Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Molecular Graphics and Modelling 22 (2003) 41–53
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

Analysis and optimization of structure-based virtual screening protocols (3). New methods and old problems in scoring function design

Ryan Smith [a,b], Roderick E. Hubbard [a], Daniel A. Gschwend [c], Andrew R. Leach [d], Andrew C. Good [e,*]

[a] *York Structural Biology Laboratory, Department of Chemistry, University of York, Heslington, York YO10 4DD, UK*
[b] *Astex Technology Ltd., 250 Cambridge Science Park, Milton Road, Cambridge CB4 0WE, UK*
[c] *ArQule, Inc., 19 Presidential Way, Woburn, MA 01801, USA*
[d] *GlaxoSmithKiline, Medicines Research Centre, Gunnels Wood Road, Stevenage SG1 2NY, UK*
[e] *Bristol-Myers Squibb, P.O. Box 5100, Wallingford, CT 06492, USA*

Accepted 14 February 2003

Abstract

Scoring function research remains a primary focus of current structure-based virtual screening (SVS) technology development. Here, we present an alternative method for scoring function design that attempts to combine crystallographic structural information with data derived from directly within SVS calculations. The technique utilizes a genetic algorithm (GA) to optimize functions based on binding property data derived from multiple virtual screening calculations. These calculations are undertaken on protein data bank (PDB) complex active sites using ligands of known binding mode in conjunction with "noise" compounds. The advantages of such an approach are that the function does not rely on assay data and that it can potentially use the "noise" binding data to recognize the sub-optimal docking interactions inherent in SVS calculations. Initial efforts in technique exploration using DOCK are presented, with comparisons made to existing DOCK scoring functions. An analysis of the problems inherent to scoring function development is also made, including issues in dataset creation and limitations in descriptor utility when viewed from the perspective of docking mode resolution. The future directions such studies might take are also discussed in detail.

Keywords: Structure-based virtual screening; Scoring functions; Stochastic optimization; Crystallographic data artifacts; Dataset construction

1. Introduction

The main thrust of structure-based virtual screening (SVS) research currently resides in the area of scoring function development [1]. Early studies concentrated on the statistical analyses of key ligand–protein interactions, with attempts being made to correlate these with experimental binding data. While such efforts continue [2,3], a noteworthy issue with this approach is that the binding data accompanying crystal structures come from a variety of disparate sources and assay conditions. It is consequently difficult to gauge the internal consistency of the data, leading to a significant source of uncertainty in the resulting calculations. A recent departure from the above methodology has come about through the exploitation of protein data bank (PDB) [4] structural data using knowledge-based approaches [5–8]. Such functions have the advantage of being able to exploit the geometric information inherent in protein–ligand complexes without resorting to explicit binding data correlation. Nevertheless, these techniques still rely explicitly on the crystal structure binding modes, which by their nature shed light primarily on why molecules bind, as opposed to why they do not. Further, the methods used thus far do not generally account for many of the inherent limitations of current SVS technology. A pertinent example of this is the assumption that the calculation will be able to reproduce the crystallographic conformation of the ligand.

Here, we present an alternative method for scoring function creation that combines PDB structural information with data derived from directly within SVS calculations. The technique utilizes a genetic algorithm (GA) to process binding property data derived from multiple SVS runs on PDB complexes using "noise" compounds together with ligands with known binding modes. The function to be optimized is the average ranking of the known ligand crystal structure for each target within its noise dataset. Consequently, the method does not rely on assay data and also

* Corresponding author.
E-mail address: andrew.good@bms.com (A.C. Good).

1093-3263/03/$ – see front matter © 2003 Elsevier Science Inc. All rights reserved.
doi:10.1016/S1093-3263(03)00125-6

has the potential to use the "noise" binding data to recognize sub-optimal interactions inherent in SVS calculations. The results of our primary efforts utilizing this technique in conjunction with DOCK [9] are presented below, together with a look at the future directions such studies should take. In addition many of the inherent problems, we encountered during scoring function development are highlighted.

2. Experimental

2.1. PDB dataset selection

The initial dataset contained 109 complexes filtered from 237 for which experimentally determined binding affinity data were available (Keske and Dixon, 1994 unpublished) and have been applied in previous scoring function design efforts [10]. In addition, those used by Böhm in his seminal efforts at scoring function design [11] were also included. Filtering required the removal of complexes that were unrefined or model-built, contained covalently bound, incompletely modeled, had macromolecular ligands, a resolution greater than 2.5 Å, or contained greater than 50 heavy atoms.

These remaining complexes were then visually inspected to identify ligands with points of crystal contact. Ligands found to make significant contacts with adjacent symmetry related protein structures were removed from the dataset. In certain cases, the ligand has been observed to makes more interactions with the symmetry related atoms than with the atoms in the deposited crystal structure (e.g. 4gr1). The problem of crystal contacts and crystal structure errors in general is perhaps an under-appreciated issue in scoring function dataset selection, although it has received attention elsewhere [12].

Given that we were interested in the creation of a generally applicable scoring function, it was deemed important to avoid bias towards particular protein targets. To this end, it was decided from the outset to select training set data for which no protein or ligand was duplicated. From the limited number of complexes remaining, 20 were selected for use as a training set, with a further 10 chosen for the test set. Six of these complexes contained proteins not present in the training dataset. The remaining four complexes, although also present in the training set, are bound to significantly different ligands (different chemotype) compared to those used in training.

Each ligand was separated from its parent complex and further processed in SYBYL [13] to assign appropriate atom types and add hydrogens, formal charges and Gasteiger and Marsilli partial atomic charges [14]. Solvent molecules were then removed with all remaining atoms, including metal ions, glycosylation sites, and other ligands and/or cofactors being treated as part of the protein. Docking site points were constructed for each protein using the DOCK SPHGEN program. All parameters used were default values as supplied the UCSF software release [9]. Visual inspection of the resulting clusters was used to remove those site points not within the active site region of the protein.

2.2. Noise dataset preparation

In addition to the ligands of the PDB dataset, a set of 100 "noise" molecules were selected for the ligand training and test sets. These were selected from the WDI [15] demo database (version 98.1) provided with DAYLIGHT software [16]. The idea behind the inclusion of this extra set of structures was to provide the GA function with data that bears similarity to that seen in a typical SVS run. This provides a tougher optimization task and the potential to pick up bad as well as good interaction patterns.

The WDI demo database initially contained 1774 structures. Molecules were then eliminated if they were also present in the PDB database, had a molecular weight greater than 500, more than eight rotatable bonds, contained any reactive groups, or generated a close contact warning on CONCORD [17] 3D structure generation. From those structures that remained 100 diverse compounds were selected with maximum dissimilarity in their Daylight fingerprints. Molecule preparation was the same as for the PDB ligands.

2.3. 3D structure and conformer generation

While DOCK has the ability to undertake flexible docking internally, in the version used (4.01) [9], some problems were found, primary among them being an inability to generate conformers rapidly (if at all) for larger more heavily substituted ligands. CONFIRM [18] was thus used to generate a set of up to 50 conformers for each of the PDB and noise ligands using the "best" algorithm and an energy threshold of 20.0 kcal/mol.

2.4. DOCKing calculations

All docking operations were performed using DOCK. The default AMBER-based parameter set supplied was replaced with those of the Sybyl force field [13], to maintain consistency in metric generation (see Section 2.5). The dimensions of each grid were defined using a box whose center coincided with the center of geometry of the site points of each protein and extended to enclose these points plus an additional "padding" distance of 5.0 Å in each direction. For the bump scoring grids, a clash overlap value of 0.65 was used, with a cutoff distance of 4.0 Å set for the contact scoring grids. Energy scoring grids were calculated using an all atom model with a 10.0 Å cutoff. The attractive and repulsive exponents of the van der Waals interaction energy were set to 6 and 12, respectively, with a distance dependent dielectric function of $4.0r$ for the electrostatic component.

DOCK was then used to dock each of the ligand conformer sets into each of the proteins. A docking protocol was sought that would produce at least a specified number of spatially diverse, low scoring orientations for each ligand

docked into each protein and that these be generated in a reasonable amount of time. The procedure finally selected was created and refined through an iterative process of trial and error. Each of the 130 (30 PDB + 100 WDI) ligand conformers was docked separately into each of the proteins in the PDB dataset. For each conformer 50,000 orientations were generated using random site point matching. Contact and energy scores were calculated for each of these orientations and the top $1000/N_{conf}$ were retained for each scoring method, where N_{conf} is the total number of conformers in the set for the current ligand. This value was rounded up to the nearest integer so that, in total, a minimum of 1000 ligand orientations were retained for each scoring method per conformer set docked per protein. The docking process resulted in approximately 1000 top contact and energy scoring orientations for each ligand in each protein. If all of the orientations of both the PDB and WDI ligands were considered, the GA would have to rank many thousands of orientations per ligand. This was too many to be considered due to CPU constraints, and so each set of generated orientations was reduced to a spatially diverse set of 200 using sphere exclusion clustering [19] based on heavy atom RMSD. The task was thus reduced to picking each crystallographic ligand orientation (CLO) out from the 23,999 ($200 \times 120 - 1$) generated "noise" orientations.

2.5. GA data generation using the TEC program

To generate the interaction data, each ligand was processed using the program "toolkit for the evaluation of candidates" (TEC). TEC has been written with empirical scoring function development in mind and as such features a diverse array of both 2D and 3D descriptors that can be flexibly combined to create custom scoring functions. The program is organized in a hierarchical fashion and controlled through an extensive set of ASCII control files. It is based primarily around the Tripos MOL2 [13] molecule file format, and uses the Tripos Force Field atom type and non-bonded parameters in metric calculations. TEC features a number of classic SVS descriptors, which are listed in Table 1.

For efficient evaluation of many of the metrics, TEC requires that the protein contribution be pre-calculated and stored in the form of scoring grids (as with DOCK). Nearest neighbor grids are also used on occasion to further reduce the computation time for metrics only dependent on short-range interactions. Calculations performed using TEC therefore proceed in two separate stages. In the first stage, scoring grids are constructed and stored for those metrics that require them. In the second stage, the relevant scoring grids are read back in and used to calculate the specified metrics for all docked molecule orientations to be processed.

A substructure query language, TSLN, is a central component of the TEC program. This language is a simplified version of SLN that can be used to assign atoms to particular chemical classes. TSLN sample definitions are shown in Table 2. TEC classes are then used to define how the interactions between certain atoms are quantified.

Table 2
Sample TSLN definitions used in the metric calculations

TSLN definition	TSLN query
Heavy atom	$\hat{H}$
Hydrophobe definitions	
Hydrophobes1	C.∗-H
Hydrophobes2	C.3(Hev)(Hev)(Hev)(Hev)
Hydrophobes3	C.2=C.2
Hydrophobes4	C.1%C.1
Hydrophobes5	C.ar
Hydrophobes6	F|Cl|Br|I
Hydrophobes7	S.3!(H)
Donor definitions	
DonorNitrogen	N.∗-H
DonorOxygen	O.3-H
Acceptor definitions	
AcceptorNitrogen	N∗ & $\hat{N}$.pl3 & $\hat{N}$.4 & $\hat{N}$.1(Hev)(Hev)!(∗)
AcceptorOxygen	O∗

Table 1
Primary descriptors calculated within the TEC program

Metric primitive	Scoring grid	Description
Clash	Yes	Number of protein/ligand atom clashes
Burial	Yes	Hydrophobic protein atoms within given radius of ligand
Electrostatic	Yes	Electrostatic interaction energy
VdwAttractive	Yes	Attractive van der Waals energy
VdwRepulsive	Yes	Repulsive van der Waals energy
LigSurface	Yes	Buried ligand hydrophobic solvent accessible surface area ($Å^2$)
RecSurface	Yes[a]	Protein Solvent accessible surface area buried ($Å^2$)
H bond	Yes[a]	Protein/ligand intermolecular hydrogen bond score
Intra H bond	No	Intramolecular hydrogen bond score for the ligand
Interaction	Yes[a]	Pair-wise interactions between the protein and the ligand
Volume	No	Ligand volume ($Å^3$)
FrozenTorsions	No	Number of rotatable bonds in ligand buried when docked
Flexibility	No	Kier flexibility index of the ligand
MolWeight	No	Molecular weight of the ligand

Metrics used in this study are highlighted in bold.
[a] These metrics make use of a nearest neighbor grid.

Grid and TLSN data can be combined with one of four function types in metric creation. These are Boolean, count, linear and Gaussian. The Boolean function simply returns 1 if the value calculated was ever between the minimum and maximum values specified, otherwise the function evaluates to 0. For example, in the context of assessing hydrogen bond distances, the Boolean function will return 1 if the distance between donor and acceptor atoms is between the minimum and maximum distance specified. The count function is used to tally the number of times a value falls within a specified range, for example the number of interactions between particular functional groups within a certain distance range.

A linear function can also be used to assess molecular interactions. Values calculated with this function lie in the range 0.0–1.0. Again, minimum and maximum values are specified for the function as well a value for which the function is an optimum (i.e. unit contribution) and the gradient of the function. This function evaluates to 0.0 if the value being assessed falls outside of the minimum and maximum range. If the value is in the range optimum to maximum then a linear function is applied, the gradient of which is defined by specifying the coordinates of an extra point. The ordinate value of this point is specified by a decay value that can either be an absolute value in the range 0.0–1.0 or as a percentage of the optimum value of 1.0. The abscissa value of the point can be specified either as an absolute value or relative to the optimum (e.g. +0.1). Suboptimal values—those in the range minimum to optimum—can be treated in one of two ways. They can either receive a full contribution (a value of 1.0) or the linear function can be reflected about the line $x_A = A_{\text{optimum}}$.

Similar parameters are specified for the Gaussian function, the exact functional form of which is

$$f(x) = \mathrm{e}^{-N(x-\text{opt})^W} \tag{1}$$

Here, N is the normalization coefficient, opt is the optimum value and W is the exponent of the function. The normalization factor is calculated as

$$N = -\frac{\ln y}{(x - \text{opt})^W} \tag{2}$$

where x is the value at which Eq. (2) will decay to y. Again, x can be specified either as an absolute value or relative to the optimum value, while y can be expressed as an absolute value or as a percentage of the maximum value of 1.0.

While we were initially interested in many of these (and numerous other) metrics, the small number of protein complexes remaining after dataset filtering forced a scaling back in our initial efforts at function method development. As a consequence only a subset of the available functions (highlighted in bold in Table 1) have been exploited to date. The clash (f_{clash}) and electrostatics (f_{electro}) terms are variants of descriptors used in DOCK, while hydrophobic surface burial (f_{hsurf}) and hydrogen bonding (f_{hbond}) metrics are derived from the work of Böhm [11]. They were chosen since they were considered to describe the core shape, hydrogen bonding and salt bridge terms that dominate many ligand–protein interactions. The matrix shown for these terms in Table 3 over the training dataset complexes illustrates the low degree of correlation. The lack of correspondence between the hydrogen bond and electrostatic terms is noteworthy and highlights the fact that, unlike hydrogens bonds, electrostatics are dominated by salt bridge interactions.

Table 3
Correlation matrix for the four TEC metrics used in these studies, calculated using all of the ligand orientations (both CLOs and generated orientations) of the WDI dataset

Metric/metric	f_{clash}	f_{hsurf}	f_{hbond}	f_{electro}
f_{clash}	1.00			
f_{hsurf}	0.41	1.00		
f_{hbond}	0.05	−0.12	1.00	
f_{electro}	0.01	0.03	−0.04	1.00

The clash metric f_{clash} provides a measure of the number of ligand and protein atoms in steric contact. The protein contribution to this score is pre-calculated and stored as a scoring grid. The clash metric can also be combined with one or more TSLN labels in order to calculate the clash score for particular substructural features of a ligand.

The LigSurface metric calculates the total buried solvent accessible surface area (SASA) for the ligand in Å^2. This calculation can be broken down into three separate stages. In the first stage, a scoring grid is constructed for the protein. In the second stage, an approximate solvent accessible surface (SAS) for each ligand is constructed. In the third stage, the constructed surface is used in conjunction with the previously scoring grid to compute the buried SASA for the ligand. Buried ligand hydrophobic surface area (f_{hsurf}) is thus obtained by combining the "hydrophobe" TSLN labels with this LigSurface metric. A probe radius of 1.4 Å has been used to construct the approximate SAS for each ligand using a point density of 5.0 points/Å^2.

The intermolecular hydrogen bond score, f_{hbond}, is based on the distance component of Böhms hydrogen bond function [11] (Eq. (3)) combined with an angular term that is derived from an analysis of the distribution hydrogen bond angles in the PDB dataset. Hydrogen bond donor and acceptor groups are defined using appropriate TSLN classes. The distance function used by Böhm has been modified so that the hydrogen bond distance is calculated between the heavy atoms of the donor and acceptor groups and not between the hydrogen and acceptor atom. Hydrogen atoms have not been used in the calculation of this metric/descriptor due to the problems of accurate placement. As a result, the ideal hydrogen bond length, as defined by Böhm, of 1.9 Å has been increase by 1.0 Å to account for the missing hydrogen atom:

$$\begin{cases} f(\Delta r) = 1.0, & \Delta r \le 0.2 \\ f(\Delta r) = 1.0 - \dfrac{\Delta r - 0.2}{0.4}, & 0.2 < \Delta r \le 0.6 \\ f(\Delta r) = 0.0, & \Delta r > 0.6 \end{cases} \tag{3}$$

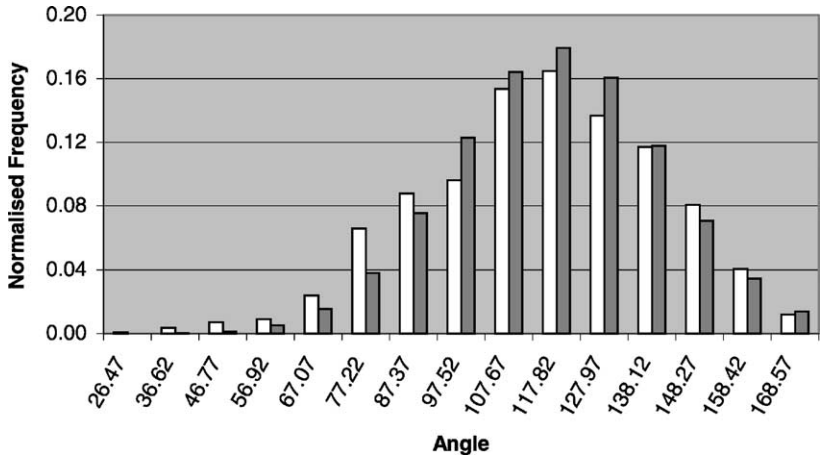

Fig. 1. Black histogram highlights the experimentally observed distribution of XÔA and DÂY hydrogen bond angles in the dataset of 144 protein–ligand complexes analyzed. White histogram shows the normal distribution fitted to this data.

where r is the deviation of the hydrogen bond length from the ideal value of 2.9 Å. In initial calculations, the hydrogen bond score was calculated using this distance term only. However, analysis of complexes not predicted well by the calibrated scoring function revealed generated ligand orientations that, although ideal in terms of hydrogen bonding distances, contained hydrogen bond angles that were not feasible. As a result, these orientations had higher scores than expected and ranked above the CLOs. An angle dependent term has therefore been added to the function. The angle function used is based on the observed distribution of XÔA and DÂY hydrogen bond angles between each ligand and protein of the entire dataset of 144 complexes studied (Fig. 1). X and Y represent any atom connected to either the donor and acceptor atom, respectively.

The experimentally observed hydrogen bond angles show an approximately normal distribution with a mean of 117.2° and standard deviation of 22.7°. A normal distribution was thus fitted to this data to create a Gaussian function, $f(\cdot)$, within TEC (Table 4) that has been combined with the distance function to penalize deviations from an ideal hydrogen bond angle of 117.2°. The overall hydrogen bond score is therefore a product of the distance and angle functions:

$$f_{\text{hbond}} = f(\Delta r, \alpha) = f(\Delta r)\, f(\alpha) \tag{4}$$

The electrostatic interaction energy f_{electro} have been determined using electrostatic grids calculated using a 10.0 Å cutoff with a distance dependent dielectric of $4.0r$.

Clash, LigSurface, and electrostatic grids were all calculated at a resolution of 0.3 Å. The clash grids have been computed using the default clash tolerance (35%). The nearest neighbor grids have been constructed at a resolution of 1.0 Å using the default cutoff radius of 5.5 Å.

Table 4
Sample TEC metric definition; in this case the angle class defined for penalizing hydrogen bond angle deviations from the ideal value of 117.2° using a Gaussian function

AngleType	AngleXDA
Hydrogen	No
Labels	Donor–acceptor
Function	Gaussian
Minimum	0
Optimum	117.2
Maximum	360
Decay	0.5 at 144.0
Exponent	2
Suboptimal	Symmetrical
End	

2.6. The genetic algorithm

The program GAOPTIMISE has been written to perform the calibration of scoring function coefficients. This program utilizes the SUGAL GA library [20] which provides an extensive range of functionality for writing GA programs and is highly customizable both by the developer and the user. The GAOPTIMISE program is essentially an interface to this library. It is responsible for reading in the metric data resulting from TEC and providing the fitness function to assess the chromosome population maintained by SUGAL. The SUGAL library handles all other aspects of the GA including selection, crossover, mutation and replacement.

The scoring function used in these studies is expressed in terms of the metrics calculated above as

$$\text{score} = a_{\text{clash}} f_{\text{clash}} + a_{\text{hsurf}} f_{\text{hsurf}} + a_{\text{hbond}} f_{\text{hbond}} + a_{\text{electro}} f_{\text{electro}} \tag{5}$$

The coefficients form the individual genes of each chromosome in the population maintained by the GA. These are optimized so that the CLO of each protein can be distinguished from the generated data based on the calculated scores. A graphical depiction of the fitness function used to achieve this shown in Fig. 2. Each chromosome in the population

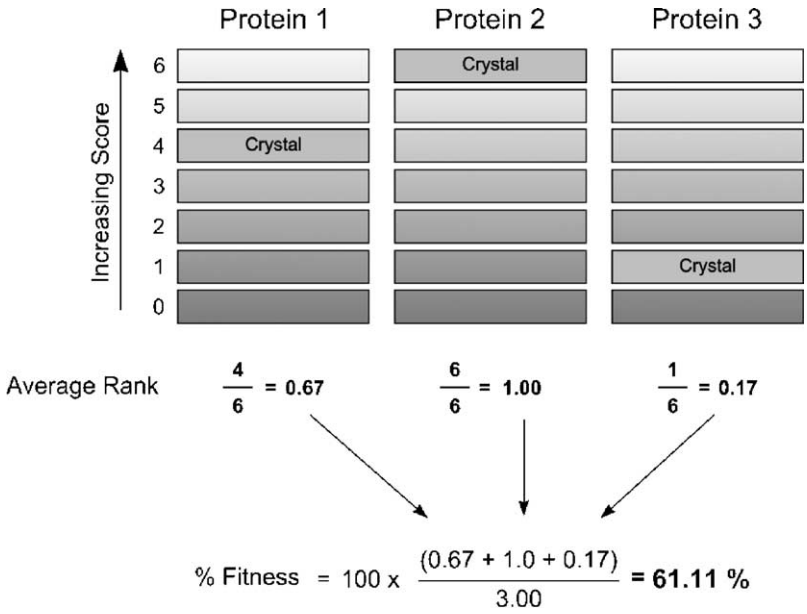

Fig. 2. A graphical representation of the GA fitness function for three proteins, each having six generated ligand orientations and one CLO. The average rank of the CLO for each of the three proteins are 4/6, 6/6 and 1/6, respectively. The overall fitness of this chromosome would therefore be the average of these three values expressed as a percentage, in this case 61.1%.

can be decoded to yield a particular set of coefficients for the scoring function in Eq. (5). Scores are then calculated for each of the generated orientations and for each of the CLOs. By sorting the orientations generated against each particular protein based on the score we can calculate the average rank for each CLO. The overall fitness of the chromosome is then calculated as the average of the average rank values for all proteins expressed as a percentage.

The same GA protocol was used to perform all of the GA calculations presented. Again, this has been developed through an iterative refinement process, although most of the parameters investigated produced essentially the same result, varying only in convergence and calculation times. In each calculation, a population of 150 chromosomes has been used. The chromosomes are represented using the native double (real) data type within SUGAL and not as a bit strings (i.e. the genes are stored using the machines native floating-point representation and not using some form of binary encoding scheme). The GA is executed until either 3000 generations have been performed or the standard deviation of the fitness values between the chromosomes of the population is less that 0.0001. In each iteration, 50% of the population are selected as potential candidates for replacing the chromosomes of the current population using a roulette wheel-based selection strategy. The candidates are then paired and undergo two-point crossover with a probability of 1.0. As the chromosomes are stored as real values and not binary strings, crossover is performed between gene boundaries. Performing intragene crossover of a real data type is likely to cause large changes in the values of the genes at the crossover point and so is likely to destroy the important information contained within the chromosome that is responsible for its fitness. Once crossover has been performed all candidates are subject to mutation. Mutation is performed out on a per gene basis with a probability of 1.0. As mentioned in the SUGAL manual, the use of the typical bit inversion mutation operator, although appropriate for use with a binary chromosome representation, is inappropriate when dealing with chromosomes represented using real data types. Instead, a Gaussian mutation operator is applied in which a random number is generated from a normal distribution with mean 0.0 and standard deviation of 3.0 (this is user definable) and added to the value of the gene.

The standard deviation decays over the course of the GA according to the following equation:

$$s_n = s_0 r^n \tag{6}$$

where n is the number generation performed, s_0 is initial and standard deviation, s_n is the standard deviation for current generation and r is the decay rate, which lies in the range 0.0–1.0. A decay rate of 0.99 has been used in all calculations. This decayed Gaussian mutation operator has the effect of introducing potentially large changes to the values of the genes at the start of the GA so that the search space is effectively explored. As the GA progresses the mutation size is gradually decreased thus helping to fine-tune the coefficients and localize the search.

The GA employs an elitist model that ensures that the best chromosome is always retained in the next generation. The child chromosomes generated by the crossover and mutation operations above replace the current members of the population if they have improved fitness based on a tournament based selection scheme. In this scheme, each candidate chromosome is compared against 10 randomly

selected chromosomes from the current population. If the current candidate chromosome has a fitness value higher than the lowest scoring chromosome in the tournament then it replaces this chromosome in the population.

It has also been observed that the crossover operator does not appear to be aiding the search as reducing the probability of crossover to 0.0 has no effect on the number of generations taken for the GA to converge or on the results produced. Therefore, it can be concluded that the search of the coefficient space is being performed by the mutation operator only. This is believed to be due to the fact that multiplying the coefficients of the scoring function by any positive value result in the same rank ordering, and hence, the calculated fitness is unaffected. Therefore, although the search space may only have one maximum in terms of calculated fitness, this can be obtained in an infinite number of ways by multiplying the coefficients of the scoring function by different positive values. The task of finding a fit solution is therefore a relatively straightforward task from the outset, as the existence of a chromosome with approximately the correct ratios of gene (coefficient) values will quickly be identified as a fit solution. The decayed mutation operator will then progressively refine the values of the genes so that the optimum ratio is obtained and the fitness is maximized. The crossover operator may help initially to generate this ratio of gene values but as the GA progresses this may have a detrimental rather than beneficial effect. For example, the two chromosomes 1, 2, 3, 4 and 2, 4, 6, 8 would produce exactly the same rank ordering of the ligands, and hence, the fitness of the two chromosomes would also be the same. Crossover (either one, two, or *N* points) anywhere within these chromosomes would destroy the correct ratio of coefficients and so is likely to reduce the fitness of the child chromosomes. In order to address this problem in future calculations, and to attempt to increase the effectiveness of the crossover operator, an alternate encoding method for the coefficient should be investigated.

More information regarding metric and GA design can be found at the following web site: http://www.ysbl.york.ac.uk/~ryan/Thesis/.

3. Results

A host of different calculations were executed during the course of these investigations. Here, we report on the results of our primary studies.

(1) In the first, the task of the GA was to optimize the rank of each CLO for the 20 training set complexes relative to the other 23,999 (100 noise and 20 training set molecules × 200 orientations) competing ligand orientations. The resulting scoring functions were then applied to rank the 10 test set complexes, with each of the 10 CLOs compared with 21,999 competing ligand orientations (100 noise and 10 test set molecules × 200 orientations). Training set results are shown in Table 5, with test set data highlighted in Table 6. GA function results are presented in Table 7 with metric contributions to the resulting scoring functions in Table 8.

Table 5
The number of orientations ranking above each CLO of the WDI training dataset based on the scoring function coefficients presented in Table 7

Complex	Run		
	1	2	3
1phd	33	33	34
3cpa	5	5	3
9aat	4	4	4
1ak3	3	3	3
2pk4	3	4	4
2tmn	3	3	5
1rnt	2	2	2
7est	2	2	2
1abe	0	0	0
1apv	0	0	0
1pph	0	0	0
1rbp	0	0	0
1snc	0	0	0
2tsc	0	0	0
3gap	0	0	0
4dfr	0	0	0
4phv	0	0	0
4sga	0	0	0
6tim	0	0	0
7cat	0	0	0
Rounded average rank	3	3	3

Table 6
The number of orientations ranking above each CLO of the WDI test dataset based on the scoring function coefficients presented in Table 7

Complex	Run		
	1	2	3
1dr1	286	286	283
1xli	39	42	29
1phg	30	31	27
2ifb	1	1	1
3fx2	1	1	1
3tpi	0	0	0
4mdh	0	0	0
5p21	0	0	0
5tmn	0	0	0
8cpa	0	0	0
Rounded average rank	36	36	34

Table 7
Normalized coefficients and percentage fitness values resulting from three CLO GA runs

Run	Coefficients				Fitness (%)
	a_{clash}	a_{hsurf}	a_{hbond}	$a_{electro}$	
1	−0.816	0.006	0.574	−0.070	99.989
2	−0.817	0.006	0.572	−0.069	99.988
3	−0.793	0.006	0.605	−0.069	99.988

Table 8
The average normalized percentage contributions of descriptors to the training set CLOs, calculated using the coefficients of Table 7 (standard deviations for each value shown in brackets)

Run	Normalized percentage contribution (S.D.)			
	$a_{clash}f_{clash}$	$a_{hsurf}f_{hsurf}$	$a_{hbond}f_{hbond}$	$a_{electro}f_{electro}$
1	−3.737 (7.828)	44.005 (24.519)	42.677 (23.722)	17.055 (18.960)
2	−3.782 (7.927)	44.055 (24.605)	42.804 (23.732)	16.923 (18.853)
3	−3.513 (7.379)	43.973 (24.631)	43.312 (23.713)	16.228 (18.141)

(2) Initial studies highlighted a number conformation and configuration SVS sampling issues that could conceivably compromise our efforts at scoring function design. As a consequence it was decided to repeat the calculations of study (1), replacing the CLO with the closest orientation (by heavy atom RMSD), generated for the known ligand during the docking stage (henceforth known as CDO). The idea behind this approach was to see whether it was possible to generate a model using ligand binding data more consistent with the docking accuracy seen in SVS calculations. Table 9 shows the heavy atom RMSDs of each CDO to its respective CLO and the corresponding number of heavy atoms for each ligand. The docking protocol was able to generated ligand orientations "close" in terms of RMSD to the CLOs in the majority of cases. Nevertheless, several CDOs are so dissimilar to their respective CLO in terms of RMSD that they cannot be considered as having adopted an orientation that approximates the correct binding mode. As such, CDOs with an RMSD of greater than 2.000 Å from their respective CLOs have not been used in the creation of the CDO models. This reduced the size of the PDB training dataset to 16 complexes and the test set to 9.

Training set results are shown in Table 10, with test set data highlighted in Table 11. GA function results are presented in Table 12 with metric contributions to the resulting scoring functions in Table 13.

(3) To permit comparison with current DOCK scoring technology, the CLO and CDO tests sets were also run using contact and energy scores. In addition the CLO and CDO derived scoring functions were run on their respective test set DOCK orientations to test the relative transferability of the scoring functions. Results are shown in Table 14.

(4) While the GA-TEC scoring functions developed here are far from maturity, it was decided to run some SVS

Table 9
CDO RMS deviations from CLO crystal structure binding mode

Complex	RMSD	Heavy atoms
4phv	10.211	46
1apv	5.072	36
7cat	3.924	48
5p21	**2.208**	**32**
4sga	2.017	34
7est	1.559	30
3cpa	1.360	17
6tim	1.328	10
4mdh	**1.150**	**44**
1xli	**1.084**	**12**
8cpa	**1.051**	**32**
3fx2	**0.967**	**31**
5tmn	**0.963**	**32**
2ifb	**0.943**	**18**
2tsc	0.901	35
2pk4	0.900	9
1rnt	0.882	24
1rbp	0.820	21
1dr1	**0.811**	**17**
4dfr	0.804	33
1ak3	0.785	23
3gap	0.780	22
1pph	0.748	30
9aat	0.739	16
3tpi	**0.669**	**16**
1abe	0.615	10
1snc	0.575	25
1phg	**0.516**	**17**
2tmn	0.395	13
1phd	0.364	11

Test set complexes are shown in bold.

Table 10
The number of orientations ranking above each CDO of the CDO training dataset based on the scoring function coefficients presented in Table 12

Complex	Run		
	1	2	3
1phd	260	260	260
7est	114	114	114
1rnt	111	111	111
2tmn	83	83	83
3cpa	66	66	66
9aat	37	37	37
2tsc	27	27	27
1abe	24	24	24
6tim	22	22	22
2pk4	16	16	16
1ak3	13	13	13
4dfr	13	13	13
3gap	9	9	9
1pph	2	2	2
1rbp	1	1	1
1snc	1	1	1
Rounded average rank	50	50	50

Table 11
The number of orientations ranking above each CDO of the test dataset based on the scoring function coefficients presented in Table 12

Complex	Run		
	1	2	3
1xli	11785	11783	11781
1dr1	9503	9501	9499
2ifb	848	848	848
1phg	99	99	99
4mdh	29	29	29
5tmn	4	4	4
3fx2	2	2	2
3tpi	2	2	2
8cpa	1	1	1
Rounded average rank	2475	2474	2475

Table 12
Normalized coefficients and percentage fitness values resulting from three CDO GA runs

Run	Coefficients				Fitness (%)
	a_{clash}	a_{hsurf}	a_{hbond}	$a_{electro}$	
1	−0.203	0.009	0.974	−0.103	99.792
2	−0.203	0.009	0.974	−0.103	99.792
3	−0.203	0.009	0.974	−0.103	99.792

scenarios to see if any additional insights could be gleaned. To test the resultant functions, two target proteins with active datasets used in previous DOCK sampling studies [21] were selected for analyses. Each active dataset was combined with ∼10,000 noise molecules from the BMS in-house database. All molecules were docked using identical protocols to the sister study [21], using a modified version of DOCK 4.01 (conformational search problems addressed). These were each docked into the target active site using DOCK contact score and site points with chemical matching and critical regions that reflected known structural biology (e.g. fatty acid binding protein ligands had to bind an acid in the acid binding region of the active site. See parent reference for more information [21]). All scoring orientations were rescored using the TEC scoring functions and the best result retained. In this way the issue of docking and ranking using different scoring functions was negated. Chemotype enrichment rates were then compared with those of the parent DOCK runs. Chemotype enrichment results are shown in Fig. 3.

Table 14
Rounded average rank DOCK contact and energy metrics when scoring CDO and CLO ligand orientations

Dataset	Contact		Energy		GA	
	CDO	CLO	CDO	CLO	CDO	CLO
Training	1957	442	17039	939	51	4
Test	2690	816	16518	6961	2476 (75)	37 (19337)

Earlier GA results are included for easy reference. In addition results in brackets highlight what happens when CDO orientations are used with the CLO function and vice versa.

Table 13
The average normalized percentage contributions of descriptors to the training set CDOs, calculated using the coefficients of Table 12 (standard deviations for each value shown in brackets)

Run	Normalized percentage contribution (S.D.)			
	$a_{clash}f_{clash}$	$a_{hsurf}f_{hsurf}$	$a_{hbond}f_{hbond}$	$a_{electro}f_{electro}$
1	−8.925 (6.896)	51.731 (31.855)	37.567 (25.988)	19.626 (17.551)
2	−8.921 (6.892)	51.729 (31.853)	37.569 (25.988)	19.622 (17.549)
3	−8.922 (6.894)	51.723 (31.857)	37.576 (25.991)	19.623 (17.549)

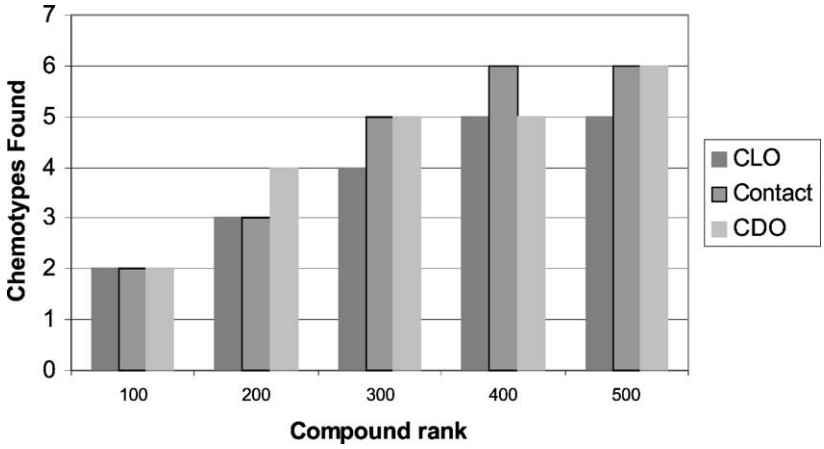

Fig. 3. Observed chemotype enrichment levels for SVS test calculations on fatty acid binding protein (eight chemotypes present in active data [21]).

4. Discussion

4.1. Results analysis

It was found that although separate GA runs did not converge to the same answer, the final functions generally bore a high degree of similarity (Tables 7 and 12). For both the CLO and CDO scoring functions, the dominant descriptors are the classic staples of scoring function development, hydrogen bonding and buried hydrophobic surface (Tables 8 and 13). The high standard deviations highlight the varying binding requirements between different active sites, however, with relative metric contributions changing widely from complex to complex.

Initial appearances would suggest that the CLO function is performing extremely well, with seemingly excellent average rank data (sub 50) for both training and test sets. The CDO model looks somewhat less impressive (Tables 10 and 11), particularly in the case of the test set, where the rank has dropped to near 2500. Analysis of poorly performing systems (Table 11, 1xli, 1dr1, 2ifb) highlights a breakdown for systems where the CDO "binding" mode is unable to maintain the ideal hydrogen bond orientations seen in the CLO.

An initial analysis ranking results also suggests that the GA functions are performing significantly better than their DOCK counterparts (Table 14). Closer data inspection highlights significant model issues, however, when the clash coefficient and the percentage contribution to the active ligand orientation score are compared (CLO Tables 7 and 8, CDO Tables 11 and 12). The CLO system assigns a far higher coefficient for the clash term relative to the CDO model, yet the contribution to binding ligand orientation score is significantly less. This is understandable in the context of the underlying data, since a CDO "binding" mode more regularly contains clashes relative to the more precise CLO scenario. As a result clash metric is used by the GA to create a partial indicator variable allowing the CLO function to differentiate the active binding orientation. The primary consequence of this is that the resulting CLO function becomes too sensitive to clashes when less precise binding data is presented to it. This can be seen when the CDO and CLO functions are run with each others test sets docked orientations (Table 14). The CDO function performs well (sub 100 rank) while the CLO model collapses (rank: >19,000). The DOCK energy score is also found to perform poorly relative to the contact metric for the same reason. In this case the r [12] repulsion term shows over-sensitivity to clashes compared to the user definable clash penalty associated with the contact score. Comparing the best GA model with the best performing DOCK score, the CDO model appears to offer some improvement (Table 14), albeit only slight in the case of the test set (rank: ∼2500 versus ∼2700).

The SVS test shows some similarity in terms of relative performance. For this study, all orientations have undergone rigid body minimization, significantly ameliorating the clash sensitivity of the CLO model. Nevertheless, CLO function enrichment is the worst performing of the scoring functions. The CDO model shows some suggestion of improvement over contact score being the only function to find four chemotypes in the top 200 ranked compounds. Performance is similar overall, however, and more work would be needed to understand if this kind of difference is truly significant and sustained.

4.2. New directions and old problems

There are numerous areas for further development in the scoring function calculations presented. A larger noise dataset is required to permit a diverse set of high scoring noise data, particularly with respect to competing hydrogen bond orientations (something that was somewhat lacking in the current noise data). The dataset should not be too large, however, given the increasing probability of finding real hits amongst the mix (∼10,000 as used in concurrent SVS sampling studies [21] should suffice). The DOCK orientations used in this study were not minimized. The results from study four highlight how minimization can partially alleviate issues regarding over-sensitivity to docking mode inaccuracies. On a more general note, replacing DOCK with a screening technique that uses an alternative (and potentially more rigorous) sampling paradigm (e.g. Prometheus [22] or GOLD [23]) would likely also prove beneficial.

Since these studies were initiated, it has been determined that the use of critical regions and chemical matching within DOCK generally illicit superior SVS results when compared with simply running DOCK unconstrained with SPHGEN generated site points [21]. Such information was not used in orientation generation, which can lead to poorer hydrogen bond mapping and over-scoring of less critical active site regions. The hydrogen bond mapping problem is indicative of perhaps the most important failing within current scoring functions, that of poor hydrogen bond description. The treatment of hydrogen bonds is often based on simple geometries without reference to their surroundings. When working on multiple structure-based design projects it quickly becomes apparent that this treatment is a gross approximation, given the significant variability evident in hydrogen bonding strengths [24,25]. This is further exacerbated in the case of salt bridges. The significance of salt bridges is such that they often constitute a critical component in the minimum pharmacophore used to define many leads [26], yet most scoring functions developed so far singly fail to grasp this importance. While the electrostatic term can go some of the way to attempting this, the inability to include polarization terms, deal satisfactorily with metal atom interactions and the use of simplistic charge models renders such an approach a partial solution at best. The complex nature of salt bridge networks [27], configuration docking mode inaccuracies and potentially significant entropic issues all contribute to a problem of considerable magnitude. One possibility to overcome some of these issues might be to create multiple scoring functions for different protein

classes. The importance of certain hydrogen bonds and salt bridges might then be better described since there would be more internal consistency in their representation. In the mean time, a more fruitful approach may be to look at descriptors that indicate the presence of repulsive hydrogen bonding and salt bridges. It is likely that any such interaction will prove problematic in binding, making its absolute value of less importance relative to the measurement of its attractive cousin. The inaccuracies in strength measurement should thus have a lessor effect on descriptor utility. Some efforts have been made in this regard with the LigScore function [3]. Until these issues are better addressed, however, the most fruitful method for incorporation of attractive hydrogen bonds and salt bridges is liable to continue to be through pharmacophoric constraints based on solid structural biology data [21].

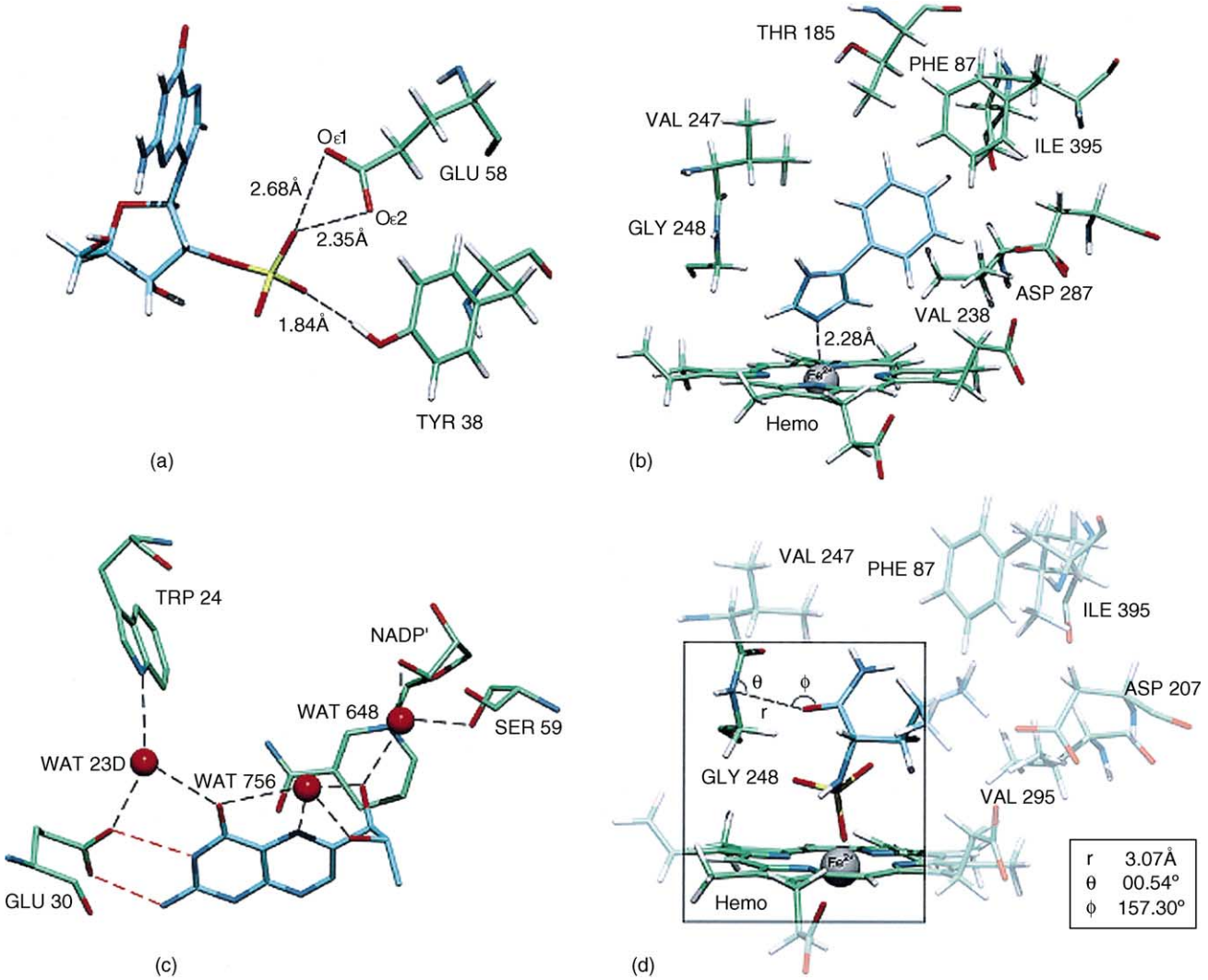

Fig. 4. Examples highlighting the complexity of dataset selection and technical/scientific issues found for complexes chosen for use in this study. (a) In general, the electrostatic interactions formed between each CLO and its protein are favorable. There are exceptions, however. A pertinent example is ribonuclease T1 complexed with 2′-GMP (add PDB three-letter code 1rnt). The unfavorable electrostatic component of the score is primarily due to the incorrect protonation state of GLU58 (this structure has been solved at a pH of 5.0 and as a consequence the glutamate is expected to be protonated). (b) The charge schemes typically used in SVS calculations are approximate and take no account of polarization effects. In many situations, particularly those involving metals where complex polarization effects are key to electrostatic interactions, these schemes are too simplistic to cope satisfactorily. An example of this is cytochrome P450cam with 4-phenyl imidazole (PDB code 1phd). (c) Hydrogen bonding via bridged water molecules is often observed but is generally ignored in scoring function design. An example of this is the crystallographically observed orientation of biopterin. As a result of the missing water molecules, only two hydrogen bonds (highlighted in red) are identified by TEC and contribute to the total score (PDB code 1dr1). PDB structure before processing and addition of hydrogens is shown. (d) Highlights a problem with the current hydrogen bond term of the scoring function. A hydrogen bond is identified between the carbonyl oxygen of the amide group from the ligand and the backbone nitrogen of GLY248 (1phd). Although both the XÔA and DÂY would result in a non-zero interaction, the backbone nitrogen is clearly not in a position to form a hydrogen bond. In cases where the donor heavy atom is free to rotate this does not cause a problem as the hydrogen atom can reorient to form the hydrogen bond. However, in this case, the backbone nitrogen is constrained by the two attached carbon atoms and is not free to do so. An additional term is required in such cases to measure the angle formed between the plane containing the three C–N–C backbone atoms and the acceptor atom.

The challenges of configuration sampling are equally significant. Increasing CPU power (for example, through GRID-based computing technology [28]) provides the possibility of increased configuration and conformation sampling to mitigate this issue. This power does not overcome the problem of missing molecular mechanic parameters that currently exist for large sections of in-house databases, however. Such holes in parameter space create a major hurdle for the calculation of accurate conformational energies in virtual screening. This represents a lessor obstacle for simple functions such as contact score, but the bar is significantly raised with more exacting scoring functions. The inability to find the exact binding mode was partially explored with our CDO calculations. The results obtained suggest that even with binding modes generally deemed correct (RMS: <2.0 Å), the possibility exists for descriptor breakdown for highly directional metrics (in this case hydrogen bonding function). Even here, however, the assumption is made that conformational sampling will find the binding conformation, something that has recently been shown to be somewhat optimistic for many ligands [29]. Finding the right resolution balance for configuration dependent descriptor parameters will continue to provide researchers with a significant challenge for some time to come. This limitation again highlights the need to bring in descriptors that measure reasons for lack of binding, one of the primary reasons for the creation of the GA scoring function paradigm. Unfortunately, the small size of the PDB dataset remaining after filtering required a major scaling back of the descriptors employed, and highlights one of the more central issues in scoring function design, that of clean dataset creation. Many new complexes have been added since the original dataset was created, and the PDB needs to be scoured for these new targets. A careful selection will be required, however, given that these and other studies [3] suggest that certain classes of molecules (sugars and multiply charged substrates are prime examples) show different binding strength behavior relative to their more “drug-like” counterparts. Further, as we have already shown, current datasets contain numerous hidden issues in PDB complexes (e.g. odd hydrogen bonding/electrostatic interactions, crystal contacts, incorrectly assigned side chains [12]) that can potentially affect the accuracy of metric construction and the resulting scoring functions (Fig. 4 contains further examples of these problems). Extensive analyses of chosen complexes are therefore likely to be crucial in ensuring optimum conditions for function development. Efforts have recently been made in this regard in an attempt to create a more universal scoring function dataset [30,31], and such research needs to be encouraged.

5. Conclusions

A new method for SVS scoring function development that permits the inclusion of data beyond that of known binders is presented together with the initial results of its application. The studies illustrate the intertwined nature of the function with other key elements of the chosen virtual screening algorithm, particularly in the context of sampling limitations. Even small deviations (RMS: <2.0 Å) from known binding modes can be enough to cause a breakdown in high-resolution descriptors (e.g. hydrogen bonding). Extensive analysis of our raw PDB data highlighted a number of important issues for obtaining clean and diverse dataset, something that has been somewhat overlooked in the field to date. These limitations are such that simpler scoring functions (e.g. DOCK contact score) will continue to play a role in future SVS calculations. Nevertheless, further work using techniques such as the one described are expected to eventually yield improvements in SVS hit rates. It is also anticipated that the ideas and challenges presented will stimulate others in the quest for more accurate virtual screening technology. (*Note*: TEC and GA program source is available to all interested parties. Please contact the corresponding author for more information. For more details on these studies, see http://www.ysbl.york.ac.uk/~ryan/Thesis/.)

References

[1] H.J. Bohm, M. Stahl, Rapid empirical scoring functions in virtual screening applications, Med. Chem. Res. 9 (1999) 445–462.
[2] C.A. Baxter, C.W. Murray, B. Waszkowycz, J. Li, R.A. Sykes, R.G.A. Bone, T.D.J. Perkins, W. Wylie, New approach to molecular docking and its application to virtual screening of chemical databases, J. Chem. Inf. Comput. Sci. 40 (2000) 254–262.
[3] M. Waldman, P. Kirchoff, J. Jiang, C.M. Venkatachalam, Progress toward a protein–ligand scoring function for fast docking. Abstr. Pap., Am. Chem. Soc. 220 (2001) CINF-1. URL: http://www.lib.uchicago.edu/cinf/220nm/final_program_220.html.
[4] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein data bank, Nucleic Acids Res. 28 (2000) 235–242.
[5] I. Muegge, Y. Martin, A general and fast scoring function for protein–ligand interactions: a simplified potential approach, J. Med. Chem. 42 (1999) 791–804.
[6] J.B.O. Mitchell, R.A. Laskowski, A. Alex, J.M. Thornton, BLEEP—potential of mean force describing protein–ligand interactions. I. Generating potential, J. Comput. Chem. 20 (1999) 1165–1176.
[7] J.B.O. Mitchell, R.A. Laskowski, A. Alex, J.M. Thornton, BLEEP—potential of mean force describing protein–ligand interactions. II. Calculation of binding energies and comparison with experimental data, J. Comput. Chem. 20 (1999) 1177–1185.
[8] H. Gohlke, M. Hendlich, G. Klebe, Predicting binding modes, binding affinities and ‘hot spots’ for protein–ligand complexes using a knowledge-based scoring function, Perspect. Drug Discov. Des. 20 (2000) 115–144.
[9] DOCK, Developed and distributed by the Kuntz group, Department of Pharmaceutical Chemistry, University of California, 512 Parnassus, San Francisco, CA 94143-0446, USA. URL: http://www.cmpharm.ucsf.edu/kuntz.
[10] D.A. Gschwend, A.C. Good, I.D. Kuntz, Molecular docking towards drug discovery, J. Mol. Recognit. 9 (1996) 175–186.
[11] H.-J. Böhm, The development of a simple empirical scoring function to estimate the binding constant for a protein–ligand complex of known three-dimensional structure, J. Comput.-Aided Mol. Des. 8 (1994) 243–256.

[12] R.W.W. Hooft, G. Vriend, C. Sander, E.E. Abola, Errors in protein structures, Nature 381 (1996) 272 (URL: http://www.cmbi.kun.nl/gv/pdbreport/).
[13] SYBYL, Developed and distributed by Tripos, Inc., 1699 South Hanley Road, Suite 303, St. Louis, MO 63144, USA.
[14] J. Gasteiger, M. Marsilli, Iterative partial equalization of orbital electronegativity: a rapid access to atomic charges, Tetrahedron 36 (1980) 3219–3222.
[15] World Drug Index, Developed and distributed by Derwent Publications Ltd., 14 Great Queen Street, London, UK.
[16] DAYLIGHT software, Developed and distributed by Daylight Chemical Information Systems, 27401 Los Altos, Suite #360, Mission Viejo, CA 92691, USA.
[17] CONCORD 4.0, Distributed by Tripos, Inc., 1699 South Hanley Road, Suite 303, St. Louis, MO 63144, USA. URL: http://www.tripos.com.
[18] CONFIRM, Part of CATALYST 3D 4.0 used in these searches, Developed and distributed by Accelrys, 9685 North Scanton Road, San Diego, CA 92121, USA.
[19] R. Wooton, R. Cranfield, G.C. Sheppey, P.J. Goodford, Physicochemical activity relationships in practice. 2. Rational selection of benzenoid substituents, J. Med. Chem. 18 (1975) 607–613.
[20] SUGAL genetic algorithm package (version 2.1), developed by Dr. Andrew Hunter at the University of Sunderland, England. URL: http://www.trajan-software.demon.co.uk/SUGAL.htm.
[21] A.C. Good, D.L. Cheney, D.F. Sitkoff, J.S. Tokarski, T.R. Stouch, D.A. Bassolino, S.R. Krystek, Y. Li, J.S. Mason, Analysis and optimization of structure-based virtual protocols. 2. Examination of docked ligand orientation sampling methodology mapping a pharmacophore for success. J. Mol. Graph. Mod., in press.
[22] Prometheus, Part of the PRO_LEADS program developed by the CADD arm of Protherics PLC (now part of Tularik), The West Heath Business and Technical Park, Runcorn, Cheshire, W47 4QF, UK.
[23] GOLD, version 1.2, Developed and distributed by the CCDC, 12 Union Road, Cambridge, CB2 1EZ, UK. URL: http://www.ccdc.cam.ac.uk/prods/gold/index.html.
[24] S.S. Abdel-Meguid, B. Zhao, K.H.M. Murthy, E. Winborne, J.K. Choi, R.L. DesJarlais, M.D. Minnich, J.S. Culp, C. Debouck, Inhibition of human immunodeficiency virus-1 protease by a C2-symmetric phosphinate. Synthesis and crystallographic analysis, Biochemistry 32 (1993) 7972–7980.
[25] C. Laurance, M. Berthelot, Observations on the strength of hydrogen bonds, Perspect. Drug Discov. Des. 18 (2000) 38–69.
[26] S.J. Teague, A.M. Davis, P.D. Leeson, T. Oprea, The design of lead-like combinatorial libraries, Angew. Chem., Int. Ed. 38 (1999) 3743–3748 (URL: http://cisrg.shef.ac.uk/shef2001/talks/davis/).
[27] S. Albeck, R. Unger, G. Schreiber, Evaluation of direct and cooperative contributions towards the strength of buried hydrogen bonds and salt bridges, J. Mol. Biol. 298 (2000) 503–520.
[28] E.K.D. Davies, W.G.R. Richards, The potential of Internet computing for drug discovery, Drug Discov. Today, in press.
[29] A.C. Good, D.L. Cheney, Analysis and optimization of structure-based virtual protocols. 1. Exploration of ligand conformational sampling techniques. J. Mol. Graph. Mod., in press.
[30] O. Roche, R. Kiyama, C.L. Brooks III, Ligand–protein database: linking protein–ligand complex structures to binding data, J. Med. Chem. 44 (2001) 3592–3598.
[31] J.W.M. Nissink, C. Murray, M. Hartshorn, M.L. Verdonk, J.C. Cole, R. Taylor, A new test set for validating predictions of protein–ligand interaction, Proteins: Struct. Funct. Genet. 49 (2002) 457–471.