Journal of Computer-Aided Molecular Design, **15:** 57–66, 2001.
KLUWER/ESCOM

FILO (Field Interaction Ligand Optimization): A simplex strategy for searching the optimal ligand interaction field in drug design

Fabrizio Melani, Paola Gratteri*, Michele Adamo & Claudia Bonaccini
Department of Pharmaceutical Science, Firenze University, Via G. Capponi 9, I-50121 Firenze, Italy

Received 26 January 2000; Accepted 21 June 2000

Key words: GRID, HIV-1 protease, ligand-receptor interactions, molecular modelling, pseudoreceptor, 3D-QSAR

Summary

A method (FILO, Field Interaction Ligand Optimization) for obtaining the optimal molecular interaction field was developed on the basis of the Simplex optimization procedure applied to a matrix of interaction energies obtained by performing a GRID computation on a suitable data set. The FILO procedure was tested on a set of nine HIV-1 protease inhibitors with known crystal structures. The results of FILO consist of the optimal molecular interaction field of a putative new ligand with optimal binding affinity. The final FILO model yields R^2 and R^2_{CV} values of 0.993 and 0.936, respectively, and finds eight negative and four positive interaction nodes for the OH probe taken as an example. The eight H bonding interactions pointed out by FILO identified well the binding site AA-residues Gly A27, Asp A29, water 501, Gly B48 and Asp A25 of HIV-1 protease.

Introduction

One of the main objectives in drug design is to find new leads and optimize the activity of compounds already known to be effective. Much effort and many strategies have been proposed in this direction in order to develop new and not yet synthesized molecules with interesting biological activity.

At present, a widely used and useful approach to 3D-characterization of compounds involved in quantitative structure-activity relationship (3D-QSAR) studies is based on calculations of energetic interaction fields between the compounds and probes [1,2].

A number of methods for building receptor site models exist in literature. Most of them deal with the receptor aspect of the problem, leading finally to the definition of pseudoreceptors or minireceptors [3–14].

The present work describes a method, FILO (Field Interaction Ligand Optimization), that deals with the problem from the ligand point of view; starting from the molecular interaction fields derived by submitting an appropriate set of molecules taken as training set to GRID computation, FILO leads to the description of the molecular interaction field of a not yet synthesized ligand with optimal binding affinity. Thus, the method can be used as a guide for the design of new and potentially more effective compounds that satisfy key interactions with the receptor protein and it seems to be a useful tool in all situations where the 3D structure of the target protein is unknown.

As any other method developed in the 3D-QSAR area, FILO leads to the definition in 3D space of a hypothesis of binding site. FILO differs from other methods because it processes and optimizes values of the molecular interaction field according to how they should result from the hypothetical optimal ligand.

However, according to 3D-procedures, FILO is dependent on the preliminary orientation of the molecules since the results are strictly influenced by the molecule alignment procedure.

Briefly, the methodology involves three steps:

I. Superimposition of the ligands.

II. Computation of the interaction energies between different probes and each of the molecules of the training set from the GRID force field. The different probes are chosen keeping in mind that the spatial information they give can be considered

*To whom correspondence should be addressed. E-mail: paola.gratteri@unifi.it

as being representative of the important chemical groups present at the active site. For example, the GRID calculation for the OH2, OH, O, etc. probes mainly accounts for the energies of the hydrogen-bonding interactions. The C3 probe yields the energies of the steric interaction between the probe and the target, and charged probes allow the electrostatic interactions to be detected.

III. Application of the FILO methodology to reduced molecular interaction fields resulting from selection node procedures according to both energy and standard deviation (SD) cut-off criteria.

Of course other variable selection procedures are possible before FILO computation, i.e, for example, GA selection [15], GOLPE/FFD variable selection [16], SRD/GOLPE method [17].

Materials and methods

Molecular modelling was performed using InsightII 97 and Sybyl [18,19] running on a RISC6000 IBM 3CT and a SGI R5000, respectively.

GRID [20] was used to calculate the energetic interactions between probes and each compound constituting the training set. The energy calculations were performed surrounding all the aligned ligands with a suitable cage and using a 1 Å spacing between grid points (nodes).

FILO procedure

1. The first step in the FILO procedure consists in defining an energy cut-off level, which is valid for all the molecular interaction fields (MF) derived from the GRID computations performed on the training set molecules. FILO then constructs a matrix, reporting the energy interaction values at specific nodes (columns) for each molecule of the training set (row). Starting from the defined cut-off energy value, FILO searches, among all the MFs of the molecules of the training set, for the nodes with the highest absolute interaction values, independent of which molecules they belong to. Once the nodes have been selected, FILO associates the corresponding energy values assumed by the molecular interaction fields of the molecules of the training set with the coordinates of those nodes. As a consequence, each selected node possesses the same coordinates for all the molecules, but not necessarily the same potential values. Generally, at this point, ca. 2000–3000 nodes per molecule are retained.
2. The second step in the FILO procedure consists of sorting the selected nodes according to decreasing values of the standard deviation (SD) of their energy interaction values. Following this pretreatment, each molecule is described by a 'reduced' molecular interaction field (MFr) with as many nodes as defined by the input condition ($E_{cut-off}$ and $SD_{cut-off}$).
3. The subsequent step of the FILO procedure is aimed at determining which nodes of reduced fields are really necessary or important for the biological activity of the optimal ligand and which are the energy values they assume. These nodes will form the optimal field, i.e., that one which will be the most similar to the MF of the highest activity molecule and the most different from the MF of the lowest activity molecule. The computation is performed according to the rules of the Modified Simplex (MS) procedure described by Nelder-Mead [21]. The sequential simplex method, introduced by Spendley et al. [22], is a highly efficient, multifactor, empirical feedback strategy that rapidly attains the experimental optimum. The modification by Nelder and Mead of the original simplex method provides for acceleration in directions that are favourable and deceleration in directions that are unfavourable. In the FILO procedure the Simplex algorithm searches for the minimum of the function g which shows how the correlation coefficient r of another function, $f(x) = \text{Log } K_i = a(CI) + b$, varies with the variation of the Carbo similarity index, CI. The Carbo similarity index, CI, is computed between the training set molecular interaction fields (P_{MF}) and the optimal field (P_{OF}) of the hypothetical optimal ligand

$$CI = \frac{\sum_{i=1}^{n} P_{MF} P_{OF}}{\sqrt{\sum_{i=1}^{n} P_{MF}^2} \sqrt{\sum_{i=1}^{n} P_{OF}^2}} \quad (1)$$

where n = number of nodes under computation.

In particular for the third step, FILO processes in a step-wise manner, evaluating node by node. In the Simplex optimization step, each molecule, defined by its reduced interaction field (MFr), is considered as a vertex of the simplex figure placed in a n-dimensional space, where n varies according to the number of nodes considered during the optimization step.

Step 3 starts considering the OF as constituted by only one node, i.e., that one having the highest SD (Step 2) and associating to each vertex of the simplex a value of the correlation coefficient *r* of the function $f(x) = \text{Log } K_i = a(CI) + b$.

In the evaluation of the first node it is important to highlight that the MFr of the training set molecules are formed only by one potential value.

Scheme 1 briefly shows the main steps of the FILO procedure.

According to Scheme 2 the MFr of each molecule is considered as the optimal field, so that as many Carbo similarity indices, CI_{Nn}, as the number N of the training set molecules are computed. CI_{Nn} are related to the biological activity (Log K_i) of the corresponding training set molecules obtaining the correlation coefficients *r* to be associated to each vertex of the simplex.

Due to Equation (1), the only possible values assumed by CI_{Nn} when $n = 1$ (one node) are $+1$ and -1. As a consequence, only two values of *r* are possible, independently of the number N of the training set molecules (Scheme 2, upper part). For the subsequent computations, FILO takes as reference that of the two *r* with the best value.

To summarize, FILO proceeds:

- considering the MFr (formed by the n nodes evaluated and included in the model plus the node under evaluation; n is always $\leq$ k) of each molecule (x) of the training set as the optimal field, and for each molecule x of the training set taken as reference
- calculating N Carbo similarity indices CI_{Nnx} (where N = number of Simplex vertices, n = number of nodes under computation, x = reference molecule)
- relating the CI_{Nnx} to $LogK_i$ of the corresponding training set molecule and, finally
- obtaining as many correlation coefficient values *r* as the number N of the training set molecules, that is as the number of the vertex of the Simplex figure. Thus, every *r* is taken as the value assumed in that vertex by the function *g* to be minimized by the Simplex.

In every Simplex iteration, each vertex is defined by the reduced molecular interaction field MFrn, i.e. only the one which comprises the selected nodes (n). Replacement of the vertex producing the worst response with a new vertex allows translation of the simplex in n-space toward the minimum value. The replacement concerns the vertex coordinates (i.e., the molecular interaction energy values) in such a way that at the end of the iterations optimal interaction energy values (OF) associated with the new vertex, and therefore to a hypothetical new ligand, are obtained but each node has maintained its original geometrical position as defined in the GRID cage. Moreover, in order to obtain energy values appropriate to the common molecular interaction fields, negative and positive thresholds for the MS function were imposed, namely $+10$ and -15 kcal/mol. Convergence is reached when the difference between *r* values of the best and the second worst vertex is lower than 0.002.

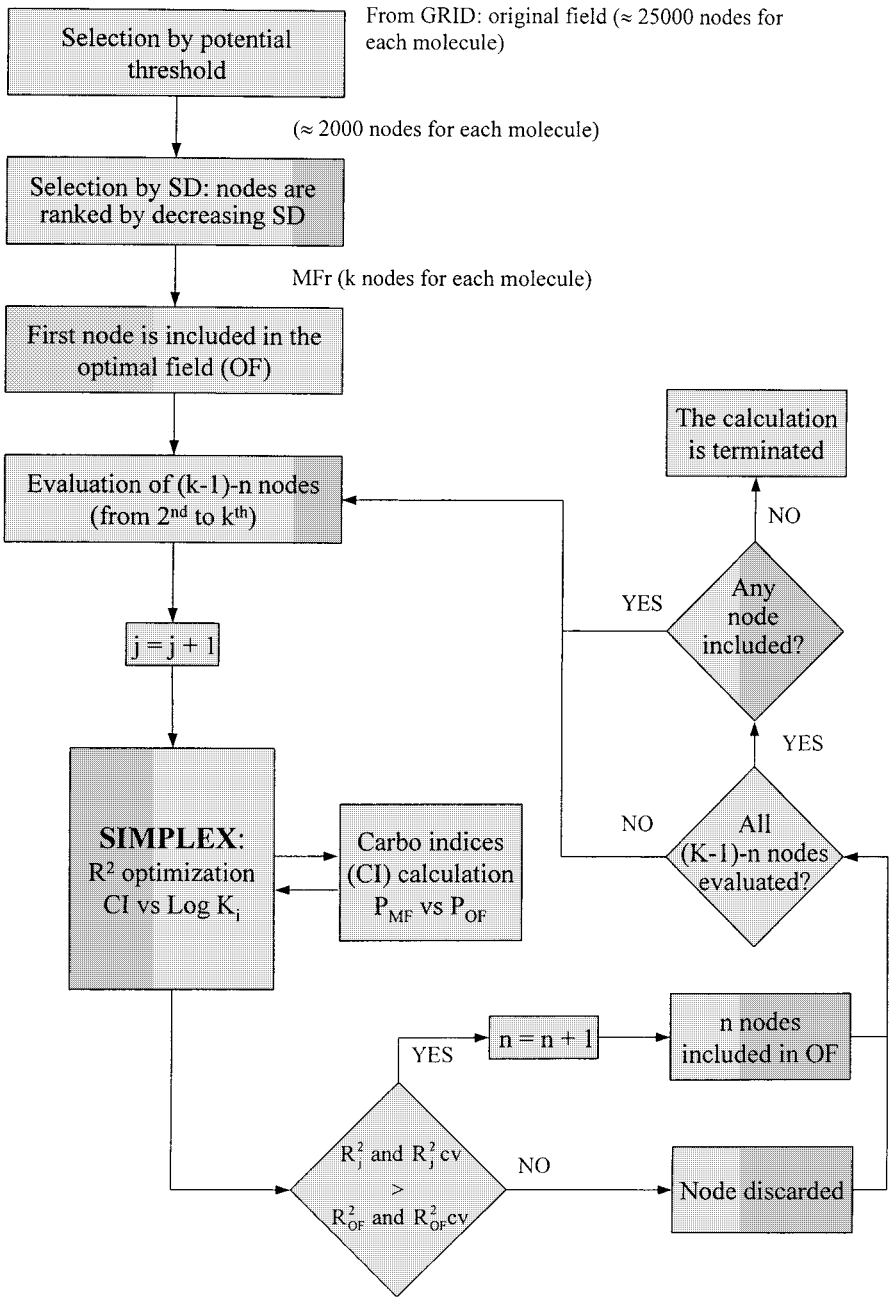
Carbo similarity indices, CI_{ON}, between OF and the MFrn of the training set molecules are calculated and then correlated with the $LogK_i$ of the corresponding molecules to obtain a R_0^2. At this point, another node is included in the computation of the optimal field (Scheme 2). It is important to note that in the evaluation of a new node, FILO starts from the original MFrn+1 of the training set molecules, thus performing the computation on a $N \times n+1$ matrix, N being the number of training set molecules and n being the number of nodes already selected at the moment. The node is included in the optimal field interaction model, only if the R_0^2 between CI_{ON} and activity is greater than R_0^2 of the previous cycle (Nxn matrix), i.e. the greatest among all the calculated. Nodes are included in the computation until a $R_0^2 = 0.65$ is reached. Then the subsequent nodes are taken only if they simultaneously increase R^2 and R_{CV}^2 (LOO leave-one-out). Alternatively, and depending on the kind of data sets, R_{CV}^2 (LTO leave-two-out, LMO leave-more-out) or external Q^2 may be used in place of R_{CV}^2 LOO, Q^2 being evaluated using the molecules of the test set.

$$Q^2 = 1 - \frac{\sum_{i=1}^{N} (y_{i\,\text{exp}} - y_{ipred})^2}{\sum_{i=1}^{N} (y_{i\,\text{exp}} - \bar{y}_{\text{exp}})^2} \qquad (2)$$

Once all nodes are evaluated, FILO reevaluates those nodes discarded in the previous run for their predictive ability (Q^2, R_{CV}^2), taking those which increase Q^2 or R_{CV}^2 of the model.

Results and discussion

The performance of the FILO procedure was tested on a data set which comprises nine structures taken from the Brookhaven National Laboratories Protein Databank (PDB) [23]. They act as HIV-1 protease inhibitors on the same receptor whose 3D structure is



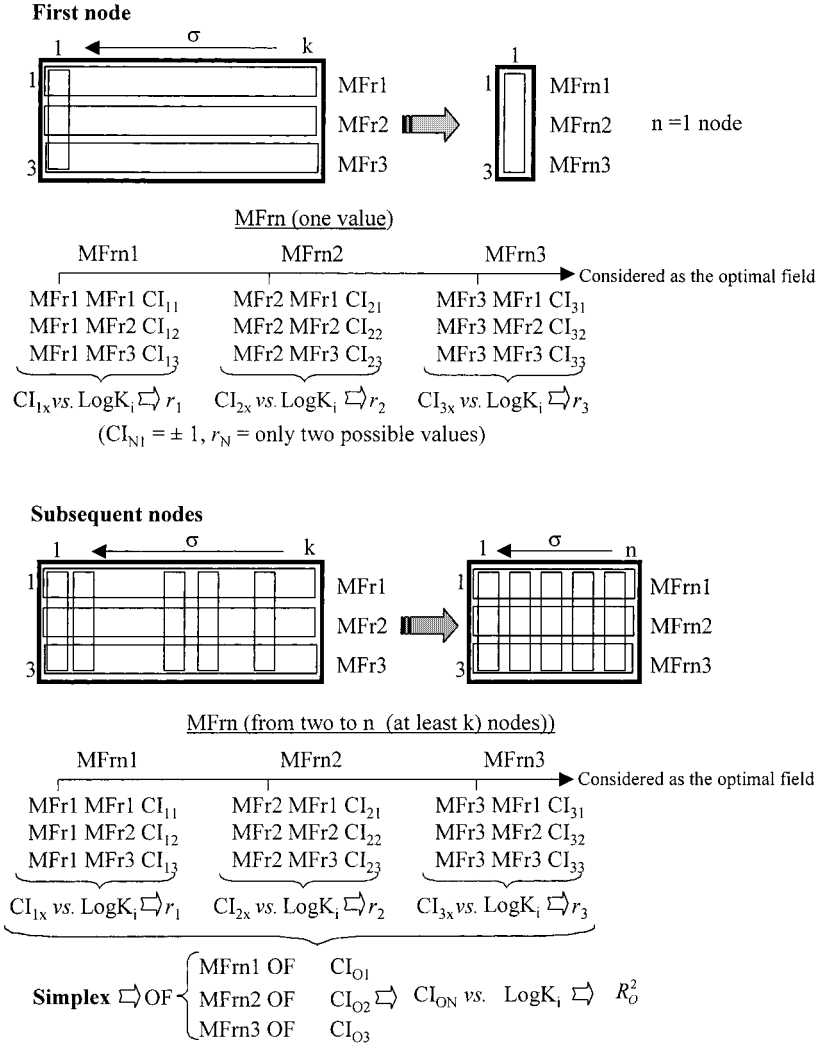
Scheme 1. Main steps of FILO procedure.

known. As a consequence the ligand alignment was achieved by superimposing the substrate binding site (Figure 1).

Furthermore, we used crystal structures because the right binding modes must be available for verifying the quality of the FILO proposed binding site.

The binding site of each of the nine molecules to HIV-1 protease is defined by X-ray crystallographic analysis at resolutions varying from 1.6 to 3.2 Å [24–29].

H bonding atoms of both compounds and protein moiety were added by the Building module of the InsightII program and the structures were further re-



Scheme 2. FILO procedure for three molecules and n nodes.

fined by minimizing the energy of the HIV-1 protease inhibitors while keeping their heavy atoms and all protein atoms at fixed positions.

HIV-1 protease inhibitors exhibit potency ranges of more than 4 log units (Table 1). Because of the small working data set, R^2_{CV} (LOO) is used in the evaluation of nodes. The probes chosen (OH and C3) in this study simulate the functional groups of the active site cavity of the protease with which the interactions may be possible. The selection ensures that electrostatic, steric and chemical properties, such as hydrogen bonding, are well considered.

OH probe. This probe represents a hydroxyl group bonded to an aromatic system capable of both donating and accepting one hydrogen bond.

GRID calculation performed with the OH probe on the HIV-1 protease data set produced 23925 interaction energies with each molecule which were reduced to ca. 2400 following the application of the energy cut-off (|energy| $\geq$ 2 kcal/mol) performed by FILO.

The coordinates of the nodes of the optimal field, the statistical parameters of the model found and its graphical representation are reported in Table 2 and Figure 2, respectively. Comparison of Figure 3 and Figure 4, showing the active site of HIV-1 protease [21–26] and the optimal attractive interaction field (i.e., negative values) obtained by FILO respectively,

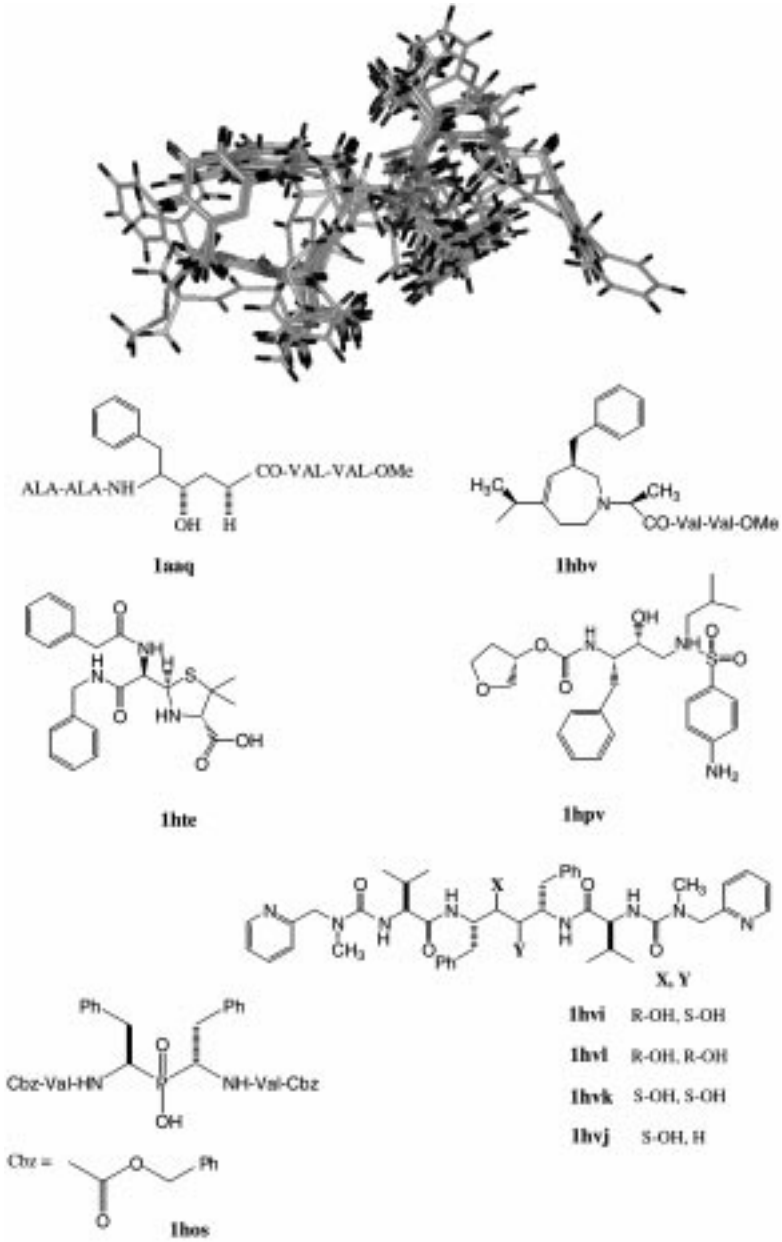


Figure 1. Ligands from nine pairs of ligand-protein crystal complexes used in the evaluation of the FILO procedure.

indicates a very good correspondence between the actual (Figure 3) and calculated (Figure 4) situation. The RMS deviation, calculated by superimposing the atoms of the interacting AA-residues of the binding site onto the attractive nodes obtained by FILO, is 0.9 Å. Attractive interactions between the binding site AA-residues, Gly A27, Asp A29, water 501, Gly B48 and Asp A25 and the ligand are well identified (regions A–E of Figures 3 and 4).

Of course, owing to the use of the OH probe, the optimal interaction field mainly points out the H bonding interactions. AA-residues which are not important for discriminating $\log K_i$ values of the data set molecules, are not found with either OH or other probes (Figure 4).

In Figure 5 the repulsive interactions (i.e., positive potential values) found by FILO are shown. The meaning of these interactions is not as easy to interpret as

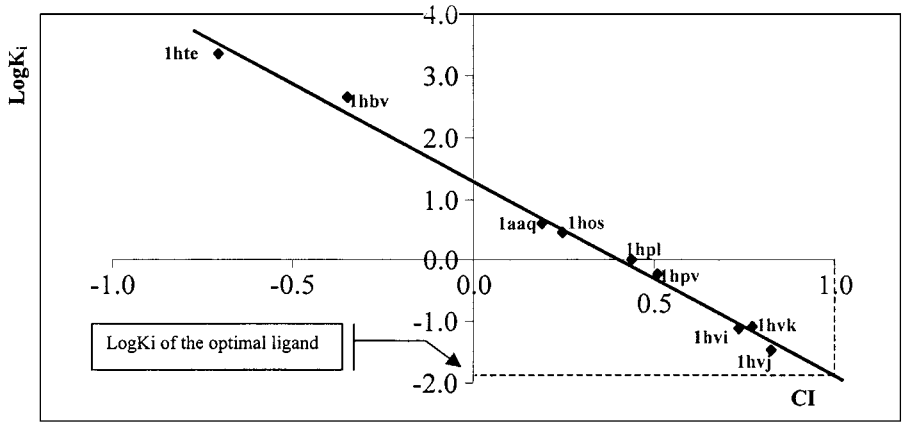


Figure 2. Graphical representation of the model obtained by the FILO procedure. On the abscissa are reported the Carbo similarity indices CI between the optimal field (OF) computed by the methodology and the molecular interaction field of each molecule of the training set. The Log Ki value corresponding to CI=1 is the one of the hypothetical optimal ligand.

Table 1. HIV-1 protease inhibitor bioactivity

PDB code	K_i (nM)	$LogK_i$
1hte	2300	3.36
1hbv	430	2.63
1aaq	4	0.6
1hos	2.8	0.447
1hvl	1	0
1hpv	0.6	−0.22
1hvi	0.084	−1.075
1hvk	0.077	−1.11
1hvj	0.035	−1.46

Table 2. Coordinates of the optimal field and statistical parameters of the model obtained by the FILO procedure (OH probe)

No. of nodes	Coordinates			Potential
	x	y	z	(kcal/mol)
1	5.000	4.000	14.000	−2.658
2	1.000	4.000	13.000	−2.075
3	6.000	4.000	14.000	−1.567
4	1.000	5.000	13.000	−5.697
5	3.000	2.000	16.000	−4.524
6	5.000	−1.000	19.000	5.12
7	8.000	−6.000	13.000	0.709
8	5.000	−3.000	12.000	−7.730
9	7.000	−7.000	14.000	−4.237
10	3.000	−3.000	13.000	10.073
11	1.000	6.000	12.000	−7.366
12	4.000	3.000	14.000	6.035
R^2	R^2_{CV}	Equation model		
0.993	0.936	$LogK_i = -3.179(CI) + 1.291$		

for the attractive areas seen in Figure 4 where a correspondence exists between nodes selected by FILO and the binding site regions where a hydrogen-bonding interaction is present.

Referring to OF repulsive nodes, their interpretation has to be understood in such a way that the hypothetical optimal ligand has to occupy the areas individualized by nodes to be repulsion possible. Thus, repulsive areas, characterized by repulsive (positive) potential values, do not correspond with the presence of binding site groups. Comparison between Figure 3 and Figure 5 clearly shows the lack of AA-residues corresponding to FILO repulsive areas (indicated by arrows in Figure 5). This is because the binding protein probably has to present a cavity into which ligands may fit.

C3 probe. This probe represents the electronic properties of an sp^3 carbon atom. The probe does not interact electrostatically and it does not form H-bonds with the target molecule. Thus, the GRID calculation accounts for the steric interaction between the target and the probe.

GRID calculation performed with the C3 probe on HIV-1 protease produces 23925 interaction energies with each molecule which were reduced to ca.

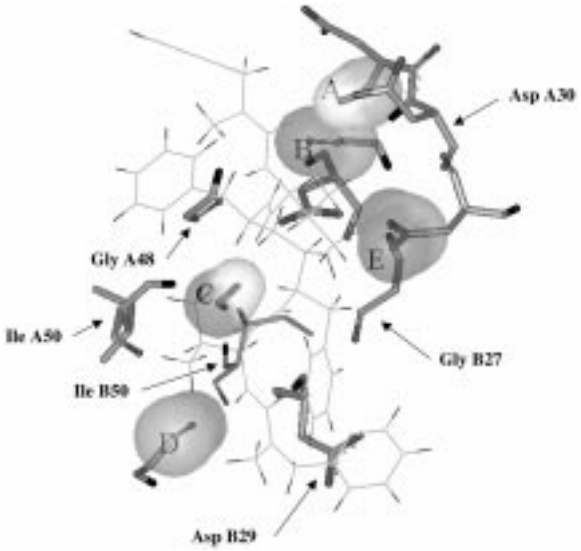


Figure 3. The active site of HIV-1 protease showing the AA-residues which interact with compound 1hvj taken as representative molecule. Spheres are representative of the binding site AA-groups derived by FILO. The arrows indicate the other binding site AA-residues of HIV-1 protease.

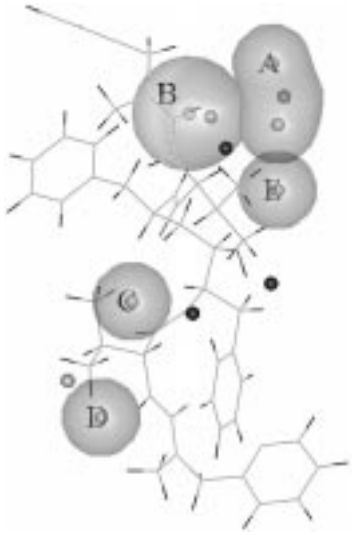


Figure 4. OH probe, HIV-1 protease. Attractive optimal interaction field regions are highlighted by spheres. Only the most active compound is represented (1hvj) for clarity.

2600 following the application of the energy cut-off (|energy| $\geq$ 2 kcal/mol) performed by FILO.

According to the properties of the C3 probe (i.e., sensibility for steric interactions between the target and ligand), only the repulsive (positive) interaction nodes found by FILO are interesting, since they define the sterically ligand-permitted areas.

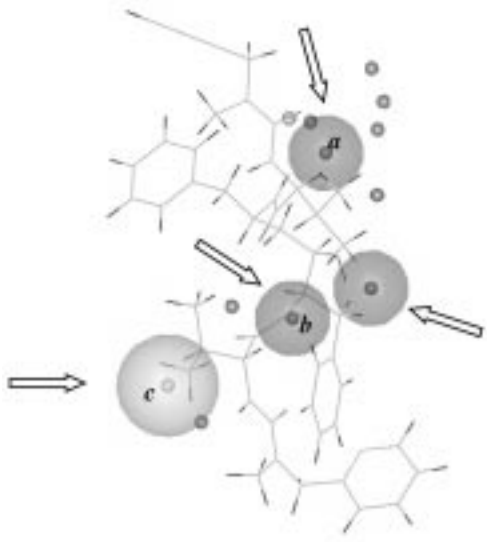


Figure 5. OH probe, HIV-1 protease. Repulsive optimal interaction field regions are highlighted by spheres. *a*, *b* and *c* indicate repulsive regions also detected by the C3 probe (Figure 6). Only the most active compound is represented (1hvj) for clarity. Arrows indicate repulsive areas.

Figure 6 shows the optimal field computed by FILO with the C3 probe. Considering the different characteristics of the used probes, the agreement between Figures 5 (OH probe) and 6 (C3 probe) is quite good. Three of the four regions reported in Figure 5 (*a*, *b* and *c*) are present in Figure 6 and their relative distances, calculated after superposition guided by the 1hvj ligand, are: *a*-*a*, 1.4 Å; *b*-*b*, 1.4 Å; *c*-*c*, 2.2 Å.

In a general way it is possible to confirm that FILO yields the optimal field of a hypothetical optimal ligand whose biological activity is equal or higher than that of the best molecule of the training set. This is a direct consequence of FILO definition and design, since the molecular interaction field of the optimal ligand is that which, at least, produces with the optimal field obtained by FILO a CI value of 1.

As a consequence and according to Figure 2, the biological activity value obtained for CI=1 is greater or equal to the value of the most active compound of the series.

Only exceptionally does the number of nodes retained with the FILO procedure exceed 20, which makes the final 3D model easy to interpret and use. Nodes retained may assume both positive and negative values.

For positive values, repulsion has to occur between the optimal ligand and the probe, that is, a bulky group situated in correspondence to nodes with pos-

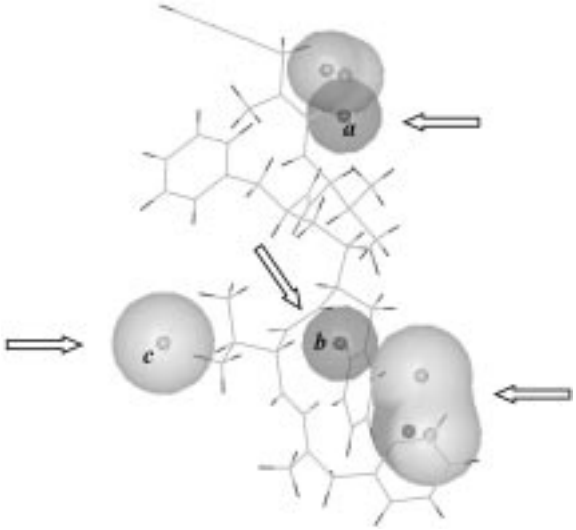


Figure 6. C3 probe, HIV-1 protease. Optimal interaction field regions are highlighted by spheres. *a, b* and *c* indicate repulsive regions also detected by the OH probe (Figure 5). Arrows indicate repulsive areas. Only the most active compound (1hvj) is represented for clarity.

itive potential values should increase the activity. As a consequence and only for no-charged probes, in these nodes the receptor should present a cavity into which the ligand may position itself.

On the other hand, nodes with negative potential values indicate that the rational design of a new ligand has to consider the presence of a group able to establish an attractive interaction with the probe. Thus, it is possible to assume the presence in the binding site of an AA-residue which is able to interact with the ligand.

Conclusions

The results obtained and presented in this paper encourage further investigations on different and more consistent data sets. Despite the number of molecules of the data set considered, the correspondence between the observed (crystallographic) and calculated (FILO) data is very promising in view of rational drug design. Regions of attractive and repulsive interactions pointed out by FILO might be conveniently considered in the development of new drugs, particularly for those systems where the receptor structure is unknown. At present, data sets having these characteristics are under study in our laboratory.

Acknowledgements

This work received financial support from the Italian Ministry of the University.

References

1. Cramer III, R.D., Patterson, D.E and Bunce, J.D., J. Am. Chem. Soc., 110 (1988) 5959.
2. Goodford, P.J., J. Med. Chem., 28 (1985) 849.
3. Snyder, J.P. and Rao, S.N., Chem. Design Autom. News, 4 (1989) 13.
4. Snyder, J.P. and Rao, S.N., Crai Channels, 11 (1990) 12.
5. Snyder, J.P., Rao, S.N., Kohler, K.F. and Pellicciari, R., In Angeli, P., Gulini, U. and Quaglia, W. (Eds.) Trends in Receptor Research, Elsevier Science Publishers, Amsterdam, 1992, pp. 367–403.
6. Snyder, J.P., Rao, S.N., Kohler, K.F. and Vedani, A., In Kubinyi, H. (Ed.) 3D-QSAR in Drug Design: Theory, Method and Applications, ESCOM, Leiden, 1993, pp. 336–354.
7. Momany, F., Pitha, R., Klimkowsky, V.J. and Venkatachalam, C.M., In Hohne, B.A. and Pierce, T.H. (Eds.) Expert Systems and Applications in Chemistry, ACS Symp. Ser. 408, 1989, p. 82.
8. Walters, D.E. and Hints, R.M., J. Med. Chem., 37 (1994) 2527.
9. Doweyko, A.M., J. Med. Chem., 37 (1994) 1769.
10. Hahn, M. and Rogers, D., J. Med. Chem., 38 (1995) 2091.
11. Hahn, M., J. Med. Chem., 38 (1995) 2080.
12. Vedani, A., Zbinden, P. and Snyder, J.P., J. Receptor Res., 13 (1993) 163.
13. Vedani, A., Zbinden, P., Snyder, J.P. and Greenidge, P.A., J. Am. Chem. Soc., 117 (1995) 4987.
14. Zbinden, P., Dobler, M., Folkers, G. and Vedani, A., Quant. Struct.-Act. Relat., 17 (1998) 122.
15. Lucasius, C.B. and Kateman, G., Chemom. Intell. Lab. Syst., 25 (1994) 99.
16. Baroni, M., Costantino, G., Cruciani, G., Riganelli, D., Valigi, R. and Clementi, S., Quant. Struct.-Act. Relat., 12 (1993) 9.
17. Pastor, M., Cruciani, G. and Clementi, S., J. Med. Chem., 40 (1997) 1455.
18. InsightII 97.0, Molecular Simulations, San Diego, CA, USA.
19. SYBYL – Molecular Modeling Software, 6.4, Tripos Incorporated, St. Louis, MO, USA.
20. GRID v. 16, Molecular Discovery Ltd., University of Oxford, England, SGI.
21. Nelder, J.A. and Mead, R., Comput. J., 7 (1965) 308.
22. Spendeley, W., Hext, G.R. and Himsworth, F.R., Technometrics, 4 (1962) 441.
23. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F. Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M., J. Mol. Biol., 112 (1977) 535.
24. Dreyer, G.B., Lambert, D.M., Meek, T.D., Carr, T.J., Tomaszek, T.A., Jr., Fernandez, A.V., Bartus, H., Caccivillani, E., Hassell, A.M., Minnich, M., Petteway, S.R., Jr. and Metcalf, B.W., Biochemistry, 31 (1992) 6646.
25. Hoog, S.S., Zhao, B., Winborne, E., Fisher, S., Green, D.W., DesJarlais, R.L., Newlander, K.A., Callahan, J.F., Moore,

M.L., Huffam, W.F. and Abdel-Meguid, S.S., J. Med. Chem., 38 (1995) 3246.
26. Abdel-Meguid, S.S., Zhao, B., Murthy, K.H.M., Winborne, E., Choi, J., DesJarlais, R.L., Minnich, M.D., Culp, J.S., Debouck, C., Tomaszek, T.A., Jr., Meek, T.D. and Dreyer, G.B., Biochemistry, 32 (1993) 7972.
27. Kim, E.E., Baker, C.T., Dwyer, M.D., Murcko, M.A., Rao, B.G., Tung, R.D. and Navia, M.A., J. Am. Chem. Soc., 117 (1995) 1181.
28. Jhoti, H., Singh, O.M.P., Weir, M.P., Cooke, R., Murray-Rust, P. and Wonacott, A., Biochemistry, 33 (1994) 8417.
29. Hosur, M.V., Bhat, N., Kempf, D.J., Baldwin, E.T., Liu, B., Gulnik, S., Wideburg, N.E., Norbeck, D.W., Appelt, K. and Erickson, J.W., J. Am. Chem. Soc., 116 (1994) 847.