PROTEINS: Structure, Function, and Genetics 49:457–471 (2002)

A New Test Set for Validating Predictions of Protein–Ligand Interaction

J. Willem M. Nissink,[1*] Chris Murray,[2] Mike Hartshorn,[2] Marcel L. Verdonk,[2] Jason C. Cole,[1] and Robin Taylor[1]
[1]*Cambridge Crystallographic Data Centre, Cambridge, United Kingdom*
[2]*Astex Technology Ltd, Cambridge, United Kingdom*

ABSTRACT **We present a large test set of protein–ligand complexes for the purpose of validating algorithms that rely on the prediction of protein–ligand interactions. The set consists of 305 complexes with protonation states assigned by manual inspection. The following checks have been carried out to identify unsuitable entries in this set: (1) assessing the involvement of crystallographically related protein units in ligand binding; (2) identification of bad clashes between protein side chains and ligand; and (3) assessment of structural errors, and/or inconsistency of ligand placement with crystal structure electron density. In addition, the set has been pruned to assure diversity in terms of protein–ligand structures, and subsets are supplied for different protein-structure resolution ranges. A classification of the set by protein type is available. As an illustration, validation results are shown for GOLD and SuperStar. GOLD is a program that performs flexible protein–ligand docking, and SuperStar is used for the prediction of favorable interaction sites in proteins. The new CCDC/Astex test set is freely available to the scientific community (http://www.ccdc.cam.ac.uk). Proteins 2002;49:457–471.**

Key words: docking test set; ligand docking; validation; drug design; flexible docking; protein–ligand data set; GOLD; SuperStar

INTRODUCTION

The concept of testing new theories and stratagems in an empirical way firmly underpins our approach to science. Especially in the field of the life sciences, where lack of data often makes testing of new algorithms difficult, the only way to make sure that a method works is by extensive application to test cases. For the testing of docking programs, several sets of protein–ligand complexes have been used in recent years. The GOLD docking program has been tested on a set of 134 protein–ligand complexes,[1] to assess its capability to predict native binding modes of ligands, a fundamental requirement of all docking algorithms. The docking program FlexX has been tested on a similarly constructed set of 200 complexes.[2] For EUDOC, a test suite of 154 complexes is reported.[3] The latter 2 test sets both include the complete GOLD test set. PRO_LEADS has been validated using a set of 70 complexes.[4] Most docking tools have been examined with distinctly smaller sets of protein–ligand complexes, e.g., a recent study describes DOCK4.0 results for 12 complexes[5]; DARWIN results are shown for 3 examples[6]; Blom and Sygusch[7] report results for 7 protein ligand complexes for a docking protocol using Fourier correlation techniques; for MCDOCK,[8] 19 protein ligand complexes are reported. Wang et al. report results for 12 complexes using DOCK3.5 in a multistep approach.[9]

Protocols that predict interaction sites in proteins or binding affinity in complexes are similarly tested on validation sets of protein–ligand complexes. The SuperStar program has been validated using the GOLD validation set of 134 complexes.[10] DrugScore has been tested on 159 complexes (taken from the FlexX validation set).[11] The PMF potentials derived by Muegge and colleagues have been validated on a set of 77 complexes.[12]

In all such cases, the validation set of protein–ligand complexes should be sufficiently large and diverse to ensure that statistically meaningful results can be obtained. Unfortunately, our knowledge-base of protein crystal structures is still relatively small, and biased in terms of diversity towards groups of protein families. Furthermore, limited resolution may preclude the use of certain structures in validation work.

In this article, we introduce the CCDC/Astex test set, a large, diverse set of protein–ligand structures that have been checked extensively for errors. The set is intended as a useful starting point for the testing and comparison of, e.g., docking algorithms and programs for binding site analysis. We discuss the criteria applied in setting up the CCDC/Astex test set, and we use the set to explore the performance of GOLD[1] (docking

Abbreviations: CSD, Cambridge Structural Database; df, degrees of freedom; PDB, Protein Data Bank; R, crystallographic resolution; RMS, root of mean squared deviations of atomic coordinates.

The Supplementary Material referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/2002/49_4/v49.457.html

*Correspondence to: Dr. J.W.M. Nissink, Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ UK. E-mail: nissink@ccdc.cam.ac.uk

Received 6 February 2002; Accepted 18 June 2002

program) and SuperStar[10] (hot-spot prediction for binding sites).

MATERIALS AND METHODS

Selection of an Initial Set of Protein-Ligand Complexes

The original GOLD test set[1] was used as the basis for the new CCDC/Astex test set. To the 134 existing entries (shown in Table I), 171 new entries were added, leading to a total number of 305 entries. Of the original GOLD entries, 3 were updated (2ack, 3mth, and 4aah), and known errors were fixed. Forty-eight entries of the Chemscore test set[4] that were not in the original GOLD set were added. Sixty-six entries of the Chemscore set are present in total, as 18 entries of the latter were already available in the original GOLD set.

The selection of a further 123 new entries (see Table I) was inspired in several ways. Topics covered at recent conferences and areas of general interest were taken as a guideline, and we attempted to cover pharmaceutically interesting protein families as well as possible. Additionally, we aimed to enrich the test set by focusing on structurally different classes of proteins. This was done by specifically searching for entries from the 58 best-populated structural classes of proteins (classification by Shindyalov[13]), and their homologues.

Relibase+[14] was used to search for and browse through families of homologous proteins and their ligands quickly. Selected complexes were rejected if their crystallographic resolution was worse than 3.0 Å, although in retrospect, some rhinovirus protein structures were added with resolutions >3.0 Å. In the case of disordered ligands, one of the orientations of the ligand was taken after visual inspection. The 305 entries of the complete CCDC/Astex set can be found in Table I.

Setting Up Protein–Ligand Complexes

Protein–ligand entries were set up starting from the original PDB data, analogous to the procedure described by Jones et al.[1] Protein and ligand structural data were separated into different files. Protein structures were reduced in some cases by deleting chains that were not near the active site, in order to limit the protein-file size; new entries have not been centered on the origin (unlike entries of the existing GOLD test set), and retain the coordinate system as given in the PDB file. The centre of the active site was determined, alongside a value for the radius to designate its extent. Water molecules were removed from the protein when present, and stored in a separate file (new entries only); however, some exceptions were made for water molecules in metal ion complexes.[1] Hydrogen atoms were added to both protein and ligand using Sybyl[15a]; the orientation of rotatable OH and NH_3 groups was not optimized. Protein flexibility has not been addressed in this test set.

Considerable care was taken over the assignment of protonation and tautomeric states of both the protein and the ligand. Aspartic and glutamic acids are generally ionized, but may be protonated on either oxygen atom in specific cases. In the case of aspartic proteases, a proton was added to one of the proximal Asp residues, in the position that results in the most favorable orientation for hydrogen bonding; if this was not possible, the protein residue that was most solvent-exposed was protonated. Particularly important is the ionization and tautomeric state of histidine residues. These can be positively charged or neutral with a proton on either N^{δ} or N^{ϵ}. Regarding the ligand, particular attention was given to charge states of basic and acidic nitrogens, and acidic oxygens. For a detailed account of common protonation states of residues in proteins, we refer to Hooft et al.[15b] In all cases, protonation states have been inferred by inspecting the immediate hydrogen- bonding environment of the groups in question.

Application of Filters for Detecting Suspect Structures

After the initial stage of preparing files and setting up the protonation states of all entries, several filters were designed. These filters were applied to identify those entries that are not reliable enough for inclusion in a data set to be used for validation purposes. Reasons to exclude protein structures can be: (1) the presence of factual and structural errors in the original PDB file; (2) an inconsistency between the electron density in the binding site and the reported position and conformation of the ligand; or, an unlikely conformation of the ligand; (3) the presence of severe clashes between protein and ligand atoms; (4) the necessity of including crystallographically related protein residues to describe the binding site correctly. These filters are described in more detail below.

Factual and structural errors in PDB files

A factual error would be an error that results in an incomplete or wrong ligand structure. An error would be structural if the actual information is more or less complete yet erroneous. A factual error, e.g., would be the inadvertent omission of part of the HETATOM fields in the PDB file, leading to an incomplete representation of the ligand structure. A structural error would be a representation of the ligand that is wrong, which led to a faulty geometry, e.g., a non-planar phenyl ring.

Unlikely ligand conformations

Ligand conformations were regarded as unlikely if (1) they were inconsistent with the electron density generated for the active site in the protein; or (2) when their geometry exhibited unlikely conformations of chemical groups that could not be justified, e.g., the presence of acyclic *cis*-ester groups.

To evaluate the match between ligand and electron density, omit maps were generated for the 70 complexes whose structure factors were available from the Protein Data Bank. The CCP4 suite of programs[16] was used to generate the omit maps. The basic procedure involved removal of all ligands from the complex. Structure factors were generated for the remaining atoms and an electron density map was calculated. The map was contoured at 2.5

TABLE I. Contents of the New Set of Protein-Ligand Complexes: Original GOLD Test Set (134 Entries), 66 Protein-Ligand Complexes From the Chemscore Set (of Which 18 Overlap With Existing Entries), and 123 New Complexes

GOLD set (134)	Chemscore set (48 + 18)	Extended entries (123)
1aaq 1abe 1acj 1acl 1acm	*Extended entries from the Chemscore set*	1a07 1a0q 1a1b 1a1e 1a28
1aco 1aec 1aha 1apt 1ase	1abf 1aoe 1apu 1dmp 1dog	1a42 1a4g 1a4k 1a4q 1a6w
1atl 1azm 1baf 1bbp 1blh	1dwb 1ebg 1epo 1ets 1ett	1a9u 1ai5 1aj7 1ake 1aqw
1bma 1byb 1cbs 1cbx 1cdg	1fax 1hpv 1htf 1hvr 1jao	1b58 1b59 1b6n 1b9v 1bgo
1cil 1com 1coy 1cps 1ctr	1jap 1mbi 1mmb 1mnc 1mtw	1b17 1bmq 1byg 1c12 1c1e
1dbb 1dbj 1did 1die 1drl	1nsd 1okl 1okm 1pgp 1phf	1c2t 1c5c 1c5x 1c83 1cf8
1dwd 1eap 1eed 1epb 1eta	1ppc 1pph 1qbr 1qbt 1qbu	1cin 1ckp 1cle 1cqp 1ctt
1etr 1fen 1fkg 1fki 1frp	1rbp 1sln 1tlp 1tnh 1uvs	1cvu 1cx2 1d01 1d3h 1d4p
1ghb 1glp 1glq 1hdc 1hdy	1uvt 2cpp 2h4n 2ifb 2r04	1dbm 1dd7 1dg5 1dhf 1dwc
1hef 1hfc 1hri 1hsl 1hyt	2tmn 2tsc 2ypi 4tpi 5abp	1dy9 1eil 1ejn 1ela 1elb
1icn 1ida 1igj 1imb 1ive	5cpp 6cpa 7cpa	1elc 1eld 1ele 1eoc 1etz
1lah 1lcp 1ldm 1lic 1lmo		1f0r 1f0s 1f3d 1fbl 1fgi
1lna 1lpm 1lst 1mcr 1mdr		1fig 1f13 1flr 1gpy 1hak
1mmq 1mrg 1mrk 1mup 1nco		1hiv 1hos 1hsb 1hti 1ibg
1nis 1pbd 1pha 1phd 1phg	*Remaining Chemscore entries (overlap*	1ivb 1ivc 1ivd 1ivq 1kel
1poc 1rds 1rne 1rob 1slt	*with GOLD set, 18 entries)*	1kno 1lkk 1lyb 1lyl 1mcq
1snc 1srj 1stp 1tdb 1tka	1abe 1cbx 1dbb 1dbj 1etr	1ml1 1mld 1mts 1ngp 1pdz
1tmn 1tng 1tni 1tnl 1tph	1hsl 1phg 1stp 1tmn 1tng	1ppi 1ppl 1pso 1ptv 1qcf
1tpp 1trk 1tyl 1ukz 1ulb	1ulb 2cgr 2ctc 2gbp 2phh	1qh7 1ql7 1qpe 1qpq 1rnt
1wap 1xid 1xie 2ack 2ada	3ptb 3tpi 4dfr	1rt2 1srf 1srg 1srh 1vgc
2ak3 2cgr 2cht 2cmd 2ctc		1vrh 1xkb 1ydr 1yds 1ydt
2dbl 2gbp 2lgs 2mcp 2phh		1yee 2Sc8 2aad 2er7 2fox
2pk4 2plv 2r07 2sim 2yhx		2mip 2pcp 2qwk 3erd 3ert
3cla 3cpa 3gch 3hvt 3mth		3gpb 3nos 3pgh 4cox 4er2
3ptb 3tpi 4aah 4cts 4dfr		4fbp 4lbd 5er1
4est 4fab 4phv 5p2p 6abp		
6rnt 6rsa 7tim 8gch		

sigma, overlaid on the original structure, and inspected (see Supplementary Material for further details).

Presence of clashes between protein and ligand atoms

Protein structures were deemed unsuitable if they contained severe clashes between protein and ligand atoms. To allow for the inaccuracy in the actual atomic positions, a clash was considered unacceptable if the protein atom-ligand atom distance was shorter than the minimum distance by 20% or more. Minimum distances were derived from small-molecule crystallograpic data by looking for the shortest contact distances in the relevant IsoStar[17] plots; e.g., for the hydroxyl-aliphatic C distance, the aliphatic hydroxyl central group was checked for methyl and methylene contacts. As short contact distances may arise from "unusual" chemical environments, they were checked visually in IsoStar. In those cases where a minimum distance could not be derived reliably due to lack of data, the sum of the non-bonded contact radii[18] was taken. Allowance was made for covalently-bonded complexes, as clashes often occur around the link between protein and ligand.

Minimum distances are displayed in Table II. In Table II, the terminology E(n) is used to designate an atom of type E that is bonded to n atoms, e.g., C(3) indicates a C atom bonded to three other atoms, such as an aromatic or olefinic carbon. Additionally, separate types have been added for charged oxygen in a carboxylic acid (O(co2)) and cationic carbon (C(cat), in guanidinium groups). Although the sub-specification of the element types in Table II may seem redundant, specific hydrogen-bonding contacts can be accounted for this way, e.g., to distinguish between oxygen and nitrogen in a salt bridge, or in an amide-amide hydrogen bond, the former of which will be much shorter.

Influence of crystallographically related protein residues

The actual binding site of a ligand in a protein can often, but not always, be determined fully by directly copying the atomic coordinates present in the PDB file. In about 10% of the cases, it is necessary to generate the crystallographically related protein chains to obtain a complete description of the binding site. We used Relibase+[14] to assess whether this was the case for all 305 entries of our set, and singled out those structures where ligand contacts to crystallographically related protein residues occur that are shorter than 4 Å.

Crystallographic Resolution

The resolution of a protein structure is directly related to the accuracy of the structural data. It is, therefore, not desirable to include structures in the test set that have insufficient resolution. Unfortunately, the availability of well-resolved, high-resolution protein structures is still limited. However, one has to keep in mind that the resolution on its own is not a good enough indicator of the actual quality or usability of the structure, as well-resolved structures may still contain badly modeled ligand

TABLE II. Minimum Distances for Selected Atom-Atom Contacts[†]

	C(3)	C(cat)	C(4)	N(2)	N(3)	N(4)	O(1)	O(co2)	O(2)	F	P(4)	S(2)	Cl	Br	I
C(3)	3.5	3.4	3.4	3.4	3.4	2.9	3.0	3.7	3.2	2.9	4.1	3.7	*3.5*	*3.6*	*3.8*
C(cat)		*3.5*	3.7	3.8	3.5	3.2	3.4	3.6	*3.3*	*3.2*	*3.5*	*3.6*	*3.5*	*3.6*	*3.8*
C(4)			3.4	3.5	3.4	3.5	3.2	3.0	*3.3*	3.0	4.0	3.7	*3.5*	*3.6*	*3.8*
N(2)				*3.2*	3.0	2.9	2.9	3.5	2.6	3.1	4.3	*3.4*	*3.4*	*3.5*	*3.6*
N(3)					3.2	3.2	2.8	2.9	2.8	3.4	3.6	*3.4*	*3.4*	*3.5*	*3.6*
N(4)						*3.2*	2.7	2.5	2.7	3.0	3.7	3.6	*3.4*	*3.5*	*3.6*
O(1)							3.2	*3.1*	2.7	3.5	3.4	3.5	*3.3*	*3.4*	*3.6*
O(co2)								*3.1*	2.4	3.0	4.0	*3.4*	*3.3*	*3.4*	*3.6*
O(2)									2.6	2.7	*3.4*	3.4	*3.3*	*3.4*	*3.6*
F										2.9	*3.3*	*3.2*	*3.2*	*3.3*	*3.4*
P(4)											*3.6*	*3.6*	*3.6*	*3.7*	*3.8*
S(2)												*3.6*	*3.5*	*3.6*	*3.8*
Cl													*3.5*	*3.6*	*3.7*
Br														*3.7*	*3.9*
I															*4.0*

[†]Contact distances are given for non-covalent interactions. Contacts were deemed unacceptable if their distance is less than the tabulated values by 20% or more. All distances are in Ångstrom. Distances are derived using IsoStar; for those cases where not enough data are available in IsoStar (italicized values), minimum distances have been estimated as the sum of non-bonded contact radii (non-bonded contact radii are taken from Rowland and Taylor, Table 3[18]; for P the Bondi radius was used, from Bondi[31]).

binding sites. On the other hand, algorithms may account for inaccuracy of the data (the use of soft potential functions in docking software, that allow slight atom-atom clashes), or may not be very dependent on data accuracy.

It was decided to provide the user with subsets of complexes at different resolutions. As a rule of thumb we recommend the use of structures with a crystallographic resolution of 2.5 Å or better.

Relation Between Set Size and Error Margins

Primarily, a test set will be used to assess the validity of predictions given by an algorithm by estimating the average performance of the algorithm. Having chosen a test set of a certain size, and given an algorithm, there are three types of error that can be expected to contribute to any prediction we make on the test set: a sampling error; an error due to data inaccuracy; and (for non-deterministic algorithms) an error due to lack of reproducibility of results.

The first type of error is linked to the limited size of the set. A test set can be regarded as being picked from a "universe" of, in our case, acceptable protein–ligand complexes. Every validation procedure that uses such a set to estimate a property (say, $\mathbf{X}$) that is calculated by a certain algorithm will yield an average value $\overline{\mathbf{X}}$. This value approximates the *real* result μ for the underlying population from which the test set was chosen. As an example, $\mathbf{X}$ could be the performance of a docking algorithm for a certain set of protein–ligand complexes, expressed as the fraction of docking solutions with RMS < 2.0 Å; $\overline{\mathbf{X}}$ would then be the estimated average performance of the algorithm; and μ would be the (hypothetical) average performance were it to be run on the whole universe of complexes. Generally, the larger the sample set is, the better $\overline{\mathbf{X}}$ will approximate μ; specifically, the standard error of $\overline{\mathbf{X}}$ will equal $s_{\overline{\mathbf{X}}}=\sigma/\sqrt{n}$, with σ being the standard deviation for the distribution of property $\mathbf{X}$, which we estimate by bootstrapping, see below, and n equal to the size of the test set.

The second type of error influences σ and is related to the inherent inaccuracy in the structural information that is present in the protein files. The error in the location of atoms is directly related to the crystallographic resolution at which the protein structure has been solved. Depending on the algorithm that is being tested, resolution might, or might not, have a drastic impact on results.

The third type of error is the variability associated with the algorithm itself that is being tested. This type of error will be present for algorithms that rely on a chance measure to obtain their solution; examples of such algorithms are genetic algorithms and simulated annealing. Ideally, such programs will yield the optimal solution every time, yet in practice, differences may occur from run to run.

Although often forgotten, we do have to take into account the standard error $s_{\overline{\mathbf{X}}}$ of our prediction $\overline{\mathbf{X}}$, if we want to decide whether results from different test sets or algorithms are different in a statistical sense. Especially for the smaller-sized test sets, the value of $s_{\overline{\mathbf{X}}}$ can be considerable, leading to large confidence intervals for our predictions. Since we do not know the variability of the underlying real distribution, the easiest way to estimate $s_{\overline{\mathbf{X}}}$ will be through the use of a bootstrapping procedure.[19] Using such a procedure has the added advantage that the three types of errors mentioned above are taken into account simultaneously.

When designing a new test set, we can influence both the first and second type of error: we try to minimize the impact of the sampling error by striving for a test set that has a large number of entries, and decrease the effect of structural errors by rejecting imprecise structures. The third type of error, irreproducibility of algorithmic results, is largely a trait of the algorithm, although it may have a component that is a result of

structural inadequacy in the test samples (if inaccuracy in the input data has a significant influence on the flow of the algorithm).

Concern for Diversity and Classification According to Protein Family

A test set used for validation should have a size that is large enough to yield a prediction of performance $\overline{\mathbf{X}}$ with a small confidence interval. Yet, such a result is rendered meaningless if the test set of structures shows low diversity in terms of both protein and ligand structures. In statistical terms, it would mean that the test set is a non-random sample from the "universe" of adequate complexes.

As our supply of protein crystal structures is still relatively small, and biased in some respect towards certain families of protein and ligand structures, it is not uncommon to see certain combinations of ligand and binding site in different crystal structures. We have strived for maximum diversity in the test set by visually checking all protein–ligand complexes, and identifying those combinations that were overrepresented. From a pair of similar structures, we retained the one with the best resolution. A classification of the test set in terms of protein family is available and will be used to analyze performance in this article.

TABLE III. List of Doubtful Entries[†]

Entry	Reason	Entry	Reason	Entry	Reason
1a07	AC	1ppl	A	1ivd	B
1a0q	A	1qbt	A	1ive	B
1a1b	A	1q17	A	1kno	B
1a1e	A	1sln	A	1lmo	B
1a4k	A	1srf	AB	1pha	B
1aj7	A	1srg	AB	2plv	B
1ake	A	1srh	AB	3gch	B
1cf8	A	1stp	A	3nos	B
1ctt	AB	1xkb	A	3pgh	B
1ela	A	1yds	A	4fab	B
1elb	A	2cgr	A	5p2p	B
1eld	AB	2er7	A	8gch	B
1ele	A	2mip	A	1b6n	C
1etz	A	3mth	A	1qh7	DE
1fbl	A	4er2	A	1dbm	DE
1jao	A	6cpa	A	1ctr	E
1lkk	A	7cpa	A	1icn	E
1mmb	A	1c2t	B	1ivc	E
1mnc	A	1ghb	B	1vrh	E
1mtw	A	1hdy	B		
1pgp	A	1hef	B		

[†]Reasons for exclusion; presence of contacts with crystallographically related chains (A); significant clashes between protein and ligand atoms (B); incorrect ligand representation (C); dubious ligand geometry (D); incongruity between ligand placement and electron density (E).

RESULTS

Application of Filters to Detect Suspect Entries

The filters were applied in the following order: (1) presence of crystallographically related protein residues near the binding site; (2) presence of severe clashes between protein and ligand atoms; (3) correct ligand structure; (4) unlikely ligand conformations; and (5) inconsistency between the electron density in the binding site and the actual conformation of a ligand. Table III shows the resulting list of 61 entries that should be regarded with care when used in a validation set.

Some examples are shown in the sections below.

Crystallographic symmetry

Inclusion of nearby crystallographically related residues, using Relibase+, shows clearly that the actual binding site of PDB entry 1mtw is only partially defined by the protein coordinates that are supplied in the PDB file. It cannot be expected that a docking program will find the right solution in this case if given only these coordinates. Indeed, GOLD has problems docking the native ligand into this structure [Fig. 1(a)], but is able to find the right solution [Fig. 1(c)] when proximal residues are included [Fig. 1(b)].

Significant clashes between ligand and protein atoms

Clashes between ligand and protein atoms occur quite frequently as a result of the PDB data being less well resolved than small-molecule CSD[20] structures. This can pose difficulties for docking algorithms, as taking into account the van der Waals repulsion will make it difficult to reproduce the actual binding pose of the ligand. This is very much the case for an entry like 1srf [Fig. 2(a)], that exhibits extremely short distances between the tert-butyl group on the ligand and an alanine (Ala50) of the protein. Another example is 8gch, which features a short contact between the carboxylate group of the ligand and the catalytic hydroxyl group of Ser195 [Fig. 2(b)]. Unlike entry 3gch, where there is a covalent link involving Ser195 (albeit in an awkward geometry), 8gch features very short distances around 2.2 Å for both ligand carboxylate oxygens.

Entries with errors and/or inconsistencies in their electron densities

As an example of an entry where there is a significant inconsistency between electron density and the ligand structure in the PDB file, we show PDB entry 1b6n in Figure 3. The crystallographic electron density extends well beyond the volume that is occupied by the ligand. On further inspection of the literature, it appeared that a phenylsulfone group is lacking from the ligand structure.[21]

The six complexes that are singled out in Table III are not necessarily erroneous, but our failure to match calculated electron density with reported ligand positions implies a lack of understanding on our part of the true binding situation. Elimination of these entries is therefore a safe option.

Details of the calculation and analysis of ligand electron density can be found in the Supplementary Material.

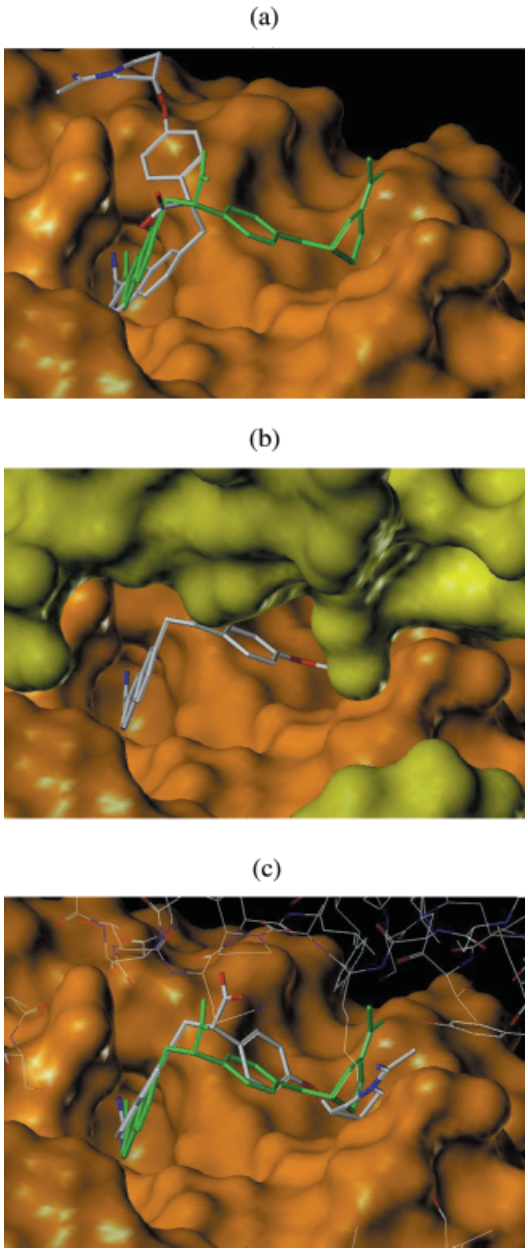

Fig. 1. GOLD results for PDB entry 1mtw. Docking of the native ligand in the binding site leads to a wrong solution (**a**, protein shown as surface, docking solution as atom-coloured sticks, reference ligand in solid color), yet inclusion of the crystallograpically related protein residues (**b**, additional protein chains shown as a surface) results in an acceptable solution (**c**, docking solution and reference ligand shown; crystallographically related protein residues shown as chain).

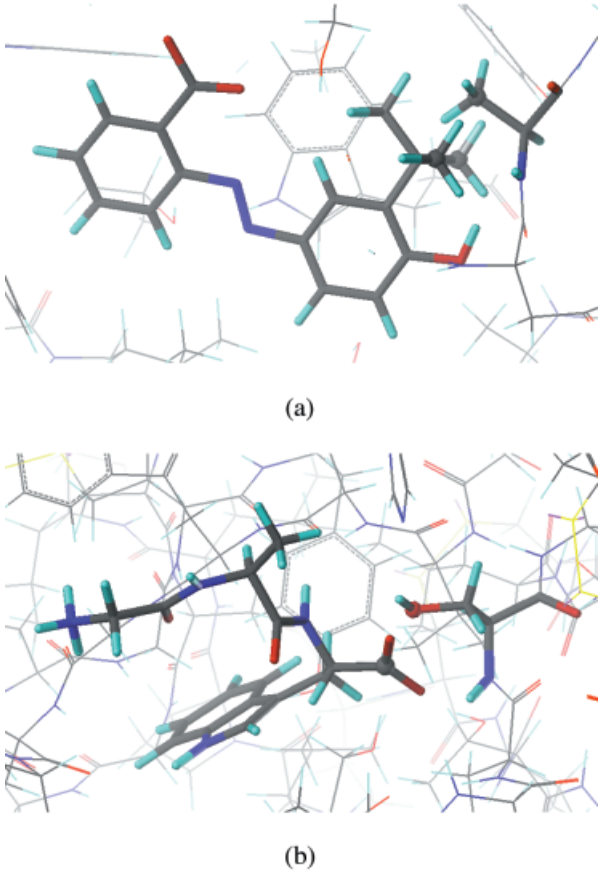

Fig. 2. Two examples of crystal structures with very short contacts between ligand and protein atoms: (**a**) PDB entry 1srf, $d_{c\text{-}c}$=2.2Å; (**b**) PDB entry 8gch, $d_{C\text{-}O}$=2.0 Å. Both ligand and protein residue are shown as capped sticks, and distances are given for the atoms highlighted as spheres.

Diversity

Diversity must be taken into account, and for that reason a classification of all entries into protein classes has been set up. The diversity of the test set was increased by identifying sets of similar protein–ligand entries, and limiting the number of their representations to alleviate redundancy. For such "doubles," or very similar entries, we chose to retain the one with the best resolution only. The PDB entries that have been removed from the set of 305 entries for reasons of diversity are shown in Table IV. Where possible, a structurally deficient entry (Table III) was picked for removal, rather than a non-deficient entry. The classification for the full set is available as supplementary material.

Clean Lists

Figure 4(a) and (b) show distributions of number of ligand atoms, number of rotational bonds, and crystallographic resolution of the protein structure for the original GOLD set of 134 entries and 66 entries of the Chemscore set. The histograms for the full set of 305 complexes are shown in Figure 4(c). The new CCDC/Astex test set contains on average heavier ligands than its predecessor, the GOLD set. There is more emphasis on ligands with a higher number of rotational bonds, and it is therefore to be expected that the new set is more difficult for docking algorithms than the original GOLD set. The number of less well-resolved protein structures has risen slightly for the new set.

Leaving out the suspicious entries from the full set results in a set of structures suitable for validation purposes (Table V). The full set of 305 entries covers 144

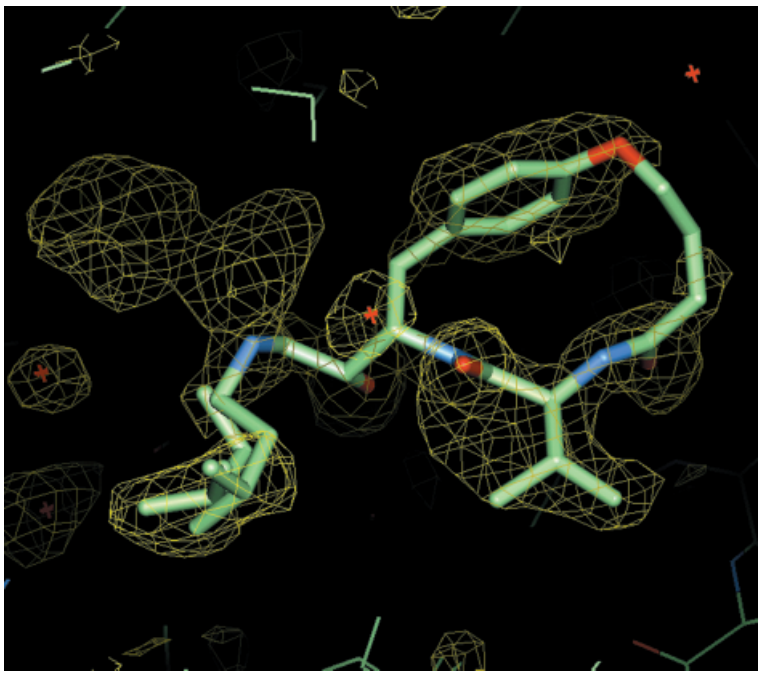

Fig. 3. PDB entry 1b6n; the ligand in the PDB lacks a phenyl sulfonamide group.

TABLE IV. PDB Entries That Were Removed From the Set for Reasons of Diversity†

1a07	*1alb*	1aha	1cin	*1dbm*	1die	*1elb*	1elc	*1eld*	1fig
1gpy	1hti	*1icn*	1igj	*1ive*	1mcr	1ml1	1nsd	1phf	1rbp
1tka	1tph	*1xkb*	2r04	2sim	3ptb	*4fab*	4tpi	6abp	

†Entries with structural deficits are shown in italics (i.e., overlapping with Table III).

different types of proteins. For the "clean" set (224 entries) obtained by excluding structurally deficient entries and the non-diverse complexes this number is 131. The "clean" sets with crystallographic resolution limited to less than 2.5 Å and 2.0 Å contain 180 entries distributed over 115 different protein types, and 92 entries over 62 types, respectively. Full lists of the protein entries present in the sets can be found in the Supplementary Material, or downloaded from the web.[22]

Most protein classes have 1 to 4 entries; classes with more representatives are (in the clean set of 224 entries): dihydrofolate reductase (5); carbonic anhydrase II (5); neuraminidases (5); trypsin (7); thrombin (α and ϵ) (9); hiv-1 protease (10).

Profiles for ligand size, number of rotational bonds, and crystallographic resolution of the clean list (224 entries) are shown in Figure 5. Profiles for the FlexX data set (200 entries) compiled by Kramer et al.[2] are largely comparable to those of the original GOLD test set, and the clean lists. The FlexX data set includes the GOLD set fully (except for one structure, 1acl), and adds 67 protein–ligand complexes. However, diversity in terms of protein–ligand complexes appears to be lower for this set of 200 entries: the FlexX set covers 99 different types of protein, and 9 of its types contain 5 or more entries (d-xylose isomerase (5); carboxypeptidase A (6); cytochrome p450-cam (6); hiv-1 protease (6); elastase (7); triosephosphate isomerase (7); trypsin (8); thermolysin (9); neuraminidase (12)). In addition, the FlexX data set contains 26 entries[23] among its 200 that we chose to omit in this study (see Table III).

Of the Chemscore data set, 66 entries are present in the new full set. The Chemscore data set contains 12 entries that we chose to neglect for structural reasons. After correction to ensure diversity, the clean list contains 49 of the Chemscore entries.[24]

VALIDATION RESULTS FOR SUPERSTAR

SuperStar

SuperStar is a program for generating maps of interaction sites in proteins using experimental information about intermolecular interactions.[10] The interaction maps that SuperStar generates are, therefore, fully knowledge-based. SuperStar retrieves its data from IsoStar,[17] which contains information about non-bonded interactions from both the Cambridge Structural Database (CSD)[20] and the Protein Data Bank (PDB).[25] For a given protein binding site and a probe group, SuperStar gathers the necessary interaction data and calculates a three-dimensional map that highlights regions in the cavity where the probe has a high probability of occurring.

Given a set of protein–ligand structures, it is possible to validate SuperStar, by assessing whether the experimentally-observed positions of ligand groups are those predicted by SuperStar. Specifically, maps are calculated for the following probes: alcohol oxygen, carbonyl oxygen, ammonium (RNH_3) nitrogen and methyl carbon. For appropriate groups in the ligands (namely hydroxyl groups, carbonyl groups, charged amino moieties, and methyls),

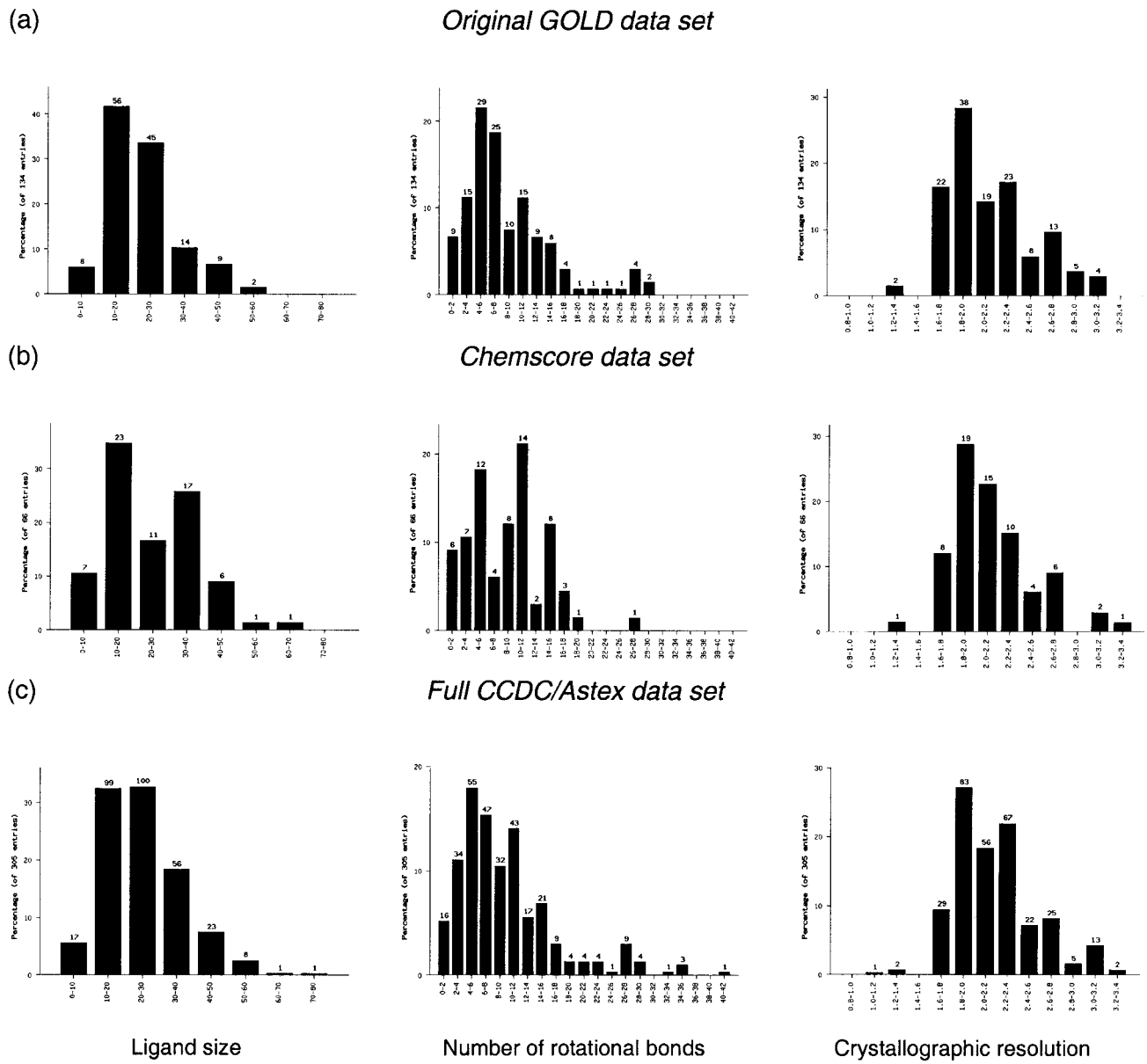

Fig. 4. Distributions of the ligand size (number of heavy ligand atoms), number of rotational bonds, and crystallographic resolution in (**a**) the GOLD test set, (**b**) the Chemscore test set, (**c**) the new CCDC/Astex test set.

the propensities of the calculated maps are checked, and a ligand group is considered to be correctly predicted if the propensity of its corresponding map is the highest of the four. The outcome of the validation is an overall success rate of prediction f_{corr} (i.e., the fraction of correctly predicted ligand groups). For further details of this procedure, we refer to Verdonk et al.[10]

SuperStar maps can be computed from raw PDB interaction data, raw CSD data, parameterized CSDdata ("fits")[26] (in which the raw data have been fitted by Gaussian representations), or a combination of parameterized and discrete CSD data ('hybrid method'). When using the latter method, SuperStar will substitute raw data with parameterized data only when a high-quality fit is available. Both the fit method and the hybrid method run considerably faster than the original CSD- based method.

Results and Discussion

Table VI shows results of a SuperStar validation using the clean lists from Table V for different methods and data sets. The set of ligand groups is considered as a whole (solvent-accessibility range 0.0–1.0), and as a subset that contains only solvent-inaccessible, i.e., buried ligand groups (solvent-accessibility range 0.00–0.02). As is clear from Table VI, rates of prediction, f_{corr} , are similar for the full list and clean lists. A general trend is observed that better prediction rates are achieved for the clean sets containing only well-resolved complexes.

TABLE V. Optimal Sets "Clean-Lists" With Different Resolution Thresholds of None, 2.5Å, and 2.0Å

No resolution threshold (224 entries)

1a28 1a42 1a4g 1a4q 1a6w 1a9u 1aaq 1abe 1abf 1acj
1acl 1acm 1aco 1aec 1ai5 1aoe 1apt 1apu 1aqw 1ase
1atl 1azm 1b58 1b59 1b9v 1baf 1bbp 1bgo 1b17 1blh
1bma 1bmq 1byb 1byg 1cl2 1cle 1c5c 1c5x 1c83 1cbs
1cbx 1cdg 1cil 1ckp 1cle 1com 1coy 1cps 1cqp 1cvu
1cx2 1d01 1d3h 1d4p 1dbb 1dbj 1dd7 1dg5 1dhf 1did
1dmp 1dog 1drl 1dwb 1dwc 1dwd 1dy9 1eap 1ebg 1eed
1eil 1ejn 1eoc 1epb 1epo 1eta 1etr 1ets 1ett 1f0r
1f0s 1f3d 1fax 1fen 1fgi 1fkg 1fki 1fl3 1flr 1frp
1glp 1glq 1hak 1hdc 1hfc 1hiv 1hos 1hpv 1hri 1hsb
1hsl 1htf 1hvr 1hyt 1ibg 1ida 1imb 1ivb 1ivq 1jap
1kel 1lah 1lcp 1ldm 1lic 1lna 1lpm 1lst 1lyb 1lyl
1mbi 1mcq 1mdr 1mld 1mmq 1mrg 1mrk 1mts 1mup 1nco
1ngp 1nis 1okl 1okm 1pbd 1pdz 1phd 1phg 1poc 1ppc
1pph 1ppi 1pso 1ptv 1qbr 1qbu 1qcf 1qpe 1qpq 1rds
1rne 1rnt 1rob 1rt2 1slt 1snc 1srj 1tdb 1tlp 1tmn
1tng 1tnh 1tni 1tnl 1tpp 1trk 1tyl 1ukz 1ulb 1uvs
1uvt 1vgc 1wap 1xid 1xie 1ydr 1ydt 1yee 25c8 2aad
2ack 2ada 2ak3 2cht 2cmd 2cpp 2ctc 2dbl 2fox 2gbp
2h4n 2ifb 2lgs 2mcp 2pcp 2phh 2pk4 2qwk 2r07 2tmn
2tsc 2yhx 2ypi 3cla 3cpa 3erd 3ert 3gpb 3hvt 3tpi
4aah 4cox 4cts 4dfr 4est 4fbp 4lbd 4phv 5abp 5cpp
5erl 6rnt 6rsa 7tim

Resolution threshold 2.5Å (180 entries)

1a28 1a42 1a4g 1a4q 1a6w 1abe 1abf 1aco 1aec 1ai5
1aoe 1apt 1apu 1aqw 1atl 1azm 1b58 1b59 1b9v 1bbp
1bgo 1blh 1bma 1byb 1byg 1cle 1c5c 1c5x 1c83 1cbs
1cbx 1cdg 1cil 1ckp 1cle 1com 1coy 1cps 1cvu 1d01
1d3h 1d4p 1dd7 1dg5 1dhf 1dmp 1dog 1dr1 1dy9 1ebg
1eed 1eil 1ejn 1eoc 1epb 1epo 1eta 1etr 1ets 1f0r
1f0s 1f3d 1fen 1fkg 1fki 1fl3 1flr 1frp 1glp 1glq
1hdc 1hfc 1hiv 1hos 1hpv 1hsb 1hsl 1htf 1hvr 1hyt
1ida 1imb 1ivb 1jap 1kel 1lah 1lcp 1ldm 1lic 1lna
1lpm 1lst 1mbi 1mdr 1mld 1mmq 1mrg 1mrk 1mts 1mup
1nco 1ngp 1nis 1okl 1okm 1pbd 1pdz 1phd 1phg 1poc
1ppc 1pph 1ppi 1pso 1ptv 1qbr 1qbu 1qcf 1qpe 1qpq
1rds 1rne 1rnt 1rob 1slt 1snc 1srj 1tlp 1tmn 1tng
1tnh 1tni 1tnl 1tpp 1trk 1tyl 1ukz 1vgc 1wap 1xid
1xie 1ydr 1ydt 1yee 25c8 2aad 2ack 2ada 2ak3 2cht
2cmd 2cpp 2ctc 2fox 2gbp 2h4n 2ifb 2pcp 2pk4 2qwk
2tmn 2tsc 2yhx 3cla 3cpa 3erd 3ert 3gpb 3tpi 4aah
4dfr 4est 4lbd 4phv 5abp 5cpp 5erl 6rnt 6rsa 7tim

Resolution threshold 2.0Å (92 entries)

1a28 1a4q 1a6w 1abe 1abf 1aec 1aoe 1apt 1apu 1aqw
1atl 1b58 1b59 1bma 1byb 1cle 1c5c 1c5x 1c83 1cbs
1cil 1coy 1d01 1d3h 1ejn 1eta 1f3d 1fen 1flr 1glp
1glq 1hfc 1hpv 1hsb 1hsl 1hvr 1hyt 1ida 1jap 1kel
1lcp 1lic 1lna 1lst 1mld 1mmq 1mrg 1mrk 1mts 1nco
1phd 1phg 1ppc 1pph 1qbr 1qbu 1rds 1rnt 1rob 1slt
1snc 1srj 1tmn 1tng 1tnh 1tni 1tnl 1tpp 1tyl 1ukz
1vgc 1wap 1xid 1xie 2ak3 2cmd 2cpp 2ctc 2fox 2gbp
2h4n 2qwk 2tmn 2tsc 3cla 3ert 3tpi 4dfr 4est 5abp
6rnt 7tim

Error analysis: Variance in f_{corr}

Bootstrapping analysis was performed for each of the clean lists (i.e., for lists of sizes 224, 180, and 92 entries) to estimate the uncertainty in the f_{corr} values (see Table VI). This analysis yields an estimate of the variance in f_{corr} that would be observed for different lists of "good" protein–ligand entries. The analysis was performed for 1,000 lists of the same size as the original list, generated by selecting protein entries from the original list randomly, and with replacement (i.e., these random lists may contain multiple instances of the same entry). Validation results were then calculated for these artificial lists, and the standard deviation was derived from the set of prediction rates that were calculated. Bootstrapping analysis was performed for both the CSD and the PDB set separately, as the underlying distributions of validation results might be different. However, analysis of the full set of 305 entries gave similar standard errors for both CSD- and PDB-based methods (s = 1.5 percent-units for all atoms, and s = 2.6 for the solvent-inaccessible ones, CSD data; s = 1.5 and s = 2.3 for PDB data). For the clean list, a similar correspondence was observed (Table VI).

Results in Table VI show that prediction rates for ligand groups disregarding solvent accessibility are better when using CSD data. Performance rates of the methods that use raw, parameterized, or hybrid CSD data do not differ statistically. When looking at the solvent-inaccessible ligand groups only (solvent-accessibility range 0.0–0.02), performance of SuperStar using CSD or PDB data is similar, in the range of 75–80%.

Does SuperStar perform differently for different proteins?

Table VII displays a breakdown of the SuperStar results according to protein type. As can be seen, filtering out doubtful and non-diverse entries reduces the size of the sets ($\langle n \rangle$) by up to a third. It is clear that omitting doubtful entries does not change the performance rates of SuperStar very much. It is difficult, however, to draw conclusions about performance rates that are derived from small sets, and therefore success rates in Table VII should be regarded as rough indications. Standard errors for sets of size $\langle n \rangle \sim 120$ are estimated to lie around 4 units, and for $\langle n \rangle \sim 60$ around 5.5 units.

In order to ascertain whether results for subgroups of certain protein types are truly different from that of the large superset (of 305 entries, or 224 for the clean list), we calculated χ^2 values as follows: each ligand probe that is used for determination of the success rate is classified as 0 if it is mispredicted, and as 1 if it is correctly predicted (the validation is carried out for four probes). χ^2 is then calculated assuming that the full set and the subset under consideration are independent; the null hypothesis states that both sets originate from the same distribution. At the 10% confidence level, the threshold for χ^2 is 2.71 (for df = 1). The sets for which we can reject the null hypothesis at this level of significance have their χ^2 value shown in italics in Table VII. We may tentatively conclude that these sets differ from the overall set (of 305 or 224 entries), and this may cause their SuperStar success rates to differ significantly (for the worse or the better) from the overall success rates of the set from which they were taken. Of course, when this many χ^2 tests are performed, it is highly

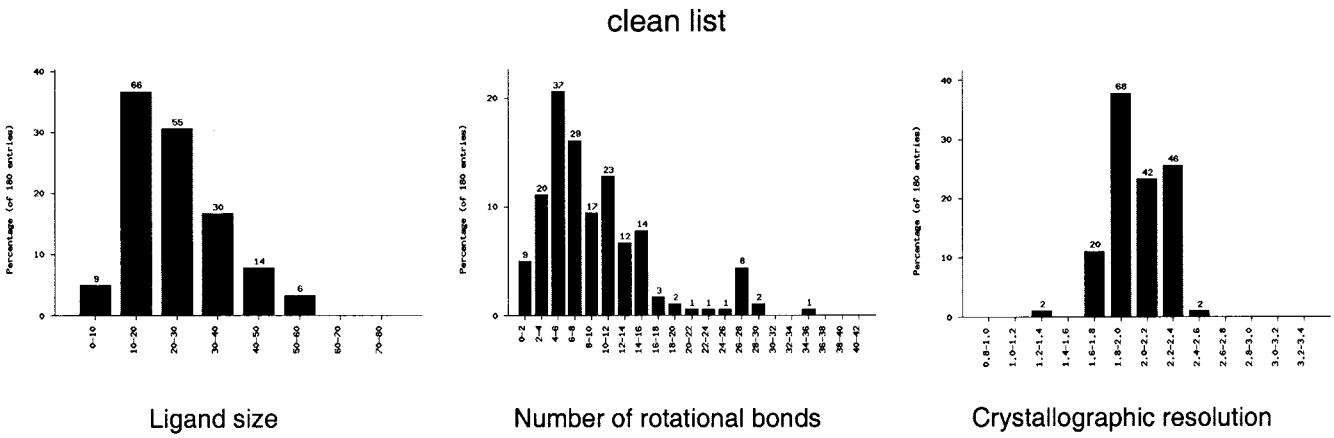

Fig. 5. Distributions of the ligand size (number of heavy ligand atoms), number of rotational bonds, and crystallographic resolution in the clean list (224 entries) specified in Table V. Histograms for the sets with crystallographic resolution less than 2.5 and 2.0 Å can be found in the supplementary material.

TABLE VI. Validation Results for SuperStar for the Clean Lists in Table V[†]

	Solvent accessibility range 0.0–1.0								Solvent accessibility range 0.00–0.02							
	Full list		Clean list, all		Clean list, R < 2.5Å		Clean list R < 2.0Å		Full list		Clean list, all		Clean list, R < 2.5Å		Clean list, R < 2.0Å	
Data type	⟨n⟩	f_{corr}	⟨n⟩	f_{corr}	⟨n⟩	f_{corr}	⟨n⟩	f_{corr}	⟨n⟩	f_{corr}	⟨n⟩	f_{corr}	⟨n⟩	f_{corr}	⟨n⟩	f_{corr}
CSD																
Raw	1,098	66	785	68	639	69	335	70	360	76	269	76	228	78	124	81
Parameterised	1,098	63	785	66	639	67	335	67	359	72	269	74	228	76	124	77
Hybrid	1,098	65	785	68	639	68	335	69	359	74	269	76	228	76	124	80
PDB																
raw	1,100	62	784	64	637	64	332	66	360	75	268	77	227	78	123	76
Bootstrapping analysis																
CSD set, s.e.		1.5		1.8		2.0		3.0		2.6		3.0		3.2		4.2
PDB set, s.e.		1.5		1.9		2.1		3.0		2.3		2.6		2.9		4.1

[†]⟨n⟩ is the number of ligand probes used, f_{corr} is the rate of prediction in [%]. Errors determined by a bootstrapping procedure are given as standard error (s.e.) in the same units.

likely that some extreme χ^2 values will occur by random chance.

In general, SuperStar's predictive performance is good when ligand groups are solvent-inaccessible, i.e., buried in the active site rather than at a surface. Aspartic proteases, serine proteases, glycosidases, isomerases, and the set of immunoglobulins (and catalytic antibodies) seem to perform less well; results for metalloproteases are above average. χ^2 values for the aspartic proteases and metalloproteases are rather high, indicating that results for these subsets are significantly different from those of the respective full or clean set. For the full set of ligand probes (solvent accessibility 0.0–1.0), the subsets of metalloproteases, lyases, and immunoglobulins yield results that differ significantly.

Conclusions

SuperStar yields similar results for the full list of 305 entries and the clean lists from which doubtful and non-diverse entries have been removed. This is not an unexpected result, as the entries that we denote as doubtful possess unacceptable features that are generally localized in a small region of the structure. The remaining part of the protein–ligand complex is not affected, and SuperStar will be well able to predict properties for that part of the structure.

On the whole, SuperStar achieves an accuracy of prediction of at least 70% for different types of proteins. Prediction rates are generally better for buried ligand atoms. For subsets of proteins, different results may be observed by chance, or arise from subsets being significantly different from the large test sets from which they originate. This can be detected using a χ^2-based test. Factors like the contents of the set in terms of type of ligand[†], or properties like

[†]An analysis of SuperStar results according to ligand size (see Supplementary Material) showed that in general, larger ligands yield worse prediction rates; results suggest that for larger ligands e.g. hydrophobic complementarity of ligand and active site is less complete than for small ligands.

TABLE VII. Breakdown of SuperStar Results[†]

	Solvent accessibility range 0.0–1.0						Solvent accessibility range 0.0–0.02					
	Full set			Clean set			Full set			Clean set		
Subset[a]	⟨n⟩	f_{corr}	χ^2	⟨n⟩	f_{corr}	χ^2	⟨n⟩	f_{corr}	χ^2	⟨n⟩	f_{corr}	χ^2
Hydrolases (132;92)	561	66	0.11	364	66	0.57	186	74	0.67	130	72	0.68
Metalloproteases (22;14)	74	80	*4.86*	45	82	*3.52*	32	94	*4.94*	18	94	*3.22*
Aspartic proteases (27;19)	217	67	0.05	145	66	0.25	77	65	*4.73*	56	63	*4.44*
Serine proteases (38;25)	103	63	0.47	56	66	0.21	34	65	2.48	20	60	2.58
Glycosidases (17;10)	100	63	0.13	65	66	0.05	19	68	0.71	17	71	0.27
Transferases & kinases (35;26)	117	61	0.52	78	69	0.08	25	80	0.69	16	88	1.09
Kinases only (12;7)	38	50	1.50	12	92	*2.86*	6	50	0.34	2	100	0.63
Lyases (12;11)	18	89	*3.74*	17	88	*2.90*	1	100	0.30	1	100	0.31
Oxidoreductases (30;25)	92	74	2.25	86	76	2.20	32	84	0.96	30	87	1.70
Immunoglobulins (34;22)[b]	108	56	*6.17*	71	58	*3.79*	21	67	1.13	15	73	0.06
Isomerases (16;11)	37	62	0.10	28	57	0.86	22	77	0.00	16	69	0.45
Lectins (6;5)	27	67	0.01	23	75	0.00	15	80	0.08	13	85	0.50
Virus proteins (5;2)	9	67	0.00	2	100	0.90	0	—	—	0	—	—

[†]Success rates (f_{corr}) are shown for subsets taken from the full list, or the clean list. ⟨n⟩ is the number of ligand probes that is used for the prediction. χ^2 values were calculated for a null hypothesis that both the full list (of 305 or 224 entries) and the protein subset come from the same distribution (see text for further details). NB: glycosidases, serine proteases, aspartic proteases, and metalloproteases are subsets of the hydrolase set. Likewise, kinases are shown as a subset of the transferases and kinases set.
[a]Number of entries of subset from full and clean set given in brackets after name.
[b]This set also contains catalytic antibodies.

TABLE VIII. GOLD Validation Results for Different Lists[†]

	$f_{RMS<0.5Å}$	$f_{RMS<1.0Å}$	$f_{RMS<1.5Å}$	$f_{RMS<2.0Å}$	$f_{RMS<2.5Å}$	$f_{RMS<3.0Å}$
All entries	14 (0.9)	44 (1.4)	59 (1.5)	68 (1.6)	75 (1.3)	80 (1.1)
Clean list	17 (1.3)	50 (1.8)	65 (1.7)	72 (1.7)	78 (1.5)	82 (1.3)
Clean list, R < 2.5Å	19 (1.5)	51 (1.9)	66 (2.0)	73 (1.9)	80 (1.7)	83 (1.5)
Clean list, R < 2.0Å	19 (2.1)	56 (2.5)	72 (2.0)	78 (1.9)	85 (1.8)	88 (1.5)

[†]All results have been averaged over 50 validation runs. All success rates *f* are given in [%]. Standard deviation given in parentheses. Results for GOLD's three-fold speed-up settings can be found in the supplementary material.

crystallographic resolution,[27] or protein side-chain flexibility may cause such differences in the observed rates of prediction for these protein subsets.

VALIDATION RESULTS FOR GOLD

GOLD

The GOLD docking program employs a genetic algorithm (GA) to optimize the position of a ligand in a binding site. During the docking process, the ligand is regarded as flexible; on the protein, OH and NH_3 groups are treated flexibly. The quality of GOLD's solutions can be tested by docking ligands into proteins for which ligand-protein crystal structures are available. The quality of the docked pose of the ligand can then be determined by comparing it to the native ligand orientation as found in the crystal structure. A good docking program should be able to reproduce such binding modes, and to do so across a broad range of different types of proteins.

Results and Discussion

GOLD validation runs were performed as follows: given a set of protein–ligand complexes, dockings were performed for all using the default settings.[28] Root of Mean Squared deviations (RMS) were then determined for the atomic positions of the first-ranked ligand pose and the native pose of the ligand as found in the crystal structure. We use the RMS value as an indicator of the quality of the solution, and report success rates, being the fraction of solutions (in the whole set) that have first-ranked poses $f_{RMS<\chi Å}$ with RMS values smaller than χÅ. Practice has shown that good docking solutions generally have an RMS of 2.0 Å or less.

GOLD results for the new test set

Table VIII collates validation results for the full validation set of 305 complexes, and for the clean lists (see Table V). Here, success rates $f_{RMS<\chi Å}$ are reported, averaged over 50 validation runs. It is clear from the results that GOLD performs better for the clean lists from which doubtful and non-diverse entries have been removed. This is expected, as these entries are the ones that will prove difficult, if not insoluble, for a docking algorithm. Performance improves further when only better-resolved structures are considered.

Error analysis: precision of GOLD validation results

Uncertainties in estimates of *f* arise from two (independent) sources: the non-deterministic nature of the GOLD genetic algorithm and the sampling error due to the fact

TABLE IX. Error Estimates for Success Rates $f_{RMS<\chi Å}$ Obtained by Bootstrapping Single Validation Runs for the Full List of 305 Entries and the Clean Lists of Table V[†]

	$f_{RMS<0.5Å}$ s	$f_{RMS<1.0Å}$ s	$f_{RMS<1.5Å}$ s	$f_{RMS<2.0Å}$ s	$f_{RMS<2.5Å}$ s	$f_{RMS<3.0Å}$ s
All entries (305)	2.0 (0.1)	2.8 (0.0)	2.8 (0.1)	2.7 (0.1)	2.5 (0.0)	2.3 (0.1)
Clean list (224)	2.3 (0.1)	3.3 (0.0)	3.2 (0.0)	3.1 (0.1)	2.9 (0.0)	2.8 (0.1)
Clean list, R < 2.5Å (180)	2.5 (0.1)	3.6 (0.0)	3.5 (0.0)	3.4 (0.1)	3.3 (0.1)	3.0 (0.1)
Clean list, R < 2.0Å (92)	3.4 (0.1)	5.0 (0.1)	5.1 (0.1)	5.0 (0.0)	4.7 (0.1)	4.3 (0.1)
Bootstrap on smaller subsets of above lists (for reasons of comparison)						
All entries (224)[a]	2.5 (0.1)	3.2 (0.1)	3.2 (0.1)	3.0 (0.0)	2.8 (0.1)	2.6 (0.1)
Clean list (180)[b]	2.8 (0.1)	3.9 (0.1)	3.6 (0.1)	3.4 (0.1)	3.1 (0.1)	2.9 (0.1)
Clean list (92)[c]	3.9 (0.1)	5.3 (0.1)	5.1 (0.1)	4.7 (0.1)	4.3 (0.2)	4.0 (0.1)

[†]Reported bootstrap standard deviations are averaged bootstrap results for 5 different validation runs. Standard deviations of these numbers given in brackets. Size of the validation set given in parentheses after set name. s: standard error. For further explanation see text. Validation runs use default settings.
[a]Bootstrapping results for a subset of 224 entries from the full list of 305.
[b]Bootstrapping results for a subset of 180 entries from the clean list of 224 (no resolution limit).
[c]Bootstrapping results for a subset of 92 entries from the clean list of 224 (no resolution limit).

that the validation set is a sample from the total universe of protein–ligand complexes.

Standard errors in Table VIII reflect the former source of error, but not the latter. We therefore used the bootstrapping protocol[19] to estimate the total standard error of f due to the two effects. Each bootstrapping run was based on 1,000 samples of size n (e.g., $n = 224$ when determining the standard error of f for the clean list) taken by sampling with replacement. For each such sample, a value of f was obtained. The distribution of f estimates for the 1,000 samples was used to obtain the standard error of f. Five replicates of this total procedure were performed; the values in Table IX are the average values obtained over the five replicates. Both sources of uncertainty are included in this standard error. The figures in parentheses in Table IX are the standard deviations of the standard errors of f; they are small, indicating that the standard errors of f have been estimated with good precision.

The standard error in the predictions f varies with the size of the set, which makes a comparison of results from, e.g., the full list and the clean list difficult. In order to assess whether the use of the clean list of structures instead of the full set has an influence on variance, we ran a bootstrap analysis using subsets of 224 taken from the full list of 305. Results (Table IX, bottom rows) show, that the actual errors observed for a validation run using 224 entries from the clean list are similar to those for a subset of 224 entries from the full list.

The clean list of 224 entries was compared with its subsets with restricted protein crystallographic resolution using the same methodology, by producing error estimates for subsets of size 92 and 180 taken from the clean list of 224 protein–ligand complexes. Errors for these sets do not differ significantly from the ones found for the R < 2.5Å and R < 2.0Å sets. This indicates that having structures with a better resolution does not improve convergence during the docking procedure per se (i.e., GOLD does not find the right pose *more confidently*). However, the number of right solutions itself *does* increase slightly with improved resolution of the protein structure (Table VIII). GOLD is able to generate roughly the same number of solutions for the less well-resolved structures as for well-resolved structures, yet ranking is better for the latter ones, leading to improved success rates for docking of well-resolved structures.

TABLE X. GOLD Success Rates for Subsets of Structures With and Without Protein-Ligand Contacts Mediated by Water Molecules[†]

	Success rates		
	$f_{RMS<1.5Å}$	$f_{RMS<2.0Å}$	$f_{RMS<2.5Å}$
No mediating waters present (55)	73 (3.9)	78 (2.7)	81 (2.4)
Mediating waters present (40)	61 (3.7)	71 (4.2)	79 (3.8)

[†]Water molecules were excluded prior to docking; docking was performed using the threefold speedup, and default settings. Values averaged over 50 validation runs (default settings); standard deviation in parentheses.

GOLD results for structures with water-mediated contacts

Waters have been removed from all complexes prior to docking. This probably deteriorates performance of the docking algorithm, as waters can mediate interactions that are essential for ligand binding. To estimate this effect, we identified a subset of structures with at least one strongly bound water molecule within 2.9 Å distance of both protein and ligand moieties. GOLD success rates for this subset (40 entries) and structures lacking mediating water molecules (55 entries) are reported in Table X. All entries are subsets of the clean list. There seems to be a trend towards lower success rates for structures that contain water-mediated contacts between ligand and protein, although the impact of leaving out water molecules is not so high as one might expect.

GOLD results for protein classes

Table XI displays GOLD results for protein superclasses. Docking was performed using the default settings and results are averaged over 50 validation runs. Focusing

TABLE XI. Success Rates for GOLD Docking (Set2, Default Settings) According to Protein Class[†]

	Success rates			χ-statistics	F-statistics	
	$f_{RMS<1.5Å}$	$f_{RMS<2.0Å}$	$f_{RMS<2.5Å}$	(χ^2)	F	P^b
Hydrolases (92)	59 (2.3)	69 (2.9)	76 (2.7)	0.32	0.47	≫0.25
Metalloproteases (14)	74 (5.8)	80 (5.5)	87 (7.0)	0.65	0.44	≫0.25
Aspartic proteases (19)	36 (6.1)	46 (7.9)	56 (8.1)	6.27	2.71	<0.025
Serine proteases (25)	55 (6.8)	65 (6.4)	71 (6.0)	0.66	0.29	≫0.25
Glycosidases (10)	72 (4.8)	76 (7.0)	80 (6.8)	0.34	0.41	≫0.25
Transferases & kinases (26)	63 (3.6)	65 (3.6)	68 (3.7)	0.63	0.21	≫0.25
Kinases only (7)	84 (4.3)	86 (0.0)	86 (0.0)	0.65	0.30	≫0.25
Lyases (11)	49 (8.2)	52 (9.0)	75 (9.4)	3.11	0.09	≫0.25
Oxidoreductases (25)	71 (4.8)	78 (4.1)	81 (3.8)	0.34	0.26	≫0.25
Immunoglobulins (22)[a]	66 (4.6)	76 (3.6)	82 (2.3)	0.15	0.19	≫0.25
Isomerases (11)	86 (4.9)	90 (2.5)	99 (2.5)	1.63	1.38	<0.10
Lectins (6)	100 (0.0)	100 (0.0)	100 (0.0)	1.94	1.49	<0.25
Virus proteins (2)	68 (32)	96 (14)	97 (12)	0.75	0.10	≫0.25

[†]Success rates are given in [%], for three different RMS thresholds. Success rates have been averaged over 50 validation runs and standard errors are given. Number of protein-ligand entries per class is shown in parentheses after name. For the χ-statistics, GOLD solutions were classified as good or wrong (df = 1) using an RMS threshold of 2.0 Å. F-statistics are given for the null-hypothesis that μ (average RMS) for subset and clean list are equal. Reported χ^2's and F's are averaged results for 15 validation runs.

[a]This set also contains catalytic antibodies.

[b]Likelihood of obtaining this result when the null hypothesis is true.

on the $f_{RMS<\chi Å}$ success rates for RMS 1.5, 2.0, and 2.5 Å, it is clear that GOLD performs similarly for most classes of proteins, with rates around 70% or better for an RMS threshold of 2.0 Å.

Performance appears to be above average for the sets of metalloproteases, kinases, isomerases, and lectins. Performance seems to be lower than expected for the aspartic protease and lyase sets. The aspartic protease set contains a high proportion of large ligands with several rotational bonds[28]; these complexes are difficult samples for docking. The lyases are difficult to dock as the set features relatively shallow binding sites and polar ligands that are partly solvent-exposed (examples are, e.g., 1aco, 2h4n); crystal waters sometimes mediate binding (e.g., in the case of 1pdz, 1okm).

However, it is extremely difficult to draw conclusions from data obtained using such small sets. When we classify GOLD solutions as "good" or "wrong" using an RMS threshold of 2.0 Å, we can use a simple χ^2 based test[30] to decide whether the observed result is really different from the success rate for the clean list. It shows that the set of aspartic proteases can be regarded as different at a confidence level of $P < 0.025$. The lyase and lectin sets have significantly different results when we allow for $P = 0.10$, and for the isomerases $P < 0.25$ applies. The results for all other sets may just differ by chance, and are not significantly different from the result for the clean list. Alternatively, we may use F statistics[19] to decide whether a subset is really different from the clean list in terms of RMS value. In this case, we calculated the F ratio using the null hypothesis that the average RMS for clean list of 224 entries and each sublist is equal. Results for F (see Table XI) indicate that only the subsets containing aspartic proteases and isomerases (and possibly the lectin set) are significantly different from the clean list, showing clearly that it is very difficult to draw meaningful conclusions from results for such small sets.

Conclusions

GOLD yields significantly better results for a validation set of complexes from which doubtful structures have been removed. Defining a good solution as one with RMS < 2.0 Å, the average success rate for GOLD was 68(1.6)% for the full set of 305 entries, and 72(1.7)% for the clean list of 224 entries. Results for a subset with better resolution gave marginally improved results (73(1.9)% for R < 2.5Å, 180 entries; 78(1.9)% for R < 2.0Å, 92 entries). Waters are removed from the binding site prior to docking, and, as expected, this has a detrimental effect on success rates. Therefore, the quoted success rates for GOLD are lower-bound estimates.

GOLD uses a genetic algorithm and, therefore, each run will produce slightly different results. The error estimates quoted in Table VIII for the average success rates tell us the uncertainty with which the average success rate has been derived. However, having a set of protein–ligand complexes that is itself a sample from a universe of possible structures means that average success rates will vary for different validation sets. As we currently have access to only one thoroughly checked set, we can only estimate this variance using a bootstrapping protocol, and a standard error of 3 units was derived (for a validation set of 224 entries).

Although it is very difficult to draw conclusions based on small amounts of data, statistical analysis reveals that success rates for aspartic proteases and, to a lesser extent, lyases and isomerases are significantly different from the set as a whole. Success rates for aspartic proteases are below expectation, probably as a result of the presence of very large ligands that contain a large number of rotatable bonds; for lyases, the low success rates can be explained by the presence

of shallow binding sites and solvent-exposed, polar ligands. Exceptionally good success rates are observed for the set of isomerases, that feature well-defined, rather hydrophobic binding sites with small to medium-sized ligands.

SUMMARY

In this article, we introduce an extensively checked test set of protein–ligand complexes for validation purposes. The full test set contains 305 entries, but is narrowed down to a smaller "clean" set by (1) taking out entries with structural or factual errors, and (2) ensuring as much diversity as possible. The resulting set contains 224 entries and covers 131 different proteins. When taking a resolution threshold into account, a set of 180 entries (115 different proteins) with resolution better than 2.5 Å remains; for a resolution of better than 2.0 Å, 92 entries can be listed (62 different proteins).

As an example of application of this new set, validation results are shown for SuperStar and GOLD. SuperStar is a program for predicting preferential sites of interaction in protein binding sites. Validation results indicate that SuperStar using CSD data has a success rate of 68(2)% for predicting 4 types of ligand groups, and 76(3)% when regarding buried ligand atoms only. CSD data yield slightly better prediction rates than PDB data: correct prediction was observed for 62(2)% of the ligand groups for the latter. Rates of prediction for buried ligand atoms only were similar when using either PDB or CSD data.

The docking program GOLD is shown to have a success rate of 72(1.7)% for yielding a first-ranked docked ligand pose with an RMS of less than 2.0 Å with respect to the ligand reference structure. Prior to docking, all waters have been removed from the active site. Doing so may have a detrimental effect on docking performance and quoted success rates are therefore minimum bounds. It is indeed observed that docking results for complexes that are observed to contain water-mediated contacts between ligand and protein are worse than those for complexes that do not contain such contacts.

Our results show clearly that higher docking success rates are observed when using a validation set of protein–ligand complexes from which suspect entries have been removed. SuperStar results, however, are not improved by removing such entries. This is expected, as the "errors" in the entries that we removed are usually localized in a small region of the structure. SuperStar is still able to predict well for the remaining part of the protein–ligand complex. GOLD, like all docking programs, relies on the integrity of the whole structure, and every inaccuracy in either ligand or protein will influence its performance.

ACKNOWLEDGMENTS

The authors thank Dr. Andreas Bergner for assistance with the calculation of crystallographically related protein chains in Relibase+, Dr. Ian Tickle at Astex Technology Ltd. for his help with the calculation of the OMIT maps, and Dr. Richard Taylor at Astex Technology for assistance with the validation runs.

REFERENCES

1. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol 1997;267:727–748.
2. Kramer B, Rarey M, Lengauer T. Evaluation of the FLexX incremental construction algorithm for protein–ligand docking. Proteins 1999;37:228–241.
3. Pang Y-P, Perola E, Xu K, Prendergast FP. EUDOC: a computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases. J Comp Chem 2001;22:1750–1771.
4. Baxter CA, Murray CW, Waszkowycz B, Jin Li, Sykes RA, Bone RGA, Perkins TDJ, Wylie W, A New approach to molecular docking and its application to virtual screening of chemical databases. J Chem Inf Comput Sci 2000;40:254–262.
5. Ewing TJ, Makino S, Skillman AG, Kuntz ID. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. J Comput-Aided Mol Design 2001;15:411–428.
6. Taylor JS, Burnett RM. DARWIN: A program for docking flexible molecules. Proteins 2000;41:173–191.
7. Blom NS, Sygusch J. High resolution fast quantitative docking using Fourier domain correlation techniques. Proteins 1997:27:493–506.
8. Liu M, Wang S. MCDOCK: A Monte-Carlo simulation approach to the molecular docking problem, J Comput-Aided Mol Design 1999;13:435–451.
9. Wang J, Kollman PA, Kuntz ID. Flexible ligand docking: a multistep strategy approach. Proteins 1999;36:1–19.
10. Verdonk ML, Cole JC, Taylor R. SuperStar: a knowledge-based approach for identifying interaction sites in proteins. J Mol Biol 1999;289:1093–1108; Verdonk ML, Cole JC, Watson P, Gillet V, Willet P. SuperStar: improved knowledge-based interaction fields for protein binding sites. J Mol Biol 2001;307:841–859; Boer DR, Kroon J, Cole JC, Smith B, Verdonk ML. SuperStar: Comparison of CSD and PDB-based interaction fields as a basis for the prediction of protein–ligand interactions. J Mol Biol 2001;312:275–287.
11. Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein- ligand interactions. J Mol Biol 2000;295: 337–356.
12. Muegge I, Martin YC. A general and fast scoring function for protein–ligand interactions: a simplified potential approach. J Med Chem 1999;42:791–804.
13. Shindyalov IN, Bourne PE. An alternative view of protein fold space, Proteins 2000;38:247–260.
14. Bergner A, Guenther J, Hendlich M, Klebe G, Verdonk ML. Use of Relibase for retrieving complex 3D interaction patterns including crystallographic packing effects, Nucl Acid Res 2002; 61:99–110.
15. (a) Sybyl molecular modelling software, Tripos Inc. 1699 South Hanley Rd, Suite 303, St. Louis, MO 63144; (b) Hooft RWW, Sander C, Vriend G. Positioning hydrogen atoms by optimizing hydrogen-bond networks in protein structures. Proteins 1996;26: 363–376.
16. Collaborative Computational Project, Number 4. The CCP4 Suite: Programs for Protein Crystallography. Acta Cryst 1994;D50:760–763.
17. Bruno IJ, Cole JC, Lommerse JPM, Rowland RS, Taylor R, Verdonk ML. IsoStar: a library of information about non-bonded interactions. J Comput-Aided Mol Design 1997;11:525–537.
18. Rowland RS, Taylor R. Intermolecular nonbonded contact distances in organic crystal structures: comparison with distances expected from van der Waals radii. J Phys Chem 100;1996:7384–7391.
19. Wonnacott TH, Wonnacott RJ. Introductory statistics, 5th ed., New York: John Wiley & Sons; 1990.
20. Allen FH, Kennard O. 3D Search and Research Using the Cambridge Structural Database. Chem Design Autom News 1993;8:1, 31–37.
21. Martin JL, Begun J, Schindeler A, Wickramasinghe WA, Alewood D, Alewood PF, Bergman DA, Brinkworth RI, Abbenante G, March DR, Reid RC, Fairlie DP. Molecular recognition of macrocyclic peptidomimetic inhibitors by HIV-1 protease. Biochemistry 1999;38:7978–7988.
22. http://www.ccdc.cam.ac.uk
23. 1ake, 1ctr, 1dbm, 1ela, 1elb, 1eld, 1ele, 1ghb, 1hdy, 1hef, 1icn,

1ivc, 1ivd, 1ive, 1lmo, 1pha, 1ppl, 1stp, 2cgr, 2plv, 3gch, 4fab, 5p2p, 6cpa, 7cpa, 8gch.
24. Viz., 1abe, 1abf, 1aoe, 1apu, 1cbx, 1dbb, 1dbj, 1dmp, 1dog, 1dwb, 1ebg, 1epo, 1etr, 1ets, 1ett, 1fax, 1hpv, 1hsl, 1htf, 1hvr, 1jap, 1mbi, 1okl, 1okm, 1phg, 1ppc, 1pph, 1qbr, 1qbu, 1tlp, 1tmn, 1tng, 1tnh, 1ulb, 1uvs, 1uvt, 2cpp, 2ctc, 2gbp, 2h4n, 2ifb, 2phh, 2tmn, 2tsc, 2ypi, 3tpi, 4dfr, 5abp, 5cpp.
25. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res. 2000;28:235–242.
26. Nissink JWM, Verdonk ML, Klebe G. Simple knowledge-based descriptors to predict protein–ligand interactions. Methodology and validation. J Comput-Aided Mol Design 2000;14:787–803.
27. See supplementary material for an overview of distributions of number of ligand atoms, number of rotational bonds and crystallographic resolutions for these sets.
28. These settings are distributed with GOLD version 1.2, October 2001.
29. See supplementary material for details.
30. Siegel S. Nonparametric statistics for the behavioral sciences. London: McGraw-Hill; 1956.
31. Bondi A. van der Waals volumes and radii. J Phys Chem 1968;68: 441–451.