Journal of Computer-Aided Molecular Design, 5 (1991) 617–628
ESCOM

J-CAMD 135

A machine learning approach to computer-aided molecular design

Giorgio Bolis[a,*], Luigi Di Pace[b] and Filippo Fabrocini[b]

[a]*Farmitalia Carlo Erba srl, Erbamont Group, R&D/CAMD, via dei Gracchi 35, I-20146 Milan, Italy*
[b]*Artificial Intelligence Group, IBM Rome Scientific Center, via Giorgione 159, I-00147 Rome, Italy*

Received 7 November 1990
Accepted 8 May 1991

Key words Artificial Intelligence; Structure–activity relationship; Thermolysin inhibitors

SUMMARY

Preliminary results of a machine learning application concerning computer-aided molecular design applied to drug discovery are presented. The artificial intelligence techniques of machine learning use a sample of active and inactive compounds, which is viewed as a set of positive and negative examples, to allow the induction of a molecular model characterizing the interaction between the compounds and a target molecule. The algorithm is based on a twofold phase. In the first one – the specialization step – the program identifies a number of active/inactive pairs of compounds which appear to be the most useful in order to make the learning process as effective as possible and generates a dictionary of molecular fragments, deemed to be responsible for the activity of the compounds. In the second phase – the generalization step – the fragments thus generated are combined and generalized in order to select the most plausible hypothesis with respect to the sample of compounds. A knowledge base concerning physical and chemical properties is utilized during the inductive process.

INTRODUCTION

The problem of structure–property relationship is central in all the applications of molecular design. In particular, in biological sciences the structure–activity relationship (SAR) has been studied for many years in order to either elucidate biological processes or develop new drugs. These studies are all based on the concept that a biological (or pharmacological) effect caused by a given molecule (drug) is a function of its chemical structure.

The task of determining the SAR between biological molecules is a challenging one, due to the complexity of chemical structures, their properties and, often, to the large number of molecules to be taken into account in the study.

The methods used to answer this problem range from the 'visual analysis', linear and non-linear

[*] To whom correspondence should be addressed.

regression analysis [1], pattern recognition techniques [2, 3] and automated structure evaluation [4] to, more recently, the use of neural networks [5].

The great difficulty in all the SAR studies is the selection of appropriate molecular properties to be used as descriptors in the various methods. The typical approach adopted in most traditional methods is the use of some sort of statistics to assess the importance of given properties in the molecules.

In order to overcome this problem, we thought of implementing a program that would be able to choose on its own the relevant properties of the compounds under study. This program would make use of artificial intelligence (AI) techniques defined as 'machine learning' which, as illustrated in the methodology section, allow a machine to learn, in some way, when given a knowledge base of molecular properties in general and a series of positive and negative examples (i.e. the active and inactive compounds).

Although the algorithm of this program could be used for several applications in molecular modeling, it was conceived in the framework of drug design and in such a context it will be illustrated. In the present work, we also limit ourselves to the consideration of molecular structures as 'two-dimensional systems', that is, at the present stage of development of the program, we do not take into account the three-dimensionality of the molecules. Further development is, however, planned: in fact, we conceived the program performing the SAR analysis as divided in two stages: the first one consists of the screening of the molecules as two-dimensional entities, and then, with the information obtained from this step, proceed to the analysis of the same molecules considered with their three-dimensional structure.

METHODOLOGY

The approach we used is an attempt to reproduce, with artificial intelligence techniques, the process that a medicinal chemist is performing when faced with the problem of visually analyzing a set of compounds for determining their structure–activity relationship. Like an expert in the field, the program has its 'perception' of a molecule and certain basic knowledge about molecules and their properties.

A molecule is represented in the program in a hierarchical way and, for this purpose, is divided into four categories or levels: at the highest level the molecule is considered as a whole, in the following one the molecule is represented as made of 'residues' (taken in a broad sense: for example, the amino acids in a peptide), at the next step the program describes the molecule with smaller molecular fragments which we called 'functional groups' (for example, a carboxyl group) and at the lowest level there is the most detailed description in terms of atomic entities. The above described elements in these categories have associated some properties like molecular weight, hydrophobicity index, point charges and so on. There is complete flexibility in the usage of the elements of the four categories and their properties and they can, in fact, be used in any combination to describe a molecular system. For example, we might want to describe a given molecule as an entity with molecular weight of 557, its second residue is a serine, the third residue is hydrophobic and the last fragment is a carboxyl group.

The properties are formalized in a dictionary independently of the program, therefore they can be added, deleted or modified at will to better describe the molecular system under consideration. As an example, a list of the properties used in this study is given in the Appendix.

The molecular system is recognized by the program as a set of atoms together with a connectivity table. The atoms are defined as 'atom types', i.e. there is an attempt to characterize their electronic state. The characterization in terms of atom types is widely used in molecular mechanics calculations; in particular, we adopted the atom type set used in the Biosym's program Discover [6]. At this stage of development of the program, reference is made to the fact that molecules are three-dimensional objects only in a very indirect way, since we use an approximate determination of the volume (expressed in cm^3 $mole^{-1}$) of molecules based on atom types [7].

AI algorithm

As mentioned above, our system makes use of machine learning techniques to deal with the SAR problem. The term 'learning', in everyday language, is used in a somewhat broad and ambiguous way which is unacceptable in any technical context. Yet, in AI terms, learning has been mainly viewed as *inducing* a general concept description from a set of positive and negative examples. Specifically, such a description should cover all the positive examples, but none of the negative ones, and it should correctly classify new examples [8]. Such a view especially stresses the inductive character of the learning process as opposed to the *deductive* character of most of human reasoning mechanisms. Consider, for instance, a program that is learning the concept of 'car'. In this case, a number of examples of vehicles (cars, trains, bicycles, trucks, etc.) might be submitted to the program. Each of them will be labelled as a positive or a negative example of the concept to be learned. The program should develop a general rule, i.e. a concept, suitable to classify correctly future examples. More formally, the task of learning from examples can be defined as follows [9]:

(a) Given:

- some positive examples, E+, and some negative examples, E−;
- a background knowledge, BK, which includes domain knowledge, general constraints and preference criteria.

(b) Find:

- a general concept description, C, which together with the background knowledge BK, implies all the positive examples, E+, and none of the negative ones, E− (C and BK $\rightarrow$ E+ and $\neg$ E−)

where the symbol $\neg$ is a logical NOT.

The background knowledge, in particular, includes the type of description language used for characterizing examples and learned concepts. A description of a concept which covers all the positive examples *(completeness)* and none of the negative ones *(consistency)* is a *candidate hypothesis*. Because the candidate hypotheses space is typically quite large, a preference criterion is used to select among the candidate hypotheses. The preference criterion is defined within the background knowledge; therefore it is a problem-specific criterion. Typical preference criteria are: hypotheses with a smaller number of descriptors, more specific hypotheses with respect to the positive examples, more general hypotheses constrained only by the negative examples, hypotheses with lower probability to occur, etc.

Generalization and specialization operators allow the exploration of the candidate hypotheses space. Let us suppose two positive examples are submitted to the program:

example_1: X = square & X = white
(X is an object with square shape and white color)

example_2: X = triangle & X = white)
(X is an object with triangular shape and white color).

Then, the application of the generalization operator will produce the following hypothesis:

hypothesis_1: X = white
(X is a white colored object)

or, if the program is provided with a background knowledge that states that triangles and squares are both polygons, it will produce the following hypothesis:

hypothesis_2: X = polygon & X = white
(X is a white colored object with polygonal shape).

Now, let us suppose that the program is provided with a negative example:

counterexample_1: X = pentagon & X = white
(X is a white colored object with pentagonal shape).

In this case, the program will apply the specialization operator at hypothesis 2 in order to make it exclude the negative example. The application of the specialization operator will produce the following hypothesis:

hypothesis_3: X = polygon & X = white & number_of_sides < 5 (X is a white colored object with polygonal shape whose number of sides is less than five)

Despite its apparent simplicity, machine learning from examples is a complex task that can be studied adopting multiple approaches. Because our system makes use of a knowledge base of chemical properties that are deductively derived during the learning process in order to construct new descriptors of the example (i.e. the compound), our approach can be classified as a *constructive* one [9].

One peculiarity of our approach is related to the manner in which objects work as examples in our application domain. In our case, the sample of active and inactive compounds used as input data is viewed as a set of positive and negative examples that allow the induction of a molecular model characterizing the interaction between the active compounds and a target biological system. However, since the experimental activity of the compounds is measured along a continuous scale, there is no clear distinction between positive and negative examples. Therefore, some examples appear more prototypical of the concept to be learned than others. Neither would a classification based on certain thresholds be reasonable because of the arbitrariness of their definition. In our approach the program automatically solves the *'labelling'* problem, i.e. the program automatically labels each of the compounds under consideration as a 'positive' or 'negative' example [10].

As we will see below, the program is guided by the differences in activity to perform this task. Therefore, when taking into account a pair of compounds, the program simply labels as 'positive' the more active and 'negative' the less active compound of the pair.

Our algorithm is based on two steps: a specialization step and a generalization step. In the first one the program learns from the inactive compounds whose structure appears very close to the structure of the active ones. Specifically, two molecules are regarded as structurally similar whenever they differ, at most, in one residue. The second step makes use of the active compounds in order to produce a general model characterizing the interaction between the active compound and its biological target.

The specialization step

In the specialization step the program identifies a number of active/inactive pairs which appear most useful in order to make the process of learning effective. The input to this step is the set of compounds with their associated experimental activities and the output is a dictionary of relevant molecular fragments.

Initially, the system seeks those pairs of molecules which, on the one hand, maximize the difference in activity, $\Delta(K)$, and, on the other, minimize the difference in structure, $\Delta(S)$. While the quantity $\Delta(K)$ is clearly defined, the determination of $\Delta(S)$ is not so straightforward. Although methods for quantifying differences in structure exist [11], at the present time our program is simply considering the difference $\Delta(S)$ as the number of different atom types present in a given pair of molecular fragments. The outlined strategy reflects the usual approach taken by the medicinal chemist during the process of drug design.

A partial matching algorithm is used to evaluate the difference between the compounds of the input data so as to find pairs of compounds which satisfy the similarity constraints described above. Specifically, the partial matching algorithm is an algorithm that maximizes the matching between two graphs.

Once the pairs of compounds which appear to be more informative are generated, the compounds in the different pairs are matched against each other. Matching allows the identification of one fragment responsible for the difference in activity between the two compounds of the pair. Our system automatically chooses the level of representation (residues, functional groups, atoms) more adequate to describe a fragment responsible for the $\Delta(K)$ in the given pair [12]. The output of this step is a dictionary of molecular fragments deemed to be very relevant from the point of view of the activity of the compounds.

Finally, the information contained in the dictionary of properties is applied to the fragments generated in the specialization step. In other words, the relevance of the fragments appearing in the dictionary is now tentatively explained in terms of their properties (hydrophobicity, polarity, volume, etc.).

The generalization step

The generalization step has the aim of deriving the most plausible model of the optimal compound, a model that would combine the generalized relevant fragments in the context of the active compounds. In order to achieve this goal the program makes use of the set of input compounds, the knowledge base and the set of the relevant fragments obtained as output of the specialization step.

The final model is generated in an iterative manner. The compound with the highest activity value is taken as the starting model. This molecule is then compared with all the other compounds of the input data with decreasing value of activity, searching for similarities in structure and properties. With each additional compound considered, the model is eventually modified in order to include all the features that cover all the active compounds and do not cover the inactive ones [13]. While performing this process, the system is faced with two decision points:

(a) the selection of the best matching between the current model and an additional compound;

(b) the selection of the best generalization for the regions of the model that appear to be different in the additional compound.

The generalization techniques applied in this step are: *dropping condition* (drop an uncommon condition), *closing intervals* (generalize values *n* and *m* into the interval *n-m*), *constructive generalization* (generate new properties about single atoms or regions in order to discover common properties by using previously defined procedures).

The result of this process produces a structure with generalized fragments. These are associated with the allowed properties and all the pointers to the actual residues, functional groups or atom types that produced the generalized fragment itself. The program also stores, for every generalized fragment, a list of residues, functional groups, atoms and properties that are forbidden at a certain position.

In this learning context, the candidate hypotheses space is the space of all the possible models. The growth of such a space is defined by the set of decision points previously described insofar as each of these points allows the generation of new possible models from the current model. In order to avoid the exponential growth of the space, the system performs a beam search: only the most promising K models are considered at each decision point, where K (the beam) is a parameter that depends on the computational capabilities of the system. A preference criterion is used to assess the quality of the model. Whenever the quality of the new model is lower than the quality of the old one, the system backtracks to the previous decision point and selects a different model. Specifically, the preference criterion is based on the number of active/inactive compounds covered by the model: the quality increases when the model covers additional active compounds and decreases when the model covers additional inactive compounds.

RESULTS AND DISCUSSION

In order to test the validity of our program we chose the well-studied system of the inhibitors of the thermolysin enzyme for which three-dimensional structural information of several complexes enzyme-inhibitor [14] is available from the Brookhaven Protein Data Bank (PDB) [15]. Thermolysin is a thermostable metalloprotease involved in several important physiological processes and, like other metalloproteases, contains a zinc ion essential for activity.

The data for inhibitory activity of a number of thermolysin inhibitors were taken from literature [16–19] and are reported in Table 1. The activity is expressed in terms of K_i values determined by Dixon plots. The input to our program consisted essentially of the data contained in Table 1 where the description of the compounds is given in terms of their atom names, atom types and connectivities.

The specialization step produced a set of 13 relevant pairs of fragments. In order to illustrate the working of this step let us consider, for instance, the pair **18**/**19** from our input data. Such a pair satisfies the similarity constraint required, since compound **18** differs from compound **19** only in one fragment. The program, therefore, identifies the residue Ibm as the one responsible for the Δ(K) in the selected pair. This means that the substitution of the Mal residue by the Ibm residue in the peptide HONH-*-Ala-Gly-NH^2, where the star denotes the variable, causes a major change in the activity of the compound. At this point the program attempts to reformulate such a difference in terms of functional groups. Both compounds **18** and **19** are automatically formalized as graphs whose nodes are functional groups and then they are matched in order to extract the differences. The program identifies in the Ibm residue the following groups: two CH_3, one CH_2, two CH and two C=O, while in the Mal residue there are one CH_2 and two C=O groups. Since the differences are found in a connected region (the boxes in Fig. 1), the program keeps the formalization of the molecular description at the functional group level.

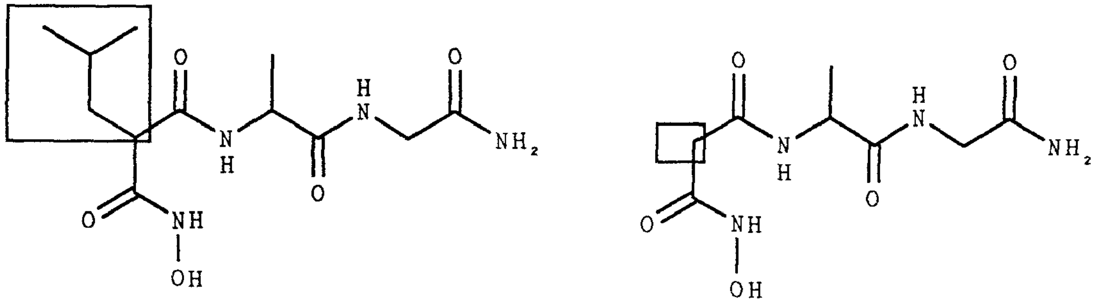
Let us consider now another pair of Table 1, namely compound **16** versus compound **20**. The difference between such compounds is a functional group, therefore the system directly reformulates such a difference in terms of atom types. Once the graphs of the corresponding compounds are matched, the system generates as difference between compounds **16** and **20** the atom type *'n'*.

Since the module performing the generalization step is presently not fully implemented, only a subset of the input compounds was used to test this part of the program. Therefore the generalized model of a 'good' inhibitor of thermolysin illustrated in this section has to be considered an intermediate or partial model of the generalization of the compounds in Table 1. The schematic

TABLE 1
K_i VALUES OF INHIBITORS OF THERMOLYSIN

No.	Compound[a]	K_i (μM)	Ref	No.	Compound[a]	K_i (μM)	Ref.
1	Z-NHNH-CS-$NHNH_2$	6700	15	**19**	HONH-Mal-L-Ala-Gly-NH_2	1100	16
2	Z-Agly-Leu-$NHNH_2$	380	15	**20**	HO-Bzm-L-Ala-Gly-NH_2	420	16
3	Z-Gly-Leu-$NHNH_2$	1100	15	**21**	CHO-HOLeu-L-Ala-Gly-NH_2	3.8	16
4	Ac-Ala-Aphe-Leu-$NHNH_2$	6500	15	**22**	Ac-HOLeu-L-Ala-Gly-NH_2	3400	16
5	Ac-Ala-Ala-Aala-Leu-$NHNH_2$	7900	15	**23**	P-NH-Et	No inhib.	17
6	L-Leu-NHOH	190	15	**24**	P-Leu-NH_2	1.3	17
7	Z-L-Leu-NHOH	10	15	**25**	P-Phe-OH	73	17
8	Z-Gly-L-Leu-NH_2	21000	15	**26**	P-Ala-Ala-OH	88	17
9	Z-Gly-L-Leu-NHOH	13	15	**27**	P-Ile-Ala-OH	0 36	17
10	Z-Gly-L-Leu-N(CH_3)OH	2230	15	**28**	P-Leu-Phe-OH	0.019	17
11	Z-Gly-L-Leu-$NHOCH_3$	No inhib.	15	**29**	Z-Phe-Gly-NH_2	350	18
12	Z-Agly-L-Leu-NHOH	2.7	15	**30**	Z-Phe-Gly	4500	18
13	Z-Gly-Gly-NHOH	940	15	**31**	Phe-Gly-NH_2	10900	18
14	Z-Gly-Gly-L-Leu-NHOH	39	15	**32**	Phe-Gly	10300	18
15	HONH-Bzm-OEt	20	15	**33**	Z-Leu-Gly-NH_2	3070	18
16	HONH-Bzm-L-Ala-Gly-NH_2	0 66	15	**34**	Z-Leu-Gly	4030	18
17	HONH-Bzm-L-Ala-Gly-OH	0.65	16	**35**	Leu-Gly-NH_2	8300	18
18	HONH-Ibm-L-Ala-Gly-NH_2	0.48	16				

[a] The abbreviations are the same as in the original articles: Z, benzyloxycarbonyl; Agly,-NHNHCO-; Aala, -NHN(CH_3); Aphe, -NHN($CH_2C_6H_5$)CO-; P, phosphoryl group $(HO)_2PO$-, Bzm, benzylmalonyl -COCH($CH_2C_6H_5$)CO-; Ibm, isobutylmalonyl -COCH($CH_2CH(CH_3)_2$)CO-; Mal, malonyl; Ac, acetyl; Et, ethyl



18 19

Fig. 1. Compounds **18** and **19** from Table 1, where the structural difference causing a large change in activity is shown in the boxes.

representation of this model is shown in Fig. 2. where the generalized inhibitor is depicted as composed of four generalized fragments. The model is the generalization of six among the most active compounds of Table 1, namely, Nos. **16**, **17**, **18**, **21**, **27** and **28**, whose activities range between 0.019 and 3.8 μM. When comparing the compounds, wherever the process of matching leads to a fragment common to all the molecules, the program represents this fragment in the model as in the original molecules; in our case the CH, NH and C = O groups (see Fig. 2). In the other cases the system gives a more descriptive representation in terms of properties of these fragments. The program however, upon request, can always display the actual fragments from which the various properties were derived.

The description of the generalized fragments present in the model is reported in Table 2. In this table a number of properties (labelled with a capital letter) are listed for each fragment, along with a number of positive and negative instances of fragments that the model covers. Among the properties, the program also lists distances, expressed in terms of number of bonds, of relevant atoms within the fragments from a reference atom outside of it.

In the generated model (see Fig. 2) the fragment R_0 is defined as hydrophilic, hydrogen-bond donor and acceptor while the fragment R_1 is defined as hydrophobic with volume between 34 and 60 cm^3 $mole^{-1}$. Both R_0 and R_1 are termed essential for activity because the deletion of these groups produces a dramatic loss of activity. Fragment R_2 is also hydrophobic while fragment R_3

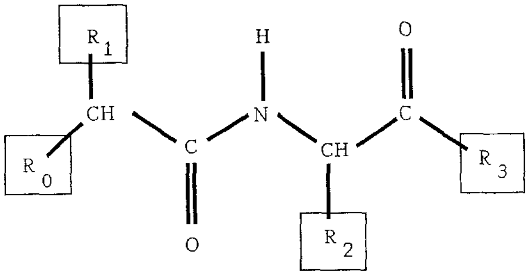


Fig. 2. Partial model of a generalized 'good' inhibitor of thermolysin obtained from the step of generalization. The model covers the compounds **16**, **17**, **18**, **21**, **27** and **28** of Table 1.

does not seem to be very relevant for any change of activity of the compounds considered. The correctness of these conclusions can be verified by looking at Table 2. The three instances of fragments (from the six active compounds analyzed) generalized in order to define R_0 are all polar and strong hydrogen-bond donor or acceptor. In this case, in fact, counterexamples are a hydrophobic group (in compound **22**, activity = 3400 μM) or a fragment that is too short (in compound **20**, activity = 420). The property of hydrophobicity for fragment R_1 is derived from the two instances of leucyl and benzyl groups in the six compounds, while the lack of this property (see compound **19**) produces a large loss of activity.

This characterization is in accordance with observations made in the crystal structures of thermolysin complexed with various inhibitors [14]. A schematic view of the active site of one of such structures, namely the one derived from the Brookhaven PDB file 5PTL, is reported in Fig. 3 [20]. From this structure it can be noticed that the moiety HONHCO (representing our generalized fragment R_0) performs hydrogen-bond interactions with residues Ala^{13} and Glu^{143} and coordinates the active site zinc ion. Furthermore, the benzyl group of the benzylmalonyl moiety (corresponding to our generalized R_1) is found in a hydrophobic pocket (defined as S_1' binding subsite) of the active site, while the following Ala side chain (corresponding to R_2) is found in the proximity of Leu^{202} and Phe^{130} of the enzyme. An analogous situation is observed looking at the crystal structure of phosphoramidon [18] bound to thermolysin (Brookhaven PDB file 1TLP): the negatively charged phosphoryl group of phosphoramidon (corresponding to our R_0) is coordinat-

TABLE 2
DESCRIPTION OF THE GENERALIZED FRAGMENTS IN THE MODEL OBTAINED

Properties of the generalized fragments		*Positive and negative instances of the generalized fragments*			
R_0	A: 21 ≤ Volume ≤ 31	R_0·	positive	1: HONHCO	**16, 17, 18**
	B: Hydrophobicity = No			2: $PO(OH)_2NH$	**27, 28**
	C: dist (HB donor atom, CH) = 1,2,3 bonds			3. CHONOH:	**21**
	D: dist (HB acceptor atom, CH) = 2,3 bonds			[exception to F: dist (O,CH) = 2 bonds]	
	E· dist (N,CH) = 1,2 bonds				
	F: dist (O [atomtype oh],CH) = 3 bonds				
	G: dist (O [atomtype o'],CH) = 2 or 3 bonds		negative	1: HOCO: ¬(H) and ¬(F)	**20**
	H: dist (furthest heavy atom,CH) = 3 bonds			2· CH_3CONOH· ¬(A) and ¬(B)	**22**
R_1	A: Hydrophobicity = Yes	R_1:	positive	1 $CH_2CH(CH_3)_2$	**18, 21, 28**
	B: 34 ≤ Volume ≤ 60			2. CH_2*Benz*	**16, 17, 27**
			negative	1· H· ¬(A) and ¬(B)	**19**
R_2	A: Hydrophobicity = Yes	R_2:	positive	1: CH_3	**16, 17, 18, 27,**
	B: 13.6 ≤ Volume			2: CH_2*Benz*	**28**
R_3	* (not relevant foı activity)				

Lowest activity of a covered compound: 3.8
Highest activity of an uncovered compound 420

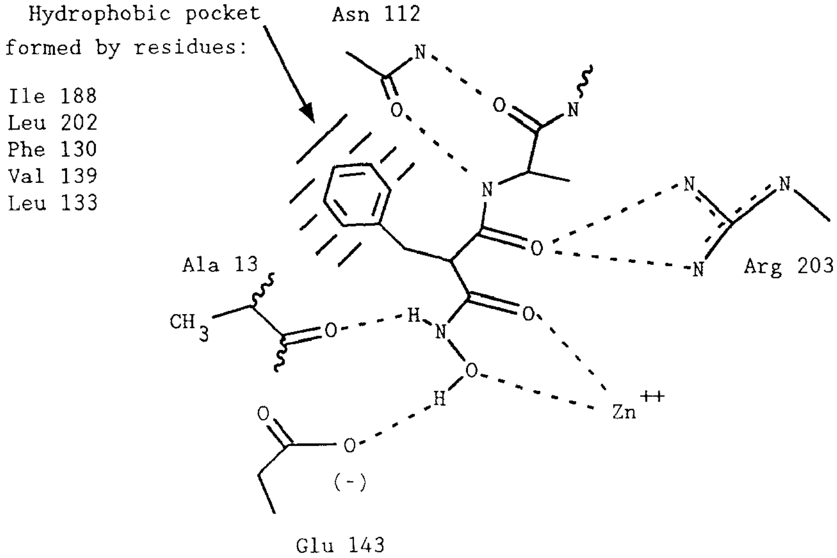


Fig. 3. Schematic view of the active site of thermolysin with a fragment of a hydroxamic acid derivative inhibitor.

ing the zinc ion, while the side chain of the leucyl residue (our R_1) fits into the hydrophobic pocket S_1'.

The comparison of the experimental data with the model automatically obtained suggests that our program is able to successfully compare and align a series of thermolysin inhibitors and from these to construct a generalized molecular model. The description of the properties of the generalized fragments of the model gives a detailed picture of the requirements that the various parts of the inhibitor must satisfy. This is of considerable help when performing SAR studies in order to classify and understand the functions of the different portions of a series of compounds.

When faced with the task of designing new inhibitors, one can envision the utilization of a large molecular fragment data base associated with a generalized model derived from a set of given inhibitors with our methodology. It would then be possible to select appropriate fragments from the data base and place them at their specific positions in the model, by imposing the constraints given in the table of the generalized fragments.

Presently the program runs under the IBM VM/SP or IBM VM/XA operating systems. In the near future we plan to move the program to an IBM RS/6000 computer under the IBM AIX operating system. IBM Prolog is the Prolog dialect used as programming language. A demonstration of the program, upon request, is presently available at the IBM Rome Scientific Center.

CONCLUSIONS

Machine learning techniques have been used in a program, which is still under development, to identify automatically the relevant fragments (described in terms of structural and other physicochemical properties) responsible for activity in a set of inhibitors of thermolysin and, furthermore, to determine a generalized model for an optimal inhibitor. The model derived is in good agreement with experimental data of crystal structure complexes of thermolysin with various inhibitors.

The advantage of the adopted approach is that a model can be derived from a small number of compounds; since statistical methods are not used, it is not very relevant that certain fragments appear frequently in the set of input molecules; on the contrary, the structural and physico-chemical properties of each compound could contribute significantly to the definition of the model. In addition, because the model is described in terms of properties and the fragments are always considered in the context of the entire compound, the interaction with the target receptor is better characterized.

This methodology, even though at present not quantitative, could already be of significant help in the task of determining the structure–activity relationship of different sets of molecules and in giving insights for designing novel compounds.

ACKNOWLEDGEMENTS

The authors would like to thank Silvia Scolastico, Alberto Guaragna and Mauro De Santis for help and discussions.

REFERENCES

1 See, for example: Martin, Y C., Quantitative Drug Design, Marcel Dekker, New York, 1978, pp. 167–213.

2 Kowalski, B.R and Bender, C.F., J. Am Chem Soc., 96 (1974) 916.

3 Chu, K.C., Feldmann, R.J., Shapiro, N B., Harard, G.F. and Geran, R.I , J. Med Chem., 18 (1975) 539.

4 Klopman, G., J Am. Chem. Soc., 106 (1984) 7315

5 Aoyama, T , Suzuki, Y and Ichikawa, H., J. Med. Chem., 33 (1990) 905

6 Biosym Technologies Inc , 10065 Barnes Canyon Road, San Diego, CA 92121.

7 Govers, H. and de Voogt, P., Quant. Struct.-Act. Relat , 8 (1989) 11.

8 Carbonell, J.G., Machine Learning, In Shapiro, S.C. (Ed.) Encyclopedia of Artificial Intelligence, Vol. 1, John Wiley, New York, NY. 1987, pp 464–488.

9 Michalski, R S., A theory and methodology of inductive learning, In Michalski, R.S., Carbonell, J.G. and Mitchell, T M (Eds.) Machine Learning Vol. 1, Tioga, Palo Alto, 1983, pp 83-134

10 Mitchell, T.M., Towards combining empirical and analytical methods for learning heuristics, In Elithorn, A. and Banerji, R. (Eds.) Human and Artificial Intelligence, North-Holland, Amsterdam, 1984, pp. 67-103.

11 Bertz, S.H. and Herndon, W.C., In Pierce, T H. and Hohne, B A (Eds.) Artificial Intelligence Applications in Chemistry (ACS Symposium Series, No. 306), 1986, pp. 169–175.

12 Flann, N. and Dietterich, T., Selecting appropriate representations for learning from examples, In Proceedings of the American Conference on Artificial Intelligence, AAAI, 1986, pp. 460–466.

13 For further details see: Bolis, G , Di Pace, L. and Fabrocini, F., Shift of bias in learning from drug compounds, In Proceedings of the European Working Session on Machine Learning, EWSL-91, Springer-Verlag, Berlin, 1991, pp. 182-194

14 The Brookhaven PDB codes for some of such structures are, for example: 4TLN, 5TLN, 7TLN, 1TLP, 6TMN.

15 a. Bernstein, F.C., Koetzle, T F., Williams, G.J., Meyer, Jr., E.E., Brice, M.D., Rodgers, J R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.
b Abola, E.E., Bernstein, F.C., Bryant, S.H , Koetzle, T.F. and Weng, J , Protein data bank In Allen, F H., Bergerhoff, G. and Sievers, R. (Eds) Crystallographic Databases – Information Content, Software Systems, Scientific Applications, Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester, 1987, pp. 107–132

16 Nishino, N and Powers, J C., Biochemistry, 17 (1978) 2846.

17 Nishino, N and Powers, J.C , Biochemistry, 18 (1979) 4340.

18 Kam, C -M., Nishino, N and Powers, J.S , Biochemistry, 18 (1979) 3032

19 Feder, J., Brougham, L.R. and Wildi, B.S., Biochemistry, 13 (1974) 1186.

20 Holmes, M.A and Matthews, B.W., Biochemistry, 20 (1981) 6912

APPENDIX

The knowledge base

Residue level

- Polarity
  - ◇ Yes (+, −)
  - ◇ No
- Hydrophobicity
  - ◇ Yes
  - ◇ No
- Volume
  - ◇ Evaluated by its subcomponents

Functional group level

- Polarity
  - ◇ Yes (+, −)
  - ◇ No
- Hydrophobicity
  - ◇ Yes
  - ◇ No
- Electron donor
  - ◇ Yes
  - ◇ No
- Electron acceptor
  - ◇ Yes
  - ◇ No
- Volume
  - ◇ Evaluated by its subcomponents

Atom type level

- H-Bond
  - ◇ Yes (donor, acceptor)
  - ◇ No
- Nucleophilicity
  - ◇ Yes (+, −)
  - ◇ No
- Acidity/basicity
  - ◇ Yes (+, −)
  - ◇ No