Pergamon

Bioorganic & Medicinal Chemistry 8 (2000) 2771–2780

BIOORGANIC &
MEDICINAL
CHEMISTRY

Linking CoMFA and Protein Homology Models of Enzyme–Inhibitor Interactions: an Application to Non-Steroidal Aromatase Inhibitors

Andrea Cavalli,[a] Giovanni Greco,[b] Ettore Novellino[b] and Maurizio Recanatini[a,*]

[a]*Department of Pharmaceutical Sciences, University of Bologna, Via Belmeloro 6, I-40126 Bologna, Italy*
[b]*Department of Pharmaceutical and Toxicological Chemistry, University of Napoli "Federico II", Via D. Montesano 49, I-80131 Napoli, Italy*

Received 2 June 2000; accepted 27 July 2000

Abstract—An approach to compare quantitatively a ligand-based (CoMFA) model and an enzyme active site model was investigated. The active site of the cytochrome P450 human aromatase was constructed by homology modeling techniques and two structurally different non-steroidal aromatase inhibitors were docked into it. A CoMFA model was then developed on a related series of non-steroidal inhibitors by correlating their inhibitory activity (expressed as $-\log IC_{50}$ values) versus only 11 steric descriptors (i.e. C_{sp3}–ligand steric interaction energies). The resulting 3D-QSAR coefficients (11) and the steric field values of the aromatase active site calculated at the same points of the CoMFA lattice (i.e. eleven C_{sp3}–protein steric interaction energies) were pair-wise compared. Specifically, when a positive coefficient was associated with a negative or low (<5 kcal/mol) value of the protein steric field or, alternatively, a negative coefficient was associated with a large positive value of the protein steric field we recorded as many matches. When a 3D-QSAR coefficient did not correspond to the protein steric potential in the sense described above we considered that point as a mis-matching point. In our view, in spite of several limitations, such a comparison represents a valuable criterion to evaluate quantitatively how convergent are the results from a 3D-QSAR CoMFA model and a homology-built protein 3D structure. © 2000 Elsevier Science Ltd. All rights reserved.

Introduction

During the past decades, the tools available to medicinal chemists for either designing new bioactive molecules or improving the old ones have grown incredibly both in number and in quality. However, considering the beginning of the computer-assisted drug design era, one can recognize two main fields from which the whole undertaking has started: Hammett's studies on linear free energy relationships[1] and the early attempts to use computer simulation and molecular graphics to model the three-dimensional properties of molecules.[2] How the two approaches could complement each other was first shown by Hansch who, during the 1980s, produced many studies demonstrating the usefulness of comparing statistical (QSAR) and 3D graphical (X-ray) models to understand ligand–enzyme interactions.[3,4]

3D-QSAR methods using ligand–macromolecule interaction energies as descriptors, such as Comparative Binding Energy (COMBINE) Analysis,[5] offer the advantage of linking tightly regression analysis with the binding site structure. This implies that accurate ligand–protein geometries should in principle afford highly predictive models. In the COMBINE approach, the ligand–protein interaction energy, calculated through a molecular mechanics force field, is partitioned into various components, each associated with a specific subpocket of the target protein. The GOLPE variable selection procedure[6] is then applied to distinguish relevant from irrelevant contributions to the binding affinities of the ligands. Differently from COMBINE, the Comparative Molecular Field Analysis (CoMFA) 3D-QSAR method[7] requires only the structures of the ligands, although knowledge of the binding site structure is extremely useful to guide molecular alignment and interpretation of the results as well.

Recently, the need of 'building bridges' between 3D protein models and 3D-QSAR studies was pointed out

*Corresponding author. Fax: +39-051-2099734; e-mail: mreca@alma.unibo.it

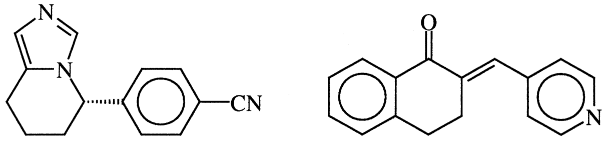
0968-0896/00/$ - see front matter © 2000 Elsevier Science Ltd. All rights reserved.
PII: S0968-0896(00)00203-0

by Kim in a very inspired article,[8] where the author showed how the two methods can act synergistically in providing useful information towards the goal of ligand design. This paper describes how homology modeling[9] and the CoMFA 3D-QSAR method[7] have been employed in combination to describe the interactions between the cytochrome P450 human aromatase active site and some non-steroidal inhibitors.

The cytochrome P450 enzymes constitute a family of heme proteins, that catalyze the metabolism of a great number of both endogenous and exogenous compounds. Although the key steps of the catalytic mechanism are thought to be similar for all the isoforms, these enzymes are able to biotransform a wide variety of substrates with high specificity and are involved in the biosynthetic pathway of several important hormones, mainly of steroidal nature. Human aromatase (CYP19, $P450_{arom}$) is the enzyme that catalyzes the conversion of androgens into estrogens, through the aromatization of the A ring of substrates like testosterone and androstenedione. Estrogens are known to be involved in the progression of the breast cancer and a block of their biosynthesis leads to a reduction of the levels of circulating hormones, which has been demonstrated to be therapeutically useful.[10,11] In consequence of that, in recent times, a considerable interest has grown around the design and the synthesis of both steroidal and non-steroidal aromatase inhibitors[12] and a large amount of structure–activity (SAR) data is now available.

The rationalization of the SAR of a class of compounds is the first necessary step towards the design of new analogues acting through the interaction with the same molecular target. With this in mind, recently, we developed CoMFA models of the aromatase inhibition by two series of non-steroidal agents represented by the lead compounds **1** (*S*-fadrozole) and **2** ((*E*)-2-(4-pyridylmethylene)-1-tetralone)[13,14] (Scheme 1). All the work was done in the absence of any information about the 3D structure of the target enzyme, although a homology-built model of aromatase proposed by Laughton et al.[15] seemed to reasonably agree with outcomes from CoMFA.

In the prosecution of our studies about the molecular mechanism of aromatase–inhibitor binding, we employed the homology building and CoMFA approaches in conjunction. Briefly, a 3D model of the aromatase active site was built, and the leads **1** and **2** were docked into it. All the other molecules were then aligned on the structures of **1** and **2**, and, from this target-based alignment, we derived a new CoMFA model expected to more closely reflect the nature of the binding site. Finally, the CoMFA and the graphic models were checked against each other on quantitative bases, and the more or less consistent results were tentatively explained in terms of the models' features.



1 **2**

Scheme 1.

Homology modeling of the aromatase active site

A number of models of human aromatase have already been reported,[15–17] and an even greater number of cytochrome P450 isozymes have also been modeled,[18–23] all of which rely on one or more of the known X-ray structures of $P450_{cam}$,[24] $P450_{terp}$,[25] $P450_{BM\text{-}3}$[26] and $P450_{eryF}$.[27] Given that the purpose of our work was not to add a new item to the list of the $P450_{arom}$ models, we referred to the work of Graham-Lorence et al.,[17] in order to perform a suitable multiple alignment of the aromatase primary sequence on those of the three P450s so far crystallized ($P450_{cam}$, $P450_{terp}$, $P450_{BM\text{-}3}$). Actually, our goal was to obtain a working model of a part of the enzyme able to illustrate some features of the binding of a specific class of non-steroidal inhibitors to the active site cavity. These inhibitors act through a competitive mechanism that involves the coordination of the heme Fe^{2+} ion by a nitrogen atom[28] and are typically represented by compounds **1** and **2** bearing an imidazole or a pyridine coordinating moiety, respectively. Consequently, we ran through the PDB looking for P450 enzymes cocrystallized with a similarly coordinated heme and found the structure of a cytochrome $P450_{cam}$ complexed with a phenylethylimidazole derivative solved at 1.60 Å resolution (PDB code: 1PHD). This structure was then chosen as a template for the coordinate transfer, because of the similarity of the cocrystallized ligand with the non-steroidal inhibitors under study.

Before the actual protein model building, the alignment of the aromatase and the $P450_{cam}$ sequences was further checked by comparing a secondary structure elements prediction for aromatase with the experimental secondary structure assignments for $P450_{cam}$ deduced from the PDB file. The aromatase secondary structure predictions were obtained through the PSIPRED protein structure prediction server,[29] that allows one to submit a protein sequence, perform a highly accurate secondary structure prediction, and receive the results of the prediction via E-mail. The coordinates of the structurally conserved regions were then directly transferred from the template protein to the one to be modeled, while the structurally variable regions were built de novo. The resulting model was checked for the correct chirality, backbone angles, and possible bad steric contacts.

After the active site region construction and prior to the optimization, the inhibitors *S*-fadrozole (**1**) and pyridyltetralone (**2**) were built into the enzyme cavity. For this purpose, the position of the phenylethylimidazole ligand cocrystallized with $P450_{cam}$ was initially considered in the orientation of **1**. To further refine the geometry of the Fe coordinating moiety of each molecule (imidazole for **1** and pyridine for **2**), the coordinates of the X-ray structures of both imidazole– and pyridine–heme complexes retrieved from the CSD (refcodes COMTED and CPOEFE01, respectively) were used.

The *p*-cyanophenyl ring of *S*-fadrozole (**1**) was oriented in such a way as to approximately reproduce the orientation of the phenyl ring of the inhibitor cocrystallized with the template $P450_{cam}$.

Results

The aromatase active site and the docking of inhibitors

Among the most conserved regions in all the P450 isozymes, there are the I, L and J helices,[18] which, in our models, are correctly aligned, although not all parts of them are included in the modeled active site. Moreover, a number of residues are invariant or highly conserved in all the P450s and they are ($P450_{cam}$ numbering) Gly60, Arg112, Gly249, Thr252, Glu287, Arg290, Arg299, Phe350, Gly353, His355, Cys357, Gly359, Ala363, and Leu375.[30] In the alignment shown in Fig. 1, all these residues, which are important for maintaining the structure and the function of the protein, are either conserved or conservatively mutated. Moreover, from Fig. 1, it results that the secondary structure element predictions for aromatase (*H* = helix; *E* = strand) match

```
                EEE          HHHHHHHHHHHHHHHHHHHHHH                    HHH
aromatase: MVLEMLNPIHYNITSIVPEAMPAATMPVLLLTGLFLLVWNYEGTSSIPGPGYCMGIG---PLI (60)
1phd     :                                    TTETIQSNANLAPLPPHV-PEHLVFDFD (27)


            HH       HHHHHHHHHHHH          EEEE  EEEEEE  HHHHHHHHH
aromatase : SHGRFLWM-GIGSACNYYNRVYG-----EFMRVWISGEETLIISKSSSMFHIMKH-NHYSSR (115)
1phd      : MYNPSNLSA----GVQEAWAVLQESNVPDLVWTRCNGGH-WIAT-RGQLIREAYED----YR (79)
                         HHHHHHHHHH     EEEEEEE EEE EEEE HHHHHHHHHHH    HH
                                                *


                       EEE       EEEE       HHHHHHH      HHHHHHHHHHHHHHHHH
aromatase: FGSKL-----GLQCIGMH-EKGIIFNNN----PELWKTTR------PFFMKALSGPGLVRMVT (162)
1phd     : HFSS-ECPFI-PREAGEAY-DFIPTSMDP-PEQRQFRALANQVVGMPVVDK------LENRIQ (132)
           HH         HHHHHHHH           HHHHHHHHHHHHHHHHHHHHH      HHHHHH
                                               *


           HHHHHHHHHHHH   HHHHH       HHHHHHHHHH  HHHHHHHH      HHHHHHHH H
aromatase: VCAESLKTHLDR---LEEVTNESGYVDVLTLLRRVML--DTSNTLFLRIPLDESAIVVKIQ-G (219)
1phd     : ELACSL-IESLRPQGQCNF–TEDYAEPFPIRIFMLLAGL----------PEE-DIPHLKYLTD (182)
           HHHHHH HHHHHHHEEEEH HHHHHHHHHHHHHHHHHHH              HHHHHHHHHH


           HHHHHHHH HHHHHH       HHHHHHHHHHHHHHHHHHHHHHHH HH        HHHHHH
aromatase: YFDAWQAL–LIKPDIFFKISWLYKKYEKSVKDLKDAIEVLIAEKRR-RISTEEKLEECMDFAT (280)
1phd     : QMT--------RPDGSM-------TFAEAKEALYDYLIPIIEQRRQK-PGT--------DAIS (221)
           HHH                     HHHHHHHHHHHHHHHHHHHHHHH            HHHH


           HHHHH         HHHHHHHHHHHHHH    HHHHHHHHHHHHH   HHHHHHHHHHHHHH
aromatase: ELILA--EKRGDLTRENVNQCILEMLIAAPDTMSVSLFFMLFLIAKHPNVEEAIIKEIQTVIG (341)
1phd     : IVAN-GQVNGRPITSDEAKRMCGLLLVGGLDTVVNFLSVFSMEFLAKSPEHRQELIE------ (276)
           HHHH EEE EEEHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
                                       *  *


                HHHH    HHHHHHHHHHH          EE    EE  EEE    EEEEE E
aromatase: ERDIKIDDIQKLKVMENFIYESMRYQPVVDLVMRKALEDDVIDGYPVKKGTNIILN-IGRMH- (402)
1phd     : -RPE---------RIPAACEELLRRFSLV-ADGRILTSDYEFHGVQLKKGDQILLPQMLSGLD (328)
                        HHHHHHHHHHHHH  E EEEEEE   EEEE EEE  EEEEEEEEEHHHHH
                               *  *         *


                                   EEEE        HHHHHHHHHHHHHHHHHH EEEEE
aromatase: -RLEFFPKPNEFTLENFAKNVPYRYFQPFGFGPRGCAGKYIAMVMMKAILVTLLRRFHVKTLQ (464)
1phd     : ERENACPMHVDFSRQKVSH-------TTFGHGSHLCLGQHLARREIIVTLKEWLTRI--PDFS (382)
                                                HHHHHHHHHHHHHHHHHHHH     E
                                       *  * * * *   *           *


                      EEEEE       EEEEEEE
aromatase: GQCVESIQKIHDLSLHPDETKNMLEMIFTPRNSDRCLEH (503)
1phd     : IAPGAQIQH-KSGIVSGVQALPLVWD---PATTKAV    (414)
           EEEEEEEEE EE EEEEEEEEEEE
```

Figure 1. Pair-wise alignment of the primary sequences of aromatase and cytochrome $P450_{cam}$ (1PHD); residues highly conserved throughout the P450 family are indicated by an asterisk; secondary structure elements predicted (for aromatase) and retrieved from the PDB file (for $P450_{cam}$) are shown above and below the aromatase and the $P450_{cam}$ sequences, respectively (*H* = helix; *E* = strand).

reasonably well with the secondary structure elements determined in P450$_{cam}$ by the X-ray analysis. This correspondence, despite the low degree of homology between the two sequences, might confer an acceptable feasibility to the aromatase model.

The inhibitors *S*-fadrozole (**1**) and pyridyltetralone (**2**) were docked in the aromatase active site as described above, and their positions after the dynamics simulations together with some surrounding residues are shown in Figure 2(a) and (b), respectively.

The primary interaction between *S*-fadrozole and the enzyme is a coordinating bond with the heme iron, but it appears that other favorable contacts are possible for the inhibitor within the active site (Fig. 2(a)). Hydrophobic interactions might occur between the tetrahydropyridine moiety of *S*-fadrozole and the side chains of Ile305, Ala306, and Thr310. The phenyl ring of **1** seems to contribute to the binding of the inhibitor by a number of possible hydrophobic contacts with the side chains of Val313, Thr310, Val369, and Val370. The cyano group appears to be H-bonded to the hydroxyl group of Ser478, thus confirming the important role ascribed to hydrogen bond acceptor functions present in the corresponding position of fadrozole-like inhibitors.[31] The relative positions of the CN and OH groups are compatible with an H-bond: the N..H distance is 2 Å, the O–H..N and the C–N..H angles are 161° and 112°, respectively.

Recently, Koymans et al.[16] modeled the position of *S*-fadrozole in the P450$_{arom}$ active site and found that the cyano-substituted phenyl ring binds in a region surrounded by Asp309, Ser478, and His480. It is satisfying that our results with the same inhibitor are in substantial agreement with this previously developed model.

The binding mode of the pyridyltetralone inhibitor **2** is shown in Figure 2(b). Again the most important feature of the interaction is the coordination bond between the pyridine nitrogen and the Fe of the heme. No additional polar interaction appears to be involved in the binding. A largely hydrophobic pocket made up by the side chains of Ile305, Val369, Val370, and Leu477 hosts the tetralinic moiety of this inhibitor. The most remarkable difference between the *S*-fadrozole and the tetralone binding modes is represented by the H-bond linking the *S*-fadrozole CN group and the Ser478 Oγ. In the tetralone–active site complex, the Ser478 side chain protrudes inside the cavity without contacting the inhibitor.

Development of a new CoMFA model derived from a target-based alignment

The model of the aromatase active site illustrated so far provides a *qualitative* description of the interactions between the enzyme and the inhibitors **1** and **2**. On the other hand, we had previously developed a CoMFA model that *quantitatively* relates the 3D properties (steric and electrostatic fields) of 49 analogues of **1** and **2** with their inhibitory activities[14] (model no. 1 in Table 1). In the absence of the 3D structure of aromatase, the molecules investigated by CoMFA had been superimposed according to a pharmacophoric hypothesis (alignment (a) in Fig. 3).

One particular aspect emerging after the docking simulations of *S*-fadrozole (**1**) and pyridyltetralone (**2**) into the aromatase active site is that they are mutually oriented in a different manner from the alignment used in the CoMFA study. With respect to *S*-fadrozole, compound **2** is partly rotated in such a way as to place the pyridine ring almost coplanar with the imidazole ring of **1**, and it also assumes a different conformation resulting in a slightly different orientation of the tetralone moiety, particularly as regards the carbonyl function (Fig. 3(b)). The original CoMFA alignment (a) was chosen a priori considering the Fe-coordinating nitrogen atoms common to both the inhibitors and on the basis of the

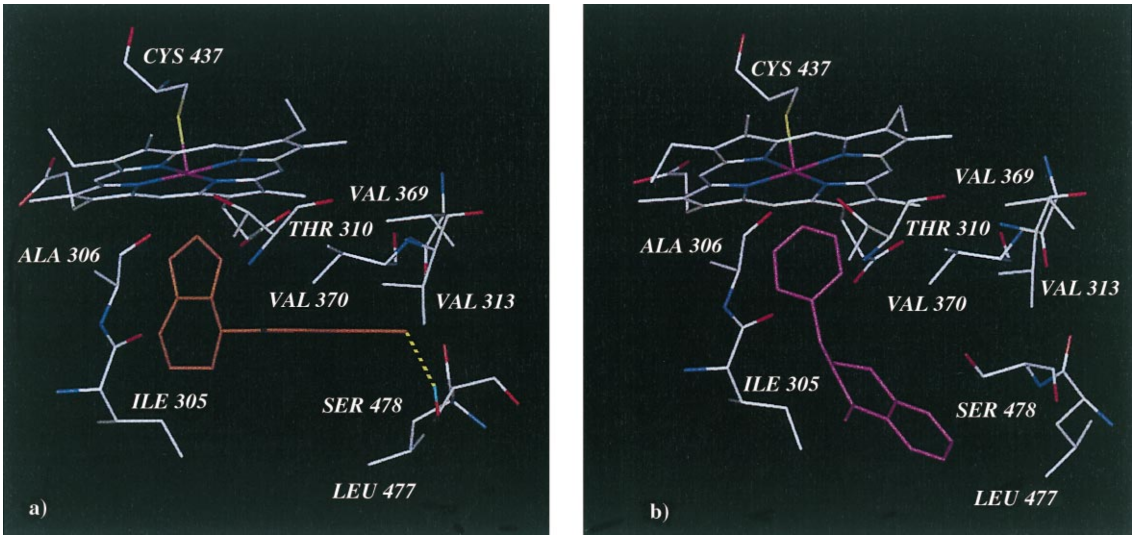


Figure 2. Aromatase active site models after molecular dynamics simulations: (a) model with fadrozole (**1**) docked in the active site: the H-bond between the inhibitor and Ser478 is shown as yellow dashed line; (b) model with the tetralone derivative (**2**) docked in the active site.

Table 1. Statistics of the CoMFA models

Model no.	Min. sigma	Number of components	q^2	s_{cross}	F	Steric contr.	Electr. contr.	Number of points
1[a]	2	2	0.860	0.555	23.266	0.44	0.56	316
2	2	2	0.848	0.578	128.488	0.50	0.50	352
3	15	1	0.826	0.612	222.605	1.00	0.00	11

[a] Ref. 14.

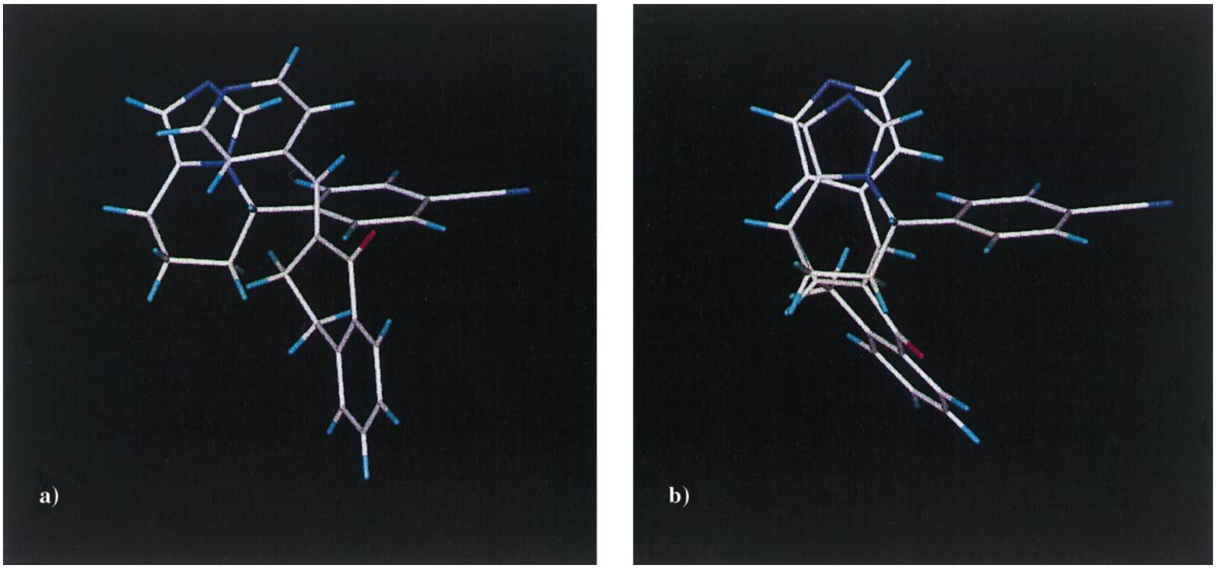


Figure 3. Alignments between the templates of the two series of inhibitors: (a) alignment used in a previous CoMFA model;[14] (b) alignment obtained after docking and molecular dynamics simulation procedures.

superimposition on steroidal aromatase ligands.[14] The alignment (b) is the result of independent dynamics simulations of the enzyme–inhibitor complexes, where the only constraint was the anchoring of the corresponding N atoms to the heme. The similarity of the alignments supports the pharmacophore approach to the atom-fitting overlay of **1** and **2**. However, it was felt that, in order to compare the CoMFA graphical outputs and the putative structure of the active site, a new CoMFA model had to be calculated from the inhibitors aligned onto the docked templates **1** and **2**. The 3D-QSAR derived from alignment (b) (model no. 2) can be compared in Table 1 with the previous 3D-QSAR (model no. 1) obtained from the pharmacophore-based alignment (a). It appears that in terms of both q^2 and s_{cross} the two models are very similar. These results are not surprising if one considers that different alignments do not necessarily lead to significantly different values of cross-validation indexes.[32–34]

Comparison between the CoMFA and the homology-built models

It is well known to all CoMFA practitioners that the coefficient contour maps can by no means be interpreted as a sort of low resolution picture of a binding site. On the other hand, one might want to relate the achievements of a CoMFA model (based purely on ligands) with the 3D structure of those portions of the target involved in the binding process (determined experimentally or derived computationally). A mutual validation of the two types of approach has been recommended,[8] such that, in case of success, the 'lateral validation' might reinforce the interpretation of the SAR under investigation. To verify whether homology building and CoMFA lead to consistent conclusions in terms of enzyme–ligand interactions for the aromatase inhibitors under study, we looked for a way to *quantitatively* compare the outcomes provided by the two techniques.

We reasoned that the CoMFA region might be taken as a spatial reference grid common to the protein and the inhibitors. In fact, by means of the probe atom, potential energy (field) values for the enzyme can be calculated at the same points where the field values of the inhibitors are calculated for the CoMFA analysis. Practically, each of the inhibitor–active site complexes was added to the CoMFA table: the docked molecules served as reference fragments for the alignment onto the corresponding molecules (**1** and **2**) already in the table, and some surrounding residue was also used in order to fit the two enzymes one on the other. The field values of the active sites (devoid of the docked inhibitors) could thus be calculated as for any other molecule of the series. Our purpose was to pair-wise compare the CoMFA coefficient values with the protein field values all over the grid points employed to derive the CoMFA model.

In order to make such a comparison practically affordable though rigorous, the number of grid points (variables) used by the CoMFA analysis had to be decreased. In fact, as reported in Table 1, the points considered in model no. 2 are 352. A way to reduce the number of points considered in a CoMFA analysis is to simply increase the minimum sigma value, i.e., the minimum value of standard deviation allowed to a variable to be included in the PLS analysis. We thus increased the minimum sigma up to a value of 15 kcal/mol (higher values dramatically worsened the q^2). The resulting model (Table 1, no. 3), based on only 11 grid points, was statistically significant (the q^2 value being slightly lower than that of model no. 2). Another remarkable feature of model no. 3 is that only the steric field contributes to statistics with all the electrostatic descriptors being rejected by the minimum sigma filter. The still good quality of the CoMFA model based on the steric field alone should be ascribed to collinearity between the two fields in model nos 1 and 2.[35]

The drastic reduction of points and the involvement of only the steric field allowed a suitable comparison of the CoMFA coefficients and the active site (steric) field values, which are reported in Table 2. It has to be noted that the different shape of the compounds docked in each of the active site models led to different energy values in some of the 11 grid points considered, due to the different position of some residues' side chains. However, given that the available volume in the active site cavity should allow both classes of molecules (**1**-like and **2**-like) to be hosted, we used only the most favorable (lowest energy) field values in the comparison with the CoMFA coefficients.

The match between the sign of the coefficient associated with the steric field in the final CoMFA equation (model no. 3) and the value of the steric field for the enzyme active site — both calculated in the same grid point — was taken as the criterion for establishing the consistency of the CoMFA and the protein modeling results. There is agreement between the two models at those points characterized by a positive CoMFA coefficient and a negative or low (< 5 kcal/mol) value of steric field for the enzyme or, alternatively, at those points where a negative CoMFA coefficient and a large positive steric energy value appear. In the first case, both models would suggest a favorable steric interaction (i.e., no steric hindrance) in the same point of the space, while, in the second case, both models would indicate a repulsive steric contact. When the values of the CoMFA coefficient and of the protein field do not correspond in the sense described above, there is a mis-match, i.e., the results of the two models at that point are not consistent.

In Figure 4, the two inhibitors *S*-fadrozole and tetralone are shown overlaid in the reciprocal orientation obtained from the dynamics simulations described above (alignment (b) of Fig. 3); some amino acid residues belonging to the aromatase active site are also reported in the positions resulting from the simulations for each of the active site models. The green and the yellow balls indicate the position of the 11 points of the grid on which model no. 3 is based (green for favorable and yellow for unfavorable increase of the steric bulk from the inhibitor side).

Examining the data of Table 2 shows up that points 3, 6, 7, 8, 9, and 10 are points of match, characterized by a positive CoMFA coefficient and a negative or < 5.0 kcal/mol value of the enzyme steric field. The CoMFA model suggests that increasing the inhibitor volume at those points of the space around the templates is favorable for the activity. Accordingly, from the target side, a probe placed at the same points gives rise to favorable interaction energies.

The remaining five points in Table 2 seem to represent as many inconsistencies between the two models. Points 2, 4, and 11 are seen as favorable by the CoMFA analysis, but the probe at these grid points gives rise to elevated interaction energies with the protein in both active site models. On the contrary, the CoMFA-negative points (**1**) and 5 are seen to be sterically unhindered specially in the active site model derived for the tetralone inhibitor. However, some observation can be pointed out as regards the reported mis-matches between CoMFA and homology models.

Table 2. Comparative analysis of the CoMFA coefficients and the steric field values at the common grid points

Point number	CoMFA coefficients model no. 3	Steric field energy (kcal/mol) fadrozole active site	Steric field energy (kcal/mol) tetralone active site	CoMFA/protein field match[a]
1	−0.0083026	5.3591752	−1.4103554	No
2	0.0083916	> 30.000000	> 30.0000000	No
3	0.0086017	2.8791974	> 30.0000000	Yes
4	0.0085612	> 30.0000000	> 30.0000000	No
5	−0.0069199	> 30.0000000	−1.6591914	No
6	0.0086984	−1.5355910	> 30.0000000	Yes
7	0.0085843	−1.8334761	> 30.0000000	Yes
8	0.0085587	> 30.0000000	3.9793310	Yes
9	0.0086514	4.9394326	> 30.0000000	Yes
10	0.0085727	5.2938356	3.9362144	Yes
11	0.0076807	> 30.0000000	> 30.0000000	No

[a] There is agreement between CoMFA coefficients and field values (Yes) when a positive CoMFA coefficient corresponds to a negative or low (< 5 kcal/mol) value of steric field, or when a negative CoMFA coefficient corresponds to a large positive steric energy value; in the contrary cases, a mismatch is reported (No). Only the most favorable (lowest energy) field values were used in the comparison with the CoMFA coefficients (see text).

For points 2 and 4, an explanation of their favorable assessment by the CoMFA analysis might be that they lie in close vicinity of H atoms belonging to flexible side chains (Val370, Ser478) which are likely to reposition themselves in the flexible inhibitor–enzyme docking process. Moreover, the Ser478 OH hydrogen lies 2.6 Å from point 4 and, as mentioned above, it is reasonable to admit that the Ser478 OH donates an H-bond to the CN nitrogen of the fadrozole-like inhibitors. In the training set, the majority of these compounds feature a hydrogen or a cyano group in the *para*-position of the pendant phenyl ring. Thus, the steric field alone is seemingly sufficient for 'recognizing' these two substituents. The contradictions relative to points 1, 5, and 11 can be explained by considering the composition of the CoMFA training set. Particularly, the region of points (**1**) and 5 (Fig. 4) is occupied by some tetralone-like inhibitors which are not truly inactive, but just less potent with respect to the potent fadrozole-like compounds: therefore this active site area is considered unfavorable even if there is space available to host the inhibitors. The area around point 11 is really a forbidden zone of the protein occupied by only less potent compounds of the fadrozole series. However, these molecules being still more active than the tetralone-like ones, the zone around point 11 is estimated as sterically favorable by the CoMFA model.

A check of the above quantitative comparison procedure was attempted by randomly rotating the protein (fadrozole-derived model), with the effect of changing the orientation of the inhibitors relative to the active site, and calculating the new steric field energy values. These new values were then compared with the CoMFA coefficients and the match/mis-match points were recorded by applying the same selection criteria used for the above analysis. The random rotation was repeated ten times and the average score of matching points was 3.3 ($s = 1.7$) with 70% of the results ≤ 4.

Discussion

As proposed by Kim,[8] CoMFA and homology building should complement each other towards the common goal of rational molecular design. While CoMFA discovers relationships between ligands' 3D properties and their binding affinities, homology building allows one to construct a plausible model of the protein 3D structure. Thus, their combined use permits us to model the ligand–protein interaction from either the ligand side or the protein side, respectively. Building a bridge between the two methods implies that the intrinsic limitations of one approach can be, in principle, compensated by the strengths of the other.

A severe limitation of CoMFA and other related 3D-QSAR techniques which use only ligand structures[36,37] is that they might fail in predicting the activity of a new compound, if many of its 3D properties reside out of the space occupied by the training set molecules, or if the predicted compound is structurally very different from those used to derive the model. When such problems are encountered, the modeler should be aware that CoMFA is of little utility and that much more information might come from the homology-built model. On the other hand, although accurate predictions of binding affinity are probably not achievable

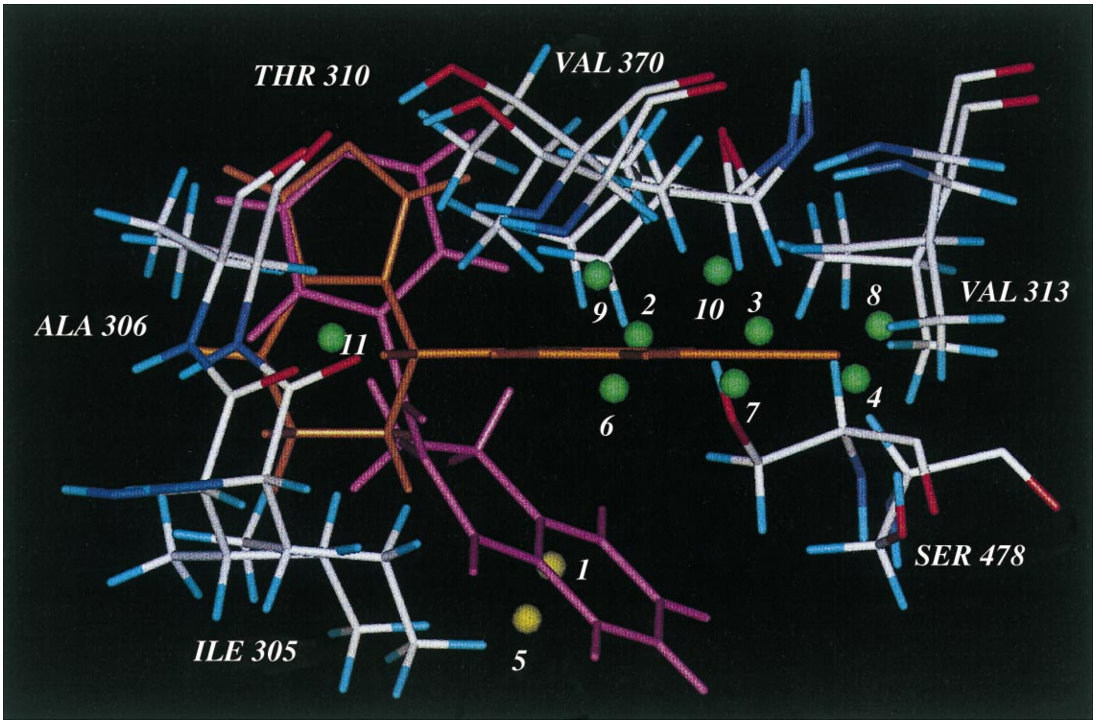


Figure 4. Inhibitors **1** (orange) and **2** (magenta) in the reciprocal orientation obtained from the molecular dynamics simulations. The 11 grid points in which the comparison between the steric CoMFA coefficient values (green positive, yellow negative; model no. 3) and the steric protein field values was carried out are also shown.

through calculations on homology-built protein–ligand complexes, docking models can nevertheless provide valuable qualitative hints for ligand design. Compared with ligand-based 3D-QSAR approaches, structure-based modeling methods are particularly attractive for their explicit representation of the interactions involved in the protein–ligand binding process. Looking at a putative 3D model of a binding site enables any medicinal chemist, including those unfamiliar with computational methods, to make hypotheses and get ideas about which molecules should be synthesized next.

One of the strengths of CoMFA is the possibility to rationalize the SAR data of a (theoretically unlimited) number of compounds. In this method, checks of self-consistency are performed through robust statistical tests such as cross-validation, data scrambling, random realignment, and so on. These procedures eliminate (or reduce) a biased evaluation of the model. Assessing the self-consistency of a homology-built model is more complicated: the modeler must take care to examine all the available experimental data to confirm or contradict the model (site specific mutagenesis, sequence hydropathic profile, SAR of ligands). Analyses of SAR are generally performed by selecting some ligands, docking them into the binding site one at a time and explaining the rank of their potencies either qualitatively (looking at favorable and unfavorable contacts) or quantitatively (one of the most challenging tasks in computational chemistry).

A direct comparison between CoMFA graphical outputs and the modeled binding site environment is fast, because PLS coefficients (or the corresponding contour maps) embody a large amount of SAR data. It is an attempt to integrate both the quantitative (CoMFA) and the qualitative (graphic) points of view of SAR analyses carried out on the same biological system. The results of such a comparison may be of worth even in the case of non-consistency of the two models. In fact, as shown above in regard to aromatase inhibitors, the mis-matching values in some of the grid points lead to a more in-depth consideration of some particular aspects of both the active site features and the CoMFA outcomes. Finally, the quantitative approach to the comparison, i.e., the use of CoMFA coefficients and field values to assess the consistency of the models, goes towards the direction of reducing as much as possible elements of bias.

Indeed, comparing PLS coefficients and probe–protein interaction energies raises several methodological questions. The accuracy of a homology-built model cannot be easily estimated before the structure of the target protein is determined experimentally. Analogously, any CoMFA model is inherently under-determined owing to the fact that it is derived from a finite number of molecules. The binding site structure is typically computationally frozen in just one conformation. Considering the low resolution of the grid used to compute molecular fields (1 or 2 Å), and the sensitivity of the Lennard–Jones potential function to the 12th power of the interatomic distance, it is not clear how to classify the values of the protein fields (i.e., which is the threshold value which discriminates between sterically accessible and forbidden regions?). Moreover, comparisons of PLS electrostatic coefficients versus corresponding probe–protein electrostatic energies have not yet been investigated.

Further research is needed to address all the above questions. As a consequence, we believe that consistency of CoMFA and homology building results over all the grid points contributing to the CoMFA model (100% of matches) is not a realistic expectation. In our case study (see Table 2), six out of 11 points were considered as matches (points 3, 6, 7, 8, 9, and 10), and the remaining five points were regarded as actual inconsistencies (points 1, 2, 4, 5, and 11). Three of the mis-matches were related to subtleties in the alignment and in the composition of the CoMFA training set: does that mean that the employed alignment was incorrect? First, it is worth noting that only the structures of the leads **1** and **2** were subjected to docking into the active site; the remaining compounds of the training set were aligned onto these two templates through an atom-by-atom fitting. While this procedure is perhaps questionable on physico-chemical grounds (low affinity ligands should not bind similarly to potent high affinity ligands), it is a common practice, within most of the 3D-QSAR methods,[36,37] to superimpose high and low potency (even totally inactive) compounds by overlapping pharmacophoric features or atoms belonging to a common scaffold. Such 'unrealistic' overlays are in fact extremely effective, from the statistical viewpoint, in highlighting those 3D features of the ligands responsible for the differences in potency. An important example was given by Klebe and Abraham, who demonstrated that experimentally derived alignments can yield CoMFA models worse than those obtained by superimposing the ligands about their common backbone.[38]

In order to evaluate whether the degree of consistency reported in Table 2 was statistically significant, the active site was randomly rotated about the aligned molecules 10 times and the number of matching points was recorded for each run of comparison. It was gratifying to observe that, throughout the 10 runs, the average rate of agreement found by chance was lower than the number of matches obtained for the original orientation of the protein. Indeed, the study of different ligand–target systems by means of the approach presented here will be necessary to fully validate the methodology. However, as a final observation, we want to point out that this work is an attempt to find a way of comparing quantitatively CoMFA and graphic models of ligand–target interactions. The search for consistency of different models describing the same phenomenon should be a primary goal for researchers using such models, both for theoretical and for practical reasons. The fact that we obtained not much more than an objective 50:50 agreement between the two models developed for the aromatase–inhibitor interactions has to be critically evaluated, and reasons for the inconsistencies have to be thoroughly investigated. However, this might pertain to the future development of the method that, as shown above, can have the ability to bring out a different point of view on the study of molecular systems of pharmacological interest.

Conclusions

We have investigated the aromatase–non-steroidal inhibitor interactions using CoMFA and protein homology modeling approaches. A quantitative comparison between CoMFA coefficients and probe–protein interaction energies was made at each point of the CoMFA lattice employed to derive the PLS equation. This comparison permitted us to evaluate how convergent were the results obtained from the CoMFA and enzyme–inhibitor docking studies. In our opinion, besides its potential use in SAR and ligand design studies, if suitably developed, this 'building bridge approach' might also be considered as a mutual validation technique for ligand-based and target-based three-dimensional models of ligand–target interactions.

Experimental

Protein homology modeling was performed on a Silicon Graphics Indigo2 workstation, using the MSI software packages INSIGHT II and DISCOVER[39] including heme29.frc.[40] The X-ray coordinates of resolved protein structures were obtained from the Brookhaven Protein Data Bank (PDB);[41] heme complex coordinates were extracted from the Cambridge Structural Database (CSD).[42] The human aromatase sequence was taken from the SWISSPROT Data Bank,[43] and the sequence alignment (Fig. 1) was generated using version 2.1 of the Homology module of Insight II.

The refinement of the two active site–inhibitor docking models was carried out by initially submitting them to 5000 steps of steepest descent minimization, then to conjugate gradient until the convergence of 0.1 kcal mol^{-1} Å^{-1}. Temperature constant molecular dynamics simulations (100 ps; $T=310$ K; time step 1 fs) were performed on both complexes. Only the distances between the heme iron and the coordinating atoms (Cys437 S, pyrrole nitrogens and inhibitors' N) were fixed, while the remaining parts of the inhibitor molecules were enabled to move in order to find the best conditions for the non-bonded interactions. The simulations were run for 40 ps with temperature control by direct velocity scaling to allow equilibration of the temperature, then for 60 ps with temperature control via coupling to a temperature bath; 600 conformations were sampled between 40 and 100 ps. From the E_{tot} versus time plots (not shown), it was possible to see that equilibration was actually achieved between 40 and 50 ps, thereby, for each model, the final conformation was obtained as the average from the last 500 ones, and it was energy minimized following the same procedure described above (conjugate gradient convergence at 0.01 kcal mol^{-1} Å^{-1}).

The CoMFA analysis was carried out by means of the SYBYL software;[44] the detailed description of molecular modeling, CoMFA column generation, and statistical treatment of the data is reported elsewhere.[14]

Acknowledgements

The authors wish to thank Yvonne C. Martin for the thoughtful suggestions during the preparation of the manuscript and for the critical reading of the final version of the paper. Scopes and limitations of comparing CoMFA contour maps and binding sites have been investigated by Y. C. Martin at Abbott Laboratories since 1989. Her research on this topic (most unpublished) has provided the theoretical background of the present work. This work was supported by grants of MURST (M. R. and A. C.), University of Bologna (M. R. and A. C.), and Regione Campania-Assessorato Ricerca Scientifica (G. G. and E. N.).

References and Notes

1. Hammett, L. P. In *Physical Organic Chemistry. Reaction Rates, Equilibria and Mechanisms*, 2nd ed.; McGraw-Hill Kokagusha Ltd.: Tokyo, 1970.
2. Levinthal, C. *Sci. Am.* **1966**, *214*, 42.
3. Hansch, C.; Klein, T. E. *Acc. Chem. Res.* **1986**, *19*, 392.
4. Selassie, C. D.; Klein, T. E. In *3D QSAR in Drug Design: Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 257–275.
5. Ortiz, A. R.; Pisabarro, M. T.; Gago, F.; Wade, R. C. *J. Med. Chem.* **1995**, *38*, 2681.
6. Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. *Quant. Struct.–Act. Relat.* **1993**, *12*, 9.
7. Cramer, R. D. III; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.* **1988**, *110*, 5959.
8. Kim, K. H. In *3D QSAR in Drug Design: Recent Advances*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer/ESCOM: London, 1998; pp 233–255.
9. Johnson, M. S.; Srinivasan, N.; Sowdhamini, R.; Blundell, T. L. *Crit. Rev. Biochem. Mol. Biol.* **1994**, *29*, 1.
10. Banting, L.; Nicholls, P. J.; Shaw, M. A.; Smith, H. J. *Progr. Med. Chem.* **1989**, *26*, 253.
11. Banting, L. *Progr. Med. Chem.* **1996**, *33*, 147.
12. O'Reilly, J. M.; Brueggemeier, R. W. *Curr. Med. Chem.* **1996**, *3*, 11.
13. Recanatini, M. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 74.
14. Recanatini, M.; Cavalli, A. *Bioorg. Med. Chem.* **1998**, *6*, 377.
15. Laughton, C. A.; Zvelebil, M. J. J.; Neidle, S. *J. Steroid Biochem. Mol. Biol.* **1993**, *44*, 399.
16. Koymans, L. M. H.; Moereels, H.; Vanden Bossche, H. *J. Steroid Biochem. Mol. Biol.* **1995**, *53*, 191.
17. Graham-Lorence, S.; Amarneh, B.; White, R. E.; Peterson, J. A.; Simpson, E. R. *Protein Sci.* **1995**, *4*, 1065.
18. Chang, Y.-T.; Loew, G. *Protein Eng.* **1996**, *9*, 755.
19. Lewis, D. F. V. In *Computer-Assisted Lead Finding and Optimization. Current Tools for Medicinal Chemistry*; van de Waterbeemd, H., Testa, B., Folkers, G., Eds.; Verlag Helvetica Chimica Acta: Basel, 1997; pp 335–354.
20. Lozano, J. J.; López-de-Briñas, E.; Centeno, E. B.; Guigó, R.; Sanz, F. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 395.
21. Burke, D. F.; Laughton, C. A.; Neidle, S. *Anti-Cancer Drug Des.* **1997**, *12*, 113.
22. Szklarz, G. D.; Halpert, J. R. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 265.
23. Höltje, H.-D.; Fattorusso, C. *Pharm. Acta Helv.* **1998**, *72*, 271.

24. Raag, R.; Li, H.; Jones, B. C.; Poulos, T. L. *Biochemistry* **1993**, *32*, 4571.

25. Hasemann, C. A.; Ravichandran, K. G.; Peterson, J. A.; Deisenhofer, J. *J. Mol. Biol.* **1994**, *236*, 1169.

26. Ravichandran, K. G.; Boddupalli, S. S.; Hasemann, C. A.; Peterson, J. A.; Deisenhofer, J. *Science* **1993**, *261*, 731.

27. Cupp-Vickery, J. R.; Poulos, T. L. *Nature Struct. Biol.* **1995**, *2*, 144.

28. Cole, P. A.; Robinson, C. H. *J. Med. Chem.* **1990**, *33*, 2933.

29. McGuffin, L. J.; Bryson, K.; Jones, D. T. *The PSIpred Protein Structure Prediction Server*; Protein Bioinformatics Group, Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK; http://globin.bio.warwick.ac.uk/psipred

30. Lewis, D. F. V.; Moereels, H. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 235.

31. Furet, P.; Batzl, C.; Bhatnagar, A.; Francotte, E.; Rihs, G.; Lang, M. *J. Med. Chem.* **1993**, *36*, 1393.

32. van Steen, B. J.; van Wijngarden, I.; Tulp, M. T. M.; Soudjin, W. *J. Med. Chem.* **1994**, *37*, 2761.

33. Krystek, S. R. Jr.; Hunt, J. T.; Stein, P. D.; Stouch, T. R. *J. Med. Chem.* **1995**, *38*, 659.

34. Tong, W.; Collantes, E. R.; Chen, Y.; Welsh, W. J. *J. Med. Chem.* **1996**, *39*, 380.

35. In order to assess the collinearity between the fields for both model nos **1** and **2**, we calculated the squared correlation coefficients between the first three latent variables (X_LATENT in the CoMFA output) obtained from the steric X-block and the first three latent variables obtained from the electrostatic X-block. The r^2 values were as follows: for the model No. 1: X_LAT1(steric) versus X_LAT1(electrost): 0.962; X_LAT2(steric) versus X_LAT2(electrost): 0.519; X_LAT3 (steric) versus X_LAT3(electrost): 0.391; for the model No. 2: X_LAT1(steric) versus X_LAT1(electrost): 0.965; X_LAT2 (steric) versus X_LAT2(electrost): 0.671; X_LAT3(steric) versus X_LAT3(electrost): 0.208.

36. Greco, G.; Novellino, E.; Martin, Y. C. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; VCH: New York, 1997; Vol. 11, pp 183–240.

37. Greco, G.; Novellino, E.; Martin, Y. C. In *Designing Bioactive Molecules. Three-Dimensional Techniques and Applications*; Martin, Y. C., Willett, P., Eds.; American Chemical Society: Washington, DC, 1998; pp 219–252.

38. Klebe, G.; Abraham, U. *J. Med. Chem.* **1993**, *36*, 70.

39. INSIGHT II (Ver. 95.0) and DISCOVER (Ver. 2.9.5), Biosym/MSI, San Diego CA, USA, 1995.

40. Kemmritz, K. Heme29 Force Field, Ph.D. Thesis, Freie Universität Berlin, Berlin, Germany, 1994.

41. Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F. Jr.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, T. *J. Mol. Biol.* **1977**, *112*, 535.

42. Allen, F. H.; Bellard, S.; Brice, M. D.; Cartwright, B. A.; Doubleday, A.; Higgs, H.; Hummelink, T.; Hummelink-Peters, B. G.; Kennard, O.; Motherwell, W. D. S.; Rodgers, J. R.; Watson, D. G. *Acta Cryst.* **1979**, *B35*, 2331.

43. Bairoch, A.; Boeckman, B. *Nucl. Acids Res.* **1994**, *22*, 3578.

44. SYBYL Molecular Modeling System (Ver. 6.4), Tripos Ass., St. Louis, MO, USA, 1997.