Journal of Computer-Aided Molecular Design **18:** 333–344, 2004.

Evaluation and application of multiple scoring functions for a virtual screening experiment

Li Xing[a,*], Edward Hodgkin[a], Qian Liu[a] & David Sedlock[b]
[a]*Tripos, Inc., 1699 S. Hanley Road, St. Louis, MO 63144, USA;* [b]*Millennium Pharmaceuticals, Inc., 35 Landsdowne Street, Cambridge, MA 02139, USA*

Received 4 December 2003; accepted in revised form 22 July 2004

Key words: binding affinity, docking, protein-ligand interactions, scoring function, virtual screening

Summary

In order to identify novel chemical classes of factor Xa inhibitors, five scoring functions (FlexX, DOCK, GOLD, ChemScore and PMF) were engaged to evaluate the multiple docking poses generated by FlexX. The compound collection was composed of confirmed potent factor Xa inhibitors and a subset of the LeadQuest® screening compound library. Except for PMF the other four scoring functions succeeded in reproducing the crystal complex (PDB code: 1FAX). During virtual screening the highest hit rate (80%) was demonstrated by FlexX at an energy cutoff of −40 kJ/mol, which is about 40-fold over random screening (2.06%). Limited results suggest that presenting more poses of a single molecule to the scoring functions could deteriorate their enrichment factors. A series of promising scaffolds with favorable binding scores was retrieved from LeadQuest. Consensus scoring by pair-wise intersection failed to enrich the hit rate yielded by single scorings (i.e. FlexX). We note that reported successes of consensus scoring in hit rate enrichment could be artificial because their comparisons were based on a selected subset of single scoring and a markedly reduced subset of double or triple scoring. The findings presented in this report are based upon a single biological system and support further studies.

Introduction

As macromolecular crystallography becomes a standard technique used by the pharmaceutical and biotechnology industries in their drug discovery projects, the demands for virtual screening of large compound collections utilizing the crystal structures of the target proteins have increased substantially [1, 2]. This is due to the fact that modern combinatorial chemistry joined with automated purification processes rapidly and immensely augments corporate high throughput screening (HTS) libraries. As a result the cost of experimental screening of the entire corporate collection against drug targets grows to be massive, let alone the limited turnaround rate and the technical complexities. Virtual screening encompasses a variety of computational techniques that allow chemists to reduce a huge compound library, in silico or virtual, to a more manageable size [2].

Docking and scoring algorithms allow users to search for structures that have the right geometric and electronic features to fit the designated sites [3, 4]. The majority of published scoring functions have been developed in association with docking methods. However, the scoring functions used to rank compounds do not have to be linked to a docking method but rather are independent and should be able to estimate the binding affinities of the receptor-ligand complexes generated by any structure-based approaches, e.g., de novo design [5]. In many cases, the scoring functions can consistently predict the binding modes of ligands of nanomolar affinity; however, they perform poorly at predicting lower affinity binders. Although the goal of quickly and accurately predicting binding affinities for an arbitrary molecule has not yet been achieved, the docking and screening procedures can

*To whom correspondence should be addressed. Current address: Pfizer, Inc., BB4I, 700 Chesterfield Parkway West, Chesterfield, MO 63017, USA. Fax: +1-636-247-7607. E-mail: li.xing@pfizer.com

select small sets of likely lead candidates from large libraries of either commercially available or 'in-house' proprietary compounds.

The most popular docking programs, DOCK [6], GOLD [7] and FlexX [8], are widely distributed. In addition to the corresponding scoring functions associated with the aforementioned docking algorithms, DrugScore [9], ChemScore [10] and PMF [11] were developed separately. DOCK and GOLD are forcefield based methods that compute the enthalpy gain or loss upon binding, while FlexX and ChemScore use empirical terms to estimate binding free energy. The coefficients of each term are obtained by performing a regression on a set of receptor-ligand complexes of known experimental binding affinities. PMF and DrugScore are knowledge-based potentials. Following the idea of an 'inverse Boltzmann' distribution it is assumed that only favorable binding modes will fit to the distribution maxima of occurrence frequencies among interatomic contacts between particular atom pairs in experimentally determined structures. One advantage of these methods lies in that they only regard non-hydrogen contacts, therefore are independent from assumptions of protonation states.

In addition to the independent scoring functions a combination of multiple methods, or the so-called consensus scoring, has been applied to several protein targets using certain chemical compositions as experimental design [12–14]. The studies typically report more or less superior performance of the consensus scoring over single scoring [12–14]. More recent development of docking algorithms attempts to incorporate information about important characteristics of ligand-protein binding modes [15] and to include the solvent effect more efficiently [16, 17]. Significant effort has also been devoted to accommodate protein flexibility in computational drug design [19–20].

In this work we describe a real application of virtual screening as an effort to identify novel inhibitor classes of a serine protease target, factor Xa. More than 500 confirmed inhibitors were mixed with about 26,000 compounds serving as candidates from LeadQuest, a general screening library from Tripos. The resulted overall hit rate was 2.06%. The compounds were docked by FlexX. Five scoring functions, including FlexX, DOCK, GOLD, ChemScore and PMF that are available from the consensus scoring (CScore®) module [21] in Sybyl [22], were evaluated in a comparative manner in terms of their abilities to reproduce the crystallographic binding mode and to increase the hit rate. The investigation on consensus scoring led to the elucidation of general mis-treatment of hit rate comparisons between single and multiple scorings. A series of potential factor Xa inhibitors was recognized from this effort.

Methods

Structure preparation

The crystal structure of factor Xa was retrieved from the Brookhaven Protein Databank (PDB code: 1FAX) [23]. The ligand molecule DX-9065a was then extracted from the complex and corrected for atom and bond typing. In particular, for the amidino group the carbon atom is assigned *C.cat* Sybyl atom type [22] and the two nitrogens *N.pl3*. After filling the valences DX-9065a was subsequently minimized by the Tripos forcefield to relieve the geometric restraints imposed by crystal packing. Due to the lack of structural water molecules around the binding site, all crystallographic waters were removed. The entire protease structure was used as the docking protein and the residues within 6.5 Å to the bound ligand were included in defining the active binding site for docking.

The recognition interaction with factor Xa is the ionic pairing with Asp189 at the bottom of the S1 site [23]. The S3 site is the aryl binding site with strong hydrophobic interactions with the side chains of Tyr99, Trp215 and Phe174. The S4 site contains a cation hole created by the backbone carbonyl oxygens of Glu97, Thr98, Ile175 and the carboxylic acid side chain of Glu97. The overall geometry of the inhibitor is characterized by an approximately 90° kink in the middle so as to direct one end of the molecule into S1 and the opposite end to S3/S4. The complete length of the S1 pocket is about 8 Å and 11.5 Å for S3/S4. The S4 site is characterized by an 8 Å width at the entrance, narrowing as the binding site progresses to the terminus.

Statistical analysis

The natural ligand DX-9065a was docked into the factor Xa active site. In order to provide a comprehensive pool of poses for subsequent scoring exercises, all the solutions yielded by FlexX were saved. This resulted in 402 plausible binding poses. These poses were then compared with the co-crystallized ligand and the average root mean squared distances (RMSD) of the corresponding heavy atoms were calculated.

A good scoring function should produce energies that correlate with the relative RMSDs of various poses to the crystallographic ligand. There is no logic in anticipating a linear relationship between the scores and RMSDs, however, the monotonicity is important. Therefore Spearman's rank correlation coefficient (r_S) was applied for this purpose. In the absence of ties, r_S is given by

$$r_S = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the rank difference for the ith observation in the two sequences. For example, if pose i is ranked number 10 by RMSD and number 15 by FlexX score, the rank difference d_i between the two rankings is 5. The summation is over all poses. For cases where there are ties, the ties are broken by assigning to each tied observation the mean rank of the rank positions for which it is tied. Two identical sequences yield an r_S of 1, while the complete reverse sequences would result in an r_S of -1.

Compound collection

LeadQuest is the screening library commercialized by Tripos, Inc. A total of 25,964 compounds were randomly selected from LeadQuest and served as candidate factor Xa inhibitors. Confirmed compounds were chosen from the Millennium Pharmaceuticals factor Xa project. These included 549 active compounds of multiple chemical series whose IC_{50}'s were equal to or less than 50 μM, and an additional 144 inactive compounds. The IC_{50}'s ranged from sub-nanomolar to micromolar affinity. Combining with the LeadQuest selection, the final data set for virtual screening consisted of 26657 compounds out of which 549 were confirmed actives, corresponding to an overall hit rate of 2.06%. All of the molecules were converted to three-dimensional structures using CONCORD in Sybyl [22] and saved in mol2 format. Protonation states were adjusted to reflect structures most predominant at physiological pH. This was done by an SPL (Sybyl Programming Language) script to add or remove protons and to correct the atom types when necessary according to the calculated pK_a values [24, 25]. The most prominent functions asking for protonation corrections are benzamidine and its bioisosteric replacements.

Docking and scoring functions

All docking calculations were performed by FlexX (version 1.9.0) in Sybyl 6.8. In order to eliminate the strong bias imposed by the salt bridge interactions with the enzyme, 'assign formal charges' was turned off. No partial charges were precomputed. In order to account for stereochemistry the sp^3 nitrogen and the R/S carbon centers were allowed to be modified during FlexX runs. The rest of the FlexX parameters were set to default. For each compound the top ranked 30 poses were saved. This pool of solutions provided a possibility for the other scoring functions to reach a different ranking list for the same molecule, thus choose a pose different from FlexX as the best solution.

When the 'assign formal charge' option was turned on, the quality of the docking poses declined considerably as indicated by a lowest RMSD of 1.76 Å. A general shift toward the higher end of the RMSD distribution further ratified the observation. Visual inspection of the poses revealed that the amidino did not form bidentate interactions employing both amino functions with Asp189. Rather one amino of the amidino, being assigned half of the unit positive charge, was pulled toward the acidic sidechain of Asp189. This suggests an inefficient treatment of the charge interaction in FlexX, especially when formal charge and/or salt bridges are involved.

Five storing functions were involved in ranking the FlexX docking results and included FlexX, DOCK, GOLD, ChemScore and PMF. These are available from the CScore module within Sybyl. DOCK score uses steric and electrostatic terms based on the AMBER forcefield [6]. In analogy the GOLD evaluation function is a sum of hydrogen bonding stabilization energy, internal van der Waals energy for the ligand conformer, and a pairwise dispersion potential between ligand and protein to describe the hydrophobic energy of binding [7]. The FlexX scoring function considers the number of rotatable bonds in the ligand, hydrogen bond interaction (including atom types and geometry), ion pairing, aromatic interactions, and the lipophilic contact energy [8]. The PMF scoring function exploits structural information of known protein-ligand complexes extracted from the Brookhaven Protein Databank and converts it into distance-dependent Helmholtz free energies [11]. The magnitude and sign of each interaction potential is based on the atom types of the interacting pair and the intervening distance. ChemScore consists of a term to estimate lipophilic contact energy, a metal-ligand

binding contribution, an empirical form for hydrogen bonds and a penalty for ligand flexibility [10]. The coefficients of each term are obtained using a regression algorithm based on a set of receptor-ligand complexes.

DOCK and GOLD scoring functions used here are not from the original programs but reproduced by Tripos based on literature descriptions [21]. Due to different implementation details from DOCK and GOLD, the outputs from Sybyl are henceforth referred to as D-SCORE and G-SCORE [21].

Results and discussion

Reproducing geometries of crystal complexes

Factor Xa locates at the convergence between the extrinsic and intrinsic pathways of the blood coagulation cascade and represents a drug target being actively pursued by the pharmaceutical industry. Its binding cavity consists of a deep and straight S1 site and relatively broad and open S3/S4 sites. The putative binding motif is the ionic interaction at the bottom of the S1 site, which for most of the potent inhibitors involves a salt bridge between a protonated basic amine and the sidechain of Asp189.

Of all the docking poses created by FlexX the lowest RMSD match to the co-crystallized ligand, DX-9065a, was 1.26 Å. Although not a grand achievement for DX-9065a of molecular weight 455, the overall binding mode as well as the key interaction motifs (i.e. the salt bridge with Asp189 and the positioning of 1-aminomethylpyrrole into the electronegative 'cation hole') were unequivocally reproduced. Four representative FlexX poses are delineated in Figure 1 in reference to DX-9065a. The best pose (RMSD 1.26 Å, magenta) matched the benzamidine moiety unequivocally. The two worse ones, RMSD 1.60 Å in green and 1.98 Å in blue, displayed deviations in the S1 and S2/3 pockets. The details of the scoring energies are tabulated in Table 1. Except for PMF, the four other scoring functions were able to promote the poses that were close to the crystal structure. Furthermore, Spearman's rank correlation coefficients (r_S) between scores and the RMSDs to the co-crystallized ligand provided information about the performances of the scoring functions in terms of discriminating good from bad solutions. Again, except for PMF the other four scoring functions yielded good correlations with the RMSDs, among which D-SCORE and G-SCORE slightly outperformed FlexX and ChemScore. G-SCORE demonstrated by a small margin the highest Spearman's rank order coefficient of 0.73. In contrast, the negative rank correlation of PMF with RMSDs was rather disappointing. Given that PMF was parameterized using existing structural information, its poor ranking of poses was unexpected.

A more accurate way to identify the statistical significance of the scores is to bin the scoring energies and calculate the average RMSDs of all the solutions in the particular bins. For a perfect scoring function the average RMSDs should increase monotonically with increasing energies within the proximity of the native binding site [21]. Because of the poor performance of PMF only FlexX, D-SCORE, G-SCORE and ChemScore are plotted in Figure 2 for comparison. The error bars are standard deviations of RMSDs. All four scores mimic monotonic distributions to some degree, given that different means of energy binning can generate slightly modified representations of the bar chart. FlexX was able to assign the lowest scores to the top best solutions. However, for higher RMSD poses its ability for discrimination became less satisfying, as indicated by the bumpy distribution of RMSDs. A similar pattern was observed for G-SCORE, which separates the bad (in high energy bins) from good solutions (in low energy bins) by a large difference in average RMSDs but fluctuates in the intermediate energy ranges. It appeared that G-SCORE ranked the best solutions in its rather next-to-best energy bins, since the RMSD of -290 to -260 kJ/mol (1.52 Å) is lower than that of -320 to -290 kJ/mol (1.93 Å). D-SCORE resembled G-SCORE in that regard. The three lowest energy bins exhibited a decreasing trend in average RMSDs. Nevertheless, for the poorer poses of greater than -180 kJ/mol, D-SCORE demonstrated a good monotonic relationship with average RMSDs. The low and somewhat constant standard deviations suggested that right poses converged tightly across the four scoring functions. Especially at the low energy end the instances of false positives were very rare.

It is worthwhile to point out that the FlexX docking algorithm is very sensitive to the protonation state of the molecule. When the amidino function is kept in its neutral representation, FlexX generated slightly higher quality poses with a lowest RMSD of 1.12 Å to the crystal ligand. However, this treatment is not considered physiologically relevant given the high experimental pK_a value of benzamidine (around 12). Therefore, even though the neutral form was favored by FlexX, it would create inappropriate molecular structures for the other scoring functions to

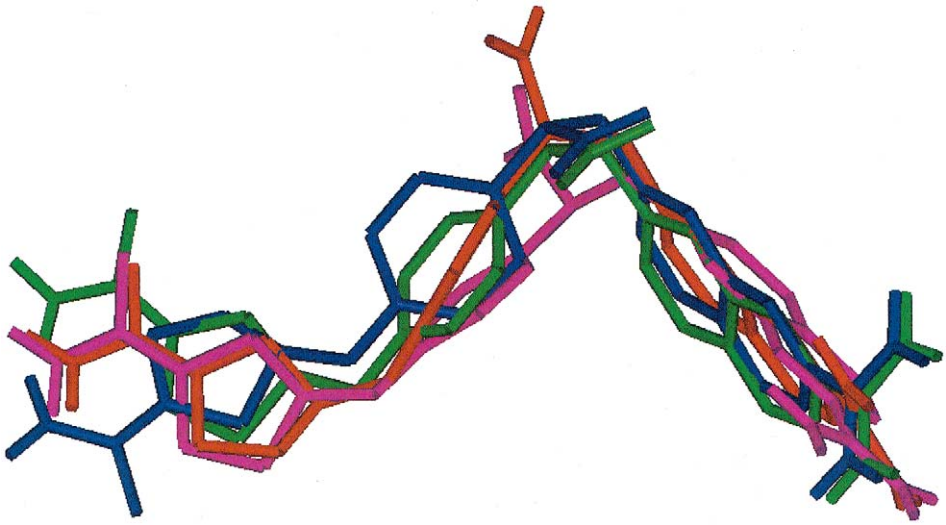


Figure 1. Representative poses generated by FlexX in reference to co-crystallized DX-9065a (red) in 1FAX. Magenta: pose of 1.26 Å RMSD to DX-9065a; green: 1.60 Å; blue: 1.98 Å.

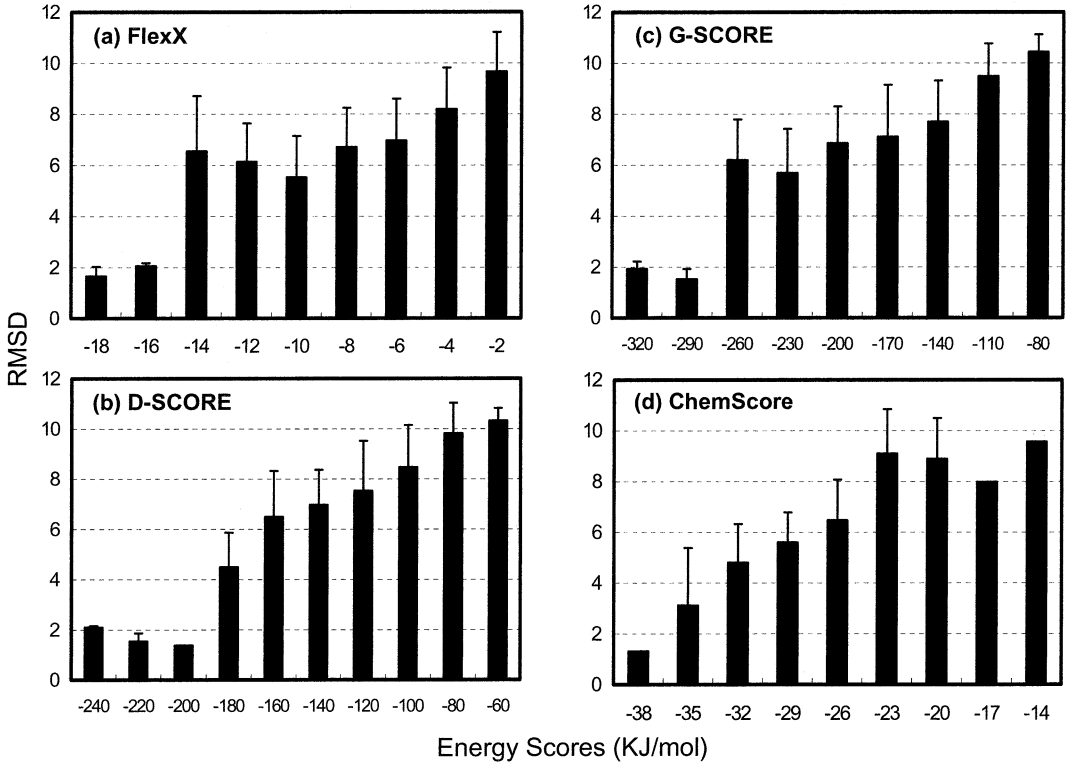


Figure 2. Distributions of average RMSDs and their standard deviations between docked and native ligand (1FAX). The error bars are standard deviations. The energy labels on the x-axis correspond to the lower end of each energy bin.

Table 1. Summary of rankings and energies by five different scoring functions of the 402 docking poses of the crystal complex 1FAX generated by FlexX. The energies are in kJ/mol.

	FlexX	D-SCORE	G-SCORE	ChemScore	PMF
Min. energy	−17.83	−234.81	−318.28	−37.39	−93.18
Max. energy	−1.29	−41.78	−55.07	−13.96	23.19
Best RMSD energy	−13.00	−217.21	−285.2	−32.2	−25.94
Rank of Best RMSD	11	4	5	6	334
r_S^*	0.63	0.72	0.73	0.62	−0.16

*r_S: Spearman's rank correlation coefficient.

evaluate. For the force field based scoring functions such as D-SCORE and G-SCORE, this should have caused deficiencies in energy calculations due to a strong dependence on the atomic representation of the molecules.

Virtual screening

FlexX was employed as the docking engine for the LeadQuest compounds and the known factor Xa inhibitors. For each compound the top ranked 30 poses generated by FlexX were saved. As such each of the four alternative scoring functions was given the alternative to pick either the same or a different best solution per compound, and its score value was subsequently used to compare to other compounds for ranking. Therefore, even though the best-scored pose was always used to represent the plausible binding of a particular compound, in many cases that pose varied from one to another as dictated by separate scoring functions. It was attempted to use the FlexX docking algorithm for generation of potential binding modes. Subsequent selection of a specific binding mode for each compound was then subject to individual scoring algorithms.

The hit rate enrichment is presented in Figure 3. Assuming that the compound screening process followed the rank sequence defined by a specific scoring function, the earlier the active compounds are screened, the fewer compounds need to be tested in order to capture most of the actives. All of the scoring functions demonstrated a certain degree of enrichment over random screening, which is the diagonal in Figure 3. The highest enrichment was achieved by FlexX. D-SCORE and ChemScore followed, with D-SCORE slightly out-performing ChemScore. PMF had a very impressive initial enrichment that surpassed D-SCORE and ChemScore. Unfortunately, it levelled off quickly and became the worst among the five scores toward the end of the screening process. G-SCORE displayed a uniform curve-up over the entire compound collection but the hit rate enrichment is comparatively moderate.

Although analyses were conducted on all scoring functions, two of these functions stood out compared to the others, FlexX and D-SCORE. The cumulative hit rate (i.e. the percent actives as functions of energy cutoff) is plotted in Figure 4. Both FlexX and D-SCORE ultimately reached the overall hit rate of 2.06% at the high energy end. FlexX achieved the highest hit rate of 80% at −40 kJ/mol, more than double the highest peak of D-SCORE (37%) at −210 kJ/mol. At the low energy end we observed another interesting difference. For FlexX the initial two peaks are of similar height, suggesting that the energy differences at −45 kJ/mol and −40 kJ/mol are basically indistinguishable for recognizing active compounds in the present experimental setting. However, for D-SCORE the second peak (37%) was much higher than the first peak (25%), demonstrating that the lowest D-SCOREs do not seize the strongest discriminating power. As a matter of fact, the hit rate at −235 kJ/mol is approximately the same as the −185 kJ/mol cutoff. Interestingly, similar oscillation was observed for D-SCORE in ranking multiple solutions of the same molecule, as discussed in the previous section. These persistent patterns inferred that there existed some weakness for D-SCORE to treat extremely tight binders. One possible reason could be the insufficient scaling of van der Waals interactions that overly penalizes the close contacts between the tight-fitting molecules and/or conformers and the protein.

One further piece of information that Figure 4 conveyed was that when used for virtual screening, an energy cutoff of −40 kJ/mol from FlexX conferred about a 39-fold increase in hit rate over random screening

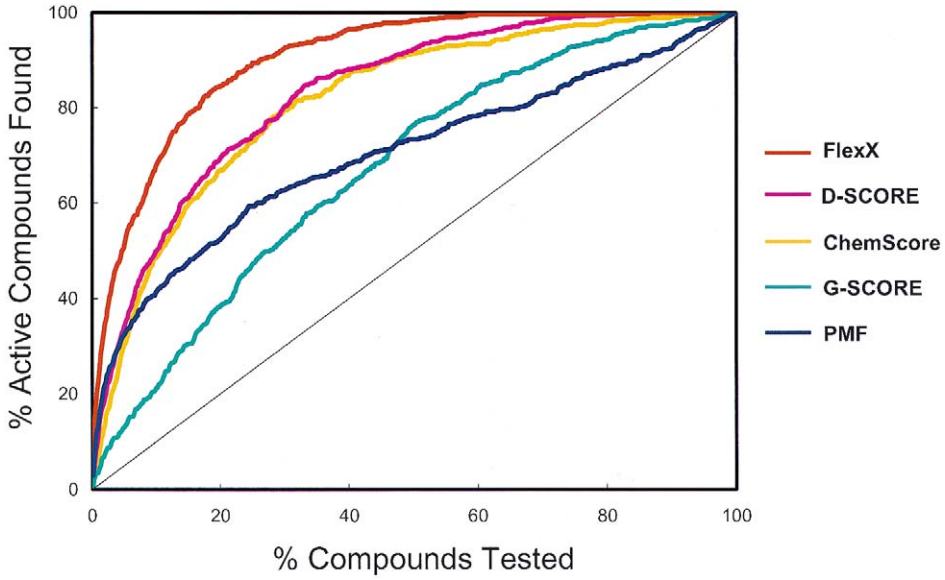


Figure 3. Hit rate enrichment for five scoring functions. The diagonal is for random screening.

(80.0% over 2.06%). When using D-SCORE, an energy cutoff of −210 kJ/mol was the best. This only served an 18-fold increase in hit rate (36.7% over 2.06%) based on this study.

In Figure 5 the histogram of the actives and inactives (including confirmed inactives and unknown LeadQuest compounds) with respect to energy is displayed for FlexX (panel a) and D-SCORE (panel b). FlexX separated the active compounds from inactives to a greater extent than D-SCORE. We observed that in order to capture more than half of the active compounds, the energy cutoffs should be −25 kJ/mol for FlexX and −150 kJ/mol for D-SCORE. Relating these back to Figure 4, the corresponding hit rate enhancements at those energy cutoffs were sixfold for FlexX and threefold for D-SCORE.

It is arguable that the number of poses saved (30 in this study) for each molecule is limited. Ideally, it would be desirable to save all possible poses in order to avoid any bias presented to the subsequent scorings. The major concern is the artificial preference to FlexX score introduced by the same FlexX docking scheme. Whereas it was not practical to save everything on the entire data set given its size, we have selected a subset of about one percent of the total molecules to assess the effect of the number of poses. Specifically, 290 inactive molecules were randomly mixed with six active compounds, maintaining approximately the same overall hit rate. For each molecule 500 poses were requested. For the vast majority of the compounds (90%) FlexX generated less than 500 poses, with a minimum number of 40 for a particular compound. The hit rates of different scoring functions were then calculated by the same protocol. The results unequivocally demonstrated that FlexX retained its leading position in hit rate enrichment among the other scores, repudiating the skepticism that its own docking engine may have run in favor of the FlexX score. Another interesting finding was that most of the scoring functions, including D-SCORE, ChemScore and G-SCORE, degraded to a small extent in performance (rather than improved) when more docking poses were presented. PMF performed more or less the same. This may seem counter-intuitive, since it is generally expected that more poses would provide more chances for the other scores to 'win'. However, that statement is based on the assumption that the scoring functions are sensitive enough to discriminate the 'bad' from 'good' poses when offered adequate opportunities. On the other hand, a large number of poses could become a challenge for the scoring functions to successfully separate out the best candidate for a single molecule, which leads to the deficiency when subsequently compared to different molecules. To that end our results showed that a larger number of poses could attenuate the ability of same scoring functions to extract the true binding mode of an individual compound and assign it a sensible score for comparison with other compounds.

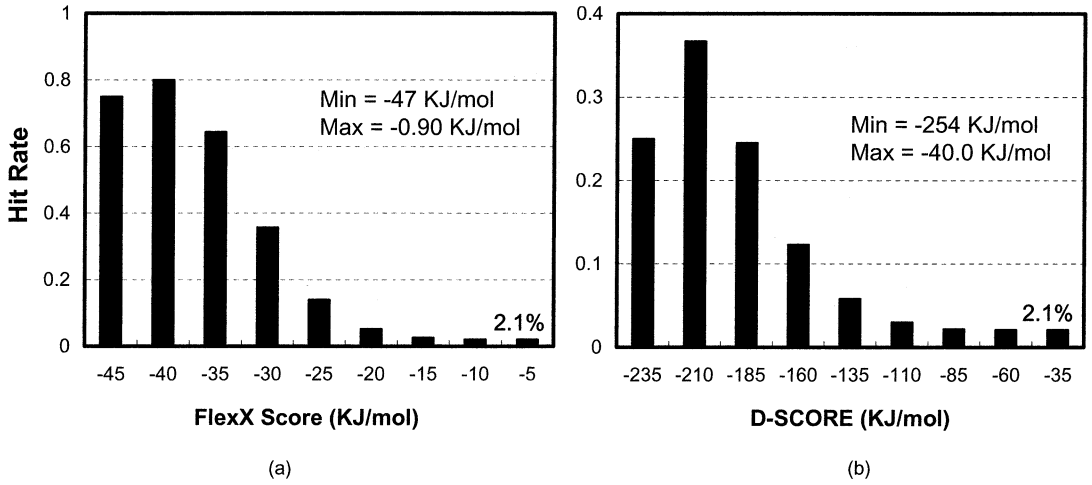


Figure 4. Hit rate at different (a) FlexX Score and (b) D-SCORE cutoffs. Both plots plateau at the overall hit rate of 2.06% at the high energy end.

Since FlexX poses resided in the local energy minima of the FlexX potential, we further minimized this subset of docked ligands in the protein environment using the Tripos forcefield. It would be ideal to relax the ligands in the respective scoring potentials; nonetheless the current practice should optimize the poses into a common frame defined by an independent forcefield engine. The poses of each molecule were scored and the best one used for ranking with other molecules. It was noted that the minimization had effectively changed the ranking of the poses as well as the absolute scores of each molecule. As a result the enrichment profile for every scoring function displayed the same fluctuation. However, the overall performances remained unchanged, with FlexX leading by an evident margin. Last but not least, this type of calculation was extremely CPU intensive (for this small set of 296 molecules it took more than eight days on an SGI R14000 processor), rendering it unsuitable for virtual screening.

Consensus scoring

In order to evaluate the usefulness of combining the scoring functions in the paradigm of consensus scoring application, the cross terms were compared and are tabulated in Table 2. Each cell contains two columns. The C_{tot} column displays the percent of compounds in common between the ranking lists generated by the two scoring functions. The C_{act} column shows the percentage of actives identified by the consensus of the two scoring functions out of the total actives. The upper row of each cell corresponds to the top 5% of the rank list while the lower row is for the top 10% of the rank list.

By definition for all the diagonal cells the C_{tot}'s are 100%, and the C_{act}'s are indeed the quantitative evaluations of the individual scoring functions themselves. We observed here as well that FlexX displayed scoring superiority, defining 51% of the total active compounds ranked in the top 5% list. In second place was D-SCORE that revealed about one-third of the actives in the top 5% ranking. Reaffirming what was observed in Figure 3 PMF exhibited a very impressive capability of finding active compounds in the top ranks, specifically 32.2% in the top 5% rank list and 41.5% in the top 10%. Furthermore, comparing the C_{act}'s of the two ranking lists we observed that doubling the number of screening compounds by extending into less favored energetics only contributed to an extra 10–20% increase in the number of active compounds identified.

For the cross terms D-SCORE and G-SCORE share the most in common, specifically 56.1% in the top 5% ranked compounds. This was expected given the fact that both algorithms estimate the enthalpic contribution upon binding using a very similar treatment of energy terms. The scoring function sharing the next most common compounds with D-SCORE and G-SCORE was ChemScore. FlexX did not overlap much of its rank list with the other functions. To that end PMF shared the least with all the other scoring functions across the board, possibly a reflection of its unique knowledge-based algorithm that is different from the rest of the scoring functions be-

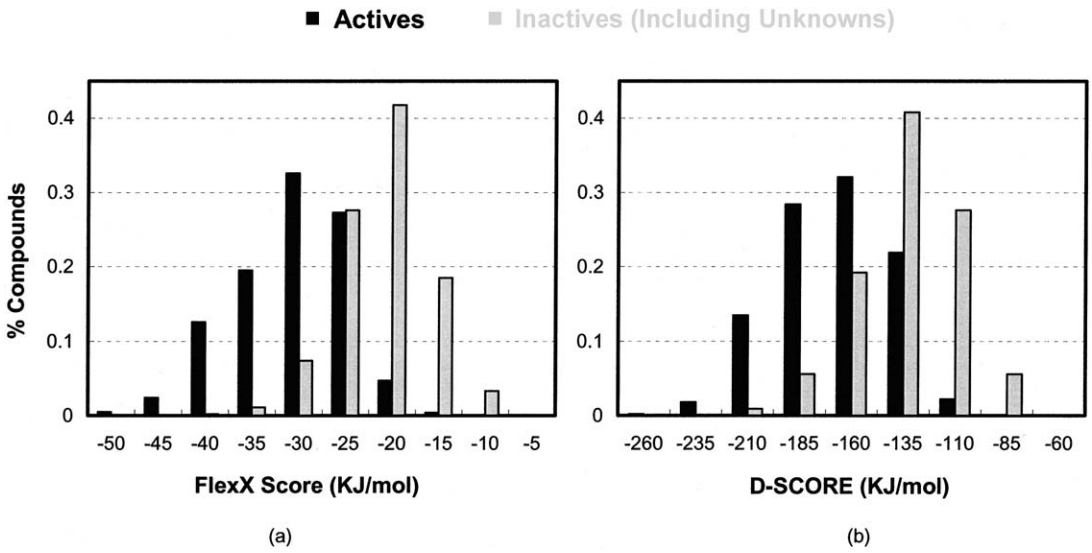


Figure 5. Histograms of actives and inactives (including unknowns) in terms of (a) FlexX Score and (b) D-SCORE. The energy labels on the x-axis correspond to the lower end of each energy bin.

ing considered. Overall the use of multiple scoring functions for consensus scoring caused dramatic loss of compounds, maintaining only one-third or less of the starting compounds in most of the cases. Similar observations have been made by previous evaluations [12, 14]. This significant loss of compounds limits the hit rate that most of the consensus scorings can achieve.

Because the number of common compounds within the top 5% or 10% rank lists might vary for different cross terms and the magnitude of C_{act} was dependent on the number of such common compounds, judging the performance of the cross terms based on hit rate by literally comparing C_{act} was not a fair approach. To address this we compared the hit rates achieved by the three most promising binary combinations, FlexX&D-SCORE, FlexX&PMF and FlexX&ChemScore, along with FlexX and PMF individually (Figure 6). According to Figure 3 FlexX was the best single scorer, and PMF was chosen because of its second-to-best performance in the early screening. As illustrated by Figure 6, albeit more complicated, the combined scoring schemes failed to compete with FlexX. A few data points would further clarify the differences: at 5% screening of entire compounds, the percent of active compounds found are 50.6% by FlexX, 41.3% for FlexX&D-SCORE, and 43.8% for FlexX&ChemScore, and the percent of actives were 84.5%, 82.6% and 77.8%, respectively, when 20% of the total compounds were screened. This was rather surprising given the general belief that consensus scoring usually yields a much higher hit rate than the individual scoring functions. However, it is worth noting that the reported higher hit rates were relative to the reduced number of compounds due to the significant compound loss caused by the consensus procedure, while the hit rates for individual scores referred to the full compound sets [12, 14]. Another ambiguity was to report the success of consensus scoring without noting the compound loss, although it presumably exists in the double or even triple scoring treatments [13]. In the real HTS setting it is essential that the hit rates are compared at the same screening stage, or in other words, using the same number of compounds screened as represented in Figure 6. The practical effect of reducing the compound set by consensus scoring could be regarded as moving toward the earlier stages of the screening process. Hence for fairness the hit rate at the same early stage needs to be compared with single scoring. Comparing hit rates for compound lists of different sizes will inevitably lead to misleading or even erroneous conclusions. Consensus scoring is not useful if it cannot effectively eliminate the false positives submitted by single scorings.

Several definitions of consensus scoring have been proposed and applied, leading to different end results. Our current method is the intersection approach, which basically takes the intersection of the top N% of each compound list sorted by individual scoring functions. The same method has been adopted in other

Table 2. Consensus scoring with pairs of five published scoring functions. C_{tot} is the percent of total compounds in common in the ranking list by the two scoring functions. C_{act} shows the percentage of actives identified by the consensus of the two scoring functions out of the total actives. The upper row of each cell corresponds to the top 5% of the rank lists while the lower row is for the top 10% of the rank lists.

	ChemScore		D-SCORE		FlexX		G-SCORE		PMF	
	C_{tot}	C_{act}	C_{tot}	C_{act}	C_{tot}	C_{act}	C_{tot}	C_{act}	C_{tot}	C_{act}
	100	30.2								
ChemScore	100	48.6								
	35.8	15.8	100	34.1						
D-SCORE	45.0	31.0	100	49.9						
	30.2	19.5	24.5	20.0	100	51.0				
FlexX	36.2	36.1	35.4	38.8	100	67.8				
	34.8	12.9	56.1	11.7	14.9	7.8	100	12.9		
G-SCORE	43.8	16.9	61.5	19.5	23.5	16.6	100	21.1		
	17.3	10.7	12.2	10.2	16.2	14.9	7.42	3.27	100	32.2
PMF	25.8	21.7	16.0	18.2	22.4	27.5	13.2	7.28	100	41.5

evaluation studies [12–14]. In the CScore module in Sybyl each scoring function casts one vote in favor of a protein-ligand complex if its score falls into a certain range of values. The CScore is the total number of votes received [21]. A recent study uses the average rank received from multiple scoring functions as the final rank of the compound [26]. In the current study we chose the intersection method, which efficiently ranks this large data set without creating large numbers of ties.

LeadQuest hits

The best ranked LeadQuest compounds were visually inspected in order to confirm their binding interactions. The very first compound selected by FlexX, 1506-07974, exemplified the binding motifs critical for interaction with factor Xa with a favorable FlexX score of −41.6 kJ/mol. The phenylpyrrazole of the molecule penetrates deeply into the S1 pocket. Next to it the urea moiety serves as the turning point for the long and linear portion of the molecule to line up with the surface residues of the S3/S4 pockets. A close to cis arrangement was observed for one of the C-N bonds of the urea. One might argue that this is an artifact created by the docking program. However, a quick search of the Cambridge Crystal Database (CSD) disclosed several instances of cis ureas, especially when the urea nitrogen was engaged in certain hydrogen bond interactions. Furthermore, the cis ureas were more frequently observed in ligand molecules in crystal complexes with proteins [27, 28]. The thiazole of 1506-07974 interacts with the hydrophobic S3 pocket. Upon a cursory glance of the top ranked docking list, several LeadQuest compounds of different chemotypes displayed appealing binding to factor Xa. One general feature similar to 1506-07974 is that they do not bear a basic amidino function as seen in most of the inhibitors targeting various stages of the blood coagulation cascade. This could be plainly interpreted as missing the recognition salt bridge with Asp189. However, vast medicinal chemistry efforts have attempted to eliminate the basic benzamidine because of the poor bioavailability and lack of oral efficacy for compounds consisting of this moiety [29, 30]. Consequently, it would be exceedingly valuable to discover amenable surrogates for benzamidine, especially in a fast and cost-effective way such as virtual screening. Candidates identified from this study could potentially displace the charge interaction if the other aspects of the S1 pocket are sufficiently complemented by the inhibitor. As a matter of fact two different crystal forms of trypsin, another proteinase closely related to factor Xa, with a choloronaphthyl piperidinylpyridine inhibitor have been reported under different pH conditions [31]. In one form under pH 7 the pyridine moiety binds with Asp189 in the S1 pocket. However, inverse binding was observed at elevated pH 8 with the hydrophobic chloronaphthyl group directed toward the S1 pocket. These observations support the notion that a basic moiety is not an absolute requirement for the inhibition of trypsin-like proteinases that possess an acidic amino acid in the specificity pocket.

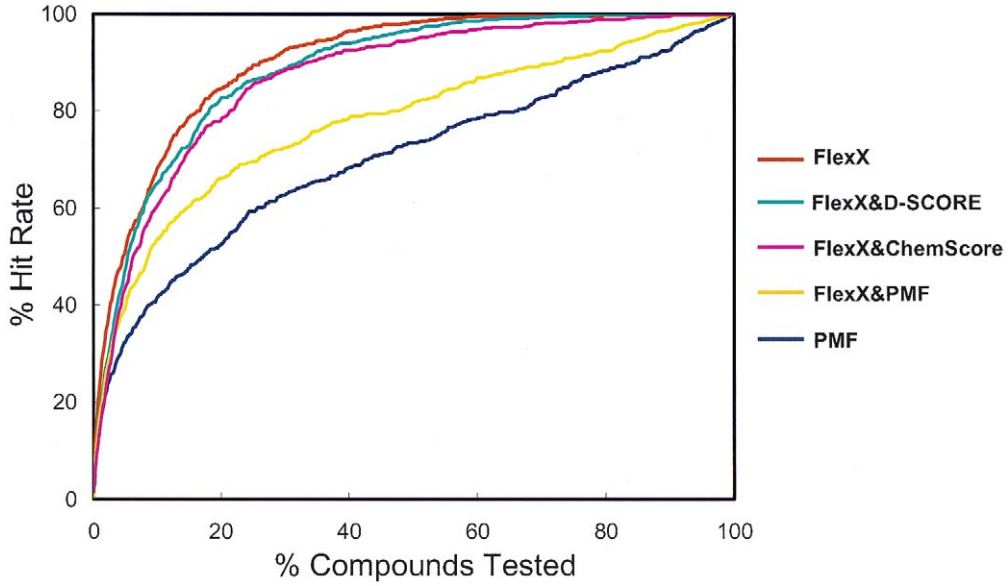
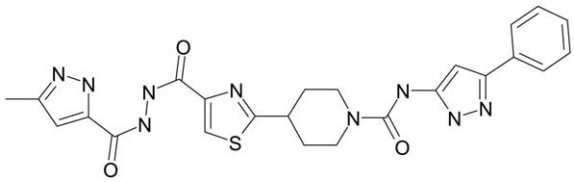


Figure 6. Hit rate comparisons for two-term consensus scorings as well as single scorings FlexX and PMF. In order of decreasing hit rate are: FlexX, FlexX&D-SCORE, FlexX&PMF, FlexX&ChemScore and PMF.



1506-07974

Figure 7. LeadQuest compound 1506-07974.

Conclusions

In this investigation we reported that four out of the five scoring functions, specifically FlexX, D-SCORE, G-SCORE and ChemScore, performed adequately in terms of ranking the docking poses according to the native ligand in the crystal structure. PMF failed to select any sensible docking poses. This was very much unexpected since PMF was parameterized on crystal complexes.

Our goal was to identify novel chemical classes of factor Xa inhibitors in the LeadQuest compound library. After docking and scoring by different methods FlexX exhibited the highest hit rate enrichment in the entire screening process, followed by D-SCORE and ChemScore. Hit rate enrichments by G-SCORE and PMF were comparatively moderate. The peak hit rate of 80% by FlexX was achieved at −40 kJ/mol energy cutoff, an enrichment factor of almost 40-fold over random screening. Preliminary results using a small data set indicated that when more poses per molecule were provided, the performance of most scoring functions degraded, including D-SCORE, ChemScore and G-Score. A series of potential factor Xa inhibitors was identified from LeadQuest with plausible capability of replacing the benzamidine moiety, leading to compounds of improved pharmacokinetic properties.

Consensus scoring by two-term combinations was evaluated. We noted that the reported successes of consensus scoring based upon the hit rate comparisons between single scoring of a specific compound set and consensus scoring of the significantly reduced set are unfair and misleading. In order to assess the true performance of any scoring scheme the hit rates have to be judged at the same stage of the screening process, otherwise using the hit rate from a smaller data set (earlier screening stage) to compare with a larger one (later screening stage) will inevitably bias the results. Based on our study none of the double scorings were able to compete with the best single

scoring FlexX result. If the same evaluation scheme was applied, the reported superiority of the consensus scoring could have vanished due to the significant loss of compounds by the consensus procedure.

We realize that the discovery we made in this work is based on a single biological target. In order to be generalized to broader applications, similar results need to be duplicated for multiple biological targets using heterogeneous compound compositions. We suggest that for productive virtual screening in certain biological settings the relative effectiveness of various single and consensus scoring systems shall be established from a case study before the most efficient scoring method is applied.

Acknowledgement

We thank Dr. Robert C. Clark and the reviewers for valuable suggestions on the manuscript.

References

1. Shoichet, B.K. and Bussiere, D.E., Curr. Opin. Drug Discov. Dev., 3 (2000) 408.
2. Abagyan, R. and Totrov, M., Curr. Opin. Chem. Biol., 5 (2001) 375.
3. Kubinyi, H., Curr. Opin. Drug Discov. Dev., 1 (1998) 4.
4. Oprea, T.I. and Marshall, G.R. Receptor-based prediction of binding affinities; Knegtel, R.M.A. and Grootenhuis, P.D.J., Binding affinities and non-bonded interacton energies; Weber, I.T. and Harrison, R.W. Molecular mechanics calculations on protein-ligand complexes, In: Kubinyi, H., Folkers, G. and Martin, Y.C. (Eds.), 3D QSAR in Drug Design, Vol. 2, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.
5. Müller, K. De novo design. In: Anderson P.S., Kenyon G.L. and Marshall G.R. (Eds.) Perspect. Drug Discov. Design, Vol. 3. ESCOM, Leiden, The Netherlands, 1995.
6. Kuntz, I.D., Blaney, J.M., Oatley, S.J., Langridge, R. and Ferrin, T.E., J. Mol. Biol., 161 (1982) 269.
7. Jones, G., Willett, P., Glen, R.C., Leach, A.R. and Taylor, R., J. Mol. Biol., 267 (1997) 727.
8. Rarey, M., Kramer, B., Lengauer, T. and Klebe, G., J. Mol. Biol., 261 (1996) 470.
9. Gohlke, H., Hendlich, M. and Klebe, G., Perspect. Drug Discov. Des., 20 (2000) 115.
10. Eldridge, M.D., Murray, C.W., Auton, T.R., Paolini, G.V. and Mee, R.P., J. Comput.-Aided Mol. Des., 11 (1997) 425.
11. Muegge, I. and Martin, Y., J. Med. Chem., 42 (1999) 791.
12. Charifson P.S., Corkery, J.J., Murcko, M.A. and Walters, W.P., J. Med. Chem., 42 (1999) 5100.
13. Bissantz, C., Folkers, G. and Rognan, D., J. Med. Chem., 43 (2000) 4759.
14. Stahl, M. and Rarey, M., J. Med. Chem., 44 (2001) 1035.
15. Hindle, S.A., Rarey, M., Buning, C. and Lengauer, T., J. Comput.-Aided Mol. Des., 16 (2002) 129.
16. Rarey, M., Kramer, B. and Lengauer, T., Proteins, 34 (1999) 17.
17. Majeux, N. and Scarsi, M., Proteins, 42 (2001) 256.
18. Broughton, H.B., J. Mol. Graph. Model., 18 (2000) 247.
19. Carlson, H.A. and McCammon, J.A., Mol. Pharmacol., 57 (2000) 213.
20. Clauben, H., Buning, C., Rarey, M. and Lengauer, T., J. Mol. Biol., 308 (2001) 377.
21. Clark, R.D., Strizhev, A., Leonard, J.M., Blake, J.F. and Matthew, J.B., J. Mol. Graph. Model., 20 (2002) 281.
22. Sybyl is a product of Tripos, Inc., St. Louis, MO; www.tripos.com.
23. ReceBrandstetter, H., Kuhne, A., Bode, W., Huber, R., von der Saal, W., Wirthensohn, K. and Engh, R.A., J. Biol. Chem., 271 (1996) 29988.
24. Xing, L. and Glen, R.C., J. Chem. Inf. Comput. Sci., 42 (2002) 796.
25. Xing, L., Glen, R.C. and Clark, R.D., J. Chem. Inf. Comput. Sci., 43 (2003) 870.
26. Wang, R., Lu, Y. and Wang, S., J. Med. Chem., 46 (2003) 2287.
27. Ikuta, M., Kamata, K., Fukasawa, K., Honma, T., Machida, T., Hirai, H., Suzuki-Takahashi, I., Hayama, T. and Nishimura, S., J. Biol. Chem., 276 (2001) 27548.
28. Internal crystallographic data.
29. Stürzebecher, J., Prasa, D., Hauptmann, J., Vieweg, H. and Wikström, P., J. Med. Chem., 40 (1997) 3091.
30. Quan, M.L., Liauw, A.Y., Ellis, C.D., Pruitt, J.R., Carini, D.J., Bostrom, L.L., Huang, P.P., Harrison, K., Knabb, R.M., Thoolen, M.J., Wong, P.C. and Wexler, R.R., J. Med. Chem., 42 (1999) 2752.
31. Klebe, G., J. Mol. Med., 78 (2000) 269.