LETTER

Comments on the Article "On Evaluating Molecular-Docking Methods for Pose Prediction and Enrichment Factors"

Emanuele Perola,* W. Patrick Walters, and Paul Charifson

Vertex Pharmaceuticals Inc., 130 Waverly Street, Cambridge, Massachusetts 02139

Received October 26, 2006

Abstract: The recent article "On Evaluating Molecular-Docking Methods for Pose Prediction and Enrichment Factors" (Chen H. et al. *J. Chem. Inf. Model.* **2006**, *46*, 401−415) contains a series of comments on a similar study we published in *Proteins* in 2004 (Perola et. al. *Proteins* **2004**, *56*, 235−249). We believe that some of those comments are misleading, and we feel that an adequate response is in order.

In the Chen paper[1] the authors compare the performance of four docking programs (ICM, Glide, GOLD, and FlexX) as tools for binding mode prediction and virtual screening. In their pose prediction study ICM reproduces 91% of the structures within 2.0 Å of the corresponding experimental structures, outperforming the second best contender (Glide) with a margin of 28%. The authors observe that their results are in stark contrast with those published in our paper,[2] in which Glide outperformed ICM on a different test set using the same metric for comparison (Glide 61%, ICM 45%). The authors suggest that problems in the receptor preparation procedure used in our study are the most plausible explanation for the poorer performance we observed for ICM. In particular, the authors state that "*Receptor preparation by Macromodel (Schrodinger) is appropriate for GLIDE but not ICM (Molsoft) as some atom type assignments generated by the former are incompatible and, therefore, unreadable by the latter*". The implication is that ICM was penalized in our study because Macromodel was used during protein preparation. They also suggest that differences in the version of the code used in the two studies might contribute to the differences in performance. To corroborate this point they mention a personal communication from ICM developer Max Totrov and one of us (E. Perola) stating that "*the Vertex study was recently repeated by the same authors using the current version of ICM, and considerable improvements in ICM docking accuracy were observed*". In this regard it should be pointed out that we had detailed exchanges with Totrov about our study, and while we endorsed this particular statement we never spoke directly with the Chen group.

* Corresponding author e-mail: emanuele_perola@vrtx.com.

In order to clarify the issues raised in the Chen paper and prevent possible misunderstandings, we provide below a full account of our efforts in regard to both studies, along with the associated findings. In particular, we would like to address the following points: (1) impact of protein preparation procedure on ICM performance, (2) magnitude of improvements in recent ICM versions, and (3) reproducibility of the Chen study.

IMPACT OF PROTEIN PREPARATION PROCEDURE

In our original study, the X-ray structures of the test complexes were converted to Macromodel format in order to perform hydrogen addition and minimization with Macromodel. Each structure was visually inspected before and after this step to assign appropriate protonation states first and ensure correct hydrogen orientations at the end. In order to generate suitable input files for ICM the structures thus prepared were then converted to pdb format. When we revisited our study in January 2005 in conjunction with Totrov, it became apparent that for 21 of the 150 structures in our test set the conversion from Macromodel to pdb had been imperfect. The problem had arisen from the presence in those 21 structures of multiple residues labeled with the same number and only differentiated by an extra letter (e.g., 151A, 151B). As a consequence of the conversion glitch, the ability of ICM to correctly read those particular input files was impaired, resulting in poor docking results for the corresponding complexes. This issue had no impact on the results obtained with Glide and GOLD, as they take Macromodel and mol2 files as inputs, respectively. The 21 problematic structures were corrected, and docking with ICM was repeated. Eight of the 21 structures were now reproduced within 2.0 Å of the experimental coordinates, compared to only 1 out of 21 in the original study. The input files for the remaining 129 structures in the data set were carefully inspected, and no additional conversion problems were found. Overall the percentage of structures reproduced by ICM within 2.0 Å of the corresponding X-ray structures increased from 45 to 49% relative to the results we originally reported. This correction does not alter the general conclusions of our original study: Glide 2.5 remains the best performer (61% structures reproduced within 2.0 Å) with significant margin over ICM 3.0 (49%) and GOLD 2.0 (48%). The relative rank order between ICM and GOLD is now inverted, but the difference in performance between the two remains marginal.

It is clear from this analysis that problems in the protein preparation procedure resulting from the use of different software packages had a minor impact on the outcome of our study and do not account in any significant way for the huge difference in ICM performance observed in the two papers. Moreover, it should be pointed out that the alleged incompatibility between Macromodel and ICM atom types is not an issue, as the atom types assigned by Macromodel have no bearing on the atom names that appear in the

10.1021/ci600460h CCC: $37.00
Published on Web 01/30/2007

Table 1. Percentage of Correct Docking Poses We Obtained on the Perola Test Set in the Original and Revised Versions of the Study[c]

program	% correct
ICM 3.0[a]	45
ICM 3.0[b]	49
ICM 3.3	61
Glide 2.5	61
Glide 3.5	61

[a] Results from the original study. [b] Results after correction. [c] Correct: RMSD ≤ 2.0 Å relative to the corresponding X-ray structure.

translated pdb file, from which ICM assigns its own atom types.

IMPROVEMENTS IN RECENT ICM VERSIONS

Following our first exchange, Totrov implemented some changes in the ICM docking code and provided us with the new executables (later incorporated in ICM 3.3). We repeated our original study using the updated code and observed a significant improvement in docking accuracy: the percentage of structures predicted within 2.0 Å of the experimental structure increased from 49 to 61%. In order to make an equitable assessment, we also repeated the study with Glide 3.5 and GOLD 2.2, the most up-to-date versions of the two programs at the time. The percentages of correctly predicted structures for these two programs were identical to those observed in our original study (61% for Glide, 48% for GOLD). In conclusion, the newly upgraded version of ICM achieved the same level of docking accuracy as the most recent version of Glide on our original test set. The performances of ICM and Glide in the original and updated versions of our pose prediction study are summarized in Table 1.

While significant progress in the accuracy of the ICM docking code emerged from these results, in no way do the improvements we observed bridge the huge gap between our original results and those reported by Chen and co-authors: the 61% accuracy we observe is still a far cry from the 91% reported in their paper on a different test set.

REPRODUCIBILITY OF THE CHEN STUDY

In order to investigate the reasons for the above discrepancy, we tried to reproduce the pose prediction study reported in the Chen paper using the same test set and the same protein preparation procedure for ICM docking. We downloaded from the PDB the 164 structures used in the Chen study and removed solvent, counterions and other small molecules located away from the ligand binding sites. Cofactors, metal ions and tightly bound water molecules were maintained in the active sites and the resulting pdb files were used as inputs for ICM (with cofactors added as separate objects to account for their bond orders). Assignment of protonation states, rotamers, and tautomeric states as well as hydrogen minimization was left to ICM, in alignment with the procedure followed in the Chen paper. The docking region was defined as a box including all the residues within 8.0 Å of the crystallographic ligand, and docking was performed using ICM 3.3 (Chen and co-authors used ICM 3.2). Using this setup ICM reproduced 56% of the structures within 2.0 Å of the corresponding X-ray structures. In order to determine whether an increased bias toward the correct solution in the docking setup could possibly enhance docking accuracy to the levels reported by Chen et al., we repeated the calculations three more times progressively reducing the size of the docking box around the crystallographic pose. The percentage of correct predictions went up to 57% when including all the residues within 5 Å of the crystallographic ligand, to 62% at 4 Å, and down to 53% at 3 Å. Additional bias was introduced by centering the initial placement of the ligand on the mass center of the crystallographic ligand and by aligning the initial orientation of the docking probe with that of the crystallographic ligand. Each of these settings was tested with docking boxes extending 8, 5, and 4 Å beyond the crystallographic ligand. In all cases the percentage of correctly docked ligands failed to exceed 61% (data not shown). We also performed docking with Glide 3.5 using the same protein preparation procedure, with the exception of hydrogen minimization, which was performed with Macromodel. Glide reproduced 59% of the structures within 2.0 Å of the corresponding X-ray structures. In order to shed some light on the discrepancy between our results and those reported by Chen and co-authors, we contacted Totrov and asked him to repeat the ICM calculations on the same test set. He repeated the docking study using our same initial setup and obtained almost exactly the same result (55% of structures reproduced within 2.0 Å). He then repeated the calculations with a smaller docking box, only including residues within 5.0 Å of the crystallographic ligand. The percentage of correct predictions went up to 59% in these conditions. He finally ran the dockings with the latter docking box using the latest version of the ICM code, which incorporates some recent improvements not yet available in the official release, and obtained 63% of correct predictions. The comparison between the results reported by Chen and co-authors and those obtained by Totrov and by us on the same test set is summarized in Table 2.

Table 2. Percentage of Correct Docking Poses Obtained on the Chen Test by Different Groups[f]

program (group)	% correct
ICM 3.2 (Chen et al.)	91
ICM 3.3 (Perola et al.)[a]	56
ICM 3.3 (Perola et al.)[b]	57
ICM 3.3 (Perola et al.)[c]	62
ICM 3.3 (Perola et al.)[d]	53
ICM 3.3 (Totrov)[a]	55
ICM 3.3 (Totrov)[b]	59
ICM 3.4 (Totrov)[b,e]	63
Glide 3.5 (Chen et al.)	63
Glide 3.5 (Perola et al.)	59

[a] Docking region defined by all residues within 8.0 Å of the X-ray ligand. [b] Docking region defined by all residues within 5.0 Å of the X-ray ligand. [c] Docking region defined by all residues within 4.0 Å of the X-ray ligand. [d] Docking region defined by all residues within 3.0 Å of the X-ray ligand. [e] Unreleased ICM upgrades used. [f] Correct: RMSD ≤ 2.0 Å relative to the corresponding X-ray structure.

It is clear from this account that both the ICM developer himself and we were unable to reproduce the results documented in the Chen paper. Based on the results summarized above, it can be reliably stated that the current versions of Glide and ICM exhibit comparable docking accuracy, as they both correctly reproduce about 60% of the structures in two broad and diverse sets of complexes when redocking ligands into their cognate receptor conformation. In order to gain a better understanding of the methodology

used in the Chen paper, we contacted the authors and asked for further details on the protocol they applied in the ICM calculations. It was explained to us that a certain degree of manual intervention was involved in the initial placement of the ligands, while the docking regions were defined using a semiautomated procedure. However, the reasons behind the large discrepancy between their results and ours remain unclear. Therefore, we thought it best to simply report these facts in a Letter, in the hopes that this would stimulate the Chen group and others to reinvestigate this curious matter. We continue to hope that the Chen results hold up, because if true, it would suggest a real scientific breakthrough that benefits the whole community; and if so, it is important for the whole community to understand the methodology so maximum value can be obtained. On the other hand, if upon further investigation the results turn out to be artifactual in nature, this is equally important to relate to the scientific community. Either way, further re-examination of the Chen study and associated results appears to be in order.

REFERENCES AND NOTES

(1) Chen, H.; Lyne, P. D.; Giordanetto, F.; Lovell, T.; Li, J. On Evaluating Molecular-Docking Methods for Pose Prediction and Enrichment Factors. *J. Chem. Inf. Model*. **2006**, *46*, 401−415.

(2) Perola, E.; Walters, W. P.; Charifson, P. S. A Detailed Comparison of Current Docking and Scoring Methods on Systems of Pharmaceutical Relevance. *Proteins* **2004**, *56*, 235−249.

CI600460H