Deb, K., A. Pratap, S. Agarwal and T. Meyarivan (2002). IEEE Transactions on Evolutionary Computation **6**(2): 182-197.