Journal of Computer-Aided Molecular Design, **16:** 825–830, 2002.
KLUWER/ESCOM

Development of biologically active compounds by combining 3D QSAR and structure-based design methods

Wolfgang Sippl*
Institute for Pharmaceutical Chemistry, Heinrich-Heine-Universität Düsseldorf, D-40225 Düsseldorf, Germany

Received 30 September 2002; Accepted 6 January 2003

Key words: AutoDock, binding affinity prediction, CoMFA, docking, GRID, GOLPE, 3D QSAR.

Summary

One of the major challenges in computational approaches to drug design is the accurate prediction of the binding affinity of novel biomolecules. In the present study an automated procedure which combines docking and 3D-QSAR methods was applied to several drug targets. The developed receptor-based 3D-QSAR methodology was tested on several sets of ligands for which the three-dimensional structure of the target protein has been solved – namely estrogen receptor, acetylcholine esterase and protein-tyrosine-phosphatase 1B. The molecular alignments of the studied ligands were determined using the docking program AutoDock and were compared with the X-ray structures of the corresponding protein-ligand complexes. The automatically generated protein-based ligand alignment obtained was subsequently taken as basis for a comparative field analysis applying the GRID/GOLPE approach. Using GRID interaction fields and applying variable selection procedures, highly predictive models were obtained. It is expected that concepts from receptor-based 3D QSAR will be valuable tools for the analysis of high-throughput screening as well as virtual screening data

Introduction

The prediction of new biologically active compounds on the basis of previously synthesized ones is one of the major challenges in today's drug design. The strategies that can be applied for this purpose fall into two major categories – the indirect ligand-based and the direct receptor-based approach. The ligand-based methods, including traditional quantitative structure–activity relationships (QSAR) and modern 3D QSAR techniques, are based entirely on experimental structure–activity relationships for enzyme inhibitors or receptor ligands. The 3D QSAR methods, especially the comparative molecular field analysis (CoMFA) [1], are nowadays used widely in drug design, since they are computationally not demanding and afford fast generation of QSARs from which biological activity of newly synthesized molecules can be predicted. However, there is a main difficulty in the application of 3D QSAR methods such as CoMFA. For a correct model, a spatial orientation of the ligands towards one another has to be found, which is representative for the relative differences in the binding geometry at the protein-binding site. The success of a molecular field analysis is therefore completely determined by the quality of the choice of the superimposition of the studied molecules [2].

On the other hand the direct structure-based methods are able to calculate fairly accurately the position and orientation of a potential ligand in a receptor binding site. This has been demonstrated by various docking studies, described in the literature (for a recent review see [3]). Most of the docking programs use empirical potential energy functions to calculate the binding energies of protein–ligand complexes. The major problem of modern docking programs is the inability to evaluate binding free energies correctly in order to rank different ligand–receptor complexes. The main problem in affinity prediction is that the underlying molecular interactions are highly complex and various terms should be taken into account to quantify

*To whom correspondence should be addressed: e-mail: sippl@pharm.uni-duesseldorf.de

the free energy of the interaction process [4, 5]. Only rigorous methods, such as the free energy perturbation or the thermodynamic integration methods are able to predict correctly the binding affinity. While these two methods clearly have the potential of providing accurate evaluation of relative binding free energies, they are very expensive in a computational sense.

Regarding the strengths of both approaches, the docking programs using protein information and the CoMFA method to develop predictive models for related ligand molecules, prompted us and others to combine both in an automated unbiased procedure (for example see [6–13]). In this context, the three-dimensional structure of a target protein, along with a docking protocol is used to guide alignment selection for comparative molecular field analysis [14, 15]. This approach allows the generation of a target-specific scoring method considering all the structure–activity data known for a distinct ligand data set. Successful studies applying a similar strategy have also been reported recently by other groups [13, 16, 17].

Computational methods

Data sets

The developed approach has been tested on three data sets: acetylcholinesterase inhibitors (AChE), estrogen receptor (ER) agonists and protein tyrosine phosphatase 1B (PTP 1B) inhibitors, respectively. The AChE data set comprises 42 compounds as training set developed by Contreras et al. [18]. The data set of ER agonists is described in detail in [15]. Comparing relative binding affinities (RBA) of compounds that are listed in more than one reference ensured the consistency of this data set. It contains 30 agonists for training and 33 novel compounds for testing. The data set of PTP 1B inhibitors was taken from a study of Malamas et al. [19]. It is the largest data set and contains 95 compounds in the training set. For all compounds standard protonation states are assumed, i.e. carboxylate are considered to be deprotonated and aliphatic amino groups are considered to be protonated.

Protein crystal structures

The protein structures of the AChE (1acl), the ER (1ere) and the PTP 1B (1ecv) co-crystallized with ligands were retrieved from the Protein Data Bank [20]. Water atoms were removed, polar hydrogen atoms were added, charges from AMBER [21] were loaded, and the protein structures were subjected to a minimization using the AMBER force field keeping all protein backbone atoms at fixed positions.

Docking simulations

The docking analysis was performed using a two-stage docking procedure applying the program AutoDock 2.4 [22]. The minimized ligand-free protein structures, as described above, were used as input structure for the docking simulations. All ligand atoms but no protein atoms were allowed to move during the docking simulation. For each ligand the simulation was composed of 100 docking runs using the standard AutoDock parameters. In a second step low-energy complexes obtained by the AutoDock calculations were re-ranked according to the interaction energy calculated with a more detailed energetic model based on the YETI force field [23]. The protein structure was held fixed during the minimization, whereas the ligand was allowed to change its conformation and position in the binding pocket.

3D QSAR analysis

The GRID/GOLPE [24] method, which is comparable to the traditional CoMFA method, was used within this study to perform 3D QSAR analyses. The alignments come for all targets directly from the top-scoring YETI orientation. For each alignment the interaction field between the ligands and a water probe was calculated using the GRID program employing a grid spacing of 1 Å. The size of the grid box used for the calculation was defined in such a way that it extended approximately 4 Å beyond each of the molecules in each dimension. The GRID calculation gave thousand of variables for each compound. The non-important variables were selected and eliminated applying D-optimal design and SRD/FFD (Smart Region Definition/Fractional Factorial Design) variable selection within the GOLPE program (for a detailed description of this approach see [25]). To form the basis for a statistical significant model, the method of partial least squares (PLS) regression was used to correlate variations in biological activities with variations in interaction fields. The optimum number of PLS components corresponding to the smallest standard error of prediction, was determined by the leave-one-out (LOO) cross-validation procedure. Using the optimal number of components, the final PLS analysis was carried out without cross-validation to generate a

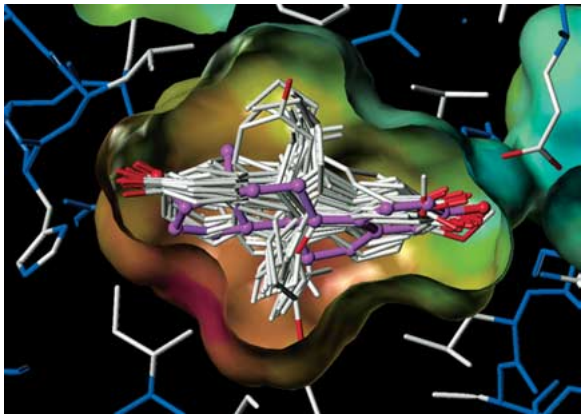

Figure 1. Alignment of 30 ER ligands within the binding pocket. The solvent accessible surface area is colored according the electrostatic potential (red: positive, blue: negative).

predictive model with a conventional correlation coefficient. In addition a second cross-validation, using five groups of approximately the same size in which the objects were assigned randomly, was performed. In this method 80% of the compounds were randomly selected and a model is generated, which is then used to predict the remaining compounds (leave-20%-out). This cross-validation technique has been shown to yield better indices for the robustness of a model than the normal LOO procedure [25, 26]. The validated 3D QSAR models obtained using different alignment strategies were then used for the prediction of the test set compounds. The quality of the external prediction is documented by the standard deviation of error prediction (SDEP).

Results and discussion

The developed methodology was applied to three different data sets for which detailed structure–activity relationships were available. Each ligand of the particular data set was automatically docked in the corresponding ligand-free protein structure as described in the methods section. For the docking the Monte Carlo simulated-annealing method implemented in AutoDock was used. In a second step the initial protein–ligand complexes generated by AutoDock were further refined using the YETI force field. The YETI protein–ligand interaction energies were then used for post-scoring the generated complexes. As an initial test of the described two-stage docking procedure, docking runs were carried out using the experimentally determined structures of the AChE com-

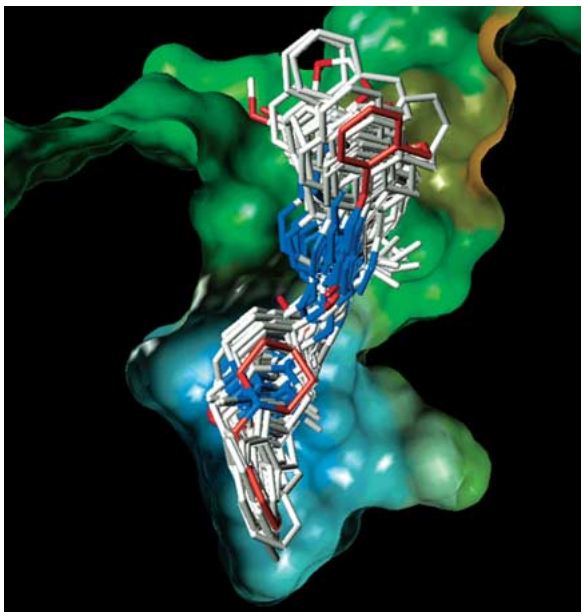

Figure 2. Alignment of 42 AChE inhibitors within the binding pocket. The solvent accessible surface area is colored according the lipophilic potential (blue: polar, brown: lipophilic).

plexed with Huperzine (1vot), tacrine (1acj), edrophonium (1ax9) and decamethonium (1acl) [14], as well as the complexes of the ER liganded with estradiol (1ere) and diethylstilbestrol (3erd) [26]. Considering the YETI interaction energies, we were able to retrieve the experimentally observed complexes among all AutoDock solutions. The ability to accurately predict the binding conformations of these complexes gave confidence that we could use the two-stage docking to evaluate the binding conformation of the remaining compounds for which no experimental structures are known. In this way all ligands of the corresponding data set were successfully docked and scored (Figures 1–3). Further support for our docking strategy came from recently solved crystal structures of the ER complexed with genistein (1qkm) and the AChE complexed with donepezil (1eve). The ligand binding conformation in both complexes has been correctly predict by the AutoDock/YETI procedure [10, 14].

Analyzing the refined protein–ligand complexes only a moderate correlation between calculated interaction energies and observed binding affinities was observed for the ER ligands (for details see [15]), whereas for AChE and PTP 1B inhibitors no significant correlation has been detected. It appeared relatively difficult to find a predictive model based on the calculated interaction energies. Several reasons for

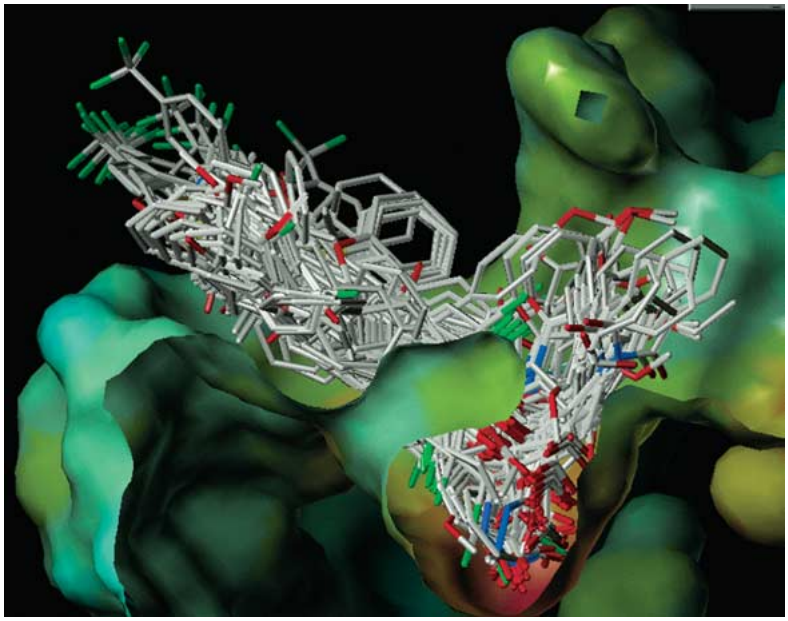

Figure 3. Alignment of 95 PTP 1B inhibitors within the binding pocket. The solvent accessible surface area is colored according the electrostatic potential (red: positive, blue: negative).

Table 1. Statistical results obtained for the three studied data sets.

Target	n (Trainingset)	LV	q^2_{LOO}	SDEP	$q^2_{L20\%O}$	SDEP	n (Testset)	$SDEP_{ext}$
AChE	42	3	0.937	0.345	0.925	0.379	6[a]	0.440
ER	30	4	0.921	0.345	0.900	0.440	33[b]	0.531
PTP 1B	95	3	0.727	0.316	0.708	0.356	35	0.477

[a]Ref. [14], [b]Ref. [15]

the low predictivity of interaction-energy based models have been stated in the literature [4]. Small errors in atomic coordinates which occur as well in well-refined X-ray structures have considerable influence on the energy landscape of ligand binding. To rank computer-generated ligands correctly the affinity estimation method must be able to correctly position novel ligands within the binding pocket and optimize their non-covalent interaction with the protein. The errors in these calculated positions are unlikely to be smaller than those of the coordinates of the protein atoms, so the estimation of affinity must be able to accommodate small deviation in atom positions. Another problem is the limited understanding of the physics and thermodynamics of ligand binding, especially solvation and entropic effects are at the moment fully neglected by interaction-energy based models. To overcome this problem more rigorous methods such as the free energy perturbation methods have to be applied. For a detailed discussion of this problem the reader is referred to the literature [4, 5].

The field of drug design knows alternative techniques to score and predict binding affinities within a set of ligands. Comparative molecular field analyses show usually a surprisingly high predictive power. Originally, the CoMFA approach was developed for ligand data sets where the structure of the target protein is unknown. However, in the last few years it was also applied to cases where the corresponding target structure has been determined [8–17]. There, the superimposition of the ligands derived from molecular docking runs is taken as starting point for a CoMFA study. In our approach we select the conformation which is top-ranked in the YETI force field for the alignment generation. For the three studied data sets the subsequent 3D QSAR analysis was carried out as described in the methods section. The optimal number of latent variables was chosen by monitoring the changes in the predictivity (q^2) index of the model

upon adding a new latent variable. To restrict the complexity of the resulting PLS model we used a maximum number of five latent variables. The statistical results for all three data sets are summarized in Table 1. In general, models with high internal predictivity have been obtained indicated by the high q^2 values (between 0.727 and 0.937) obtained by the leave-one-out cross-validation. The models are also robust, indicated by the high correlation coefficients (between 0.708 and 0.925) obtained using the leave-20%-out cross-validation procedure (see Table 1 for details).

Besides the internal validation, the models were also checked for their predictivity of novel ligands. 33 recently developed ER agonists were selected as external test set [15]. The prediction of the binding affinities on the basis of the generated receptor-based 3D QSAR model yielded an external SDEP value of 0.531 which is close to the SDEP value obtained for the internal validation. Comparable statistical results have been obtained for the AChE inhibitors. Six novel inhibitors which were designed on the basis of the GRID/GOLPE model have been successfully predicted [14]. The external prediction shows an SDEP value of 0.440 indicating the applicability of the method for the design of novel biologically active compounds.

Conclusions

A key goal of this work was the development of a simple and robust method for the prediction of the binding affinity of novel biomolecules. We suggest that 3D QSAR methods are an important complement to structure-based design methods. If one already has a series of molecules and their corresponding binding affinities, then 3D QSAR equation may provide a valuable method to forecast affinity of further analogs. Knowledge of the structure of the binding pocket would guide the further modelling and should prevent unwarranted extrapolation of such equations. At the moment, the observed structure–activity relationships of ligands provide a more sensitive measure of ligand-receptor affinity than do protein-based affinity prediction methods [26–29]. An interesting novel strategy which might overcome the problem of neglecting the protein information in a 3D QSAR analysis has been very recently published by Gohlke et al. [29]. It is expected that novel concepts from 3D QSAR will be valuable tools for the analysis of high-throughput screening as well as virtual screening data.

Acknowledgements

The author wishes to thank Dirk Classen-Houben and Prof. Dr. Hans-Dieter Höltje, Heinrich-Heine Universität Düsseldorf, for their assistance and support of the project.

References

1. Cramer, R.D., III, Patterson, D.E. and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.
2. Klebe, G. and Abraham, U., J. Med. Chem., 36 (1993) 70.
3. Taylor, R.D., Jewsbury, P.J., Essex, J.W., J. Comput. Aided Mol. Des., 16 (2002) 151.
4. Tame, J.R.H., J. Comput. Aided Mol. Des., 13 (1999) 99.
5. Kollman, P., Chem. Rev., 93 (1993) 2395.
6. Ortiz, A.R., Pisabarro, M.T., Gago, F. and Wade, R.C., J. Med. Chem., 38 (1995) 2681.
7. Holloway, M.K., Wai, J.M., Halgren, T.A., Fitzgerald, P.M., Vacca, J.P., Dorsey, B.D., Levin, R.B., Thompson, W.J., Chen, L.J. and deSolms, S.J., J. Med. Chem., 38 (1995) 2280.
8. Cho, S.J, Garsia, M.L., Bier, J. and Tropsha, A., J. Med. Chem., 39 (1996) 5064.
9. Tokarski, J.S. and Hopfinger, A. J., J. Chem. Inf. Comput. Sci., 4 (1997) 792.
10. Sippl, W., Contreras, J. M., Rival, Y. and Wermuth, C.G., In: K. Gundertofte, F.S: Jorgensen (Eds.), Molecular Modelling and Predicting of Bioactivity, Plenum Press, New York, pp. 53–58 (1998).
11. Vaz, R.J., McLEan, L.R. and Pelton, J.T., J. Comput. Aided Mol. Des., 12 (1998) 99.
12. Bursi, R. and Grootenhuis, P.D. J. Comput. Aided Mol. Des., 12 (1999) 341.
13. Lozano, J.J., Pastor, M., Cruciani, G., Gaedt, K., Centeno, N.B., Gago, F. and Sanz, F., J. Comput. Aided Mol. Des., 13 (2000) 341.
14. Sippl, W., Contreras, J. M., Parrot, I., Rival, Y. and Wermuth, C.G., J. Comput. Aided Mol. Des. 15, (2001) 395.
15. Sippl, W., Bioorg. Med. Chem., 10 (2002) 3741.
16. Vieth, M. and Cummins, D.J. J. Med. Chem., 43 (2000) 3020.
17. Costantino, G., Macchiarulo, A., Camaioni, E. and Pellicciari, R., J. Med. Chem., 44 (2001) 3786.
18. Contreras, J. M., Parrot, I., Sippl, W., Rival, Y. M. and Wermuth, C.G., J. Med. Chem., 44 (2001) 2707.
19. Malamas, M.S., Sredy, J., Moxham, C., Katz, A., Xu, W., McDevitt, R., Adebayo, F.O., Sawicki, D.R., Seestaller, L., Sullivan, D., Taylor J.R., J. Med. Chem., 43 (2000) 1293.
20. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E., Nucleic Acids Research, 28 (2000) 235.
21. Singh, U. C. and Kollman, P. A., J. Comput. Chem., 5 (1984) 129.
22. Goodsell, D. S., Morris G. M. and Olson, A. J., J. Mol. Recognit., 9 (1996) 1.
23. Vedani, A. and Huhta, D. W., J. Am. Chem. Soc., 112 (1990) 269.
24. Baroni, M., Constantino, G., Cruciani, G., Riganelli, D, Valigi, R. and Clementi, S., Quant. Struct.-Act. Relat., 12 (1993) 9.
25. Pastor, M., Cruciani, G. and Watson, K., J. Med. Chem., 40 (1997) 4089.

26. Sippl, W., J. Comput. Aided Mol. Des., 14 (2000) 559.
27. Liu, H., Huang, X., Shen, J., Luo, X., Li, M., Xiong, B., Chen, G., Shen, J., Yang, Y., Jiang, H. and Chen, K., J. Med. Chem., 45 (2002) 4816.
28. Huang, X., Xu, L., Luo, X., Fan, K., Ji, R., Pei, G., Chen, K. and Jiang, H. J. Med. Chem., 45 (2002) 333.
29. Gohlke, H. and Klebe, G., J. Med. Chem., 45 (2002) 4153.