ISSN 0233–7657. Biopolymers and Cell. 2013. Vol. 29. N 5. P. 418–423 doi: 10.7124/bc.000833

BIOINFORMATICS

UDC 577.322.23

Geometric filters for protein–ligand complexes based on phenomenological molecular models

O. O. Sudakov[1], O. M. Balinskyi[1], M. O. Platonov[2], D. B. Kovalskyy[3]

[1]Taras Shevchenko National University of Kyiv
64, Volodymyrs'ka Str., Kyiv, Ukraine, 01601

[2]Institute of Molecular Biology and Genetics, NAS of Ukraine
150, Akademika Zabolotnoho Str., Kyiv, Ukraine, 03680

[3]Department of Biochemistry, University of Texas Health Science Center
7703 Floyd Curl Drive, San Antonio, TX 78229-3900, USA

saa@univ.kiev.ua

Molecular docking is a widely used method of computer-aided drug design capable of accurate prediction of protein-ligand complex conformations. However, scoring functions used to estimate free energy of binding still lack accuracy. ***Aim****. Development of computationally simple and rapid algorithms for ranking ligands based on docking results.* ***Methods****. Computational filters utilizing geometry of protein-ligand complex were designed. Efficiency of the filters was verified in a cross-docking study with QXP/Flo software using crystal structures of human serine proteases thrombin (F2) and factor Xa (F10) and two corresponding sets of known selective inhibitors.* ***Results****. Evaluation of filtering results in terms of ROC curves with varying filter threshold value has shown their efficiency. However, none of the filters outperformed QXP/Flo built-in scoring function Pi . Nevertheless, usage of the filters with optimized set of thresholds in combination with P_i achieved significant improvement in performance of ligand selection when compared to usage of P_i alone.* ***Conclusions****. The proposed geometric filters can be used as a complementary to traditional scoring functions in order to optimize ligand search performance and decrease usage of computational and human resources.*

Keywords: drug design, molecular modeling, docking, scoring, geometric filtering.

Introduction. Nowadays, computer-aided drug design is a widely used technique. It is mostly based on molecular docking and scoring approach [1]. Docking is the procedure of protein (target) and small molecule (ligand) complexes geometry optimization aimed at finding the global energy minimum of the system. According to thermodynamics, the most likely configuration of the complex corresponds to the Gibbs free energy minimum. Energy is usually estimated using certain force field model and its minimization is performed by various methods. Typically, a large collection of small molecules is docked against the protein active site. For each optimized complex, different characteristics (scores) are calculated to estimate binding free energy. Ligands with the highest scores are filtered for further testing in biophysical, biochemical and/or cell-based screening assays.

Although there are many publicly available and commercial tools for molecular docking and filters for scoring, some problems still exist. While docking usually can provide adequate results for optimized geometry prediction, scoring is a tricky thing and requires human intrusion like visual inspection of three-dimensional molecular complex structures by a drug discovery expert [2].

As the mechanisms of intermolecular interaction in a protein–ligand complex exceed the classical mechanics limits, accurate prediction of binding energy needs quantum mechanical calculations, which boost requirements for memory size and floating point calculations speed by orders of magnitude. Furthermore, flexibility of both protein and ligand molecules causes an increase

of freedom degrees of a typical system up to hundreds or thousands for simulations with implicit solvent models and even tens of thousands in case of explicit solvent.

As a result of these complications, precise virtual screening of large collections of small molecules becomes practically impossible, which forced us to use simplified models with empirical scoring algorithms.

In this paper, we introduce geometric filters, which are designed to select protein-ligand complexes from the database of molecular docking results. The filters use the molecular geometry of protein–ligand complex as a main filtering characteristic as opposed to approximated potentials of inter-atomic interaction or other loosely defined and computationally expensive functions.

The main idea behind this approach is based on the fact that molecular docking can predict the molecular geometry stationary point rather accurately, which has been proved by numerous X-ray structural analysis experiments [3]. All mentioned above makes the proposed filters robust and quick for interactive usage.

Materials and methods. In this study, four types of geometry-based filters were introduced: nearest atom filter (NA), center of mass filter (COM), out coefficient filter (OUT) and hydrogen bond filter (HB). Description of the filters is provided below.

Nearest atom filter finds atom of the ligand that is the nearest to the given atom of protein in the current complex. Ligand passes the filter if this distance is less than the specified value:

$$\min_l \left| \vec{r}_l - \vec{r}_p \right| < R_{\min},$$

where l is ligand atom index, p is the given protein atom index, $\vec{r}_n$ is position of the n-th atom, $R_{\min}$ is the specified minimal distance. This filter has complexity of $O(n)$ and can select ligands that are partially located close to the given atom of the protein active site and evidently screens it from solvent. Filtering results may be modified by considering only ligand atoms of certain type.

Center of mass filter finds the distance from ligand center of mass to the given protein atom. Ligand passes the filter if this distance is less than the specified value:

$$\left| \frac{\sum_l \vec{r}_l m_l}{\sum_l m_l} \right| < R_{\min},$$

where m_n is mass of the n-th atom. This filter can select ligands that are located close to the specified atom of the protein active site and evidently screen it.

Out coefficient filter calculates numerical characteristic which approximates the probability of destroying the given protein-ligand complex. The following model is used. The complex will be destroyed if ligand binds to some external molecule and bonds with protein are destroyed. The probability of this event, P, may be approximated as

$$P = \frac{N_e}{N_l},$$

where N_e is average number of ligand atoms that may bind to the external molecule and N_l is total number of ligand atoms. The less P, the more stable the complex . Number N_e is estimated as

$$N_e = \sum_k^{N_l} p_k^e,$$

where k is ligand atom index and p_k^e is the probability that k-th ligand atom will bind to the external molecule. Probability p_k^e depends on the number of protein atoms that bind to the k-th ligand atom and shield it from outside. Probability of shielding may be described by the Markov field model with the Gibbs distribution:

$$p_k^e = e^{-n_k},$$

where n_k is average number of protein atoms which shield the k-th ligand atom. Number n_k may be estimated in the same way:

$$n_k = \sum_p^{N_p} b_p^k,$$

where b_p^k is the probability that p-th protein atom binds to the k-th ligand atom. Probability b_p^k in turn also may be described by Markov field model:

$$b_k^p = e^{-\frac{\left| \vec{r}_l - \vec{r}_p \right|}{R_{kp}}},$$

where R_{kp} is characteristic length of chemical bond between k-th ligand atom and p-th protein atom. The final expression for P is thus

$$P = \frac{\sum_{k}^{N_l} \exp\left(-\sum_{P}^{N_p} e^{-\frac{\left|\vec{r}_l - \vec{r}_p\right|}{R_{kp}}} \right)}{N_l}.$$

This filter requires $O(n^2)$ operations. To simplify the case, R_{kp} is defined to be the same for all atom pairs and equal to 1.1 C. In this case, the exact value influences only the value of filtering function but not its behavior.

Hydrogen bonds filter calculates estimated number of hydrogen bonds between ligand and protein atoms. Each hydrogen bond is characterized by strength coefficient that may be estimated as

$$p_H = A e^{-\frac{\left|\vec{r}_{A1} - \vec{r}_H\right|}{R_H}} e^{-\frac{\left|\vec{r}_{A2} - \vec{r}_H\right|}{R_H}} e^{-\cos\varphi},$$

where $\vec{r}_{A1}$ is position of the 1-st acceptor, $\vec{r}_{A2}$ is position of the second acceptor, $\vec{r}_H$ is position of hydrogen atom, R_H is characteristic length of hydrogen bond and φ is the bond angle.

To optimize the filters parameters and verify their effectiveness, we conducted a cross-docking study. Human serine proteases thrombin (gene *F2*) and factor Xa (gene *F10*) were selected as targets. X-ray crystal structures of the protein catalytic sites were retrieved from RCSB Protein Data Bank [4], entries 1oyt and 1f0s respectively. Two sets of selective small molecule inhibitors containing 244 compounds for thrombin and 331 compounds for factor Xa were retrieved from MDDR database [5]. After generation of stereoisomers and ionization using LigPrep software [6], compounds were docked into the three-dimensional protein active site structure using QXP/Flo software [7] with 100 steps of SDOCK+ routine. 10 lowest energy complex structures were selected for each compound structure, which resulted in a total of 10,580 and 7,410 complexes for thrombin and factor Xa inhibitor sets respectively. Filters were applied to the complexes, and their performance was evaluated in terms of receiver operating characteristics (ROC).

Table 1
Summary of all filters applied in the study. Residue numbering according to the crystal structure, PDB ID 1oyt

Filter ID	Filter name	Protein atom
Pi	QXP/Flo built-in scoring function P_i	–
OUT	Out coefficient	–
HB	Hydrogen bonds	–
NA182	Nearest atom	Leu99 CG
NA314	Nearest atom	Asp189 OD1
NA63	Nearest atom	Tyr60 CD1
COM	Center of mass	Gly216 HN

Description of all filters is provided in Table 1. Arbitrary atoms of thrombin active site which were selected for the nearest atom filters and center of mass filter are shown in Fig 1.

Results and discussion. To evaluate efficiency of the filters, we conducted a virtual screening study where two sets of small-molecule inhibitors of two serine proteases were docked against two sets of their selective inhibitors. This resulted in a set of protein–ligand complexes with both «native» and «wrong» inhibitors. For each filter, a receiver operating characteristic (ROC) was built. For each of the two proteins, the compounds which have at least one protein-ligand complex passed through a filter were considered positives. Out of positives, essentially, the compounds from one protein's inhibitor set were considered true positives (TP), while compounds from another protein's inhibitor set were considered false positives (FP). Additionally, a ROC was built for the docking software QXP/Flo+ built-in scoring function P_i (Fig. 2). For the two protein crystal structures, we compare only filters which are independent of arbitrary protein atom selection: OUT, HB and P_i.

It is clear from the ROCs, that the filters can be used efficiently for selection of inhibitors, except the hydrogen bond filter in case of factor Xa. However, the QXP/Flo built-in scoring function outperforms any of the proposed filters.

Next, we focused on selection of thrombin inhibitors with introduction of atom-specific filters NA and

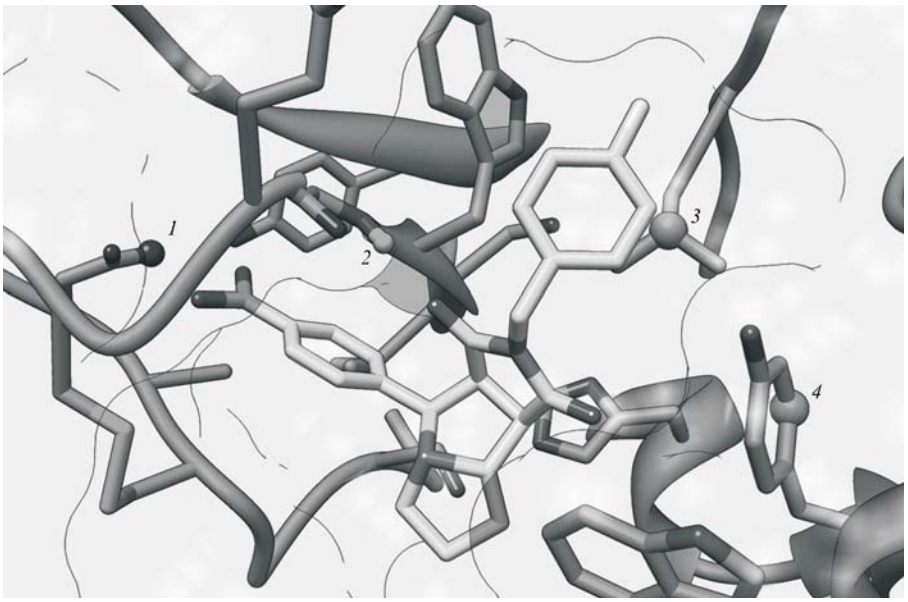

Fig. 1. Three-dimensional structure of human thrombin catalytic site in complex with small-molecule inhibitor retrieved from PDB entry 1oyt. Atoms selected for filtering are shown: *1* – atom 314, Asp189 OD; *2* – center of mass atom, Gly216 NH; *3* – atom 182, Leu99 CG; *4* – atom 314, Tyr60 CD1

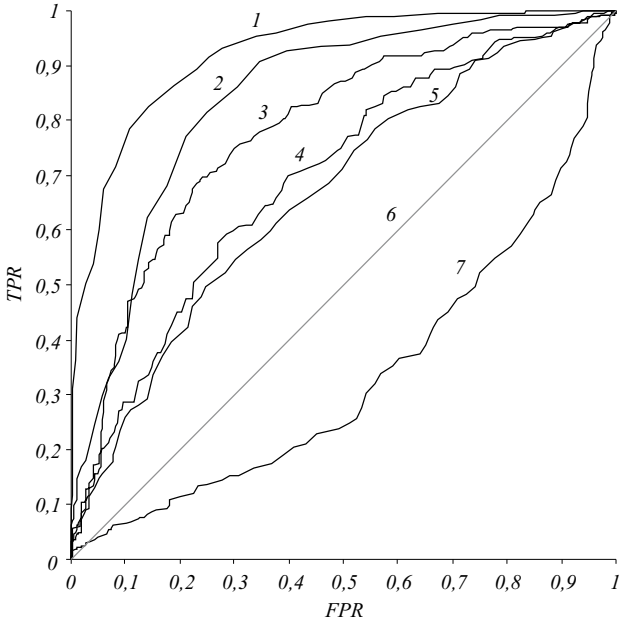

Fig. 2. Receiver operating characteristics for thrombin and factor Xa, for the filters which are independent of protein atom selection: *1* – FXa PI; *2* – Thrombin PI; *3* – Thrombin HB; *4* – Thrombin OUT; *5* – FXa OUT; *6* – Random guess; *7* – FXa HB

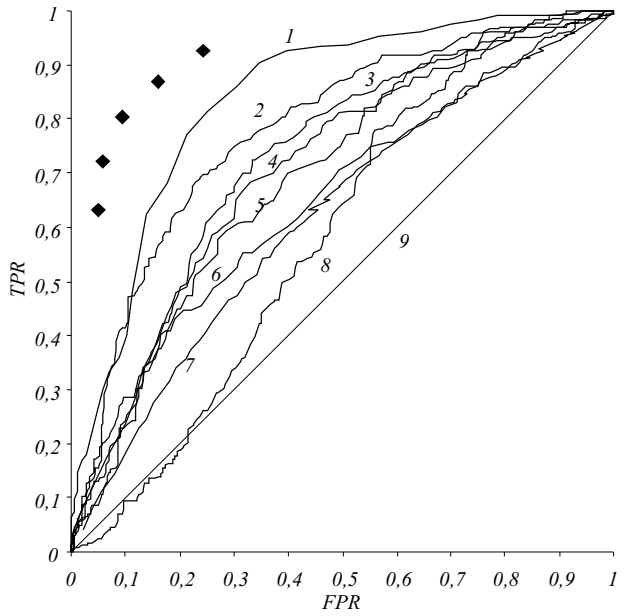

Fig. 3. Receiver operating characteristics for different filters and their combination. ROC curve for random selection of protein-ligand complexes is included; *C* – Combination; *1* – PI; *2* – HB; *3* – COM; *4* – NA314; *5* – OUT; *6* – NA182; *7* – Random guess (complex based); *8* – NA63; *9* – Random guess (compound based)

COM. ROCs for all filters for this case are provided in Fig. 3.

As mentioned in Materials and methods, the number of compounds in the two sets is different. Furthermore, as 10 complexes were generated for every stereoisomer and every possible ionization state, different compounds also have different number of complexes in the docking output. As a result, the number of thrombin inhibitor complexes (true positives) is about 25 % higher than that of factor Xa inhibitors (false positives). To investigate an impact of this inequality, in addition to obvious compound-based random guess ROC (TPR = FPR), a complex-based random guess ROC curve was built (Fig. 3). At this point, all complexes had equ-

Table 2
Best-scoring filter cut-off value combinations. Values for nearest atom and center of mass filters are in angstroms (C)

TPR	FPR	P_i	OUT	HB	NA182	NA314	NA63	COM
0.926	0.242	3.1	0.77	1.0	4.6	9.1	4.8	5.8
0.869	0.160	3.4	0.75	1.0	4.7	9.8	4.7	5.8
0.803	0.094	3.5	0.73	0.9	4.7	8.9	4.6	6.3
0.721	0.057	3.7	0.75	0.9	4.7	8.4	4.5	6.5
0.631	0.048	3.7	0.75	0.9	4.7	8.1	4.3	6.4

al probability to pass a «random guess filter», and this probability was considered the filter parameter, varying along the ROC curve.

As one can see from the Fig. 3, all ROC curves are located above the compound-based random guess line, which proves the filters efficiency. Furthermore, all of them are located above complex-based random guess curve, except those for the nearest atom filters NA182 and NA63, which are only effective in certain ranges. However, none of the filters outperformed the built-in scoring function P_i. DesPi te that, use of Pi alone would not be a proper choice. Really, let us consider that in a tyPi cal docking setup, we screen a set of about 50,000 compounds to obtain a docking library of no more than 5,000 compounds, which are going to be tested experimentally. As we do not expect more than a few percent of true binders in the initial set, size of the docking library can be estimated as size of initial set multiplied by false positive ratio (FPR), which in this case should be 10 % at maximum. As one can see from the ROCs, even the best-performing at this FPR filters, P_i and HB, give only about 40 % of true positives, which is generally not acceptable as it means loss of more than half of potentially active compounds at the very first stage of drug development. To address this issue, we carried out multiple filtering, in which all protein-ligand complexes were sequentially conducted through all 7 filters, including the built-in scoring, thus applying logical conjunction to the filter conditions. In this computation, both TPR and FPR are the functions of 7 variables, which are filter cut-off values. The values of TPR and FPR were sampled in a broad range of filter parameters to optimize filtering performance. The resulting ROC data points are plotted in Fig. 3, and filter cut-off values for them are provided in Table 2.

Conclusions. The proposed geometric filters for protein–ligand complexes have shown their efficiency for selection of specific inhibitors in a cross-docking study for serine proteases thrombin and factor Xa. However, their efficiency in terms of receiver operating cha- racteristics is lower than that of QXP/Flo+ native sco- ring function. Nevertheless, the filters can significantly improve virtual screening performance when used in combination with the scoring function. When compa- red to usage of the scoring function alone, target-spe- cific tuning of filtering parameters achieved an incre- ase of TPR from 40 % to 80 % at 10 % FPR, and from 30 % to 65 % at 5 % FPR.

Acknowledgements. Software development and computations were performed using Ukrainian National Grid Infrastructure and computing cluster of Taras Shevchenko National University of Kyiv [8, 9].

О. О. Судаков, О. М. Балінський, М. О. Платонов, Д. Б. Ковальський

Геометричні фільтри для комплексів білок–ліганд на основі феноменологічних молекулярних моделей

Резюме

Молекулярний докінг є широко застосовуваним обчислювальним методом пошуку лігандів біомолекул, здатним до достатньо точного передбачення конформацій комплексів білок–ліганд. У той же час скоринговим функціям, що використовують для оцінки сили зв'язування, бракує точності. ***Мета****. Розробка обчислювально простих та швидких алгоритмів для вибору потенційних лігандів з комплексів, отриманих у результаті докінгу.* ***Методи****. Створено обчислювальні фільтри, засновані на геометричних співвідношеннях у комплексі білок–ліганд, ефективність яких перевірено крос-докінговим дослідженням із застосуванням кристалічних структур людських серинових протеаз тромбіна (F2) і фактора 10а (F10), а також двох відповідних наборів відомих селективних інгібіторів за допомогою програмного забезпечення QXP/Flo.* ***Результати****. Оцінено результати застосування фільтрів у термінах ROC-кривих із змінними пороговими значеннями та показано їхню ефективність. Проте жоден з фільтрів не пе-*

*ревершив за ефективністю вбудовану скоринговy функцію P_i програми QXP/ Flo. Тим не менш, використання фільтрів з оптимізованими пороговими значеннями у комбінації з P_i дозволило значно збільшити ефективність порівняно із застосуванням лише P_i. **Висновки**. Розроблені геометричні фільтри можуть слугувати доповненням до традиційних скорингових функцій для оптимізації пошуку лігандів і зменшення залучення обчислювальних та людських ресурсів.*

Ключові слова: комп'ютерна розробка ліків, молекулярне моделювання, докінг, скорингова функція, геометричні фільтри.

А. А. Судаков, А. М. Балинский, М. О. Платонов, Д. Б. Ковальский

Геометрические фильтры для комплексов белок–лиганд на основе феноменологических молекулярных моделей

Резюме

*Молекулярный докинг – широко используемый вычислительный метод поиска лигандов биомолекул, способный довольно точно предсказывать конформацию комплекса белок–лиганд. В то же время скоринговые функции, используемые для оценки силы связывания, недостаточно точны. **Цель**. Разработка вычислительно простых и быстрых алгоритмов для выбора потенциальных лигандов из комплексов, полученных в результате докинга. **Методы**. Созданы вычислительные фильтры на основе геометрических соотношений в комплексе белок–лиганд, эффективность которых проверена кросс-докинговым исследованием с применением кристаллических структур человеческих сериновых протеаз тромбина (F2) и фактора 10а (F10), а также двух соответствующих наборов известных селективных ингибиторов с помощью программного обеспечения QXP/Flo. **Результаты**. Оценены результаты применения фильтров в терминах ROC-кривых с переменными пороговыми значениями и показана их эффективность. Однако ни один из фильтров не превзошел по эффективности встроенную скоринговую функцию P_i программы QXP/Flo. Тем не менее, использование фильтров с оптимизированными пороговыми значениями в комбинации с P_i позволило существенно увеличить эффективность в сравнении с применением только P_i. **Выводы**. Разработанные геометрические фильтры могут служить дополнением к традиционным скоринговым функциям для оптимизации поиска лигандов и уменьшения привлечения вычислительных и человеческих ресурсов.*

Ключевые слова: компьютерная разработка лекарств, молекулярное моделирование, докинг, скоринговая функция, геометрические фильтры.

REFERENCES

1. *Hurmach V. V., Balinskyi O. M., Platonov M. O., Borysko P. O., Prylutskyy Y. I.* Molecular docking method involving SH2-domains // Biotechnology.–2012.–**5**, N 2.–P. 31–40.
2. *Cole J. C., Murray C. W., Nissink J. W., Taylor R. D., Taylor R.* Comparing protein-ligand docking programs is difficult // Proteins.–2005.–**60**, N 3.–P. 325–332.
3. *Wang R., Lu Y., Wang S.* Comparative evaluation of 11 scoring functions for molecular docking // J. Med. Chem.–2003.–**46**, N 12.–P. 2287–2303.
4. *Berman H. M., Westbrook J., Feng Z., Gilliland G., Bhat T. N., Weissig H., Shindyalov I. N., Bourne P. E.* The protein data bank // Nucleic Acids Res.–2000.–**28**, N 1.–P. 235–242.
5. *Schuffenhauer A., Zimmermann J., Stoop R., van der Vyver J.J., Lecchini S., Jacoby E.* An ontology for pharmaceutical ligands and its application for *in silico* screening and library design // J. Chem. Inf. Comput. Sci.–2002.–**42**, N 4.–P. 947–955.
6. *LigPrep* User manual, Version 2.3.–New York: Schrodinger, 2009.–116 p.
7. *Mcmartin C., Bohacek R. S.* QXP: Powerful, raPi d computer algorithms for structure-based drug design // J. Comput. Aided Mol. Des.–1997.–**11**, N 4.–P. 333–344.
8. *Zynovyev M., Svistunov S., Sudakov O., Boyko Y.* Ukrainian Grid Infrastructure: Practical Experience // Proc. 4-th IEEE Workshop IDAACS 2007 (September 6–8, 2007, Dortmund, Germany).–Dortmund, 2007.–P. 165–169. doi:10.1109/IDAACS.2007.4488397
9. *Salnikov A. O., Sliusar I. A., Sudakov O. O., Savytskyi O. V., Kornlyuk A. I.* MolDynGrid Virtual Laboratory as a part of Ukrainian Academic Grid infrastructure // Proc. IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (21–23 September 2009, Rende (Cosenza), Italy.–Rende, 2009.–P. 237–240.

Received 02.04.13