Prediction of Enzyme Binding: Human Thrombin Inhibition Study by Quantum Chemical and Artificial Intelligence Methods Based on X-ray Structures[#]

G. Mlinsek,[†] M. Novic,[‡] M. Hodoscek,[†] and T. Solmajer*[,†,§]

Laboratory of Molecular Modeling and NMR Spectroscopy and Laboratory of Chemometrics, National Institute of Chemistry, Hajdrihova 19, POB 660, 1001 Ljubljana, Slovenia, and Lek, d.d., Research and Development, Celovska 135, 1526 Ljubljana, Slovenia

Received July 30, 1999

Thrombin is a serine protease which plays important roles in the human body, the key one being the control of thrombus formation. The inhibition of thrombin has become a target for new antithrombotics. The aim of our work was to (i) construct a model which would enable us to predict Ki values for the binding of an inhibitor into the active site of thrombin based on a database of known X-ray structures of inhibitor−enzyme complexes and (ii) to identify the structural and electrostatic characteristics of inhibitor molecules crucially important to their effective binding. To retain as much of the 3D structural information of the bound inhibitor as possible, we implemented the quantum mechanical/molecular mechanical (QM/MM) procedure for calculating the molecular electrostatic potential (MEP) at the van der Waals surfaces of atoms in the protein's active site. The inhibitor was treated quantum mechanically, while the rest of the complex was treated by classical means. The obtained MEP values served as inputs into the counter-propagation artificial neural network (CP-ANN), and a genetic algorithm was subsequently used to search for the combination of atoms that predominantly influences the binding. The constructed CP-ANN model yielded Ki values predictions with a correlation coefficient of 0.96, with Ki values extended over 7 orders of magnitude. Our approach also shows the relative importance of the various amino acid residues present in the active site of the enzyme for inhibitor binding. The list of residues selected by our automatic procedure is in good correlation with the current consensus regarding the importance of certain crucial residues in thrombin's active site.

1. INTRODUCTION

Thrombin is one of the key enzymes in the blood coagulation system.[1] It plays many important roles, but the key one is the control of thrombus formation. The inhibition of thrombin has become a prime target for antithrombotics control of thrombosis-linked pathological states. New and better antithrombotics may contribute significantly to improving the health of the population. The criteria that an ideal thrombin inhibitor should fulfill are inhibition of thrombosis without affecting haemostasis, a long half-life, oral bioavailability, no serious side effects, and a large therapeutic range. A variety of inhibitors have been synthesized,[2] but only some of them are beyond Phase I clinical trial. Although many have very good affinity constants, they do not pass other tests such as bioavailability, toxicity, etc.

The active site cleft of the thrombin surface consists of the catalytic triade (His 57, Asp 102, and Ser 195) and three specificity pockets S1, S2, and S3 which are schematically represented in Figure 1. These pockets ensure a high specificity for the substrate that thrombin cleaves. The S1 site is lined by the residues Ala 190 and Gly 226 with the Asp 189 positioned at the bottom of the specificity site. The latter is responsible for the high specificity of thrombin for the cleavage of substrates after Arg and Lys residues in the sequence. The S2 site is constructed by the residues Tyr 60a−Trp 60d which outline a small lipophilic pocket where proline and other small hydrophobic groups of the inhibitor nestle. The S3 specifity site is modeled by the chain segments Trp 96−Leu 99 and Arg 173−Ile 176 with addition of hydrophobic side chain of Trp 215.

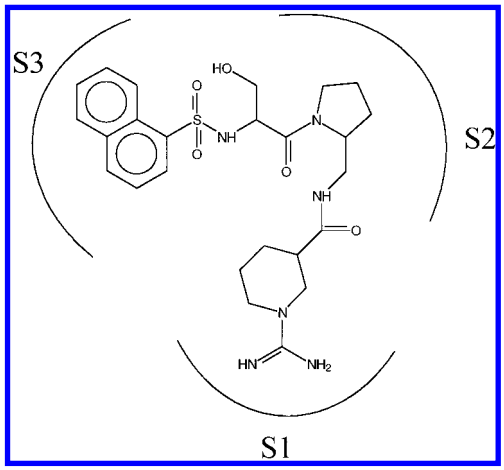

Figure 1. Schematic representation of thrombin active site with bound inhibitor BMS-186282 (ref 25) and labeled sites of interaction on the protein surface S1−S3.

* Corresponding author phone: +386-01-4760-277; e-mail: tom.solmajer@ki.si.

[†] Laboratory of Molecular Modeling and NMR Spectroscopy, National Institute of Chemistry.

[‡] Laboratory of Chemometrics, National Institute of Chemistry.

[§] Lek, d.d., Research and Development.

[#] This paper is dedicated to Professor M. Randic (Des Moines) in honor of his 70th birthday.

10.1021/ci000162e CCC: $20.00
Published on Web 05/04/2001

Table 1: List of Thrombin−Inhibitor Complex Structures Used in This Study with References and Binding Constants of the Compound to Thrombin Ki

no.	PDB label	Ki (nM)	refs
1	7KME	4.60	20
2	8KME	3.90	
3	1BB0	0.64	21
4	1BA8	0.	
5	1CA8	−0.49	
6	1TMB	2.26	22
7[a]	1A2C	−0.30	23
8	1A4W	3.08	24
9	1A5G	−2.15	25
10	1A46	3.30	
11	1A61	1.38	
12	1B5G	1.00	
13	1BMN	0.56	26
14	1BMM	1.90	
15	1DWB	2.48	27
16	1DWC	1.59	
17	1DWD	0.82	
18	1HDT	1.24	29

[a] IC_{50} value available for this compound only.

Structure-based drug design has become a well-established approach for the discovery of novel bioactive molecules.[3] The crystal structures of the ligand−macromolecular receptor complex are central to structure-based drug design. In the present work we explore the possibility of an in-depth computational analysis of experimental X-ray structures which have become increasingly available in structural database.[4] To test our general hypothesis that the use of chemometric methodology could yield the missing element in the puzzle of correlating biological activity with known three-dimensional molecular structure we applied the method to the specific case of a series of thrombin−inhibitor complexes. A thorough search of the protein database for thrombin in the complex with an inhibitor has given us an opportunity to perform a balanced study of the possibility of predicting the binding constant of the ligand since the binding constant data extends over 7 orders of magnitude. Approaches classified as artificial intelligence such as artificial neural network (ANN) algorithms[5,6] and genetic algorithm (GA)[7] were also used (see Methods section for detailed description) to select the amino acid residues of the protein that play the most important role in the binding of the inhibitor. Such an algorithmic methodology enables an objective and more quantitative detection of properties important for binding and lends itself as a viable alternative and complementary method for use in rational structure-based drug design. Other methods to predict binding affinity have been reported in the literature: docking calculations using various scoring functions,[8] 3D QSAR calculations,[9] COMFA/COMSiA analyses,[10,11] computational approaches employing thermodynamic integration approaches,[12,13] etc. The approach described in this work makes use of the chemical information that defines the most important enzyme−ligand interactions (i.e. the properties of ionic bonds, hydrogen bonding and hydrophobic sites on the molecular structure are represented in the relative electrostatic field at the van der Waals surface[14−17]). This information is subsequently used to reduce the structural representation of the three-dimensional enzyme−ligand complex at atomic resolution given by the atomic coordinates of the complex inhibitor−enzyme and link it to the biological activity data. The counter-propagation artificial neural networks models which are generated use MEP data as inputs and yield target biological activity as outputs.[18,19] By using such carefully selected MEP data computed at the van der Waals surface interface of the experimentally determined ligand−enzyme complex the surface points of every ligand molecule can be mapped onto the same reference surface. Thus it is possible to compare the interactions between the ligand and enzyme active site equally and to describe them in a quantitative way. We show that this method faithfully reproduces the same important amino acid residues which were previously derived by X-ray protein crystallography, kinetic studies, and site directed mutagenesis.

There is a conceptual parallelism between the results obtained by the widely used COMFA method[10,11] which employs PLS formalism and the artificial intelligence approach (ANN+GA) used in the present work to filter MEP data. Both methods use the MEP representation of the ligand−enzyme interaction and subsequently use statistics ranking and selection to determine the key parts of this interaction. The main advantages of our algorithm are as follows: the possibility of using a more precise QM/MM approach for MEP computation without the need for any parametrization such as scaling or "cut off" parameters; the introduction of additional molecular descriptors such as logP are not necessary; and important inhibitor−enzyme interactions are obtained in an automatic fashion.

Previously, Schramm et al.[18] used a similar approach consisting of MEP computation and subsequent ANN training to select the features of the enzyme−inhibitor contact surface. They introduced a spherical reference surface with a diameter slightly larger than the largest inhibitor molecule and used it to map the ligand−enzyme interactions in order to ensure that similar regions on different molecules enter the same region on the neural network. We hypothesized that our use of experimentally determined contact surfaces of the enzyme−inhibitor complex would provide a superior database compared to other mapping procedures.

However, in addition to the novel mapping procedure we show that the structural representation by MEP values at all points at the van der Waals surface of the complexed ligand in the could be further reduced and optimized by the use of a genetic algorithm. Thus, a relatively small number of points chosen by the automatic procedure on the contact surface in the enzyme−ligand complex can be generated which efficiently and precisely correlates the biological activity with structure of the inhibitor molecule for subsequent use in the design of novel inhibitors (work in progress in this laboratory).

2. METHODS

The method described below consists of the following steps:

(a) Data Set Preparation. A complete search of the Brookhaven PDB for thrombin structures was first carried out. Thirty enzymatic structures which present thrombin complexed with an inhibitor were found, and of these 18 were found to be reversibly bound inhibitors.[20−29] The rest of inhibitors were covalently bound and thus were omitted from this study in order to exclude the variance in the underlying biochemical mechanism of the series.

Binding affinity constants were available for 17 of the enzyme−inhibitor complex strutures, while only the IC_{50} was available in the single remaining structure. The chemical structures of the inhibitors were heterogeneous in terms of size, with between 18 and 99 heavy atoms. The complexes did not belong to a homologous series either. However, by careful analysis of the X-ray structural data of the series it was found that all inhibitors were bound to the same site on the thrombin enzymatic surface. The inhibitors structures were quite diverse although the majority of them were peptidomimetics containing three or four structural units one of which was a cyclic peptide derivative.

To validate (i) the role of different crystallization procedures used in the experimental data and (ii) how the inhibitor's binding influences the geometry of the active site in the inhibitor−thrombin complex structure, we superimposed all enzyme structures onto the reference structure (structure 1BMN, ref 25 was arbitrarily chosen as a reference structure). The superposition criterion was that only heavy atoms of amino acids forming beta sheets that we found to be the most burried and thus least vulnerable to the influence of the solvent surroundings were superimposed. RMS values obtained for all heavy atoms in the superposition list (85 residues and 686 heavy atoms in total) ranged from 0.51 to 0.77 Å. All inhibitors were at least in part anchored to the thrombin's active site pockets S1−S3 and were similarly oriented in the active site cavity (some of inhibitors with a more complex structure were partly protruding out of the bounds of these pockets). Thus, on the basis of RMS changes introduced by inhibitor binding it was possible to conclude that binding of reversible inhibitors to the active site of the thrombin enzyme exerts only a slight influence on the geometry of the active site and that the possibility of a significant conformational change was found to be almost negligible.

First, all atoms on the active site surface of the protein that were less than 10 Å away from any atom of any on the inhibitors were selected. Atoms at the protein surface are more important for interactions with the inhibitor than other atoms. The van der Waals surface of these atoms was determined and about 30−40 points per atom on the surface selected. The points per atom on the van der Waals surface were selected automatically and randomly.[30] Such point density was found to be sufficient for the smooth representation of MEP values for each atom. The set of atoms on the enzyme surface are selected as a mathematical union of key atoms for each individual inhibitor interaction with the enzyme. This also fulfills the requirement that each MEP vector has the same length as that needed for the AI procedure. There were $N = 322$ atoms in this exhaustive list of contacts between the inhibitors and enzyme in the database and therefore between 11.000 and 12.000 coordinate points were determined to express the contact surface for each thrombin−inhibitor complex.

In our computational scheme we used the following computer program packages: Charmm/Gamess/Insight[30−32] and in house CP-artificial neural network and genetic algorithm packages.[6,7] The computational procedure which was implemented consisted of the following steps.

(b) MEP Computation. A standard QM/MM method[34] was utilized to compute the molecular electrostatic potential at surface points obtained as described above. All inhibitor atoms were treated quantum mechanically by the ab initio approach. Standard 6-31G basis set had been previously shown to be sufficently accurate to represent MEP values. The enzyme's electrostatic interactions were treated classically by including the parametrized atomic charges of the enzyme atoms in the one electron Hamiltonian of the complex for the Hartree−Fock-Roothan iterative computation of the inhibitor wave function/electron density. Thus, the enzyme is perturbing this density in a different way for each inhibitor. We did not go beyond this approximation due to computational expense.

(c) Preselection of the Computed MEP Values. The computed MEP values were ordered by size for each of the 30−40 values derived from $N = 322$ atoms and for each of the inhibitor−enzyme complexes. To reduce the number of input parameters for the artificial neural network to a practical size without losing out on accuracy of representation the points chosen for each of the atoms give a valid description of the computed electrostatic property at the contact surface. This was determined as follows: if all the MEP values for a given atom were of the same sign (positive or negative) the two points with the highest value were chosen; if for a given atom the MEP values were of different signs (positive and negative) the highest positive and the highest negative value were chosen. This choice of MEP representation for a given complex inhibitor−enzyme is somewhat arbitrary. However, the method in its present form appears rather robust since it operates on sufficiently small portions of the molecular surface where the variation in MEP values is not terribly large. Nonetheless, further experimental work is needed to optimize the choice of representative MEP values at the enzyme−inhibitor complex contact surface.

This procedure can be compared to the choice of representative points of Schramm et al.[18] These authors perform a uniform outward mapping in space of each inhibitor in which the MEP is computed—onto a spherical reference surface in order to obtain a sufficiently reduced number of points after the mapping. To reduce the number of points these authors introduced an adjustable parameter. In such a way information about the interactions at the inhibitor−enzyme contact surface is also retained.

(d) Application of an Artificial Neural Network Yields a Nonreduced Inhibitor−Enzyme Interaction Model. ANN Training Step. The MEP values for the vector, which represented 644 points at the contact surface with each atom being represented by two points served as inputs for a counter-propagation artificial neural network[33] (CP-ANN). This method was demonstrated previously to give better results than the similar back-propagation ANN if the number of items in the database is small as is the case with a small number of data for inhibitor−thrombin complex geometries in our database. It is important to reiterate that while the inhibitors were of varying molecular sizes they all had the same "size" contact electrostatic surface.

During the leave-one-out test the network was trained using the data for n-1 inhibitors and the inhibition constant value for the nth inhibitor was predicted. This procedure was repeated n times yielding the list of n predictions. To carry out the subsequent genetic algorithm (GA) procedure the inhibitors were then divided into three groups: (i) a training set comprising 9 inhibitors, (ii) a test set of 6 inhibitors, and (iii) a prediction set of 3 inhibitors. The dataset of 18

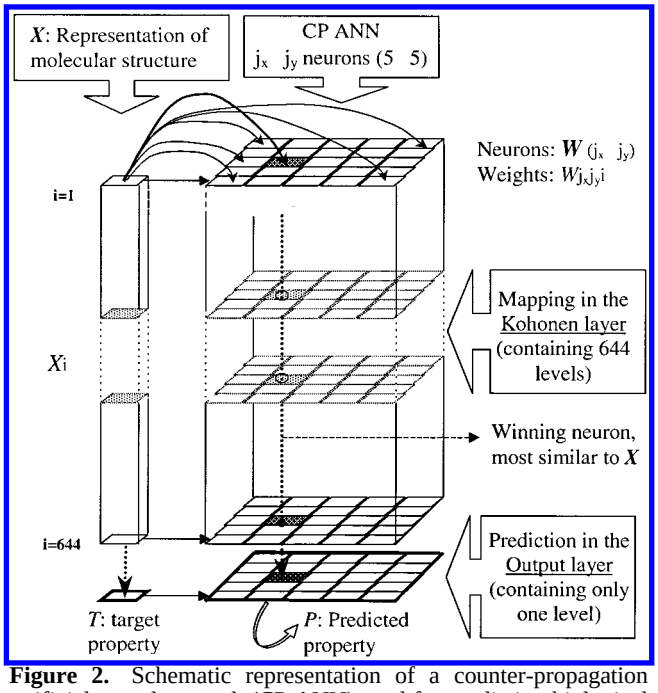

Figure 2. Schematic representation of a counter-propagation artificial neural network (CP ANN) used for predicting biological property (target *T*) from molecular structure (*X*) represented with 644 values of MEP as described in the text above. During the training the weights (W) of the winning neuron (the winning neuron is determined with the position (j_x, j_y) in the Kohonen layer according to the similarity between the object (X_i, $i = 1...644$) and the neuron's weights (Wj_xj_yi, $i = 1...644$)) are adapted step by step, in learning cycles, to the compounds of the object.

compounds is, in fact, a set of pairs of structure representation vectors (X_i, $i = 1, n$) and property values called targets (T).

Initially, three inhibitors were removed at random from the data set of 18 compounds in order to use them as a prediction set. Such a prediction set enabled us to test the quality of the final model. Furthermore, it should not influence either the determination of CP ANN model parameters such as number of training epochs, number of neurons, maximal and minimal learning rates, etc., or the next step of the procedure in which the selection of variables with GA is performed (step (e) below).[35] Once the compounds for the prediction set were removed, the remaining 15 compounds were divided into a training set and a test set taking into account that the training set should contain as many compounds as necessary to encompass the entire information space, i.e., all structural and property coordinates. In the present work the selection of compounds in the training set is based on the results of leave-one-out cross-validation (LOO CV), which measures the quality of the model itself but is too time-consuming to be performed several thousand times in a GA procedure.

The CP-ANN procedure was performed on the training set of nine molecules. The inhibition constants in the test set of $N = 6$ compounds were predicted by the network as described below. During the training the weights of the winning neuron (for definition see legend to Figure 2) and close neighbors were corrected in small steps in such a way that in the last cycle they are completely adapted to the input object. The correction of weights in the output layer yield similarity to the objects' target values in the training cycles. Thus, by training the network we obtained a model whose prediction ability is determined by checking it with the test objects.

ANN Testing Step. The testing procedure is as follows: each test object *X* first places (Figure 2) in the upper (Kohonen) layer the most similar neuron. The corresponding weight at the position j_x, j_y in the lower (output) layer gives the predicted target value. The difference between target *T* and the predicted property *P* is squared and summed over all objects of the test set to give the PRESS (predicted residual errors sum of squares) measure of error. The regression equation linking theoretical and predicted properties is also determined. The correlation coefficient *R* derived from this equation represents the quality measure of the proposed model.

(e) Representation Reduction by Use of GA Approach. In the last step of the present procedure the complete contact surface for interaction of all inhibitors in the data set with the enzyme was reduced to its most relevant part by the use of a genetic algorithm. A genetic algorithm consists of three basic processes mimicking Darwinian evolution: crossover, mutation, and selection. In the crossover step, new chromosomes are generated by mixing those of the parents. In the mutation step, individual bits of the chromosome are exchanged at random, and in the selection step the best chromosomes are identified for the next round. This last step should yield the lowest possible number of points on the enzymatic contact surface which gives a satisfactory prediction for Ki for each individual inhibitor in the data set. The number of input parameters (644) representing the electrostatic interaction with 322 atoms of the enzyme determines the length of the chromosome in bits.

The following selection procedure was implemented: each bit in the allele could take a value of 0 or 1. A value of 1 was used to assign the MEP value at a point in the model, while the value of 0 described the lack of taking into account the MEP value in a point on the contact surface.

A pool of 100 chromosomes was tested using a random choice of the points on the surface. For each chromosome a CP-ANN model was obtained and simultaneously the reduction of vector length was performed. These vectors with reduced length in the training set of inhibitor−enzyme complexes were used for training the CP ANN. The quality of each such chromosome representation was again determined by PRESS in the test set. The best chromosomes were crossed over and mutated. In the new pool of 100 chromosomes the new pattern representation was tested for quality. This procedure was repeated 900 times, and thus in each generation a more representative pattern of contact points on the surface was obtained. Finally, the resulting optimized model CP-ANN incorporates the ability to predict the binding affinity of the inhibitor.

3. RESULTS AND DISCUSSION

The interactions of the inhibitor with the enzyme were computed for a series of inhibitors at points on their common surface in contact with the enzyme. Thus, the most important parts of the electrostatic, hydrophobic, and hydrogen bonded interactions of the inhibitor were mapped onto the active site surface of thrombin.

In Figure 4 the MEP plotted at the van der Waals surface of thrombin in complex with the inhibitor including the 3D structure as determined in ref 20 is shown for each of the points on the surface.

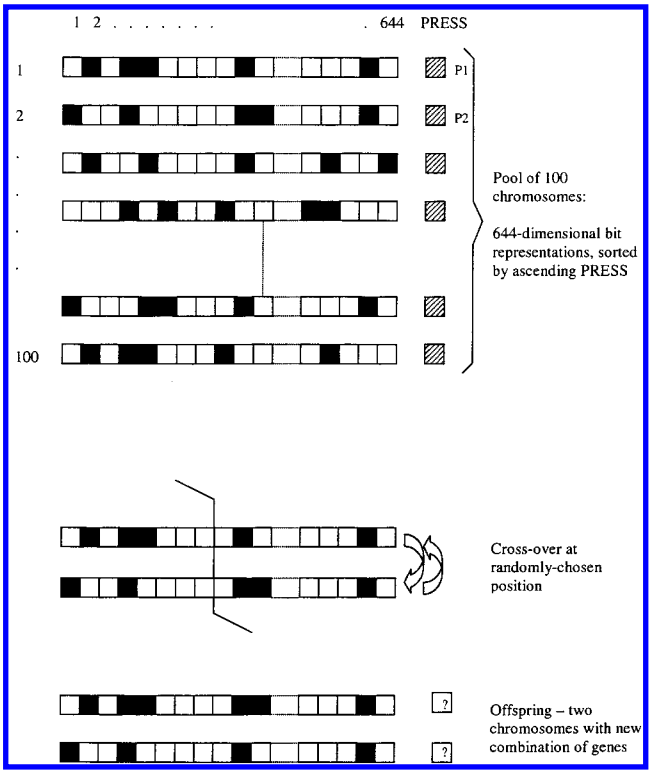

Figure 3. Schematic representation of GA procedure. The 644-dimensional chromosomes are ordered according to the optimization criterion obtained from a CP-ANN model (PRESS of p*K*i of six test compounds). The crossover event is shown on the first two chromosomes. Black and white squares indicate values 1 or 0 for individual choromosomes' compounds (genes). See text for additional explanation.

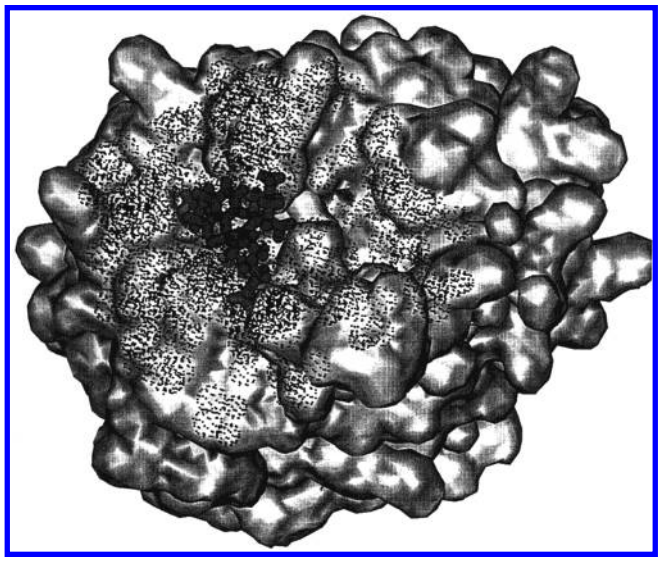

Figure 4. Molecular electrostatic potential map of an inhibitor on the van der Waals surface of thrombin (X-ray structure 1BMN, ref 25). The solvent accessible surface of thrombin is shown as solid light gray surface. The inhibitor position in the active site cleft is given by solid dark gray sticks. The MEP values are computed in the points shown in black.

The selection of MEP variables for the thrombin−inhibitor structure representation using a combination of ANN and GA algorithms (Figure 4) has two goals: (i) to reduce the dimension of the representation vectors (in our case, from 644 to 10−60) and (ii) to correlate the selected variables with a possible biochemical mechanism underlying the X-ray structure data which were used in this work.

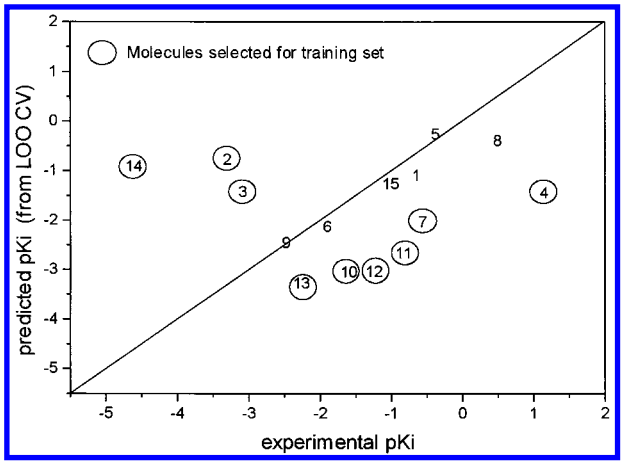

Figure 5. Selection of the training set on the basis of CV-LOO (leave one out) on the set of 15 molecules.

Following this hypothesis, the MEP values at these points were used in a subseqent chemometrical approach to yield the computed biological inhibition constant. The accuracy of the results shows that MEP values offer a valuable description of the relationship between structure and inhibition.

Model Optimization using the ANN Approach. Figure 5 shows how selection of the training set on the basis of CV-LOO (leave one out) was performed on a set of 15 molecules. To select the compounds for the training set in an optimal way, the plot of the predicted versus experimental inhibition constant of 15 compounds resulting from the LOO CV procedure was analyzed. Each predicted Ki was obtained from a different model as defined by the LOO CV procedure: an individual model was trained with *n*-1 compounds, namely all the compounds except the one whose property was being predicted. If the prediction error obtained is large, it can be assumed that the training set with *n*-1 compounds does not describe the properties of the excluded compound well. Consequently, a compound causing a large prediction error in LOO CV should be included in the training set.

The final model is thus sensitive to the elimination of such unique compounds from the training set. According to the prediction errors from LOO CV, nine compounds which produce the largest prediction errors if omitted from the training set (shown by circles in Figure 5) were put into the training set, while six compounds with smaller prediction errors were excluded and put into the test set in order to test the model during the process of training and determining CP ANN parameters. The same division of the compounds into training and test sets was used to calculate the optimization criterion in the GA step which was applied in order to select variables from the 644-dimensional structure representation vectors.

The results of training the CP-ANN model using molecules 1−9 are presented in Figure 6, the regression results for the test set of the six molecules in Figure 7, and the predicted activity of three compounds of the prediction set using the CP ANN model in Figure 8.

Thus, this intermediate CP ANN model generated from the nonreduced 644 dimensional structure representation gives predicted affinity constant values, Ki, for inhibitors in the test set with the reasonable correlation coefficient $R = 0.81$ and standard deviation of 0.48 (Figure 7).

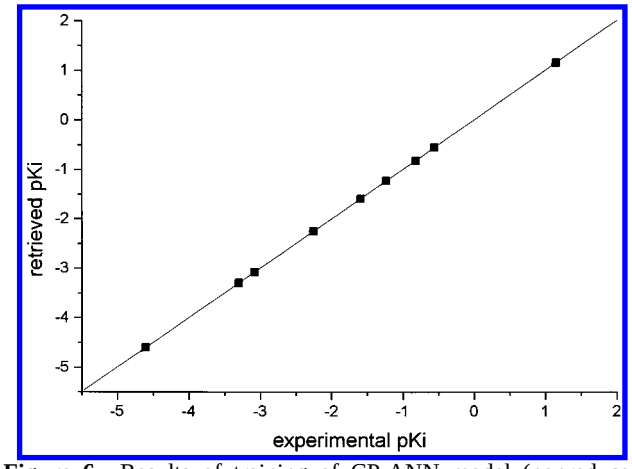

Figure 6. Results of training of CP-ANN model (nonreduced representation by vector length 644) using molecules 1−9.

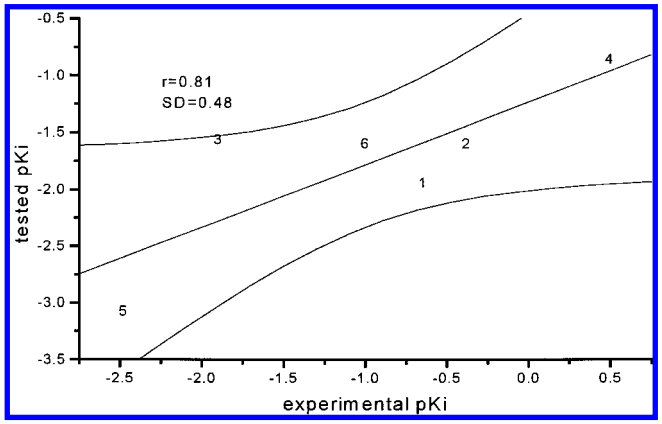

Figure 7. Regression results for the test set of molecules 10−15 (as in Figure 6).

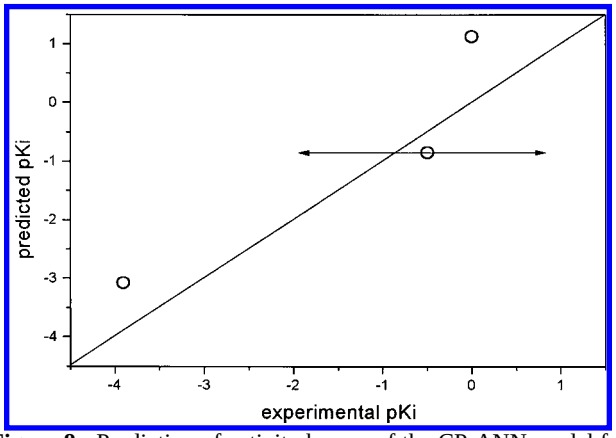

Figure 8. Prediction of activity by use of the CP-ANN model for molecules 16−18 (as in Figure 6).

It has to be stressed that the results shown in Figures 6−8 are not describing the final CP ANN model. They are given here to substantiate the suitability of the CP ANN model based on the nonreduced representation of the training and test set of compounds to yield the optimization criterion for the final CP ANN model obtained during the GA variable selection procedure. Consequently, the ANN parameters were not completely optimized in this intermediate step what is evident from comparison of Figures 6 and 7. The errors of the computed Ki values of compounds from the training set are too small in comparison with the errors of the test set of compounds. However, the ANN parameters are further changed during the GA procedure, and these parameters then describe the final CP ANN model for the reduced representation.

Reducing the Representation using GA. The second result is given by the genetic algorithm procedure which selects the protein amino acid residues with the largest influence on the prediction ability of the procedure, i.e., those having the largest influence on binding of the inhibitor to the enzyme's active site (Figure 9).

The optimization criterion in the GA algorithm was the correlation of the experimental Ki of six test compounds with those predicted by a CP ANN model which was trained with nine compounds. As shown in the legend to Figure 9, 10−57 variables (denominated as "bits", see also description of a genetic algorithm procedure in the Methods section) out of 644 were selected and are indicated by eight different symbols in the plot. The corresponding correlation coefficients are also shown in the legend. The best correlation was obtained if the molecular structure was represented by the 56 variables labeled with black circles in the sixth line.

The procedure is further illustrated with the diagram showing the propagation of GA from a random origin up to the 900th generation. The optimization criterion for determining the quality of the reduced structure representation, i.e., the correlation (R and RMS) between the experimental Ki of test compounds and those predicted by CP ANN, are shown for 900 generations. The convergence of the procedure is shown in Figure 10.

The last improvement in RMS is observed at the 773rd generation. Using the resulting reduced representation the final correlation results are given in eq 1. The correlation of experimental biological activity with computed bioactivity is given by the linear regression for the training set ($N = 9$) shown in Figure 11 and for the test set ($N = 6$) in Figure 12. The regression equation is

$$Y = 0.85086\ (\pm 0.10923)X - 0.16483\ (\pm 0.15149) \quad (1)$$

and the regression coefficient is $R = 0.97$ and $S_d = 0.26$.

The possibility of overtraining is excluded what is evident from errors in the training set (Figure 11) and test set (Figure 12).

The predictions for the nonbiased three compounds prediction set are given in Figure 13.

The reduced representation is expressed as a chromosome with the best fit value given by the PRESS value. In our case the chromosome with the best fit value of PRESS = 0.12 has 56 genes turned to 1, which means that 56 of the selected variables contain a large amount of information for predicting the biological activity of the compounds under investigation. These 56 variables were analyzed and found to originate from atoms present in 32 protein residues (out of the 259 present in the human α thrombin chain B). Not all residues determined by GA from the reduced representation were represented with equal frequency. Those that appeared the most frequently are the following: Tyr 60A, Pro 60B, Pro 60C,Trp 60D, Asn 60G, Phe 60H, Arg 73, Asn 95, Trp 96, Asn 98, Ser 171, Ile 174, Arg 175, Asp 189, Ala 190, Glu 217, and Lys 224.

The schematic representation of the thrombin amino acid residues list shown in Figure 14 and ascertained by our algorithmic procedure matches the consensus active site description given in relevant literature and in the Introduction well.[1,2] It should be stressed that in the list of selected

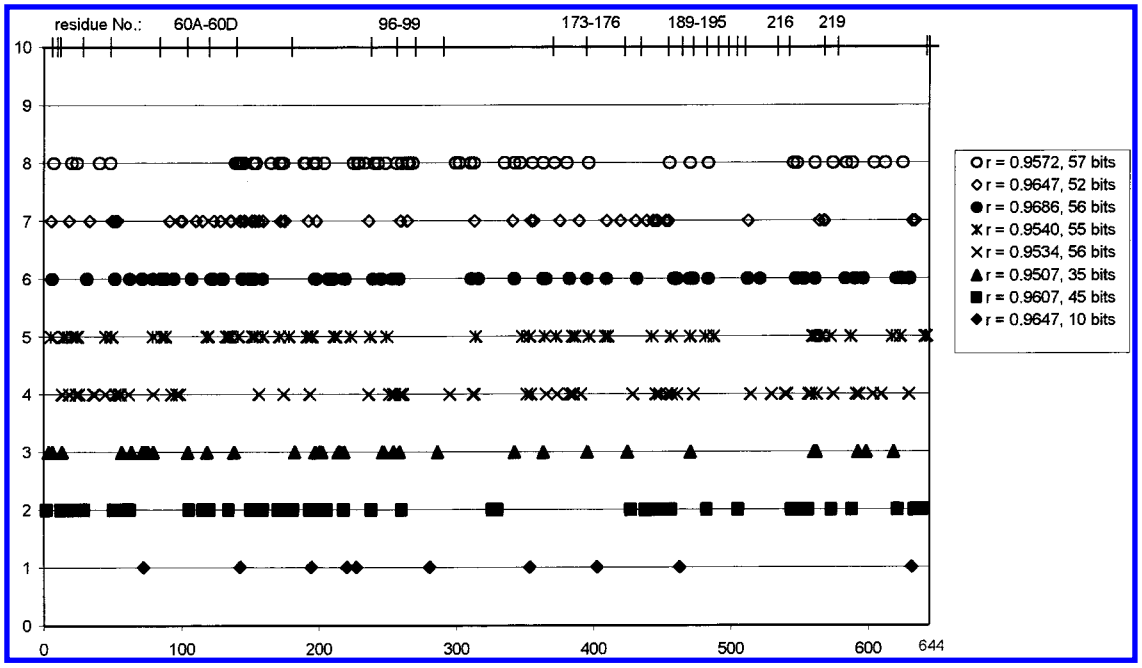

Figure 9. Results of representation reduction by the GA process: eight possible molecular structure representations with variables proposed by a GA started from eight different random origins.

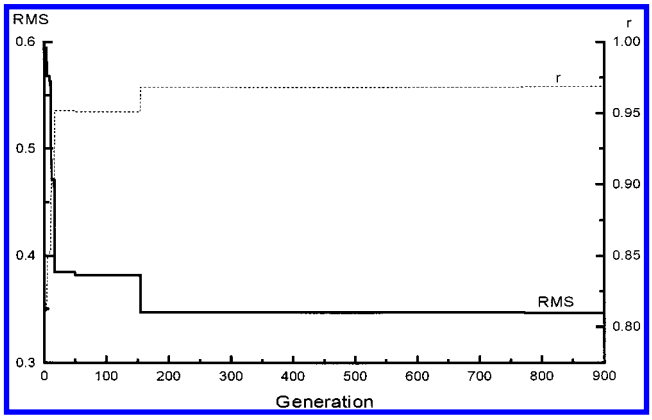

Figure 10. The propagation of GA started from one of the random origins up to the 900th generation. The optimization criterion for quality of reduced structure representation, i.e., the correlation (R and RMS) between experimental p*K*i of test compounds and those predicted by CP ANN, are shown for 900 generations. The last improvement of RMS is observed at 773rd generation.

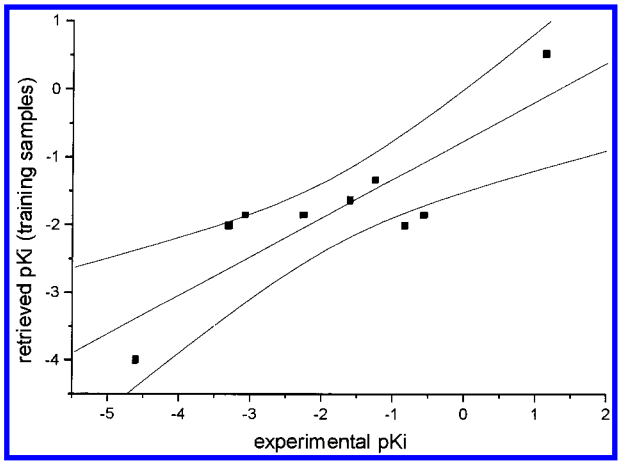

Figure 11. Results of training of reduced representation model on the training set molecules 1−9.

residues all three specificity pockets S1 to S3 are well represented.

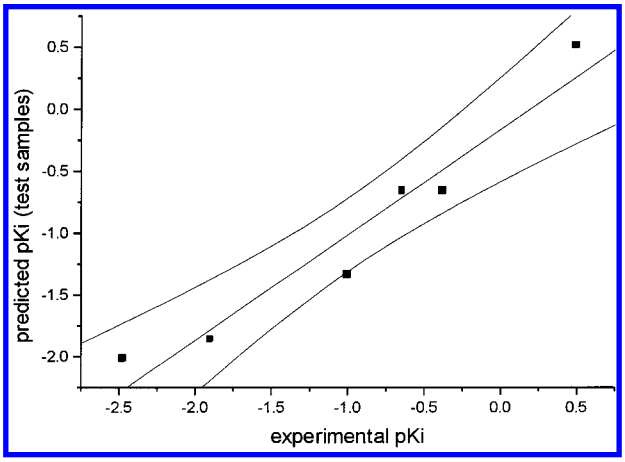

Figure 12. Regression results of reduced representation model for the test set of the molecules 10−15.

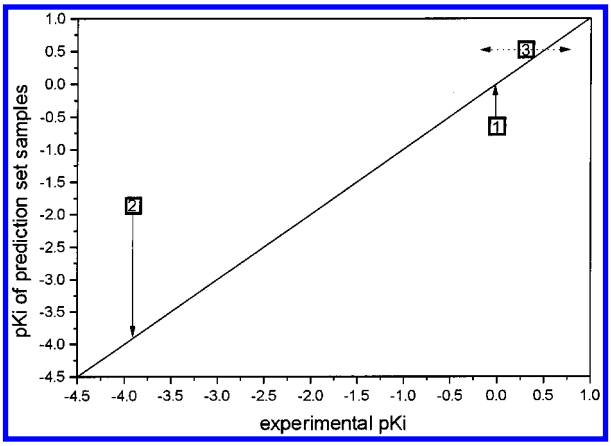

Figure 13. Prediction of activity of reduced representation model for molecules 16−18.

Furthermore, such a procedure appears to provide a model for correlating the binding affinities of the inhibitors with the structure of the enzyme−inhibitor complex.

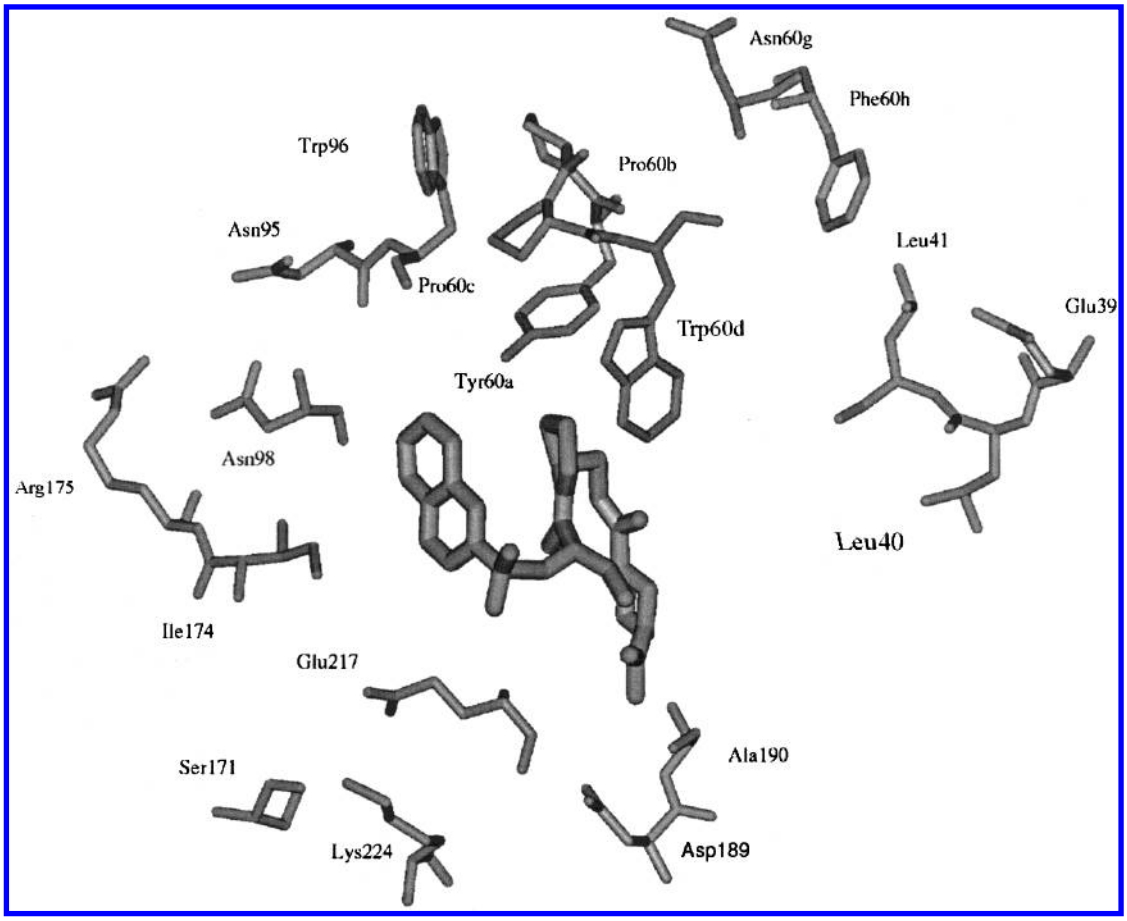

Figure 14. Amino acid residues in thrombin active site which were selected by the CP-ANN and GA algorithm to have most influence on the binding of inhibitors in the series (compare the list of residues given in refs 1 and 2).

CONCLUSIONS

The main objective of this work was to present a model which would enable a correlation of the three-dimensional structure of an enzyme−inhibitor complex with its inhibition constant. In the specific case of thrombin inhibition a key step in the blood coagulation cascade we tried to elucidate the structural features (amino acid residues in the primary sequence) of the enzyme's active site which determine the strength of inhibitor binding that could subsequently lead to more refined inhibitor structures.

These features enable one to make a priority list of electrostatic and structural requirements that need to be fulfilled in order to provide for good inhibition. We hypothesized that the electrostatic potential calculated at the surface of the protein and inhibitor includes sufficient information on forces contributing to the enthalpy of the free energy of binding. Our method is a departure from the method of Schramm et al. in that it eliminates the necessity of introducing a spherical reference surface[18] in order to enable the synchronized mapping of MEP values. Instead we opted to project the MEP values onto the same, carefully chosen points on the contact surface of inhibitor−enzyme complex. The constructed CP ANN model yielded predicted Ki values with a correlation coefficient of 0.96. Although the number of inhibitor−enzyme complexes used in the present study is limited we believe that the intrinsic quality and quantity of these data (we use experimentally determined three-dimensional atomic coordinates of ligands in the enzymatic active site environment only) has allowed reliable statistical conclusions to be made. Our work also indicates the importance of the algorithmically selected residues present in the active site of thrombin for the binding of the inhibitor. The list of residues obtained by our approach was found to be in good correlation with the current hypothesis[1,2] on the importance of amino residues at the molecular surface of thrombin's active site.

ACKNOWLEDGMENT

The Ministry of Science and Technology of Slovenia is thanked for financial support to G.M. We are grateful to Ms. Charlotte Taft for critical reading of the manuscript.

REFERENCES AND NOTES

(1) Stubbs, M.; Bode, W. A player of many parts: the spotlight falls on thrombin's structure. *Thrombosis Res.* **1993**, *69*, 1−58.
(2) Rewinkel, J. B. M.; Adang, A. E. P. Strategies and progress towards the ideal orrally active thrombin inhibitor. *Curr. Pharmac. Des.* **1999**, *5*, 1043−1075.
(3) *Structure-based Ligand Design;* Gubernator, K., Bohm, H-.J., Eds.; Wiley-VCH: 1998.
(4) Bernstein, F. C.; Koetzle, T. F.; Williams, G. J. B.; Meyer, E. F.; Brice, M. D.; Rodgers, J. R.; Kennard, O.; Shimanouchi, T.; Tasumi, M. Protein Data Bank−Computer-based Archival File for Macromolecular Structures. *J. Mol. Biol.* **1977**, *112*, 535−542.
(5) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*; Wiley-VCH: Weinheim, 1999.
(6) Zupan, J.; Novic, M.; Ruisánchez, I. Kohonen and Counterpropagation Artificial Neural Networks in Analytical Chemistry. *Chem. Intell. Lab. System* **1997** *38*, 1−23.
(7) Zupan, J.; Novic, M. Optimisation of structure representation for QSAR studies. *Anal. Chim. Acta* **1999**, *388*, 243−250.

(8) Bohm, H. J. Prediction of binding constants of protein ligands: A fast method for the prioritization of hits obtained from de novo design or 3D database search programs. *J. Comput. Aid. Mol. Des.* **1998** *12*, 309−323.
(9) Kubinyi, H. Hamprecht, F. A. and Mietzner, T.; Three-Dimensional Quantitative Similarity-Activity Relationships (3D QSiAR) from SEAL Similarity Matrices. *J. Med. Chem.* **1998**, *41*, 2553−2554.
(10) Bohm, M.; Sturzebecher, J.; Klebe, G.; Three-Dimensional Quantitative −Structure−Activity Relationship Analysis Using Comparative Molecular Field Analysis and Comparative Molecular Similarity Indices Analysis to Elucidate Selectivity Diffferences of Inhibitors Binding to Trypsin, Thrombin and Factor Xa. *J. Med. Chem.* **1999**, *42*, 458−477.
(11) Cramer, R. D.; DePriest, S. A.; Patterson, D. E.; Hecht, P. The Developing Practice of Comparative Molecular Field Analysis. In *3D QSAR in Drug Design Theory, Methods and Applications*; Kubinyi, H., Ed.; ESCOM: Leiden, 1993; pp 443−485.
(12) Jones-Hertzog, D. K.; Jorgensen, W. L. Binding affinities for sulfonamide inhibitors with human thrombin using Monte Carlo simulations with a linear response method. *J. Med. Chem.* **1997**, *40*, 1539−1549.
(13) Aquist, J.; Medina, C.; Samuelsson, J. E. A new method for predicting binding affinity in computer-aided drug design. *Prot. Eng.* **1994**, *7*, 385−391.
(14) Davis, A. M.; Teague, S. J.; Hydrogen bonding, Hydrophobic Interactions, and Failure of the Rigid Receptor Hypothesis. *Angew. Chem.* **1999**, *38*, 736−749.
(15) Oblak, M.; Randic, M.; Solmajer, T. Quantitative Structure−Activity Relationship of Flavonoid Analogues: 3. Inhibition of $p56^{lck}$ Protein Tyrosine Kinase. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 994−1001.
(16) Kuntz, I. D.; Chen, K.; Sharp, K. A.; Kollman, P. A. The maximal affinity of ligands. *Proc. Natl. Acad. Sci. U.S.A.* **1999**, *96*, 9997−10002.
(17) Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne W. A. Conformational changes of small molecules binding to proteins. *Bioorg. Med. Chem. Lett.* **1995**, 3, *4*, 428−441.
(18) Kline, P. C.; Schramm, V. L.; Pre-Steady-State Analysis of the Hydrolytic reaction Catalyzed by Purine Nucleoside Phosphorylase. *Biochemistry* **1995**, *34*, 1153−1162.
(19) Braunheim, B. B.; Miles, R. W.; Schramm, V. L.; Schwartz S. D. Prediction of inhibitor binding free energies by quantum neural networks. Nucleoside Analogues binding to Trypanosomal Nucleoside Hydrolase. *Biochemistry* **1999**, *38*, 16076−16083.
(20) Mochalin, I.; Tulinsky, A. Structures of thrombin retro-inhibited with SEL2711 and SEL2770 as they relate to factor Xa binding. *Acta Crystallogr.* **1999**, *D55*, 785−793.
(21) Krishnan, R.; Zhang, E.; Hakansson, K.; Arni, R. K.; Tulinsky, A. Highly selective mechanism-based thrombin inhibitors: structures of thrombin and trypsin inhibited with rigid peptidyl aldehydes. *Biochemistry* **1998**, *37*, 12094−12103.
(22) Maryanoff, B. E.; Qui, X.; Padmanabhan, K. P.; Tulinsky, A.; Almond, H. R.; Andrade-Gordon, P.; Greco, M. N.; Kaufman, J. A.; Nicolaou, K. C.; Liu, A.; Brungs P. H.; Fusetani, N. Molecular basis for the inhibition of human alpha thrombin by the macrocyclic peptide cyclotheonamide A. *Biochemistry* **1993**, *32*, 8048−8052.
(23) Steiner, J. L. R.; Murakami, M.; Tulinsky, A. Structure of thrombin inhibited by Aeruginosin 298-A from blue-green alga. *J. Am. Chem. Soc.* **1998**, *120*, 597−598.
(24) Matthews, J. H.; Krishnan, R.; Costanzo, M. J.; Maryanoff, B. E.; Tulinsky, A. Crystal structures of thrombin with thiazole-containing inhbitors: probes of the S1 binding site. *Biophys. J.* **1996**, *71*, 2830−2839.
(25) Charles, R. S.; Matthews, J. H.; Zhang, E.; Tulinsky, A. Bound structures of novel P3−P1′ beta strand mimetic inhibitors of thrombin. *J. Med. Chem.* **1999**, *42*, 1376−1383.
(26) Malley, M. F.; Tabernero, L.; Chang, C. Y.; Ohringer, S. L.; Roberts, D. G. M.; Das, J.; Sack, J. S. Crystallographic determination of the structures of human alpha thrombin complexed with BMS-186282 and BMS-189090. *Prot. Sci.* **1996**, *5*, 221−228.
(27) Banner, D. W.; Hadvary, P. Crystallographic analysis at 3.0A resolution of the binding to human thrombin of four active site directed inhibirors. *J. Biol. Chem.* **1991**, *266*, 20085−20093.
(28) Hilpert, K.; Ackermann, J.; Banner, D. W.; Gast, A.; Gubernator, K.; Hadvary, P.; Labler, L.; Muller, K.; Schmid, G.; Tschopp, T. B.; van de Waterbeemd, H. Design and synthesis of potent and highly selective thrombin inhibitors. *J. Med. Chem.* **1994**, *37*, 3889−3901.
(29) Tabernero, L.; Chang, C. Y.; Ohringer, S. L.; Lau, W. F.; Iwanowicz, E. J.; Han, W. C.; Wang, T. C.; Seiler, S. M.; Roberts, D. G.; Sack, J. S. Structure of a Retro binding peptide inhbitor complexed with human alpha thrombin. *J. Mol. Biol.* **1995**, *246*, 14−20.
(30) InsightII Version 97.0; MSI Inc.: San Diego, CA, 1997.
(31) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan; Karplus, M.: Charmm − A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.
(32) Schmidt, M. W.; Baldridge, K. K.; Boatz, J. A.; Elbert, S. T.; Gordon, M. S.; Jensen, J. H.; Kotseki, S.; Matsunaga, N.; Nguyen, K. A.; Su, S. J.; Windus, T. L.; Dupuis, M.; Montgomery, J. A. Gamess, *J. Comput. Chem.* **1993**, *14*, 1347−1363.
(33) Hecht-Nielsen, R.; Counterpropagation networks. *Appl. Optics* **1987**, *26*, 4979−4984.
(34) Lee, Y. S.; Hodoscek, M.; Brooks, B. R.; Kador, P. F. Catalytic mechanism of aldose reductase studied by the combined potentials of quantum mechanics and molecular mechanics. *Biophys. Chem.* **1998** *70*, 203-216.
(35) Smith, B. M.; Gemperline, P. J.; Wavelength selection and optimization of pattern recognition methods using the genetic algorithm. *Anal. Chim. Acta* **2000**, *423*, 167−177.

CI000162E