GARD: A Generally Applicable Replacement for RMSD

J. Christian Baber,*[,†,||] David C. Thompson,[†,#,||] Jason B. Cross,[‡,⊥] and Christine Humblet[§]

Chemical Sciences, Wyeth Research, 200 Cambridge Park Drive, Cambridge, Massachusetts 02140 and 865 Ridge Road, Princeton, New Jersey 08543, and Chemical Sciences, Wyeth Pharmaceuticals and Research Headquarters, 500 Arcola Road, Collegeville, Pennsylvania 19426

Received March 18, 2009

The root-mean-squared deviation (rmsd) is a widely used measure of distance between two aligned objects − often chemical structures. However, rmsd has a number of known limitations including difficulty of interpretation, no limit on weighting for any portion of the alignment, and a lack of normalization. In this work, a **G**enerally **A**pplicable **R**eplacement for rms**D** (GARD) is proposed. In this implementation atomic contributions are weighted by their relative importance to binding, as determined statistically by Andrews et al.,[1] and as such this method is 'chemically aware'. This novel measure is normalized and does not have many of the failings of traditional rmsd. It is, thus, perfectly suited for a wide variety of uses, including the assessment of the quality of poses produced from molecular docking programs and the comparison of conformers. Rmsd and GARD are compared in their ability to assess docking software and multiple examples of the use of GARD to rescue essentially correct poses with a high rmsd are presented.

INTRODUCTION

The root-mean-squared deviation (rmsd) is a metric used to assess the differences between a set of values predicted by a model and actual observed values for the system under investigation. It is an ubiquitous measure employed in a wide range of quantitative fields, such as economics, meteorology, and technology. It is also heavily used within the computational life sciences, from assessing the quality of protein models[2,3] through to measuring the quality of docking poses.[4] It is this latter use of rmsd, as applied to small molecule docking pose assessment, which will form the basis of the present work.

The cycle of experimental structure determination, computational prediction, chemical synthesis, and subsequent biological testing is used by virtually all pharmaceutical and biotechnology companies when optimizing novel, "lead-like" chemical matter into something "drug-like".[5,6] Protein structures are usually derived from X-ray crystallographic, or possibly nuclear magnetic resonance (NMR), experimental data, and a model is built representing the interactions between a protein and a bound ligand. It is worth noting that such structures are always models built on experimental data and, thus, have significant limitations in cases where the resolution is poor, the ligand occupancy is low, or multiple binding modes are present. However, this type of structural data is generally the best available and, together with a simulation of proposed compounds docking to the experimentally derived protein model, often forms the basis of computational lead optimization exercises.

The first step in any computational prediction is a validation of the proposed docking software to ensure that it is competent to reproduce the experimentally observed binding mode. The metric of rmsd, based on the displacement of atoms in a docked pose compared to the reference structure, is typically employed and is calculated using the following expression:

$$\mathrm{rmsd} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2} \qquad (1)$$

where d_i is the Euclidean distance between N pairs of equivalent atoms i. In structure-based computational chemistry, when evaluating docked poses, a rmsd of no more than 2 Å for a pose is usually considered acceptable.

There are a large number of docking programs commonly in use from a variety of academic and commercial sources, and there are a range of papers comparing docking programs.[7−12] Such comparisons are usually focused on: (i) the ability of a program to successfully regenerate the known crystallographic binding modes of a range of protein complexes,[13−16] and (ii) the ability of a program to enrich virtual screening hit lists.[17−27] It is widely acknowledged that comparing docking programs in a meaningful way is a nontrivial exercise,[4] and, in correctly assessing (i), rmsd is used extensively as a measure of success.

Typically, when assessing a docking program one would look first at the top scoring pose and associated rmsd. If this pose has a rmsd less than 2 Å relative to the X-ray conformation (a somewhat arbitrary cutoff often used in docking assessment), then it is generally considered to constitute a docking success.[10] Docking failures come in several forms with different causes.[28] Failures in scoring occur when the energy function is unable to score a low rmsd pose better than a high rmsd pose. In these cases, sampling produced a good pose, but errors in the scoring function mean

* Corresponding author. E-mail: baberj@wyeth.com.
† Chemical Sciences, Wyeth Research, Cambridge, Massachusetts.
‡ Chemical Sciences, Wyeth Pharmaceuticals and Research Headquarters.
§ Wyeth Research, Chemical Sciences, Princeton, New Jersey.
# Current address: Boehringer Ingelheim Pharmaceuticals, Inc., 900 Ridgebury Road, Ridgefield, CT 06877.
⊥ Current address: Cubist Pharmaceuticals, Inc., 65 Hayden Avenue, Lexington, MA 02421.
|| The first two authors contributed equally to this work.

10.1021/ci9001074 CCC: $40.75
Published on Web 07/20/2009

that it was not the best scoring one. Alternatively, failures may occur when the sampling is insufficient to generate any low rmsd poses for the ligand, and only poses with a high rmsd (and often poor score) are returned. This can result from either incomplete sampling of poses or the use of a scoring function that is unable to recognize and score the native ligand conformation appropriately. Finally, complete docking failures are those cases where the docking program is unable to return any ligand poses at all due to a deficiency either in sampling or in generating poses which score worse than some internal cutoff.

This type of docking evaluation scenario is problematic for a number of reasons. First, the statistics describing the method, and ultimately used for comparing across methods, are penalized by the fact that anything from 2 Å to an infinitely poor rmsd is generally considered incorrect.[10,29,30] Given such a coarse cutoff it is very likely that useful information is being discarded. Second, such comparative statistics can be skewed by even a very small number of very bad poses to the interpreted detriment of the whole method. Third, it is possible to get solutions that have a good rmsd but form completely different interactions with the protein than those observed experimentally.[31,32] This can be particularly problematic when comparing small or nearly symmetrical ligands, wherein even random placement in a small active site might result in a much better than expected proportion of 'good' rmsd poses. Finally, the inverse problem can occur when a poor rmsd hides an essentially correct binding mode (a binding mode which preserves the core interactions with the protein). This situation might occur through a highly flexible portion of a ligand, which is grossly misplaced and skews the rmsd unfavorably.

There are several examples in the literature of methods that try to work around some of the issues associated with rmsd. The relative displacement error (RDE) method overcomes the problem of skew in aggregate statistics by reducing the importance of large deviations; statistics compiled using the RDE measure are less dominated by very bad docking poses.[33] Although, such a method would still fail to recover poses that are essentially correct and that form all of the appropriate interactions. One method that does account for this failing is the interaction-based accuracy classification (IBAC) method, which has been compared with rmsd on several test docking poses.[31] However, this method is highly subjective and not easily automated. Further approaches such as the real space R-factor (RSR) method have sought to include experimental information into a replacement for rmsd in assessing protein−ligand docking accuracy.[34] This measure is still 'unbounded' for poor poses: a RSR measure of 1 corresponds to a docking pose that is as good as the crystallographer's refined model of the ligand, while a RSR of greater than 1 corresponds to a pose that fits less well than the model built by the crystallographer. It is interesting to note that none of these approaches overcomes all of the problems with rmsd in one method.

In this present work, a **G**enerally **A**pplicable **R**eplacement for rms**D** (GARD) is introduced. It addresses many of the fundamental flaws of rmsd, and it is easily extensible and automated. The method is described in detail in the Methodology section. Several examples are then presented which highlight its utility as compared to rmsd. GARD is first compared to rmsd in the assessment of the quality of poses generated by a wide variety of different docking programs. This section also includes a number of specific examples, which represent a wide range of protein families and ligand chemotypes, and highlight instances in which GARD overcomes many of the problems associated with rmsd. This is followed by a more general discussion describing possible future developments and alternative uses for the GARD methodology. Concluding remarks are then offered.

Table 1. Functional Group Contributions to Binding Energy as Determined by the Analysis of Andrews and Co-Workers[1]

no.	functional group	energetic contribution (kcal/mol)
1	DOF[a]	−0.7
2	$C(sp^2)$	0.7
3	$C(sp^3)$	0.8
4	N^+	11.5
5	N	1.2
6	CO_2^-	8.2
7	OPO_3^{2-}	10.0
8	OH	2.5
9	C═O	3.4
10	O, S	1.1
11	halogen	1.3

[a] Degrees of internal conformational freedom.

METHODOLOGY

Before introducing the GARD method, two of the main problems associated with the use of rmsd when assessing docking poses are reiterated and summarized below:

1. The skewing of aggregate statistics by a small number of very poor docking poses. This occurs through averaging rmsd values, which themselves are formally bound on the interval $[0, \infty)$.

2. Essentially correct information about ligand binding being discarded through high rmsd values due to small proportions of the ligand being very poorly aligned. This might manifest itself in the correct placement of core functional elements being obscured in the rmsd measure by a flexible, unimportant, group.

To overcome the first issue associated with rmsd, GARD is renormalized to the unit interval [0,1], such that failures are given a low value (close to zero), and good poses are assigned values close to one. The shifting of bad poses to close to zero ensures that they can be correctly incorporated into aggregate statistics for fair comparisons across methods.

To address the second issue, the maximum effect that any individual atom can have on the GARD measure is limited through the use of a weighting scheme, which mirrors the importance of the correct placement of atoms and functional groups essential for binding. The weighting scheme used here is that of Andrews and co-workers,[1] wherein a regression analysis was performed on 200 drugs and enzyme inhibitors, resulting in a linear relationship estimator between commonly occurring functional group moieties and the binding free energy of the protein−ligand association event. It was observed that, in the data set used, charged groups contributed to binding more strongly than polar groups, which, in turn, contributed more strongly than nonpolar groups. The binding energy contributions of the 10 functional groups determined by Andrews are presented in Table 1. The use of the Andrews weighting in the GARD methodology incorporates an element of 'chemical awareness'. The use

of a statistically derived weighting is also an attempt to remove subjectivity in assessing protein−ligand interactions; the correct placement of key groups will be emphasized, while the incorrect placement of unimportant groups will not. The use of such a statistical set carries its own problems, but the issues associated with this are well-known and manageable. Care has been exercised in the implementation of the GARD algorithm and its treatment of protonated nitrogen atoms. A careful examination of the Andrews and co-workers data set for deriving their initial free energy decomposition suggested that only four-coordinate positively charged nitrogen atoms should be given the weighting +11.5 kcal/mol. From a GARD perspective, this means that a protonated nitrogen atom in an unsaturated ring system is counted as a neutral nitrogen atom. This did not occur in any of the docking poses examined in this work. This would also mean that nitrogen atoms in the guanidinium functionality would be considered neutral (assigned a weight of +1.2 kcal/mol) as would the nitrogen in the nitro functionality. Specifically, for the nitro functionality, the nitrogen atom would have the weight of a neutral nitrogen species, while each oxygen atom would be described using the hydroxyl oxygen weight of +2.5 kcal/mol. Neither of these functionalities were sufficiently represented in the Andrews set to receive a direct weighting themselves and this represents a limit upon how to account for these functionalities. To ensure that all atoms are considered, any atom type not explicitly assigned a value in the Andrews scheme is assigned a weighting of 0.5 in the GARD algorithm.

Although the Andrews analysis is somewhat dated it is widely used[35−38] and, to the best of the authors' knowledge, has not been superseded. The GARD design allows for any weighting scheme to be used, and an updated analysis of the type performed by Andrews is something being actively considered.

With these considerations in mind, the GARD score is calculated as follows:

$$\mathrm{GARD} = \frac{\sum_{i=1}^{N} \delta_i w_i}{\sum_{i=1}^{N} w_i} \tag{2}$$

where δ_i is a measure of the alignment of atom i in the pair of structures being compared, and w_i is the weight of atom i. The parameter δ_i represents an alignment score and is related to the Euclidian distance d between pairs of atoms i, such that:

$$\delta_i = \begin{cases} 1 & d_i \le d_{\min} \\ \left(\dfrac{d_i - d_{\min}}{d_{\max} - d_{\min}}\right) & d_{\min} \le d_i \le d_{\max} \\ 0 & d_i \ge d_{\max} \end{cases} \tag{3}$$

The formulation of δ_i represents a linear interpolation between what is considered acceptable ($d_{\min}$ Å) and unacceptable ($d_{\max}$ Å) deviation from a reference position and is illustrated in Figure 1. While these cutoffs are somewhat arbitrary, we feel they represent a balanced assessment of what is considered a 'generally good alignment' versus a 'generally bad alignment'. In all of the calculations reported below, $d_{\min} = 1.0$ Å and $d_{\max} = 2.5$ Å, so that atoms within 1.0 Å of their correct position, are assigned an alignment score of 1, and atoms further than 2.5 Å away from their ideal positions are assigned a score (value of δ_i) of 0, regardless of exactly how large the deviation. Varying the values of $d_{\min}$ and $d_{\max}$ may well be useful for other applications of GARD, or in cases where there is some doubt over the precise position of the atoms being considered (such as with low resolution crystal structures), where $d_{\max}$, and possibly $d_{\min}$, could be increased, although this was not examined in the present work.

The w_i is the Andrews weightings described above (Table 1). However, this weighting scheme could be replaced with anything from $w_i = 1, \forall i$, or $w_i = 1$ ($i \in$ heavy atoms) to much more complex schemes where the weighting associated with an atom is based on the error present in the position of the known structure.

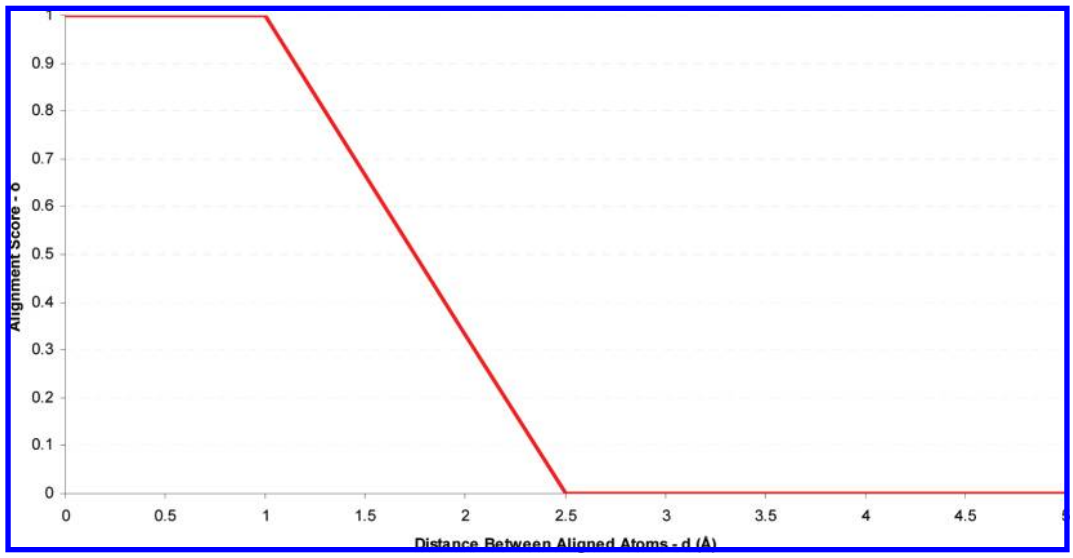

Figure 1. Form of the alignment scoring function used to assess docking poses.

This functional form, while cleanly including chemical information, also has the benefit of being normalized to the unit interval. To emphasize again, this is a nontrivial aspect of this metric and ensures that bad poses are 'mapped' to zero. It also ensures that aggregate statistics built around GARD (e.g., average GARD values) do not suffer from the positive skew problems associated with any outlying large deviations. Given that the GARD score is fundamentally a weighted measure of the proportion of the structure that aligns, a value of 0.5 has been used in this work to separate well aligned structures from poorly aligned ones. However, as is the case with rmsd, where poses with an rmsd between 2 and 3 Å are generally considered "nearly correct", this should not be seen as a hard limit. In some cases poses with GARD scores as low as 0.25 that capture at least some of the key features of binding have been observed, and it is recommended that this lower cutoff be used when attempting to rescue such poses.

The algorithm detailed above has been implemented using the OEChem Python library[39] from OpenEye Scientific Software. Symmetry in the molecule is accounted for through use of the GetAutomorphism function, and the minimum rmsd automorphism is determined at the outset of GARD to ensure correct symmetry adjusted atom−atom correspondences are used in calculating atom−atom distances. Furthermore, as this method uses atomic information, and Andrews' initial decomposition was in terms of functional groups, atomic contributions are determined via an analysis of the bonding and connecting patterns of each atom *i*, such that each atom is assumed to contribute equally to the total energetic contribution for each group. For example, each atom in the OPO_3^{2-} functionality is assumed to carry an equal 2.0 kcal/mol units of its total 10.0 kcal/mol value.

RESULTS

The GARD algorithm is designed to be versatile and extensible, with drop-in functions for measuring the quality of the alignment for each atom as well as assigning a weighting to each atom. In this paper, the algorithm has been applied to the problem of docking pose assessment (across a wide range of docking programs) and includes both an overall assessment of programs and poses as well as specific examples. Through the use of cutoffs and a linear interpolation scheme for the alignment function (eq 3), the aim is to mimic the thought process of an experienced modeler who would generally consider alignments as either *correct* or *incorrect* with only a small gray region between the two. The use of the Andrews scores for weighting is designed to ensure that those parts of the compound most important for binding have the greatest effect on the GARD score. However, the high Andrews weighting for charged groups may be an issue when dealing with solubilizing groups, which are generally not important for binding and may not be well-defined in crystal structures. In the complexes that we have examined this has not been found to be a problem but could potentially have an effect on overall performance assessment. This is discussed in more detail in one of the examples below.

Evaluation of Docking Programs. The docking poses used in this study were originally generated for a recent evaluation of docking programs.[12] The data set of X-ray structures used for docking consisted of 68 protein−ligand complexes taken from the Astex high-resolution data set with additional kinase and nuclear receptor targets included. Cognate ligands from 68 X-ray complexes were docked into their respective protein structures using several docking programs, including DOCK[40−42] (version 6.1), FlexX[43] (version 2.0.3), GLIDE[44,45] (version 4.5), ICM[46,47] (version 3.5−1), PhDOCK,[48,49] and Surflex[50−52] (version 2.1). Default parameters were used for the docking programs with the exception of Surflex, which was used in both the default mode and with the ring flexibility turned on (denoted Surflex Ringflex). For the GLIDE program, both the standard and extra precision algorithms (denoted GlideSP and GlideXP, respectively) were used. Initial ligand coordinates used as input for the docking were generated using CORINA,[53] with multiple stereoisomers and flexible ring conformations in the output rather than simply a single conformation. This was done to mimic "real world" conditions where the bound conformation and absolute stereochemistry of a compound may not be known. The 10 top scoring poses (or fewer, depending on the specific output for a particular X-ray complex/docking program combination) were retained for analysis. For additional details, please refer to Cross and coworkers.[12]

These poses were then evaluated using both the GARD and rmsd measures. The first thing that was noted was that in a number of cases programs failed to generate any poses for some targets, resulting in complete docking failures. In these cases, no rmsd value could be calculated, and so the target was simply left out of the calculation of average rmsd, resulting in an artificially low average error. This is not a problem for the GARD measure where, by definition, a failure to generate any pose is automatically assigned a score of 0, indicating that no atoms were correctly aligned. In the case of the rmsd, an attempt was made to correct for this problem by assigning a value of 21 Å (slightly higher than the maximum calculated rmsd of 20.4 Å for the top scoring pose generated by any method across all targets) to any complete docking failures. This allows values to be assigned to each target and is, thus, a more accurate measure of overall performance but does significantly penalize methods for not identifying any poses, something that is often better than identifying poor poses. As can be seen in Table 2 (which also shows the number of such complete failures for each method), the inclusion of docking failures changes the rank order of the docking methods compared to uncorrected rmsd. However, when the mean rmsd is considered the aggregate values and new rankings, are clearly dependent on the actual value of rmsd assigned to complete docking failures, which is undesirable. In order to correct this and to prevent the unbounded nature of rmsd from having too large an effect on the overall average rmsd, it was decided that medians rather than means should be used when comparing methods and calculating rankings. This results in a corrected rmsd ranking that is not dependent on the actual value assigned to failures, as long as such values are higher than for any other pose and the method successfully generates poses for at least 50% of the targets being considered. This is a clear illustration of one of the key failings of rmsd but is something that is trivially handled through GARD or, indeed, any measure bounded on the unit interval.

Table 2. Complete Comparison of RMSD and GARD Scores for the Top Scoring Pose Across 68 Different Targets for a Variety of Docking Methods[a]

method	DOCK	FlexX	GlideSP	GlideXP	ICM	PhDOCK	Surflex	Surflex Ringflex
number of complete failures	1	4	3	2	0	1	0	0
mean rmsd	3.28	4.27	2.08	1.97	1.86	4.57	2.93	2.26
mean corrected rmsd	3.54	5.26	2.92	2.53	1.86	4.81	2.93	2.26
mean GARD	0.57	0.46	0.69	0.70	0.73	0.40	0.56	0.68
rank mean rmsd	6	**7**	**3**	**2**	1	**8**	5	**4**
rank mean corrected rmsd	6	*8*	*4*	*3*	1	*7*	5	*2*
rank mean GARD	***5***	**7**	**3**	**2**	1	**8**	***6***	**4**
median rmsd	1.30	2.59	1.16	1.11	0.98	3.66	1.66	1.19
median corrected rmsd	1.32	3.09	1.22	1.15	0.98	3.70	1.66	1.19
median GARD	0.75	0.40	0.89	0.83	0.88	0.22	0.72	0.85
rank median rmsd	5	7	**3**	2	1	8	6	**4**
rank median corrected rmsd	5	7	*4*	2	1	8	6	*3*
rank median GARD	5	7	***1***	***4***	***2***	8	6	*3*

[a] Docking poses in this analysis are from Cross and co-workers.[12] Ranks that are different from the corresponding rank calculated using RMSD (as used in the original assessment) are shown in italic. Ranks that are different from the corresponding rank calculated using corrected RMSD (using a value of 21 Å for complete docking failures) are shown in **bold**.

As shown in Table 2, both assigning a high rmsd to docking failures and using the GARD methodology results in a lower assessment for programs that fail to generate any poses for some of the targets. Tables 2−4 also illustrate how scoring using the GARD algorithm affects the overall ranking of docking methods. Table 2 is the most important of these since it shows the performance of the pose scored most highly by the docking program (Top Scoring). In this case, it can be seen that correcting the rmsd for cases when no poses were identified does increase the median rmsd in most cases, significantly so for FlexX, which moves a median rmsd from 2.59 to 3.09 Å. However, the rank ordering of the methods is not changed by this adjustment. This is not the case for the GARD scores. If the methods are ranked by median GARD scores, then the overall order is substantially changed with only three of the eight methods retaining their original ranks. In most cases, the change is only one position, although this is not the case for GlideSP, which is ranked fourth by rmsd but comes top of the rankings using GARD. Closer examination shows that, while the rank orderings change substantially, the median GARD scores for the top four methods (GlideSP, ICM, Surflex Ringflex, and GlideXP) are very close (0.89, 0.88, 0.85, and 0.83). This is particularly notable for the top two methods (GlideSP and ICM), which have effectively the same GARD scores (0.89 and 0.88) but median corrected rmsds of 1.22 and 0.98 Å, respectively, in fourth and first place in terms of corrected rmsd.

When the pose with the lowest rmsd (Table 3) or highest GARD score (Table 4) rather than the pose with the best docking score is considered, there are far fewer differences in rank order. This is primarily due to the fact that poses with a very low rmsd will also have a high GARD score, and poses with a very high GARD score are also likely, but not guaranteed, to have a low rmsd. This relationship can be seen in more detail in Figure 2, which shows the correlation between the GARD measure and rmsd for up to the 10 highest scoring poses across 68 targets and 8 docking methods (a total of 4 725 points). In general, there is a very high degree of correlation ($R^2 = 0.59$), although it is not simply a linear relationship. It appears that a given GARD score has a minimum associated rmsd, but there does not seem to be any clearly defined maximum rmsd meaning that the range of rmsd values associated with any given GARD score increases as the GARD score falls. To some extent this is by design: the GARD measure cannot be skewed by a large error in a small part of the structure as is possible with rmsd, but such a large error is only possible as long as some portion of the structure is not correctly aligned. Another design feature is noticeable for GARD scores at, or near, zero, which displays a very large range of rmsds. A GARD score of zero indicates that the pose is essentially *incorrect* with every atom at least 2.5 Å away from its correct position. However, poses with a GARD score of near zero have a corresponding rmsd of between 3.0 and 24.9 Å. This large range of rmsd (which is theoretically unbounded) is the cause of the skewing of aggregate statistics, which does not arise with GARD.

Figure 3 shows the subset of data with GARD scores of at least 0.75, a total of 1 469 points. Here it is possible to see that some poses with high GARD scores have a very poor rmsd, up to 4.85 Å. These are poses for which a large proportion of the atoms in the structure are correctly aligned, particularly those likely to interact directly with the protein, yet have some atoms that are significantly displaced. In total there are 58 poses with a GARD score over 0.75 and a rmsd of at least 2 Å (3.95% of the poses). These poses are likely to be considered 'poor' using rmsd, yet the high GARD scores (well above the standard 0.5 pass mark) indicate that the majority of atoms in each of the structures have at least a reasonable alignment, particularly those atoms likely to be involved in binding. Two examples of outliers from Figure 3 are shown in Figures 4 and 5 and are described in more detail below.

Examination of Docking Poses. A number of docking poses have been examined in more detail and have been compared to published crystallographic structures. It is important to note that the goal of the GARD algorithm is to determine if two structures are well aligned and not, directly at least, to determine if a given pose is correct. The use of GARD to assess docking pose quality is based upon the assumption that the ligand present in the crystal structure is itself in the correct pose, which may not always be the case for low-resolution structures such as 2ITY (discussed below).[54] This is exactly the same assumption used with rmsd and is likely to be valid in the vast majority of cases.

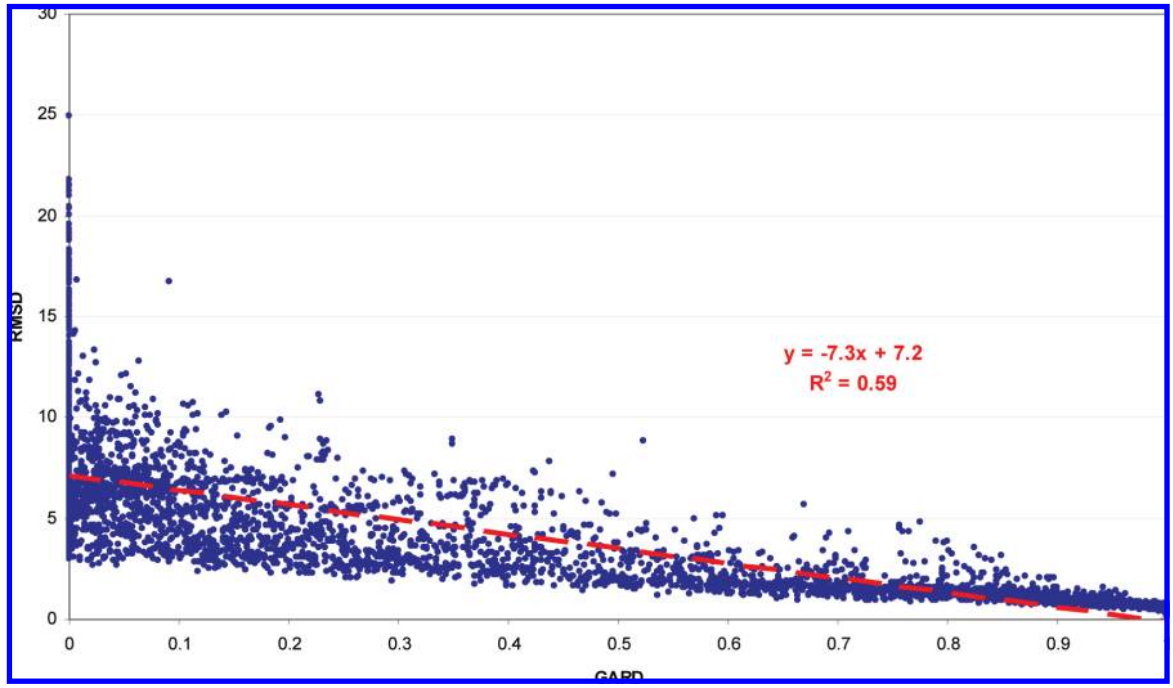

Figure 2. Correlation between GARD scores and rmsd across the top 10 poses of compounds from 68 different targets and 8 docking methods (4 725 points).

Table 3. Comparison of Median RMSD and GARD Scores for the Lowest RMSD Pose Across 68 Different Targets for a Variety of Docking Programs[a]

Method	DOCK	FlexX	GlideSP	GlideXP	ICM	PhDOCK	Surflex	Surflex Ringflex
Median rmsd	1.09	1.81	0.68	0.76	0.71	2.48	1.38	1.08
Median Corrected rmsd	1.12	2.13	0.72	0.79	0.71	2.56	1.38	1.08
Median GARD	0.85	0.57	0.96	0.94	0.97	0.38	0.75	0.87
Rank Median rmsd	5	7	***1***	3	***2***	8	6	4
Rank Median Corrected rmsd	5	7	**2**	3	**1**	8	6	4
Rank Median GARD	5	7	**2**	3	**1**	8	6	4

[a] Ranks that are different from the corresponding rank calculated using RMSD (as used in the original assessment) are shown in **red**. Ranks that are different from the corresponding rank calculated using corrected RMSD (using a value of 21 Å for complete docking failures) are shown in **bold** and ***italic***.

Table 4. Comparison of Median RMSD and GARD Scores for the Pose with the Highest Gard Score Across 68 Different Targets for a Variety of Docking Programs[a]

Method	DOCK	FlexX	GlideSP	GlideXP	ICM	PhDOCK	Surflex	Surflex Ringflex
Median rmsd	1.14	1.81	0.72	0.78	0.71	2.64	1.39	1.08
Median Corrected rmsd	1.17	2.13	0.75	0.79	0.71	2.67	1.39	1.08
Median GARD	0.85	0.57	0.97	0.94	0.97	0.43	0.76	0.88
Rank Median rmsd	5	7	2	3	1	8	6	4
Rank Median Corrected rmsd	5	7	2	3	1	8	6	4
Rank Median GARD	5	7	**1**	3	**2**	8	6	4

[a] Ranks that are different from the corresponding rank calculated using RMSD (as used in the original assessment) are shown in **red**. Ranks that are different from the corresponding rank calculated using corrected RMSD (using a value of 21 Å for complete docking failures) are shown in **bold** and ***italic***.

1A4Q[55] *– Neuraminidase with Dihydropyran-phenethyl-proply-carboxamide Inhibitor.* Figure 4 shows a neuraminidase inhibitor pose generated using Surflex Ringflex that is a clear outlier in Figure 3. In this case, the docked pose clearly picks up the majority of the hydrogen-bonding interactions in the binding site and has a GARD score of 0.78. However, the two hydrophobic side chains of the ligand are substantially misplaced resulting in a large rmsd

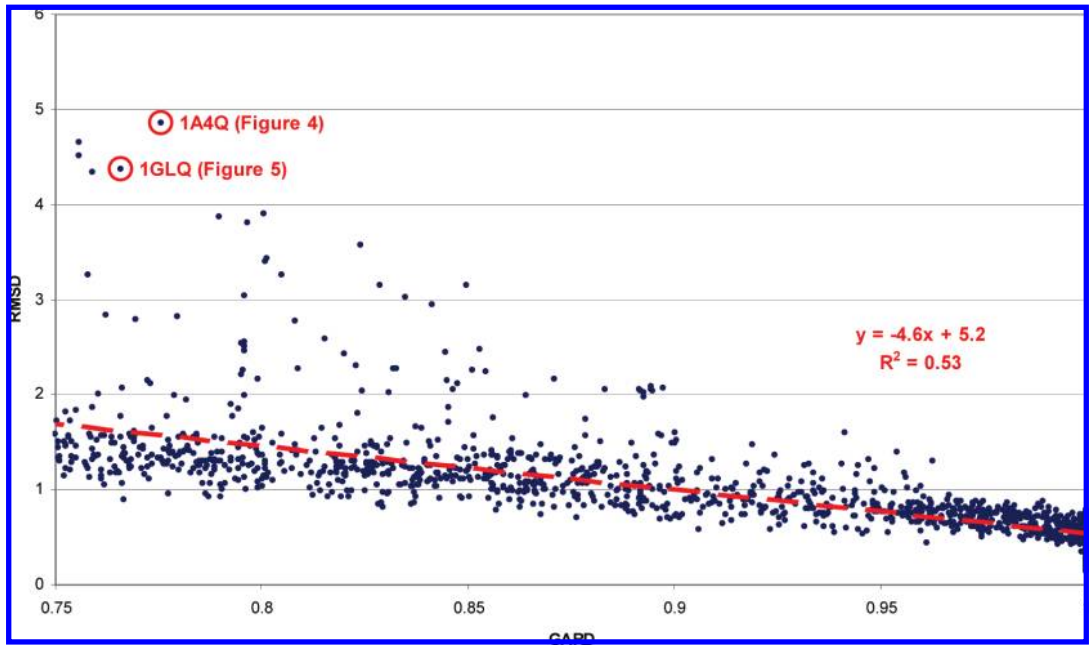

Figure 3. Correlation between GARD scores and rmsd for those poses with a GARD score of at least 0.75 across the top 10 poses of compounds from 68 different targets and 8 docking methods (1 469 points).

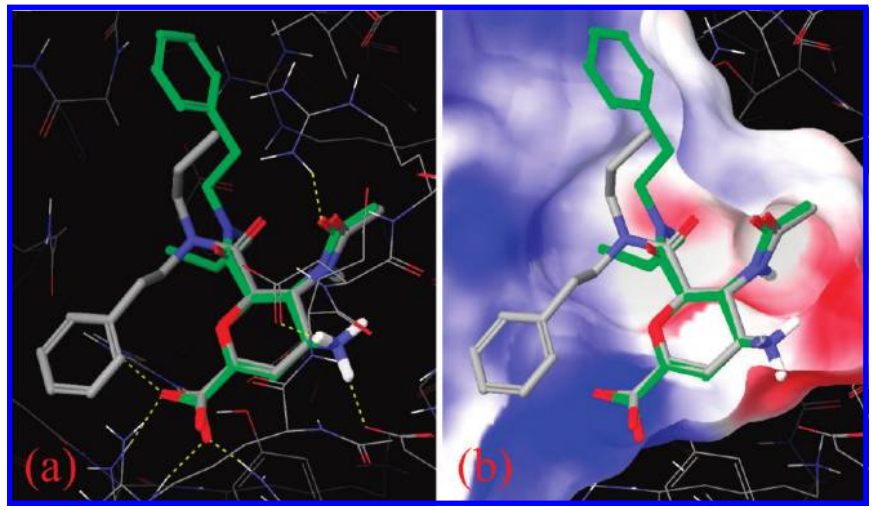

Figure 4. 1A4Q crystal structure[55] with ligand shown in green tubing. The pose shown in gray is an outlier from Figure 3 with an rmsd of 4.9 Å and a GARD score of 0.78. (a) Shows the structures with hydrogen bonds between the crystal structure ligand and protein shown in yellow. (b) Shows the same view with a (clipped) surface colored by electrostatic potential.

(4.9 Å). This is a typical example where a large error in a relatively small portion of the structure can result in a very high rmsd, yet has a much smaller impact on the GARD score.

The key interactions between the ligand in 1A4Q and the influenza A sialidase receptor consist of strong charge−charge interactions and hydrogen bonds, along with some highly specific hydrophobic contacts. The carboxylic acid makes charge−charge interactions with a cluster of three arginine residues (118, 292, and 371), while the amide carbonyl oxygen makes a favorable hydrogen bond to the side chain of Arg-152; the methyl is involved in hydrophobic interactions with residues Trp-178 and Ile-222. The *N*-phenethyl-*N*-propylamide substituent fits into the so-called 'glycerol pocket', which is created via a major rearrangement in the conformation of the active-site Glu-276.[55] The propyl chain of the bridge occupies a partially lipophilic region of the protein defined by the side chains of the residues Ala-246, Asn-249, Glu-276, and Arg-292. This compound has an IC_{50} of 2 nM against Sialidase A[55], while the compound without the hydrophobic side chains and only two methyl substituents on the amide have an IC_{50} of 2400 nM,[56] indicating the importance of the hydrophobic contacts described above. This may be a weakness of the GARD methodology (which by design down-weights hydrophobic interactions), although it is clear from a docking perspective that the hydrogen-bonding interactions are what cause the core to be correctly placed. This serendipitous interaction between the docking scoring function and GARD help recover this pose from being incorrectly assigned as a docking failure, even though it has an rmsd of 4.9 Å.

1GLQ[57] *− Glutathione-S-transferase with p-Nitrobenzyl Glutathione.* The docking pose generated through ICM (Figure 5) identifies an additional possible hydrogen-bonding interaction not present in the crystal structure that leads to

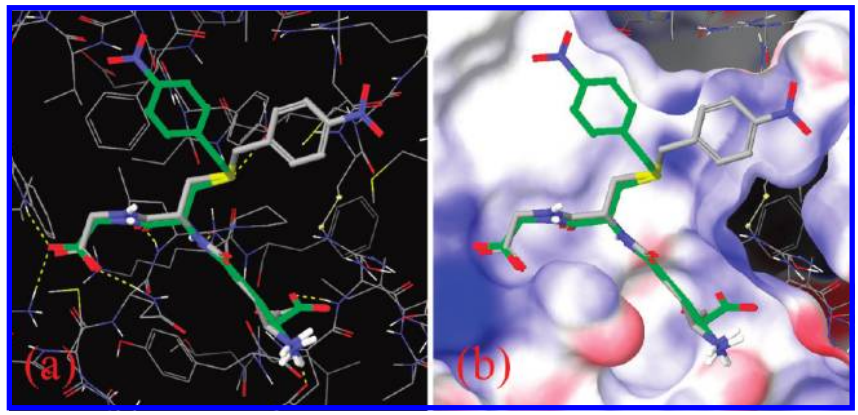

Figure 5. 1GLQ crystal structure[57] with ligand shown in green tubing. The pose shown in gray is an outlier from Figure 3 with an rmsd of 4.4 Å and a GARD score of 0.77. (a) Shows the structures with hydrogen bonds between the crystal structure ligand and protein shown in yellow. (b) Shows the same view with a (clipped) surface colored by electrostatic potential.

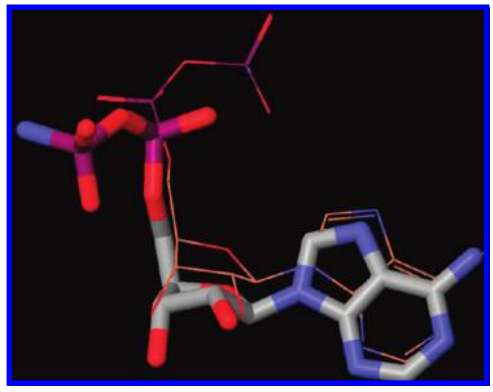

Figure 6. 1QPC crystal ligand[58] is shown in gray tubing. The top scoring pose is shown in brown wire. This is also the pose with the best rmsd and GARD score (rmsd = 2.54 Å, GARD = 0.63) among the 10 GlideSP poses generated.

the flipping of the nitro-benzyl side chain. The rest of the structure and almost all of the hydrogen-bonding interactions are maintained, but the flipping of this side chain results in a high rmsd of 4.4 Å. The GARD score of 0.77 does not reflect the fact that substantially fewer atoms are misaligned, as compared to the previous 1A4Q example (8 atoms rather than the 11 atoms misaligned in Figure 4), due to higher weighting of the nitro group, which effectively increases the size of the incorrectly positioned side chain. As with Figure 4, the high GARD score indicates that the majority of the pose is correctly positioned, but the high rmsd tells us that those parts of the structure that are not correctly positioned are significantly out of place.

1QPC[58] *− LCK with AMPPNP.* Figure 6 shows the top scoring pose for the 1QPC ligand docked back into the crystal structure using GlideSP. This pose has both the lowest rmsd and highest GARD score of the 10 poses generated. Manual examination of the pose shows that the core of the structure is docked correctly with very little deviation between the docked and crystal poses, particularly in the hinge binding region that is important for kinase ligands. The high rmsd (2.54 Å, large enough to be considered a docking failure) is caused by the error in the position of the solvent exposed side chain, which may be very mobile and is not as well-defined as the rest of the ligand in the crystal structure. Even with the large weighting assigned to polar groups by the Andrews analysis, GARD does not unduly suffer from this large deviation in position and correctly identifies this pose as good with the majority of the atoms aligned well and assigns a score of 0.63. As with the examples above, the GARD score indicates that some of the atoms in the structure are not correctly aligned but the majority, particularly those weighted highly in the Andrew's analysis, are.

A number of illustrative examples not in the docking evaluation set were also examined in some detail and are described below.

1HPX[59] *− HIV Protease with KNI-272 Inhibitor.* Figure 7 shows the lowest rmsd (rmsd = 1.89 Å, GARD = 0.63) and highest GARD scoring (rmsd = 2.35 Å, GARD = 0.75, also the top ranked by GlideSP) poses of the 1HPX ligand redocked to the crystal structure using GlideSP. While the highest GARD score pose (shown in brown) has a worse rmsd, it captures more of the hydrogen-bonding interactions than the lowest rmsd pose. This can be seen in Table 5, which shows the deviation in atom positions between the best rmsd and best GARD poses compared to the crystal pose for those atoms involved in hydrogen bonds to the receptor (as identified by the RCSB PDB Ligand Explorer[60] and shown in Figure 8).

The pose with the lowest rmsd misaligns the hydrogen-bonding groups, particularly amide 5 which hydrogen bonds to Asp-25 and is flipped in the docked pose relative to the crystal structure, and the central benzyl side chain but places the isoquinoline group in approximately the correct region, albeit with the wrong orientation. As well as identifying the majority of hydrogen bonds, the highest scoring GARD pose aligns the central benzyl group very well, while the terminal isoquinoline is grossly incorrect. The relatively hydrophobic end of the isoquinoline is exposed to solvent in the crystal structure, however, in the top scoring pose (both by GARD and GlideSP), the isoquinoline is tucked into a somewhat hydrophobic region of the protein. Here the limited contribution from each atom and normalized alignment factor, included in the GARD algorithm, are beneficial as a further movement of the isoquinoline out of position (making the alignment worse and further increasing rmsd) does not reduce the GARD score. This allows the correct alignment of the key hydrogen-bonding groups and benzyl side chain deep in the pocket to drive the overall GARD score assigned to this pose.

2ITY[54] *− EGFR Kinase with Iressa.* In the original work of Andrews and co-workers, charged groups, particularly basic nitrogens, are weighted very highly. This was initially

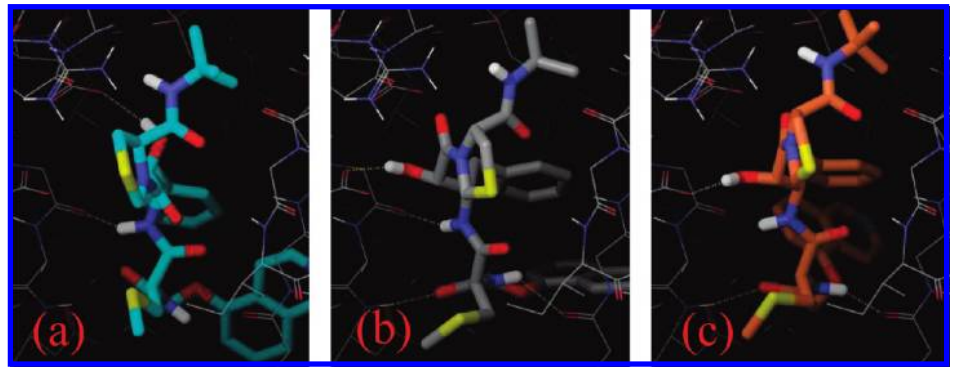

Figure 7. (b) 1HPX crystal pose[59] shown in gray tubing. (a) The docking pose shown in cyan corresponds to the pose with the best rmsd of 1.89 Å (GARD = 0.63, ranked 10/30 using GlideSP). (c) The docking pose shown in brown is the pose with the best GARD score (0.75) and is also top scoring pose using GlideSP (rmsd = 2.35 Å).

Table 5. Deviations in Heavy Atom Positions for Key Hydrogen-Bonding Atoms Identified by the RCSB PDB Ligand Explorer[60] between the Two Docking Poses Shown in Figure 7 (a and c) and the 1HPX Crystal Structure Ligand (Figure 8)[a]

Hydrogen Bonding Atom	Best GARD Pose Error (Å)	Best RMSD Pose Error (Å)	Difference
1	0.29	1.15	0.86
2	0.32	0.97	0.65
3	0.62	0.68	0.06
4	1.35	2.99	1.64
5	1.04	4.38	3.34
Average	0.72	2.03	1.31

[a] The best GARD pose has a GARD score of 0.75 and a rmsd of 2.35 Å compared to a GARD score of 0.63 and a rmsd of 1.89A for the pose with the lowest rmsd.

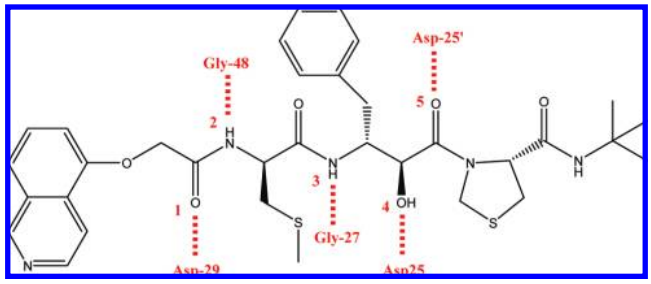

Figure 8. Ligand from 1HPX crystal structure[59] with the five hydrogen-bonding regions identified by the RSCB PDB Ligand Explorer[60] indicated in red.

a cause of concern since such groups may be used to improve the solubility of drug candidates and have little to do with binding. Since the aim of the GARD scheme is to reflect the importance of those parts of the structure that interact with the protein, assigning a high weighting to solubilizing groups, which are generally solvent exposed and flexible, and thus, may not have a well-defined position, was thought to be undesirable. However, examination of the results from a number of complexes has shown that GARD is actually less affected by solubilizing groups than rmsd. This is due to two features of GARD: (1) the measure is fundamentally the proportion of atoms that are well aligned, and solubilizing groups tend to be relatively small compared to the structure as a whole, and (2) the alignment factor in this implementation of the GARD algorithm is normalized with the minimum score reached at 2.5 Å, so large deviations in the position of solubilizing groups do not have the overwhelming effect that they tend to have on rmsd.

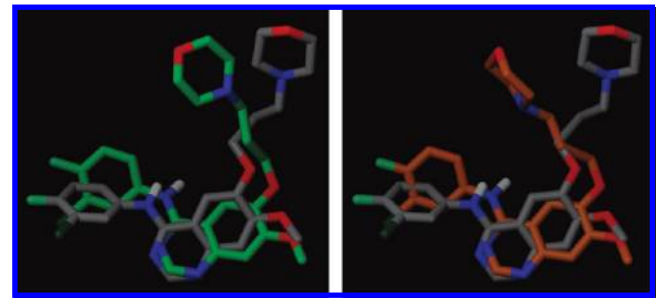

Figure 9. 2ITY crystal ligand[54] shown in gray tubing. The two top scoring docking poses are shown in green and brown. Both of these poses would be considered rmsd failures with rmsds of 2.36 (in green) and 2.74 Å (in brown) but GARD successes with values of 0.52 and 0.50, respectively.

This is illustrated in the case of 2ITY (Figures 9 and 10), where the two top scoring docking poses both align well with the ligand crystal structure except for the morpholine solubilizing group. Both of these poses would be considered docking failures with rmsds of 2.36 and 2.74 Å but score well using GARD with values of 0.52 and 0.50, respectively. Examination of the poses in the context of the protein clearly shows that the morpholine is exposed to solvent, and that the key binding site interactions are generally maintained. The high weightings assigned to the morpholine results in a GARD score slightly lower than would be expected from the proportion of the structure that is poorly aligned but does not completely overwhelm the score. Examining poses with GARD scores greater than 0.5 would, therefore, allow these poses to be identified even though a number of highly weighted atoms are misaligned, effectively rescuing them.

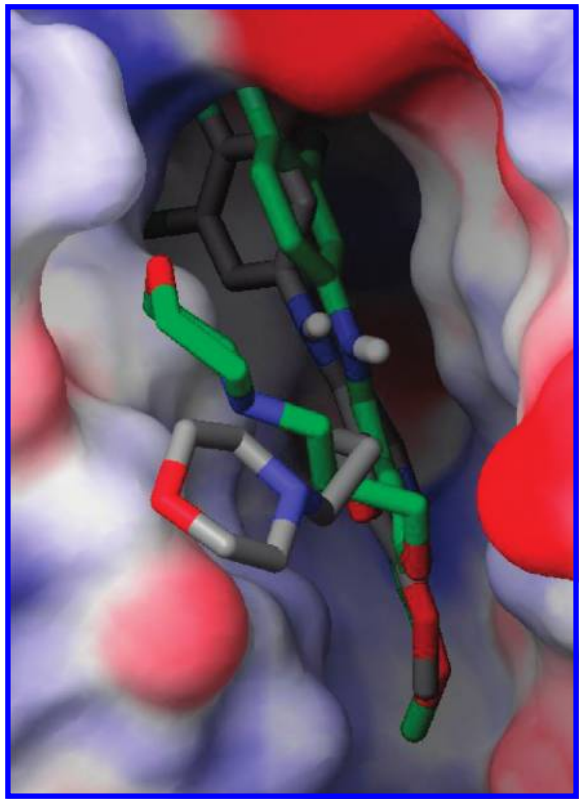

Figure 10. 2ITY crystal ligand[54] shown in gray tubing, top scoring pose shown in green (rmsd = 2.36 Å, GARD = 0.52). The poor rmsd is due to the large error in position of the solvent exposed morpholine group, even though the majority of interactions with the protein are maintained.

DISCUSSION

The authors believe that the basic GARD methodology introduced in this paper has been well tested in the context of the analysis of docking poses and found to be sound. When the Andrews weightings were first included, there was concern that the high value assigned to polar groups would result in overweighting the importance of solubilizing groups, which would not be desirable due to their flexibility. However, since such groups generally only make up a small proportion of the structure, the overall effect on the GARD score is relatively small. The algorithm is designed to be extensible, and different alignment scoring and weighting schemes could be included. The analysis performed by Andrews and co-workers could be updated using more recent data, and it would be trivial to calculate different sets of weights for specific protein classes. The scoring scheme could also be customized for individual proteins or crystal structures, for example, when comparing docked poses with crystal structures, the crystallographic confidence in ligand location could be taken into account, potentially reducing the weight of flexible ligand side chains. Furthermore, interactions with the protein could explicitly be taken into account, as long as they could be clearly identified. This could be as simple as decreasing the weighting of atoms that have a low level of van der Waals contact with the protein or as involved as attempting to identify explicit interactions, which would receive an increased weighting in the GARD algorithm.

The progressive scoring scheme used in these examples assumes that any difference in position of less than 1 Å is considered correct and any difference greater than 2.5 Å is considered unacceptable (with a linear interpolation between the two). It may be more appropriate to use alternative values depending on the resolution of the crystal structure being considered, for example, increasing both d_{min} and d_{max} for structures with low resolution and, thus, greater uncertainty in the bound position of the ligand. It could well be the case that in other applications a different alignment scoring scheme would be more appropriate. Since the basic GARD algorithm is designed to be able to use any alignment scoring scheme, countless variants could be included with little difficulty.

Although not discussed in detail in the present work, the use of GARD to identify distinct docking poses or conformers has been suggested and is thought to be beneficial. In these applications rmsd is used to identify poses or conformers that are different enough from those previously identified to be of interest. Basing the measure of fit on the proportion of atoms aligned, rather than allowing a relatively small number of very badly aligned atoms to skew the score, should reduce the problem of flexible side chains producing disproportionately larger numbers of poses or conformers, since each atom has a maximum contribution to the score no matter how well, or poorly, it aligns to other structures. Including GARD in such a process would also allow variable weighting to be assigned to atoms. Using the Andrews weighting scheme discussed above would allow changes in those parts of the structure likely to contribute to binding (those identified as important in the Andrews analysis) to be weighted more highly than the atoms less likely to make a significant contribution to binding affinity. This would result in oversampling of poses or conformers with different hydrogen-bonding positions (often in the core) and under sampling of side chains unlikely to be involved in binding, something that could be useful in pharmacophore modeling. It would also be possible to devise different weighting schemes to address issues specific to the application intended for the conformers.

One potential drawback of the GARD approach is that for complexes that are considered incorrect, those with a GARD score of zero, some information is effectively discarded. All that is known is that every atom is at least d_{max} Å out of position with no reduction in score possible for larger displacements. In these cases, it may be useful to consider both the rmsd and GARD score together when examining alignments. It would also be possible to modify the alignment factor (δ_i from Equation 3) to an alternative form, such as a sigmoid function. Such a change would allow improved differentiation between badly aligned structures, but would result in a loss of the simple cutoff based nature of the alignment score in the current GARD implementation and so was not used in this assessment.

By construction the GARD measure can be used in any of the settings where rmsd is currently applied; either as a more 'chemically aware' alternative or as an addition, or postscreen, to an rmsd analysis to prevent false negatives polluting the final data set. The benefits of GARD over rmsd are of a fundamental nature, incorporated by design, and, as such, are completely transferable. It is the author's hope that the basic GARD algorithm, potentially with different weightings and alignment functions, will be found useful and will be applied to a range of problems in the field.

CONCLUSIONS

An alternative to the standard rmsd metric has been proposed. This measure, dubbed GARD, has been shown to be versatile and is conceptually simple. By construction, GARD overcomes many of the key problems associated with rmsd. First, it is normalized to unity. This ensures that poorly docked poses are assigned a low GARD score and do not skew resulting aggregate statistics. This, in turn, allows for a fair comparison among statistics generated through a variety of sources without, unfairly, penalizing (sometimes in a crippling fashion if the data point is extremely egregious) outliers. Second, through an inclusion of a 'chemical awareness', use of GARD can prevent the discarding of perfectly good data; an essentially correct binding pose in which all of the small molecule functional elements are in the correct position is not overlooked through having too large an rmsd driven solely by the misplacement of a small portion of the structure, which may not significantly contribute to the binding. Third, this methodology is completely automated with few user specified input parameters. Indeed, in this implementation the use of variable d_{min} and d_{max} allows for the 'tweaking' of the GARD measure. When assessing data in the context of crystallographic information, more rigid cutoffs may be applicable for highly resolved experimental data, while the cutoffs may need to be relaxed for use with poorer data sets. The automated nature of the GARD score also obviates the need for subjective manual assessment and allows large numbers of alignments to be assessed quickly. Finally, the GARD measure has the added advantage of being conceptually simple to interpret; although the alignment function and atom weighting means that it is not a trivial relationship, at its base, the GARD score is a measure of the proportion of atoms correctly aligned. From a practical perspective, given that it is no more expensive to compute both rmsd and GARD, than rmsd alone, one could envisage using GARD in conjunction with rmsd to avoid the loss of false negatives in a docking experiment.

At its heart, GARD is a practical solution to a difficult problem and, while not completely solving all of the issues inherent in rmsd, is clearly a promising approach.

ACKNOWLEDGMENT

The authors thank the other members of the docking research team, YongBo Hu, Kristi Fan, and Brajesh Rai, who took part in the original assessment of docking program performance, and Jack Bikker for his help with the manuscript and insightful comments.

REFERENCES AND NOTES

(1) Andrews, P. R.; Craik, D. J.; Martin, J. L. Functional group contributions to drug-receptor interactions. *J. Med. Chem.* **1984**, *27* (12), 1648–57.
(2) Maiorov, V. N.; Crippen, G. M. Significance of root-mean-square deviation in comparing three-dimensional structures of globular proteins. *J. Mol. Biol.* **1994**, *235* (2), 625–34.
(3) Damm, K. L.; Carlson, H. A. Gaussian-weighted RMSD superposition of proteins: a structural comparison for flexible proteins and predicted protein structures. *Biophys. J.* **2006**, *90* (12), 4558–4573.
(4) Cole, J. C.; Murray, C. W.; Nissink, J. W. M.; Taylor, R. D.; Taylor, R. Comparing protein-ligand docking programs is difficult. *Proteins: Struct., Funct., Bioinf.* **2005**, *60* (3), 325–332.
(5) Joseph-McCarthy, D. Computational approaches to structure-based ligand design. *Pharmacol. Ther.* **1999**, *84* (2), 179–191.
(6) Lang, P. T.; Aynechi, T.; Moustakas, D.; Shoichet, B.; Kuntz, I. D.; Brooijmans, N.; Oshiro, C. M. , Molecular docking and structure-based design. In *Drug Discovery Research: New Frontiers in the Post-Genomic Era*; Huang, Z., Ed; John Wiley & Sons, Inc.: Hoboken, NJ, 2007; pp 3−23.
(7) Bursulaya, B. D.; Totrov, M.; Abagyan, R.; Brooks, C. L., III. Comparative study of several algorithms for flexible ligand docking. *J. Comput.-Aided Mol. Des.* **2004**, *17* (11), 755–763.
(8) Kellenberger, E.; Rodrigo, J.; Muller, P.; Rognan, D. Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins: Struct., Funct., Bioinf.* **2004**, *57* (2), 225–242.
(9) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins: Struct., Funct., Bioinf.* **2004**, *56* (2), 235–249.
(10) Warren, G. L.; Andrews, C. W.; Capelli, A.-M.; Clarke, B.; LaLonde, J.; Lambert, M. H.; Lindvall, M.; Nevins, N.; Semus, S. F.; Senger, S.; Tedesco, G.; Wall, I. D.; Woolven, J. M.; Peishoff, C. E.; Head, M. S. A Critical Assessment of Docking Programs and Scoring Functions. *J. Med. Chem.* **2006**, *49* (20), 5912–5931.
(11) Onodera, K.; Satou, K.; Hirota, H. Evaluations of Molecular Docking Programs for Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47* (4), 1609–1618.
(12) Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of several molecular docking programs: Pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.* **2009**, *49*, (6), 1455−1474.
(13) Wang, R.; Lu, Y.; Wang, S. Comparative Evaluation of 11 Scoring Functions for Molecular Docking. *J. Med. Chem.* **2003**, *46* (12), 2287–2303.
(14) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., III. Assessing Scoring Functions for Protein-Ligand Interactions. *J. Med. Chem.* **2004**, *47* (12), 3032–3047.
(15) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of Docking Performance: Comparative Data on Docking Algorithms. *J. Med. Chem.* **2004**, *47* (3), 558–565.
(16) Englebienne, P.; Fiaux, H.; Kuntz, D. A.; Corbeil, C. R.; Gerber-Lemaire, S.; Rose, D. R.; Moitessier, N. Evaluation of docking programs for predicting binding of Golgi a-mannosidase II inhibitors: a comparison with crystallography. *Proteins: Struct., Funct., Bioinf.* **2007**, *69* (1), 160–176.
(17) Bissantz, C.; Folkers, G.; Rognan, D. Protein-Based Virtual Screening of Chemical Databases. 1. Evaluation of Different Docking/Scoring Combinations. *J. Med. Chem.* **2000**, *43* (25), 4759–4767.
(18) Stahl, M.; Rarey, M. Detailed Analysis of Scoring Functions for Virtual Screening. *J. Med. Chem.* **2001**, *44* (7), 1035–1042.
(19) Schulz-Gasch, T.; Stahl, M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *J. Mol. Model.* **2003**, *9* (1), 47–57.
(20) Cummings, M. D.; DesJarlais, R. L.; Gibbs, A. C.; Mohan, V.; Jaeger, E. P. Comparison of automated docking programs as virtual screening tools. *J. Med. Chem.* **2005**, *48* (4), 962–976.
(21) Kontoyianni, M.; Sokol, Glenn S.; McClellan, Laura M. Evaluation of library ranking efficacy in virtual screening. *J. Comput. Chem.* **2005**, *26* (1), 11–22.
(22) Yang, J.-M.; Chen, Y.-F.; Shen, T.-W.; Kristal, B. S.; Hsu, D. F. Consensus Scoring Criteria for Improving Enrichment in Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45* (4), 1134–1146.
(23) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of Topological, Shape, and Docking Methods in Virtual Screening. *J. Chem. Inf. Model.* **2007**, *47* (4), 1504–1519.
(24) Zhou, Z.; Felts, A. K.; Friesner, R. A.; Levy, R. M. Comparative Performance of Several Flexible Docking Programs and Scoring Functions: Enrichment Studies for a Diverse Set of Pharmaceutically Relevant Targets. *J. Chem. Inf. Model.* **2007**, *47* (4), 1599–1608.
(25) Deng, W.; Verlinde, C. L. M. J. Evaluation of Different Virtual Screening Programs for Docking in a Charged Binding Pocket. *J. Chem. Inf. Model.* **2008**, *48* (10), 2010–2020.
(26) Kellenberger, E.; Foata, N.; Rognan, D. Ranking Targets in Structure-Based Virtual Screening of Three-Dimensional Protein Libraries: Methods and Problems. *J. Chem. Inf. Model.* **2008**, *48* (5), 1014–1025.
(27) Sheridan, R. P.; McGaughey, G. B.; Cornell, W. D. Multiple protein structures and multiple ligands: effects on the apparent goodness of virtual screening results. *J. Comput.-Aided Mol. Des.* **2008**, *22* (3−4), 257–265.
(28) Verkhivker, G. M.; Bouzida, D.; Gehlhaar, D. K.; Rejto, P. A.; Arthurs, S.; Colson, A. B.; Freer, S. T.; Larson, V.; Luty, B. A.; Marrone, T.; Rose, P. W. Deciphering common failures in molecular docking of

ligand-protein complexes. *J. Comput.-Aided Mol. Des.* **2000**, *14* (8), 731–751.
(29) Jain, A. N.; Nicholls, A. Recommendations for evaluation of computational methods. *J. Comput.-Aided Mol. Des.* **2008**, *22* (3−4), 133–139.
(30) Jain, A. N. Bias, reporting, and sharing: computational evaluations of docking methods. *J. Comput.-Aided Mol. Des.* **2008**, *22* (3−4), 201–212.
(31) Kroemer, R. T.; Vulpetti, A.; McDonald, J. J.; Rohrer, D. C.; Trosset, J.-Y.; Giordanetto, F.; Cotesta, S.; McMartin, C.; Kihlen, M.; Stouten, P. F. W. Assessment of docking poses: Interactions-based accuracy classification (IBAC) versus crystal structure deviations. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 871–881.
(32) Pang, Y.-P.; Perola, E.; Xu, K.; Prendergast, F. G. EUDOC: a computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases. *J. Comput. Chem.* **2001**, *22* (15), 1750–1771.
(33) Abagyan, R. A.; Totrov, M. M. Contact area difference (CAD): a robust measure to evaluate accuracy of protein models. *J. Mol. Biol.* **1997**, *268* (3), 678–685.
(34) Yusuf, D.; Davis, A. M.; Kleywegt, G. J.; Schmitt, S. An Alternative Method for the Evaluation of Docking Performance: RSR vs RMSD. *J. Chem. Inf. Model* **2008**, *48* (7), 1411–1422.
(35) Keseru, G. M.; Makara, G. M. The influence of lead discovery strategies on the properties of drug candidates. *Nat. Rev. Drug Discovery* **2009**, *8* (3), 203–212.
(36) Reynolds, C. H.; Tounge, B. A.; Bembenek, S. D. Ligand Binding Efficiency: Trends, Physical Basis, and Implications. *J. Med. Chem.* **2008**, *51* (8), 2432–2438.
(37) Bartoli, S.; Fincham, C. I.; Fattori, D. Fragment-based drug design: combining philosophy with technology. *Curr. Opin. Drug Discovery Dev.* **2007**, *10* (4), 422–429.
(38) Zhao, H. Scaffold selection and scaffold hopping in lead generation: a medicinal chemistry perspective. *Drug Discovery Today* **2007**, *12* (3−4), 149–155.
(39) *OEChem*, Version 1.4.0 OpenEye Scientific Software, Inc.: Sante Fe, NM, 2006.
(40) Shoichet, B. K.; Bodian, D. L.; Kuntz, I. D. Molecular docking using shape descriptors. *J. Comput. Chem.* **1992**, *13* (3), 380–97.
(41) Ewing, T. J. A.; Kuntz, I. D. Critical evaluation of search algorithms for automated molecular docking and database screening. *J. Comput. Chem.* **1997**, *18* (9), 1175–1189.
(42) Moustakas, D. T.; Lang, P. T.; Pegg, S.; Pettersen, E.; Kuntz, I. D.; Brooijmans, N.; Rizzo, R. C. Development and validation of a modular, extensible docking program: DOCK 5. *J. Comput.-Aided Mol. Des.* **2006**, *20* (10/11), 601–619.
(43) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins: Struct., Funct., Genet.* **1999**, *37* (2), 228–241.
(44) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739–1749.
(45) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: A new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47* (7), 1750–1759.
(46) Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM - a new method for protein modeling and design: applications to docking and structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15* (5), 488–506.
(47) Totrov, M.; Abagyan, R., Flexible protein-ligand docking by global energy optimization in internal coordinates. *Proteins: Struct., Funct., Genet.* 1998, Suppl. 1, (Suppl. 1), 215−220.
(48) Joseph-McCarthy, D.; Thomas, B. E. I. V.; Belmarsh, M.; Moustakas, D.; Alvarez, J. C. Pharmacophore-based molecular docking to account for ligand flexibility. *Proteins: Struct., Funct., Genet.* **2003**, *51* (2), 172–188.
(49) Joseph-McCarthy, D.; McFadyen, I. J.; Zou, J.; Walker, G.; Alvarez, J. C., Pharmacophore-based molecular docking: A practical guide. In *Virtual Screening in Drug Discovery*; Alvarez, J. C., Shoichet, B., Eds.; CRC Press: Boca Raton, FL., 2005; pp 327−347.
(50) Jain, A. N. Surflex: Fully automatic flexible molecular docking using a molecular similarity-based search engine. *J. Med. Chem.* **2003**, *46* (4), 499–511.
(51) Jain, A. N. Surflex-Dock 2.1: Robust performance from ligand energetic modeling, ring flexibility, and knowledge-based search. *J. Comput.-Aided Mol. Des.* **2007**, *21* (5), 281–306.
(52) Pham, T. A.; Jain, A. N. Parameter Estimation for Scoring Protein-Ligand Interactions Using Negative Training Data. *J. Med. Chem.* **2006**, *49* (20), 5856–5868.
(53) *CORINA*, Version 1.82; Molecular Networks GmbH: Erlangen, Germany, 1997.
(54) Yun, C.-H.; Boggon, T. J.; Li, Y.; Woo, M. S.; Greulich, H.; Meyerson, M.; Eck, M. J. Structures of lung cancer-derived EGFR mutants and inhibitor complexes: mechanism of activation and insights into differential inhibitor sensitivity. *Cancer Cell* **2007**, *11* (3), 217–227.
(55) Taylor, N. R.; Cleasby, A.; Singh, O.; Skarzynski, T.; Wonacott, A. J.; Smith, P. W.; Sollis, S. L.; Howes, P. D.; Cherry, P. C.; Bethell, R.; Colman, P.; Varghese, J. Dihydropyrancarboxamides Related to Zanamivir: A New Series of Inhibitors of Influenza Virus Sialidases. 2. Crystallographic and Molecular Modeling Study of Complexes of 4-Amino-4H-pyran-6-carboxamides and Sialidase from Influenza Virus Types A and B. *J. Med. Chem.* **1998**, *41* (6), 798–807.
(56) Smith, P. W.; Sollis, S. L.; Howes, P. D.; Cherry, P. C.; Starkey, I. D.; Cobley, K. N.; Weston, H.; Scicinski, J.; Merritt, A.; Whittington, A.; Wyatt, P.; Taylor, N.; Green, D.; Bethell, R.; Madar, S.; Fenton, R. J.; Morley, P. J.; Pateman, T.; Beresford, A. Dihydropyrancarboxamides Related to Zanamivir: A New Series of Inhibitors of Influenza Virus Sialidases. 1. Discovery, Synthesis, Biological Activity, and Structure-Activity Relationships of 4-Guanidino- and 4-Amino-4H-pyran-6-carboxamides. *J. Med. Chem.* **1998**, *41* (6), 787–797.
(57) Garcia-Saex, I.; Parraga, A.; Phillips, M. F.; Mantle, T. J.; Coll, M. Molecular structure at 1.8 Ang. of mouse liver class Pi glutathione S-transferase complexed with S-(p-nitrobenzyl)glutathione and other inhibitors. *J. Mol. Biol.* **1994**, *237* (3), 298–314.
(58) Zhu, X.; Kim, J. L.; Newcomb, J. R.; Rose, P. E.; Stover, D. R.; Toledo, L. M.; Zhao, H.; Morgenstern, K. A. Structural analysis of the lymphocyte-specific kinase Lck in complex with non-selective and Src family selective kinase inhibitors. *Structure (London)* **1999**, *7* (6), 651–661.
(59) Baldwin, E. T.; Bhat, T. N.; Gulnik, S.; Liu, B.; Topol, I. A.; Kiso, Y.; Mimoto, T.; Mitsuya, H.; Erickson, J. W. Structure of HIV-1 protease with KNI-272, a tight-binding transition-state analog containing allophenylnorstatine. *Structure (London)* **1995**, *3* (6), 581–90.
(60) *RCSB PDB Ligand Explorer*, Version 3.4; Research Collaboratory for Structural Bioinformatics: Piscataway, NJ, 2009.

CI9001074