J Mol Model (2006) 12: 965–972
DOI 10.1007/s00894-006-0109-z

ORIGINAL PAPER

Michel E. B. Yamagishi · Natália F. Martins ·
Goran Neshich · Wensheng Cai · Xueguang Shao ·
Alexandre Beautrait · Bernard Maigret

A fast surface-matching procedure for protein–ligand docking

Received: 5 January 2005 / Accepted: 1 December 2005 / Published online: 4 May 2006

Abstract A very simple, fast, and efficient scheme is proposed for performing preliminary protein–ligand docking as the first step of intensive high-throughput virtual screening. The procedure acts as a surface-complementarity filter that first calculates the 2D-contour maps of both the protein cavity and of the ligands using a spherical harmonics description of the associated molecular surfaces. Next, the obtained 2D-fingerprint images are compared to detect their complementarity. This scheme was tested on three typical cases of protein cavities, namely, a well-closed pocket, a small open pocket, and a large open one. For that purpose, for each case, a sample of 101 ligand conformers was generated (the X-ray one and 100 different conformers generated using simulated annealing), and these conformational samples were ranked according to the complementarity with the protein cavity surface. Compared to traditional docking procedures such as FRED (considered as typical of a very fast rigid body docking algorithms) and GOLD (considered as typical of the more accurate flexible docking algorithms), our procedure was much faster and more successful in detecting the right X-ray conformation. We did, however, identify a certain weakness in the case of the very large pocket where results were not as expected. In general, our method could be used for incorporating indirectly flexibility in protein–ligand docking calculations as such a scheme can easily handle several conformational states of both the protein and the ligand.

Keywords Spherical harmonics · Fingerprints · Surface matching · Docking

Background

Recent development in combinatorial chemistry and high-throughput screening (HTS) has significantly increased the number of compounds for biological essays [1]. However, when screening hundreds of thousands of new chemical entities or natural-product extracts, the cost of HTS reagents and equipment could hamper the HTS process. Practical realities of compound-handling capabilities and HTS costs and the limitations of a particular chemistry scheme often come to bear on the question of finding appropriate supplementary approaches to reduce the costs while exploring a larger chemical diversity [2].

Virtual screening methods (VHTS) were therefore developed to handle large sets of compounds and to improve the "hit-rate" of the discovery programs. In this respect, virtual screening, or in silico screening, is an attractive approach able to increase levels of interest in the pharmaceutical industry as a productive and cost-effective technology in the search for novel lead compounds [3]. However, while virtual screening is potentially able to select valuable hits among chemical libraries of millions of compounds by using common docking algorithms, it is also costly in terms of central processing unit (CPU) time

MB proposed the fingerprint-matching algorithm, carried out the fingerprint-matching experiments and helped to draft the manuscript. NM carried out the docking experiments with FRED and GOLD. GN participated in the coordination and helped to draft the manuscript. WC generated the fingerprints using MSSH. XS generated the fingerprints using MSSH. BM participated in the design of the study, coordination and helped to draft the manuscript. All authors read and approved the final manuscript.

M. E. B. Yamagishi (✉) · G. Neshich
Embrapa Informática Agropecuária,
Caixa Postal 6041, Av. Dr. André Tosello, n° 209,
Barão Geraldo-Campinas, (SP)-CEP 13083-886, Brazil
e-mail: michel@cbi.cnptia.embrapa.br

N. F. Martins
Embrapa-Genetic Resources and Biotechnology,
CP 02372, Brasília-DF, Brazil

W. Cai · X. Shao
Department of Chemistry,
University of Science and Technology of China,
Hefei, Anhui, 230026, People's Republic of China

A. Beautrait · B. Maigret
UMR CNRS/UHP 7565, eDAM group,
H. Poincare University, BP 239,
54506 Vandoeuvre les Nancy Cedex, France

and efficiency so that its use “as is” is questionable for performing high-throughput on millions of compounds or more. Therefore, the question is now how to improve the speed and the quality of docking algorithms to perform large scale and efficient virtual screening [4].

Within this line of research, we have introduced the funnel strategy for docking, consisting of different levels of filtering within the docking algorithms. This procedure starts with the matching of geometrical surfaces of the target and the candidate ligands (the fastest filtering step) and then goes to the more sophisticated free-energy calculations (more time-consuming filtering step). The first geometrical filter should be able to handle, in a fast way, millions of molecules, filtering in a “crude” way the relevant compounds to be passed to the next filters. The latter one could be achieved by considering only a few tenths of molecules for the free-energy calculations.

In docking calculations, filtering by surface matching between the protein target and putative ligands has already been used and has proved to be useful [5]. The main problem in this aspect is the matching algorithm used for ranking the screened molecules according to a protein pocket that usually considers both 3D-partners to measure their possible surface complementarity. We have recently proposed a fingerprint concept that is able to translate the 3D-information into a 2D-map able to describe the whole 3D-patterns using spherical harmonics [6]. This concept was originally applied to several examples of existing protein–ligand complexes found in the Protein Data Bank (PDB), but the previous application was limited as we only used a sample of X-ray ligand conformations for the target/ligand fingerprint comparison.

Here, we extend the previous concept by filtering large samples of different conformations to see if the method is able to detect among all these conformations the ones able to fit best within the target pockets. For this purpose, we selected three representative cases of target pockets, namely, a well-defined closed pocket, an open mouth pocket, and a large opened pocket. Except for the last case, the method was able to detect the right conformation when compared to the X-ray data, proving that, and with some additional information in the case of large pockets, the method is really accurate and very fast for filtering large databases of candidate compounds.

Results

The procedure flowchart is as follows. First, the Molecular Surface using Spherical Harmonics (MSSH) program, [6]) generates the kinemage-format (http://kinemage.biochem.duke.edu/kinemage) files to store the fingerprints obtained for each conformation of each ligand. Each of these images was next stored as an extended vector, so that, using the cosine scoring function, we were able to rank all pairs of images further. Our results are shown in Table 1a–c:

(1) Table 1a describes the results for the easiest case: the well-defined cavity observed in the **1ie9** system. Our results show that the cosine scoring function was not only able to give a high score to fingerprint pairs corresponding to the configuration described in the PDB file, but also rank it in the first place. Figure 1a shows the fingerprints of the first three best-scored conformations of the ligand considered.

(2) Table 1b describes the results for an open cavity (**1qf0** system) and shows that, as in the previous case, the right pairs of fingerprints corresponding to the configuration described in the PDB file and positioning of the ligand within the protein cavity were also found with the highest cosine score. Figure 1b shows the fingerprints of the first three best scored conformations of the associated ligand.

(3) Table 1c shows the results obtained for the most difficult case, i.e., when the open mouth (**1npo** system) was much larger than in the last one. In this case, the cosine scoring function was not able to identify the right pair (the one identified by X-ray data). Indeed, the right pair appeared in the 40th position in the rank even using 30 contour maps, as shown in Fig. 1c. This can be explained by the fact that because the open mouth was too large, the inflation procedure applied during spherical-harmonic generation and afterwards used to obtain the cavity’s fingerprint, extends too much outside the cavity, so that the fingerprint obtained protrudes too much and is unable to represent the surface of the cavity with good accuracy.

Discussion

To compare this approach with more “classical” docking methods, we have applied both rigid body docking and flexible docking programs to the same sets of data as used above in our method. Considering extensive discussions about the merits vs failures of several docking procedures [7, 8], we have retained the FRED method (http://www.eyesopen.com) which is considered as an accurate and very fast rigid-body docking algorithm [9] and the GOLD method (http://www.accelrys.com) considered as one of the top for performing the flexible docking approach [10]. For both methods, we have used several scoring functions, namely, shape, chemscreen, PLP, and screenscore for FRED, goldscore, and screenscore for GOLD. For GOLD, we have also proceeded to two runs, one considering only one possibility for each of the 100 conformations of each ligand, the second one considering 50 possibilities for the X-ray conformation of the ligand. This second run was used to see how the GOLD algorithm behaved in the case of multiple-conformation searches.

From the results obtained (shown partly in Table 1a–c and on Fig. 2a–c), it appears that both FRED and GOLD, whatever the scoring function used is, are able to predict the X-ray conformation of the ligand in the **1ie9** system as the top-ranked conformation for the docking. For the second case, named the **1qf0** system, GOLD was unable to detect the X-ray conformation of the ligand in the ten top-ranked solutions among the 100 conformers sampled: for

Table 1a Results of the **1IE9** PDB complex. The conformation identification numbers are given to illustrate the difference in the ranking between the 3 methods used, and should not be related to similarity between conformations. RX means a conformation similar to the structure of the ligand as found in the X–ray data

Cos rank	Conf. *N*°	Cos score	FRED rank	Conf. *N*°	FRED score$^{\&}$	GOLD rank	Conf. *N*°	GOLD score$^{\#}$
1	RX	0.1232	1	RX	−27.829	1	RX	79.49
2	2	0.1066	2	12	−26.850	2	17	76.38
3	77	0.1055	3	43	−25.845	3	23	75.13
4	75	0.1017	4	54	−25.838	4	94	74.87
5	15	0.0996	5	90	−25.715	5	20	74.59
6	68	0.0989	6	52	−25.351	6	54	74.53
7	76	0.0968	7	84	−24.956	7	45	74.23
8	69	0.0919	8	32	−24.826	8	85	73.21
9	61	0.0910	9	41	−23.258	9	46	73.06
10	8	0.0909	10	39	−22.981	10	87	72.90

&: Scoring function: screenscore
#: Scoring function: goldscore

Table 1b Results of the **1QF0** PDB complex. The GOLD score for the RX conformation, not found within the “best 10” docked possibilities in that case, is only 74.56

Cos rank	Conf. *N*°	Cos score	FRED rank	Conf. *N*°	FRED score$^{\&}$	GOLD rank	Conf. *N*°	GOLD score$^{\#}$
1	RX	0.0996	1	RX	−61.057	1	51	93.75
2	91	0.0909	2	22	−38.771	2	9	92.50
3	5	0.0909	3	17	−33.973	3	3	92.25
4	42	0.0855	4	27	−33.865	4	98	91.87
5	95	0.0840	5	59	−33.825	5	84	91.70
6	15	0.0831	6	41	−33.577	6	35	91.55
7	41	0.0815	7	77	−32.391	7	64	91.37
8	38	0.0813	8	74	−31.910	8	17	91.10
9	79	0.0785	9	63	−31.794	9	80	90.90
10	48	0.0784	10	93	−31.646	10	16	90.65

&: Scoring function: chemscore
#: Scoring function: goldscore

Table 1c Results of the **1NPO** PDB complex. The COS and GOLD scores for the RX-like conformation which was not found within the "best 10" are 0.1478 and 47.57 respectively. No results were given for FRED as the program was crashing in this case

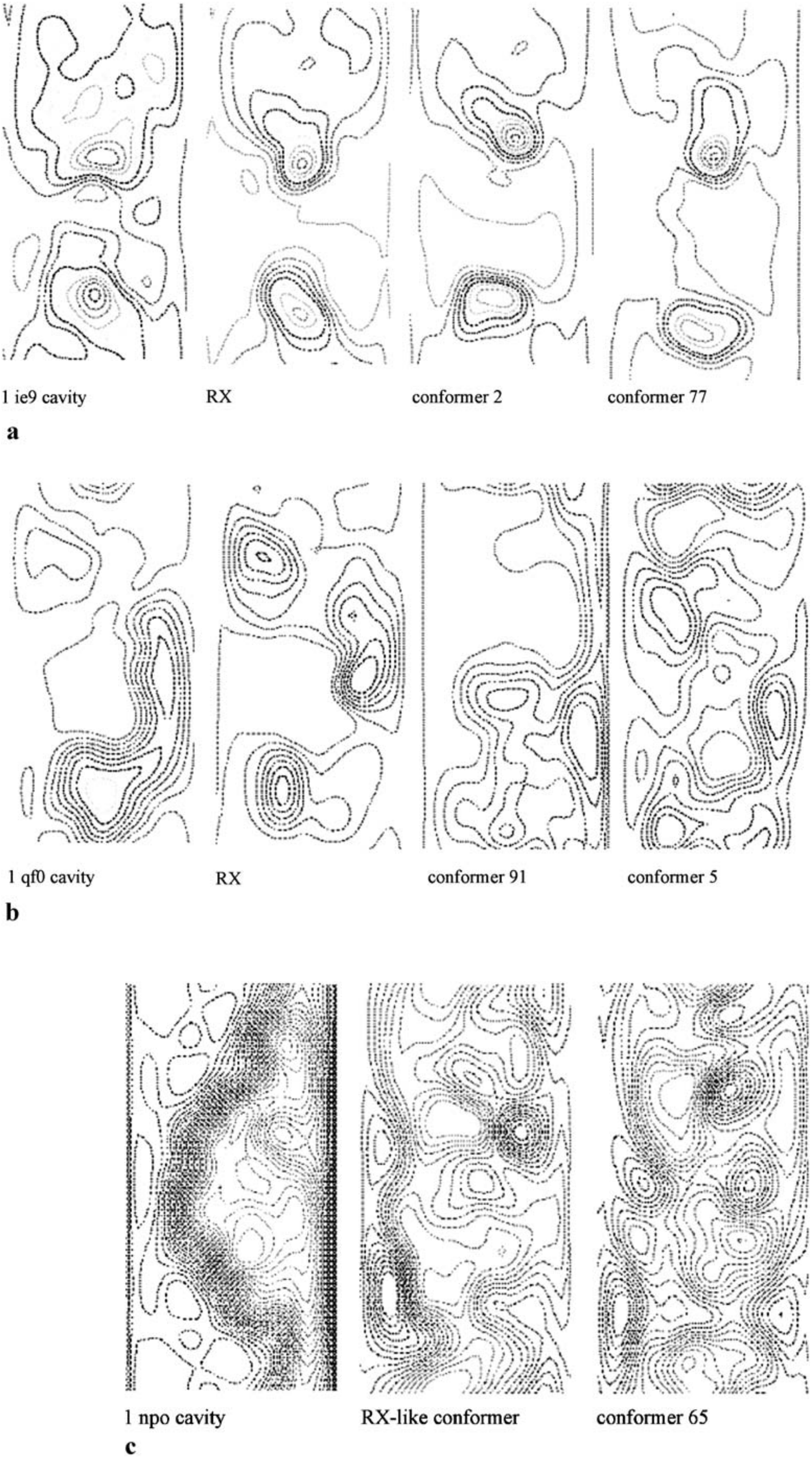
Cos rank	Conf. *N*°	Cos score	GOLD rank	Conf. *N*°	GOLD score$^{\#}$
1	100	0.2545	1	70	74.22
2	65	0.1810	2	62	71.51
3	4	0.1795	3	69	71.29
4	79	0.1739	4	37	70.76
5	14	0.1694	5	31	68.70
6	55	0.1671	6	27	67.20
7	66	0.1661	7	8	66.39
8	38	0.1659	8	65	65.90
9	2	0.1642	9	78	65.27
10	77	0.1641	10	51	64.88

#: Scoring function: goldscore

FRED results, the situation is better as FRED is able to detect the X-ray conformation among the ten top-ranked, whatever the scoring function used is, and rank the X-ray conformer as the top one when the “screenscore” function is used. The situation is worse for both FRED and GOLD in the case of the **1npo** complex as the FRED program crashed in that case, whatever the conditions used are, while GOLD finds solutions far away from the X-ray one.

The price, in terms of CPU time between our approach and the FRED and GOLD ones, is definitively in favor of

Fig. 1 The contour maps represent the variation of r as a function of (θ, ϕ) for any point $M(r, \theta, \phi)$ on the molecular surfaces expressed in terms of spherical coordinates. The x–y coordinates of the maps represent the variations of θ and ϕ from 0–180° and 0–360°, respectively. **a** Fingerprints for the **1IE9** system. **b** Fingerprints for the **1QF0** system. **c** Fingerprints for the **1NPO** system



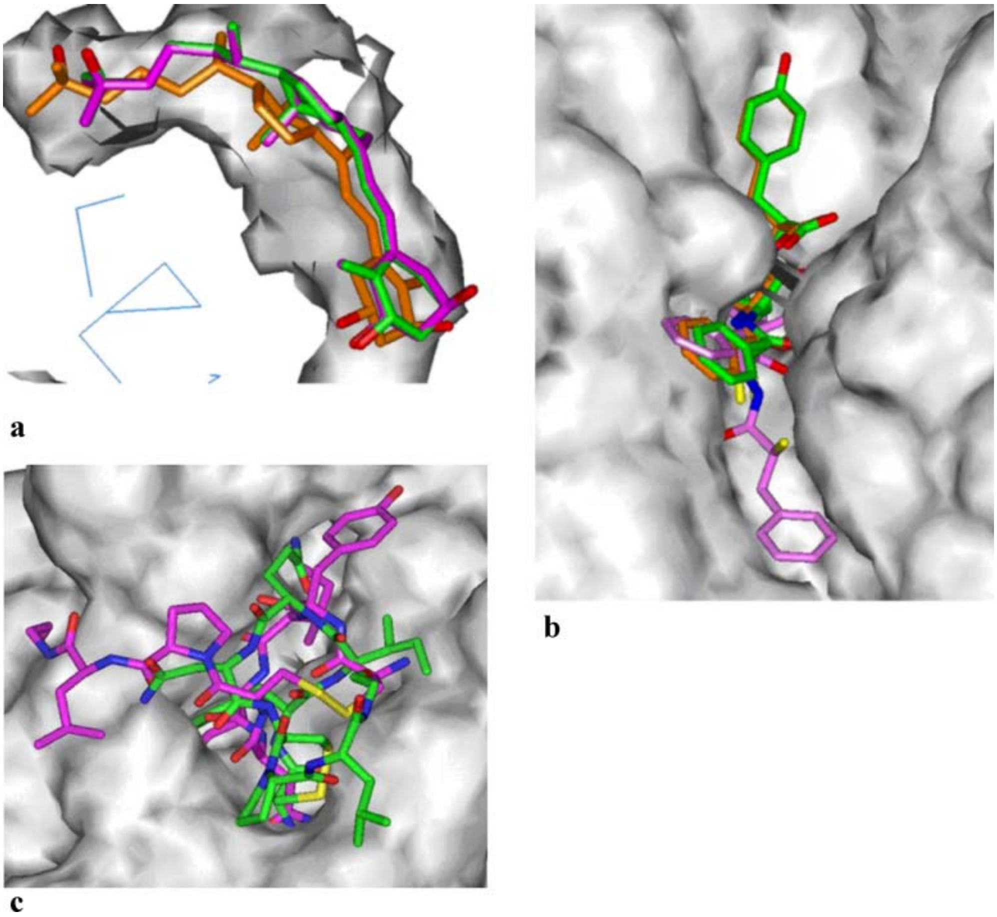


Fig. 2 **a** Position of the ligand in the **1IE9** system according to the best ranks obtained from Table 1a: *COS* result is depicted in *green*, *GOLD* result in *purple* and *FRED* result in *orange*. **b** Position of the ligand in the **1QF0** system according to the best ranks obtained from Table 1a: *COS* result is depicted in *green*, *GOLD* result in *purple* and *FRED* result in *orange*. **c** Position of the ligand in the **1NPO** system according to the best ranks obtained from Table 1a: *COS* result is depicted in *green*, *GOLD* result in *purple* and *FRED* result in *orange*

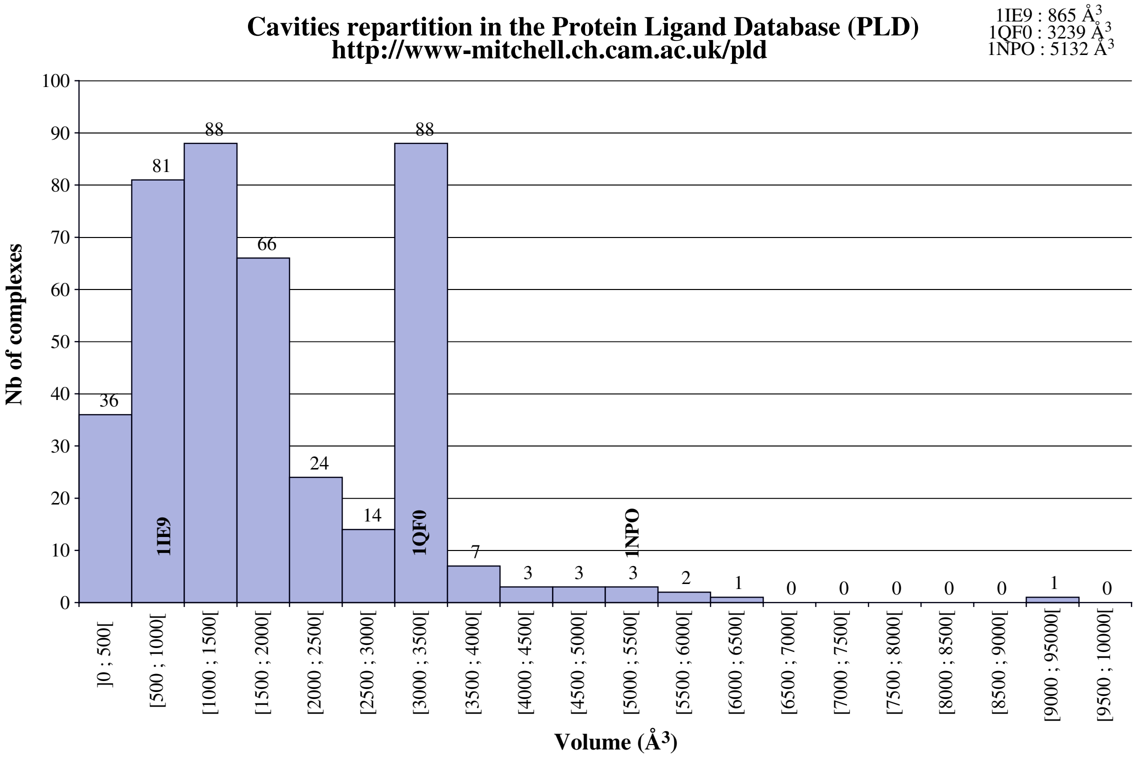


Fig. 3 Number of complexes vs pockets volume

our method. Our method required only a few seconds on a single PC computer with Pentium 4, 2.2 GHz (after obtaining the fingerprint images) for matching the 2D-contours with the target function. At the same time, several minutes are necessary for FRED (using the same performance PC computer) and even a few hours for GOLD (around 6,000 s using eight processors on an SGI O3800 machine).

Conclusion

These results highlight the usefulness of using a simple surface-matching to retrieve interesting ligand conformers that are able to complement a target pocket spatially and therefore, to detect possible candidates for performing more elaborate docking calculations. The surface matching presented here therefore appears as a very simple criterion to compare molecular surfaces efficiently. Associated to the 2D-contour maps obtained from our spherical-harmonics molecular surface algorithm, this fingerprint-matching strategy may be considered as an efficient preliminary filter for eliminating a large part of the candidate molecules in virtual screening, opening room for high-throughput facilities. Moreover, this strategy allows an indirect consideration of both protein and ligand flexibility as it allows considering many conformers of both partners (for example: several side-chain positions for the protein part, and a large body of conformers for the ligands). This method can also be improved easily by incorporating physical–chemical properties associated to the fingerprints by painting them according to these properties. Work is in progress along these lines in the associated laboratories.

Materials and methods

To test the validity of the proposed approach, while not applying it to a huge number of protein/ligand complexes, we first analyzed the geometrical characteristics of the binding sites in a public-domain database of 485 protein–ligand complexes [11]. This analysis, performed with the MSSH program [6], revealed that the binding pockets could be classified roughly into three categories (see Fig. 3) according to their volumes and surface areas. Therefore, we selected only three representative cases of protein–ligand complexes within this database according to their pocket characteristics as representative of the whole complexes, namely, the **1IE9**, **1QF0** and **1NPO:**

1. **1IE9** describes the Vitamin D receptor complexed to superagonist 20-Epi ligands and presents an almost closed pocket in which the ligand is entirely embedded;
2. **1QF0** describes the binding of α-mercaptoacyldipeptides in the thermolysin active site and shows a narrow but still open pocket;
3. **1NPO** is the neurophysin–oxytocin complex presenting a large ligand bound in a large and open pocket.

To filter ligand conformations and to detect the most adapted ones to the protein cavity, for each system, we have proceeded into three successive steps:

1. conformational sampling of the ligands
2. calculation of the fingerprints for the target and for the ligand (one fingerprint per conformation)
3. matching the target vs ligand fingerprints.

Conformational sampling

For each of the three systems considered here, we have investigated the conformational possibilities of each ligand by intensive MD simulated annealing, producing 100 different minimum-energy conformations suitable for docking inside the receptor groove (the RMSDs between all heavy atoms of all conformers are ≥2 Å). We checked that the sampling was large enough to include at least one conformer similar to the one found in the X-ray complex. In all these conformational sampling calculations, we used the Accelrys InsightII software (http://www.accelrys.com):

The following procedure was used to sample the ligands' conformational spaces

1. A starting conformation of each ligand, found in the protein–ligand complex described in a corresponding PDB file, was subjected to 10,000 steps of energy minimization using a conjugate-gradient algorithm. The consistent valence force field (CVFF) was used in the Discover molecular mechanics and dynamics program.
2. These starting conformations were then used for a pseudo-simulated annealing conformational sampling. Simulated annealing involves a temperature increase of the system, followed by a slow cooling to avoid local minima, thereby, trying to locate the global minimum region of the energy function. This conformational sampling was performed using molecular dynamics (MD) in a vacuum (over 40,000 steps with a time step of 1 fs. The equations of motion were integrated using the Leapfrog version of the Verlet algorithm. The (N, V, T) ensemble was used at a fixed value of T with the Berendsen algorithm. This method allows maintaining the temperature at fixed value, by means of a coupling to an external bath. The simulated annealing-like method used here consists of at least one thousand loops of slow cooling, each one leading to a low energy conformation. Each loop begins by fixing the temperature to 1,000 K, followed by 5,000 steps of MD. The temperature was then decreased by steps of 100 K, decreasing the temperature by 100 K every 5,000 steps, so that after 40,000 steps, the temperature of the system corresponds to 300 K.
3. The final conformation obtained at the end of this process was energy-refined using again the conjugate-gradient algorithm and compared to all the previous ones (including the starting X-ray conformation). If the RMSD value between this working conformation and

all the previous ones is larger or equal to 2 Å, the working conformation was stored and the process was continued using this conformation to start a new simulation at high temperature with a slow cooling stage, as described above. If the working conformation was similar to a previous one, the process was repeating starting from a randomly defined conformer. We checked, therefore, that this sampling process was independent of the starting conformation, while large enough, as most of the minima obtained differed significantly from the X-ray conformation, which was nevertheless obtained in some cases.

This procedure was repeated until 100 different energy-minimized conformations were produced for each of the three ligands considered here. The dielectric constant used in all these calculation was distance-dependent (ε=1×r) to simulate roughly the electrostatic shielding due to the solvent.

Fingerprint calculations

We used the procedure described previously [6] consisting of a deflation step for the ligand conformation and an inflation one for the protein cavities. Concerning a ligand molecular surface, we first defined a uniform triangular mesh on an ellipsoid embracing the molecule, such a mesh being next "deflated" step by step to obtain an approximation of the molecular surface. The "deflated" ellipsoid can then be used to obtain a description of the molecular surface based on spherical-harmonic expansions. In the case of a protein cavity, the inverted mapping procedure starting from a mesh on a sphere placed at the center of the cavity is achieved by inflating this mesh step by step with a similar technique, the "inflated" sphere being next described by a spherical-harmonics expansion. To obtain the protein cavities fingerprints, all water molecules were removed from the PDB files and the center of the inflating sphere was obtained from the center of mass of the ligand as positioned within the cavity in the X-ray data.

Using such descriptions, we were able to describe any point on both ligand and cavity molecular surfaces in terms of spherical coordinates (r,θ,φ). Mapping the variations of r in terms of è and φ variations provides us the fingerprint maps expressed as contour lines of r in the (θ,φ) space.

Consider a fingerprint denoted by A. Mathematically, the fingerprints are contour lines in a spherical coordinate system (r,θ,φ), and can be represented as a matrix with 180 columns and 360 rows. Each element a_{ij}∈A, where 0≤i<360 and 0≤j<180, is either the value of the radius on the contour line at θ$_i$ and φ_j, where $\theta_i = \frac{2\pi i}{360}$ and $\frac{\pi j}{180}$, or a_{ij}=0. A more suitable way to represent the fingerprint images in our application is to store them as extended vectors in $\Re^{n+m}$ space, where m is the number of rows and n is the number of columns.

Fingerprint matching

Given the fingerprint of a cavity target in a protein and the fingerprints of hundreds of ligand conformations, the objective is to retrieve a small number of ligand fingerprints that could be matched optimally to the cavity fingerprint. This very problem is found in data mining [12] and image retrieval problems [13], where the optimality is defined in terms of a scoring function. There are more then 20 scoring functions that could be used to rank the matches. In our case, there are two special particularities that should be taken in account:

1. fingerprint images are contour lines, and we want to use just the few of them;
2. the images are sparse in the sense that there are large parts of them with completely null pixels

According to Tan et al. [14] the best scoring function to this kind of problem is the cosine function because of its null invariance property. This property of the cosine function is useful for comparison purposes involving sparse images, where the co-presence of non-null pixels is more important than the co-absence.

The cosine scoring function is defined as follows. Given two non-null vectors $A\in\Re^n$ and $B\in\Re^n$, the cosine function is defined by

$$Cos(A,B) = \frac{\langle A,B\rangle}{\|A\|\|B\|},$$

where $\langle\bullet,\bullet\rangle$ denotes the usual inner product in $\Re^n$, i.e., $\langle A,B\rangle = \sum_{j=1}^{n} a_j b_j$, and $\|\bullet\| = \sqrt{\langle\bullet,\bullet\rangle}$ its associated norm.

The algorithm to match the fingerprints uses the cosine scoring function to rank all the possible pairs. We used only ten contour lines in the fingerprint images, as the cosine scoring function still produced the best results for this choice. Only the first k best scored pairs, where k is a predefined number depending on the application, are selected as possible matches. The complexity of this algorithm is the same as that for matrix multiplication and can be implemented in an efficient way because of the sparsity of the fingerprints [15].

Acknowledgements This work was supported by the PRA B02-04 of the French–Chinese Association for Scientific and Technical Research (AFCRST), FAPESP 01/08895-0, FINEP 1945/01, and CNPq 401695/2003-4. BM was recipient of a CAPES visiting researcher fellowship while this work was performed. We thanks the CINES computer center in Montpellier for providing access to the GOLD program on their O3800 SGI machine, and OPENEYE for providing free access to all their software according to an academic license.

References

1. Murray CW (2004) Principles and practice of high throughput screening. Blackwell Science

2. Stahura FL, Bajorath J (2004) Comb Chem High Throughput Screen 7:259–269
3. Lengauer T, Lemmen C, Rarey M, Zimmermann M (2004) Drug Discov Today 9:27–34
4. Fradera X, Mestres J (2004) Curr Top Med Chem 4:687–700
5. Singh J, Chuaqui CE, Boriack-Sjodin PA, Lee WC, Pontz T, Corbley MJ, Cheung HK, Arduini RM, Mead JN, Newman MN, Papadatos JL, Bowes S, Josiah S, Ling LE (2003) Bioorg Med Chem Lett 13:4355–4359
6. Cai W, Shao X, Maigret B (2002) J Mol Graph Model 20:313–328
7. McConkey BJ, Sobolev V, Edelman M (2002) Curr Sci 7:845–856
8. Kontoyianni M, McClellan LM, Sokol GS (2004) J Med Chem 47:558–565
9. Schulz-Gasch T, Stahl M (2003) J Mol Model 9:47–57
10. Verdonk ML, Cole JC, Hartshorn MJ, Murray CW, Taylor RD (2003) Proteins 52:609–623
11. Puvanendrampillai D, Mitchell JBO (2003) Bioinformatics 19:1856–1857
12. Demiriz A (2004) Data Mining and Knowledge Discovery 9:147–170
13. Kubo M, Aghbari Z, Oh KS, Makinouchi A (2003) IEICE Trans Inf Syst 8:1406T–1415
14. Tan PN, Kumar V, Srivastava J (2004) Inf Syst 29:293–313
15. Golub GH, Loan CFV (1996) Matrix computation, 3rd edn. John Hopkins University Press