BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

A global optimization algorithm for protein surface alignment

Paola Bertolazzi[1], Concettina Guerra[2,3*], Giampaolo Liuzzi[1]

Abstract

Background: A relevant problem in drug design is the comparison and recognition of protein binding sites. Binding sites recognition is generally based on geometry often combined with physico-chemical properties of the site since the conformation, size and chemical composition of the protein surface are all relevant for the interaction with a specific ligand. Several matching strategies have been designed for the recognition of protein-ligand binding sites and of protein-protein interfaces but the problem cannot be considered solved.

Results: In this paper we propose a new method for local structural alignment of protein surfaces based on continuous global optimization techniques. Given the three-dimensional structures of two proteins, the method finds the isometric transformation (rotation plus translation) that best superimposes active regions of two structures. We draw our inspiration from the well-known Iterative Closest Point (ICP) method for three-dimensional (3D) shapes registration. Our main contribution is in the adoption of a controlled random search as a more efficient global optimization approach along with a new dissimilarity measure. The reported computational experience and comparison show viability of the proposed approach.

Conclusions: Our method performs well to detect similarity in binding sites when this in fact exists. In the future we plan to do a more comprehensive evaluation of the method by considering large datasets of non-redundant proteins and applying a clustering technique to the results of all comparisons to classify binding sites.

Background

The function of a protein typically depends on the structure of specific binding sites located at the surface of the protein where the interaction with a ligand takes place. The identification of protein binding sites, their classification and analysis is of much interest for drug design and treatment of diseases. Binding sites recognition is generally based on geometry often combined with physico-chemical properties of the site since the conformation, size and chemical composition of the protein surface are all relevant for the interaction with a specific ligand.

In this paper we address the problem of optimally aligning protein surfaces, i.e. of finding atom pairs on two protein surfaces that occupy spatially equivalent positions. Our computational method integrates geometry with chemical properties of the matched atoms. It can be applied to the comparison of binding sites as well as of any other surface patches, such as cavities, that may be of interest.

Although the literature in protein surface alignment is not as vast as the one on complete structure or fold alignment, nevertheless several matching strategies have been designed for the recognition of protein-ligand binding sites and of protein-protein interfaces. They include hashing techniques [1,2], graph theoretic methods [3-6], descriptors based on moments [7] and moment invariants [8], shape descriptors such as spin images [9-11]. A few web servers have recently become available [12-16].

Most of the proposed methods require the solution to a 3D matching problem which is a well-studied problem also in computer vision and robotics. It can be formulated as follows: given two sets A and B of points, find two possibly large subsets A' of A and B' of B with high degree of *similarity*. There are various ways of defining the similarity between two point sets in 3D space leading to the proposal of different distance functions and

* Correspondence: guerra@cc.gatech.edu
[2]Dipartimento di Ingegneria Informatica, Universitá di Padova, Via Gradenigo, 6a, 35100 Padova, Italy
Full list of author information is available at the end of the article

associated algorithms; they include the root mean square distance, the closest point distance [17], the Hausdorff distance [18], the bottleneck distance [19].

An important aspect of the matching is the choice of a suitable surface representation; in the literature common ways of representing a surface are Connolly's representation [20], alpha-shapes [21] and pseudo-vertices [2]. In our approach, we represent the surface as a cloud of points, each corresponding to a surface atom. Thus, the protein surface alignment problem is the same as the aforementioned 3D matching problem.

One possible way to solve the surface alignment problem is by using the well-known Iterative Closest Point (ICP) algorithm [17] from which we draw our inspiration. The ICP algorithm, originally introduced in the area of computer vision for image registration, has been used in bioinformatics [22] for the alignment of complete protein structures. Indeed, we take a similar approach for surface alignment, namely we search for the isometric transformation which best superimposes two given protein structures.

Our main contribution is in the adoption of a different, more efficient global optimization approach along with a new dissimilarity measure. The global optimization algorithm we design belongs to the class of controlled random search methods [23-25]. These methods, although heuristic in nature, are very efficient and reliable for the global minimization of nonlinear multivariate functions of several variables. In the past years, controlled random search algorithms have been successfully used to solve many real world problems, see for instance [26-31]. The dissimilarity measure we propose is based on the solution to an "*Asymmetric Assignment Problem*" on a bipartite graph associated to the matching problem. Our method is capable of generating very accurate local alignments. We benchmark it on various sets of protein structures from the PDB [32], and compare its performance with that MolLoc [12].

Notations and Assumptions

In this section we introduce some notations and assumptions that will be used throughout the paper. Given two protein structures $\mathcal{P}$ and $\mathcal{Q}$, let us denote by P and Q the two finite sets of points corresponding to the atoms of the active sites of the two structures $\mathcal{P}$ and $\mathcal{Q}$, respectively. We let

$$n = |P| \quad \text{and} \quad m = |Q|$$

and assume, without loss of generality, that $n \le m$. The set P is conventionally representative of a query shape while Q defines a reference model shape.

An isometric transformation in three-dimensional space can be defined by a unit quaternion $a_r = (a_0, a_1, a_2, a_3)^\top \in \Re^4$ ($||a_r|| = 1$) and by a translation vector $a_t \in \Re^3$. Let $a^\top = (a_r^\top a_t^\top)$ be the transformation defining vector and denote by T_a the corresponding transformation, so that

$$\gamma = T_\alpha(x) = R(a_r)x + a_t$$

for every $x \in \Re^3$, where R (a_r) is the rotation matrix defined by the unit quaternion a_r as follows:

$$R(a_r) = \begin{pmatrix} R_{11} & R_{12} & R_{13} \\ R_{12} & R_{22} & R_{23} \\ R_{13} & R_{23} & R_{33} \end{pmatrix}$$

where

$$\begin{aligned} R_{11} &= a_0^2 + a_1^2 - a_2^2 - a_3^2, \\ R_{12} &= 2(a_1a_2 - a_0a_3), \\ R_{13} &= 2(a_1a_3 + a_0a_2), \\ R_{22} &= a_0^2 + a_2^2 - a_1^2 - a_3^2, \\ R_{23} &= 2(a_2a_3 - a_0a_1), \\ R_{33} &= a_0^2 + a_3^2 - a_1^2 - a_2^2 \end{aligned}$$

Let $\Theta \subset \Re^7$ be the set of all vectors $a \in \Re^7$ defining an isometric transformation in $\Re^3$. Given a transformation vector $a \in \Theta$, let $T_a(P) = P_a$ denote the set of points obtained by applying the transformation T_a to every point of P, that is

$$T_a(P) = P_a = \{\gamma : \gamma = R(a_r)p + a_t, \forall\, p \in P\}.$$

Let: ψ: $P \to Q$ denote a point to point mapping that associates to every point of P a point of Q. Since, as assumed above, P and Q are finite sets, the class Ψ of all mappings ψ has finite cardinality given by $|\Psi| = m^n$.

Let $\psi \in \Psi$ be a given mapping and a be a vector defining an isometric transformation, then the mean square error function between P and Q is the following

$$f(\psi, a) = \frac{1}{n} \sum_{p \in P} ||\psi(p) - R(a_r)p - a_t\ ||^2. \tag{1}$$

The surface alignment problem consists in finding a mapping $\psi^* \in \Psi$ of points in P to points in Q and an isometric transformation a^* such that

$$f(\psi^*, a^*) \le f(\psi, a),$$

for all $\psi \in \Psi$ and $a \in \Theta$.

The problem can also be formulated in terms of the following definition of an assignment. A function $\phi \in \Psi$ is an *assignment* from P to Q if, by definition, it is injective, that is for every $p_1, p_2 \in P$, $p_1 \ne p_2$ implies $\phi(p_1) \ne \phi(p_2)$.

Let us denote by $\Phi \subseteq \Psi$ the class of all possible assignments from P to Q. Obviously, since P and Q are finite sets, Φ is finite as well and its cardinality is $|\Phi| = m\,(m-1)\ldots(m-n+1)$.

Results

Algorithm

A well-known algorithm for shape alignment is the Iterative Closest Point Algorithm [17]. This algorithm stems from the idea that, once a mapping $\bar{\psi} \in \Psi$ is fixed, it is possible to compute the isometric transformation $a \in \Theta$ that minimizes the function $f(\bar{\psi}, a)$ (a closed-form expression for $a(\bar{\psi})$ has been given in [33] where we refer the interested reader for the relevant details). Let $a(\bar{\psi})$ be the minimizer of $f(\bar{\psi}, a)$, that is

$$a(\bar{\psi}) = \arg\min_{a\in\Theta} f(\bar{\psi}, a).$$

Hence, the problem implicitly considered by the ICP Algorithm is the following two-level optimization problem

$$\begin{aligned} &\min_{\psi} f(\psi, a) \\ &s.t. \psi \in \Psi \\ &\quad a = \arg\min_{a\in\Theta} f(\psi, a). \end{aligned} \tag{2}$$

As it is stated in [17], where ICP has been originally proposed, the method converges to a solution which is a local minimum of the two-level Problem (2). Futher, in [17] it has been shown that the final transformation $\bar{a}$ and mapping $\bar{\psi}$ obtained by Algorithm ICP heavily depend on the initial relative positioning of sets P and Q.

In this section we discuss the use of a continuous global optimization algorithm for the solution of the shape alignment problem. To this aim, it is necessary to reformulate the shape alignment problem in a complementary way with respect to Problem (2). More in particular, the inner-level problem becomes the one defining the mapping function ψ (instead of the transformation a) once the transformation vector $a \in \Theta$ is fixed in the outer level.

Namely, we consider the following two-level optimization problem

$$\begin{aligned} &\min_{a} f(\psi, a) \\ &s.t. a \in \Theta \\ &\quad \psi = \arg\min_{\psi\in\Psi} f(\psi, a). \end{aligned} \tag{3}$$

Problem (3) can be reduced to a one-level optimization problem by considering that for every vector $a \in \Theta$, the inner-level problem of (3) admits a globally optimal solution, which we denote by

$$\psi(a) = \arg\min_{\psi\in\Psi} f(\psi, a) \tag{4}$$

and represents the closest point mapping. Hence, Problem (3) can be equivalently stated as

$$\min_{a\in\Theta} g(a) \tag{5}$$

where $g(a) = f(\psi(a), a)$. Every global solution a^* of (5) is, by definition, a solution such that $f(\psi(a^*), a^*) \le f(\psi(a), a)$, for all $a \in \Theta$.

Observe that the computation of function g requires the computation of the optimal mapping $\psi(a)$, that is, the solution to Problem (4). This latter problem can be solved with a time complexity $O(nm)$ in the worst case [17] which can be relevant for n and m large. Moreover, due to its definition, $g(a)$ is a non-smooth (Lipschitz) continuous function and its derivatives are not available. Indeed, for the minimization of function $g(a)$ we can neither directly use its derivatives nor approximate them through finite differences since this would require too much time and produce numerical derivatives which are unreliable because of the non-smoothness of function g.

On the basis of the above observation we propose the use of a controlled random search method for the solution to Problem (5). In the following we briefly recall the global optimization algorithm that we use and which was originally proposed in [25] and successively improved in [24]. It is a population based algorithm in the sense that, through-out the entire optimization process, a population of points is maintained and iteratively updated in such a way that they cluster around the global minima of the objective function. Roughly speaking, the method is composed of two distinct and consecutive phases: a global phase and a local phase. During the global phase an initial population of points (defining roto-translations in three-dimensional space) is generated by randomly sampling a sufficiently large set of points over some feasible domain. Then, at every iteration of the local phase, a new point is generated and the population is updated if this new point improves on the worst point of the population. More in details, the algorithm can be described by the following steps.

1. **initialization** Let $N = 6$ and choose an integer $M \gg N$. The objective function is sampled on a set S of M points $a \in \Re^7$ randomly chosen within the feasible domain Θ strictly containing the global minimizer.
2. **stopping criterion** If the maximum and minimum values of the objective function over S are sufficiently close to each other, namely

$$g_{\max} - g_{\min} < \epsilon,$$

where

$$g_{\max} = \max_{a \in S} g(a), \quad g_{\min} = \min_{a \in S} g(a)$$

then STOP.

3. **search phase** $N+1$ points are randomly chosen in the set S. Then,

(a) the *weighted* centroid a_c of the $N + 1$ points is computed;

(b) the new trial point $\tilde{a}$ is computed by doing a *weighted reflection* of the centroid onto the worst point among the selected $N + 1$ points.

Namely, let $a^{\dagger}$ be the worst point, then

$$\breve{a} = (1+\alpha)a_c - \alpha a^{\dagger}, \quad \text{and} \quad \tilde{a} = W\breve{a}$$

where $\alpha \in (0,1)$ is a reflection parameter and

$$W = \begin{pmatrix} 1/\xi & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & 1/\xi & 0 & \dots & \dots & \dots & 0 \\ 0 & 0 & 1/\xi & 0 & \dots & \dots & 0 \\ 0 & \dots & 0 & 1/\xi & 0 & 0 & 0 \\ 0 & \dots & \dots & 0 & 1 & 0 & 0 \\ 0 & \dots & \dots & \dots & 0 & 1 & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 & 1 \end{pmatrix},$$

$$\xi = ||(\breve{a}_0\ \breve{a}_1\ \breve{a}_2\ \breve{a}_3)^{\top}||$$

The normalization matrix W is introduced to ensure that the first four components of the resulting point represent a unit quaternion (i. e. a rotation).

The parameter α is iteratively updated during the optimization process [24] in such a way that its value tends to zero as the iteration count increases and the difference $g_{\max}$ - $g_{\min}$ decreases.

4. **updating phase** If the objective function value on the new point $\tilde{a}$ improves on the maximum function value over S, then the set S is updated by adding the new point and discarding the worst one. Otherwise the set S is left unchanged and the new point is discarded. The algorithm continues iterating through steps 2-4.

The algorithm starts by randomly choosing $M \gg N$ points over the feasible set Θ. In the literature, a typically accepted value of M is $25N$ [25,29]. This value is able to convey to the algorithm sufficient ability to find the global minimum point without excessively slowing down the convergence.

A new dissimilarity measure

In this section we propose a new dissimilarity measure between two given sets of points of two proteins. This measure is based on a distance other than the closest point distance.

In particular, it can be noted that, using the closest point distance, it can happen that different points of set $T_a(P)$ are mapped to the same point of set Q. This, in turn can yield a distance value which is small just because many points are all mapped to the same closest point.

In order to avoid this undesirable effect, we consider the function $f(\phi, a)$ defined in (1) where $\phi \in \Phi$ is an assignment function and let, for every a, $\phi(a)$ be a global solution to problem

$$\min_{\varphi \in \Phi} f(\varphi, a). \tag{6}$$

Problem (6) can be formulated as a 0, 1-optimization problem and is, indeed, the combinatorial optimization problem known as the *Asymmetric Assignment Problem* (AAP).

In particular, let $G(P, Q, E)$, $E \subset P \times Q$, be the bipartite directed graph characterized by the two sets of nodes P and Q and by the edges between all pairs of nodes, one of P and the other of Q. Then, for every pair $e = (p, q) \in E$, define

$$c_e = \left\| q - T_a(p) \right\|^2.$$

Let $s \in \{0, 1\}^{|E|}$ be the edge incidence vector and consider the following minimum cost assignment problem

$$\begin{aligned} &\min_s c^{\top} s \\ &\sum_{e \in \delta^+(p)} s_e = 1, \quad \forall p \in P \\ &\sum_{e \in \delta^-(q)} s_e \le 1, \quad \forall q \in Q \\ &s \in \{0,1\}^{|E|}, \end{aligned} \tag{7}$$

where, $\delta^+(p)$ and $\delta^-(q)$ are the sets of edges leaving node p and, respectively, entering in node q.

Note that the constraints of Problem (7) require each node $p \in P$ to be assigned to exactly one node $q \in Q$ and each node $q \in Q$ to be assigned to at most one node $p \in P$, which is why Problem (7) is known as *Asymmetric Assignment Problem*.

Clearly, it is

$$f(\varphi(a), a) = c^{\top} s^{*}$$

where s^* is the optimal solution to Problem (7).

Problem (7), and hence (6), can be solved very efficiently by *ad-hoc* codes that have time complexity $O(\sqrt{nm}\log(n\mathcal{C}(T_a)))$ where $\mathcal{C}(T_a) = \max_{p\in P, q\in Q}\{||q - T_a(p)||\}$, see for instance [34].

We are now able to define our new dissimilarity measure, that we call *matching distance*.

Definition 1 *Given an isometric transformation* $a \in \Theta$ *and two distinct sets of points P and Q, the matching distance between* $T_a(P)$ *and Q is given by* $f(\phi(a), a)$.

Reasoning as in the preceding section, we can now search for a global solution to problem (5) where now $g(a) = f(\phi(a), a)$.

Integration of physico-chemical properties

Up to this point, the discussed approach is based on geometry only. However, as is well known in biology, there are other properties that affect the binding of molecules. For instance, electrostatic as well as hydrophobic-hydrophilic properties play an important role in protein-protein and protein-ligand interactions. Thus, we consider a variant of our approach in which we integrate physico-chemical properties. Specifically in the graph $G(P, Q, E)$ we assume that the edge $e = (p, q)$ is present only if the two atoms $p \in P$ and $q \in Q$ have the same physico-chemical properties. According to [35], we say that p and q have the same physico-chemical properties if they are both Acceptor (ACC), Donor (DO), Acceptor/Donor (AD), Aliphatic (ALI) or Aromatic (PI). Furthermore, we assume that, for every $p \in P$ at least a node $q \in Q$ exists such that $(p, q) \in E$.

Testing

We applied our method, referred to as Continuous Optimization (CO) method in the following, to the comparison of binding sites of proteins. We integrated physico-chemical properties in our method, as discussed in the previous section. The structures of the proteins in complex with specific ligands are taken from the PDB [32]. The binding sites are extracted by a simple algorithm that finds all protein atoms within a certain distance (4.0 Å) from an atom of the ligand. We run our algorithm on pairs of binding sites producing in output the list of matched atoms on the two binding sites, the rigid transformation that best superimposes them, and the RMSD after superposition.

We benchmarked CO on a dataset of 100 proteins in complex with 9 ligands that differ in chemical composition as well as in size and shape. The results of all-to-all pairwise comparisons are visualized by means of a distance matrix and by the ROC curves. The goal is to evaluate the ability of CO in assigning a binding site to the correct group of proteins, i.e. those binding the same ligand.

We then present more detailed results on a set of 19 binding sites of proteins in complex with the ligand ATP with the goal of judging the quality of the alignments. For each comparison we report the number of aligned atoms as well as their RMSD after superposition. The results on this dataset are compared with those of another method, MolLOC [12], which derives the same two measures, i.e. number of aligned atoms and RMSD.

Classification of proteins according to their bound ligand

In the first experiment we perform all-to-all comparisons on a dataset of 100 proteins in complex with one of 9 ligands: AMP, ATP, FAD, FMN, GLC, HEME, NAD, PO4, and Steroid. This dataset was used in [36] for an analysis of shape variation in protein binding sites. The proteins were carefully selected, with a number of criteria, so that the dataset is non-redundant and the binding sites are not evolutionary related.

The results of all-to-all comparisons are illustrated by means of the distance matrix of Figure 1. An entry of the matrix corresponds to a protein pair and contains a value related to the number of aligned atoms of the binding sites of the pair. Namely, in the matrix we report

$$2\left(\frac{\text{num. aligned atoms}}{n+m}\right)$$

where n and m are the numbers of atoms of the two binding sites. The proteins are listed along the rows and columns of the matrix so that proteins binding the same ligand are grouped together. Horizontal and vertical black lines on the matrix separate different groups of proteins. The matrix is color-coded from 0 to 1, with red corresponding to high number of aligned atoms and therefore high similarity in the shape of the binding sites and blue to the lowest degree of similarity. A good classification of sites based on bound ligands implies the presence of mostly red areas around the main diagonal, corresponding to pairwise comparisons within the same group of proteins, i.e. in complex with one specific ligand. This can be in fact observed in the image matrix although with different degrees for the different groups of proteins. As it is known [36], ligand PO4 tends to be rigid, exhibiting little conformational variability in the binding. Not surprisingly, the corresponding area is the one showing the highest degree of similarity. The method CO appears to perform well also in distinguishing the PO4 group from any other group, as PO4

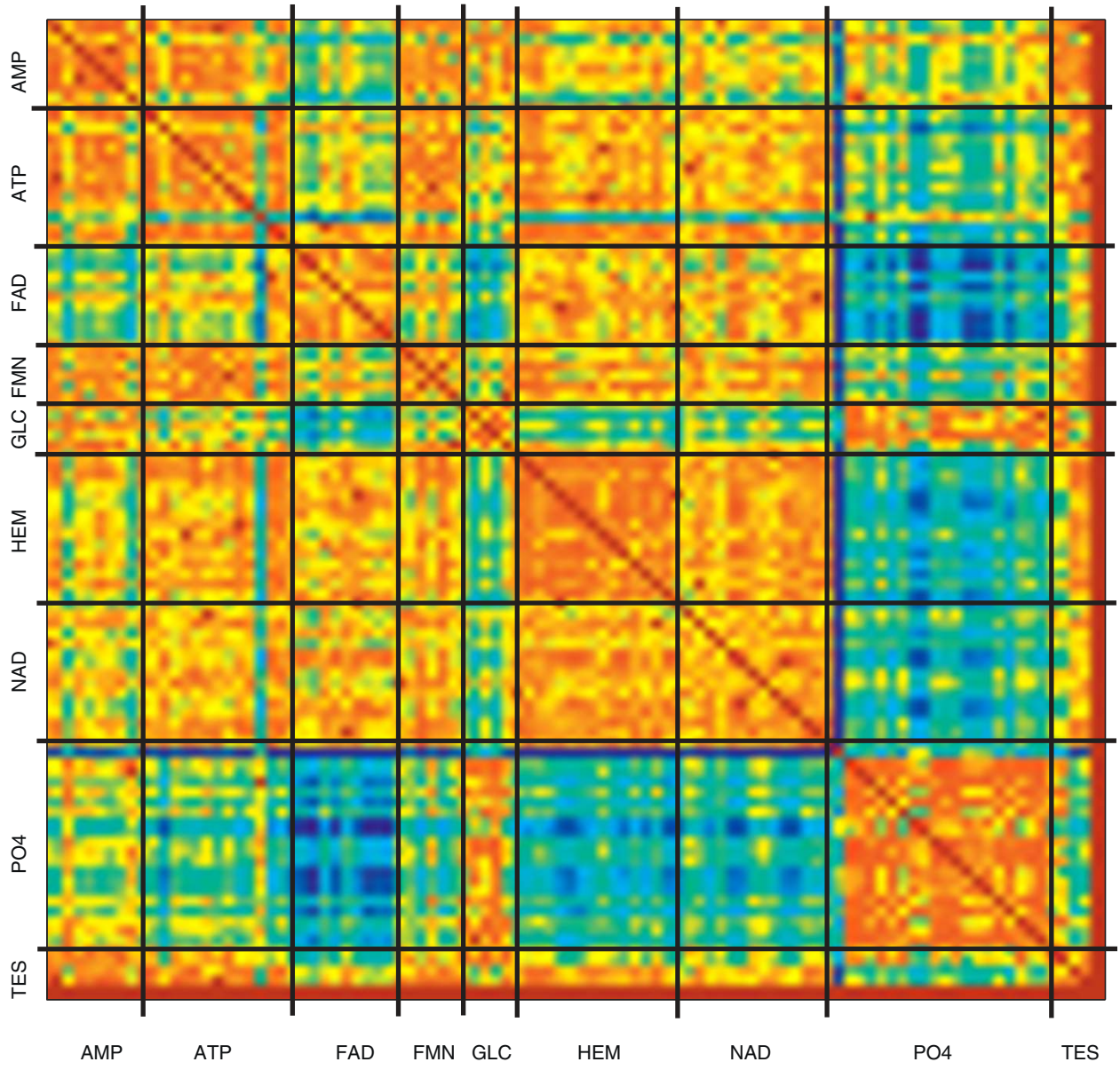

Figure 1 Distance matrix. The matrix shows the results of all-to-all comparisons. The 9 ligands are indicated along the rows and columns and the proteins binding each ligand are grouped together. The grid of horizontal and vertical black lines separates different groups of proteins. The matrix is color-coded from 0 (blue) to 1 (red), with red corresponding to high number of aligned atoms and therefore high similarity in the shape of the binding sites and blue to the lowest degree of similarity.

binding sites are more similar to themselves than to binding sites of other groups. Similar considerations apply to steroid and GLC. A good performance is also obtained for the HEME group, although the discriminating power with the NAD group is not clear. As noted in [37], ligand ATP has great variation in its conformation when binding different proteins: it can be in an extended conformation or in a compact one, resulting in different sizes and shapes of the binding regions. This is reflected in our experiments, as can be seen from the distance matrix where blue or green areas are present.

An important aspect of an alignment method is its ability to retrieve, for a given query binding site, those proteins of the dataset binding the same ligand. To evaluate CO in this task we resort to ROC curves. The results of the comparisons of the query with all other proteins are ranked from the best to the worst in terms of the number of aligned atoms. A pairwise comparison in the ranked list is considered correct or true positive if

the other protein of the pair binds the same ligand as the query. The results are summarized by the receiver operating characteristic (ROC) curves in Figure 2 that display the fraction of true positives or correct answers vs. the fraction of false positives for all positions of the ranked solutions. The best possible prediction results would yield a curve through the point in the upper left corner or coordinate (0, 1) of the ROC space. A completely random guess would give a point along a diagonal line from the left bottom to the top right corners. We repeated this experiment with each protein of the dataset as query. Each curve in Figure 2 shows the average values obtained on all query proteins of a group. As expected, the curve corresponding to PO4 (in green) deviates the most from the diagonal line being the closest to the top-left corner of the ROC square. Thus CO has a good success in predicting a PO4 binding site. By contrast, the worst performance is achieved for NAD binding proteins with the associated curve in dotted black.

From the above sets of experiments we can conclude that CO has a good accuracy in the retrieval of similarity information: for a given query binding site the highest scoring solutions are generally the binding sites of the dataset in complex with the same ligand as the query. Furthermore, when a good similarity in the

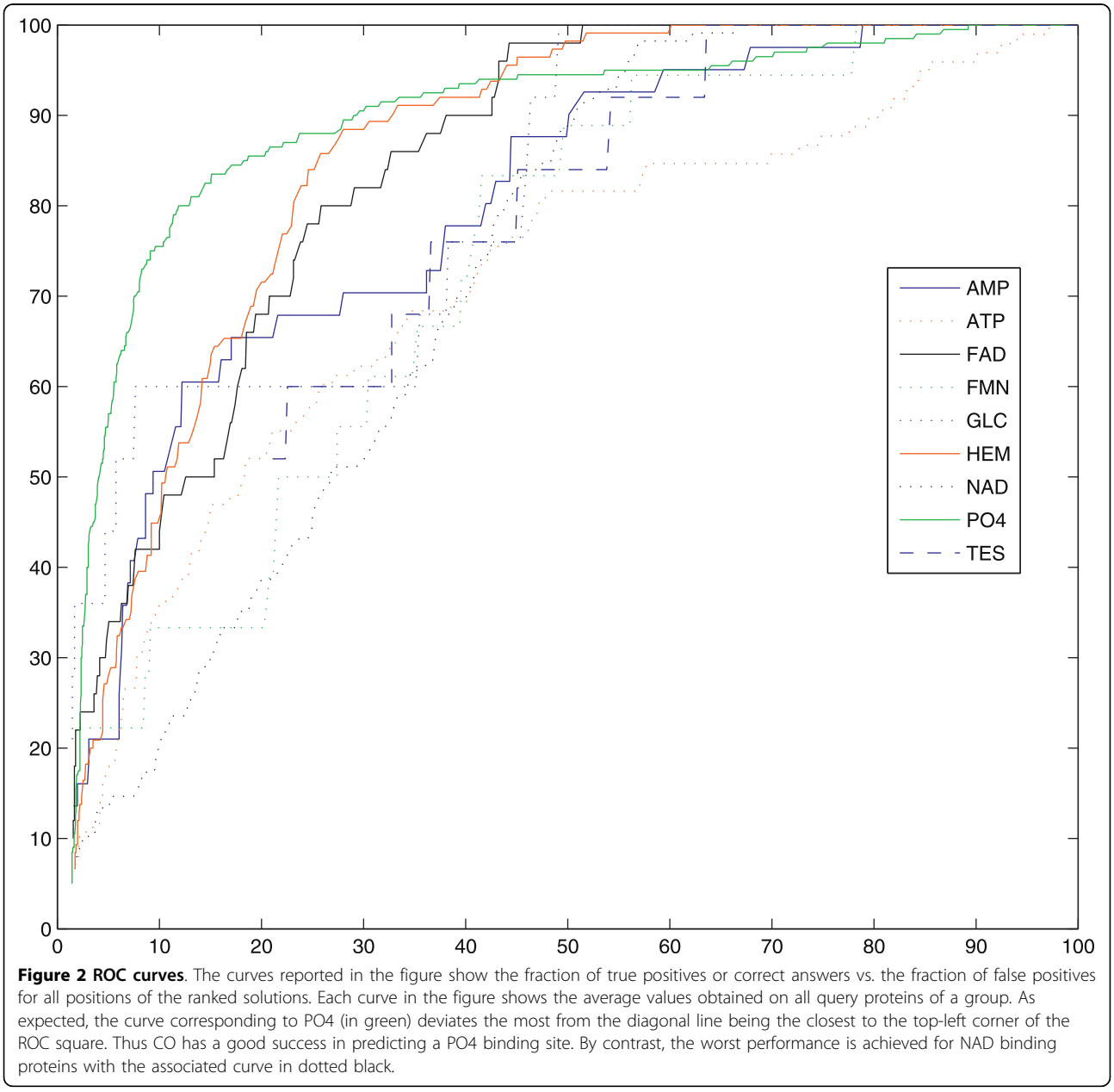

Figure 2 ROC curves. The curves reported in the figure show the fraction of true positives or correct answers vs. the fraction of false positives for all positions of the ranked solutions. Each curve in the figure shows the average values obtained on all query proteins of a group. As expected, the curve corresponding to PO4 (in green) deviates the most from the diagonal line being the closest to the top-left corner of the ROC square. Thus CO has a good success in predicting a PO4 binding site. By contrast, the worst performance is achieved for NAD binding proteins with the associated curve in dotted black.

binding is expected because of the relative rigidity of the ligands, CO is able to capture such a similarity, as shown in the distance matrix.

Comparing CO with other alignment methods on ATP binding sites

Several studies have been conducted to evaluate and compare different methods for determining the structural similarity of proteins. For instance, a comprehensive assessment of structural alignment methods is presented in [38] where six publicly available programs are evaluated on almost 9 million pairs of proteins. However, a similar large-scale experiment is not available for the related problems of aligning protein surfaces and binding sites, despite the growing number of methods and web servers available. There are several factors that contribute to the difficulty of the comparison. First, different methods solve different instances of the matching problem: some methods compare binding sites, while others recognize binding sites in cavities or even entire surfaces. Second, the methods differ in the input representations and scoring functions. For instance, in CO the input points are the atom centers, in Multibind a reduced set of points, the pseudo-centers. In [36] the points are the spherical sample points derived from the atomic coordinates. MolLoc, on the other hand, uses Connolly's [39] points and a richer surface representation based on local shape descriptors of surface points. As for the scoring function, although most methods produce the RMSD of the superimposed structures, some methods have a different native scoring function that cannot be easily derived by other methods.

As a comprehensive evaluation of all the techniques is beyond the scope of this paper, only MOlLoc will be considered in comparison with CO. The reason for choosing MolLoc is that both methods judge the quality of the alignments by the number of aligned atoms and their RMSD after superposition. Such measures are available at MolLOC website. As Multibind does not report the RMSD of two aligned structures at its website it will not be considered here. Moreover, the method in [36], based on spherical harmonics and benchmarked on the same dataset of 100 proteins, is not used in our evaluation because it computes a measure of similarity of two shapes without an alignment.

As observed in [38], although the ROC curves are a valid tool for assessing the quality of a classification approach they are often of limited value in comparing different methods; in fact such curves take into account only the ranking of the alignments not their quality. For this reason, since we want to assess the quality of the alignments we choose the geometric measure SAS [38,40,41]. Clearly, a better match has a higher number of aligned atoms and smaller RMSD. Since the two measures are not independent SAS combines them into a

Table 1 Comparison of CO with MolLoc

		CO			MolLoc		
Rank	**Protein Pair**	**N. corresp atoms**	**RMSD**	**SAS**	**N. corresp atoms**	**RMSD**	**SAS**
1	atpE-1hck	62	1.2	1.94	45	1.3	2.89
2	1atpE-1phk	57	0.91	1.6	63	0.9	1.43
3	1atpE-1csn	50	1.18	2.36	55	0.9	1.64
4	1atpE-1nsf	34	2.11	6.21	11	1.4	12.73
5	1atpE-1j7k	25	1.81	7.24	25	1.6	6.4
6	1atpE-1e8xA	24	1.74	7.25	20	1.7	8.5
7	1atpE-1f9aC	21	2.17	10.33	18	1.6	8.89
8	1atpE-1kay	20	1.9	9.5	8	1.7	21.25
9	1atpE-1yag	20	1.92	9.6	17	1.6	9.41
10	1atpE-1a82	19	2.02	10.63	13	1.9	14.62
11	1atpE-1jjv	18	1.76	9.78	10	1.8	18
12	1atpE-1gn8A	17	2.37	13.94	14	1.6	11.43
13	1atpE-1b8aA	16	2.05	12.81	10	2	20
14	1atpE-1mjhA	16	2.28	14.25	14	1.9	13.57
15	1atpE-1e2q	15	1.39	9.27	5	1.8	36
16	1atpE-1kp2A	13	1.51	11.62	15	1.9	12.67
17	1atpE-1ayl	12	1.21	10.08	16	2	12.5
18	1atpE-1g5t	7	2.26	32.29	8	1.6	20
			avg.SAS	10.04		avg.SAS	12.88

Pairwise comparisons of the binding site of protein 1atp with other 18 proteins all binding ATP (columns 2). The results of CO (columns 3-5) and MolLoc (columns 6-8). For a de_nition of SAS see the text. The comparisons are ranked based on the number of corresponding atoms in CO (column 3).

single expression: SAS = (RMSD × 100)/(num. aligned atoms).

We run both programs on the set of 19 proteins used in [42] for a related although different problem, that is binding site recognition within a cavity. The proteins all bind ligand ATP and are from different families according to the structural classification SCOP [43].

We performed pairwise comparisons of the active site of the Catalytic Subunit of cAMP- dependent Protein-Kinase (pdb code 1atp, chain E) with each of the remaining proteins of the input data set. Of the set of proteins only three belong to the same SCOP family as 1atp, namely 1phk, 1csn and 1hck. In Table 1 for each comparison we report the number of aligned atoms along with the RMSD obtained by CO (columns 3-4) and MolLoc (colums 6-7). The entries of the table are listed and ranked according to the number of corresponding atoms obtained by CO (column 3). We observe (see Table 1) that both methods correctly rank at the top three positions the proteins in the same family as 1atp, that is 1phk, 1hck and 1csn. Furthermore for the same three proteins the RMSD is typically very low (approx. 1.5 Å). Lower scores are obtained for distantly related proteins, as for instance 1g5t. The table also reports the SAS measure for CO (column 5) and MolLoc (column 8) and their average at the bottom of the same two columns. As a lower SAS value indicates a better match, it follows that CO on average achieves a better quality than MolLoc with respect to this measure.

Figure 3 shows an example superimposition of the binding sites of ligand ATP of proteins 1atp and 1hck after the computed rototraslation is applied.

We conclude this section by reporting that the execution times of CO on average on all 18 pair-wise comparisons considered in this experiment was 14.06 s for a total of 253.1 s on an Intel Pentium IV processor running at 2.66Ghz with 1Gb main memory. As we mentioned before, the low computational complexity of our proposed approach is one of the key points of our design. We do not report the execution times of MolLoc since they are not available from the web server interface.

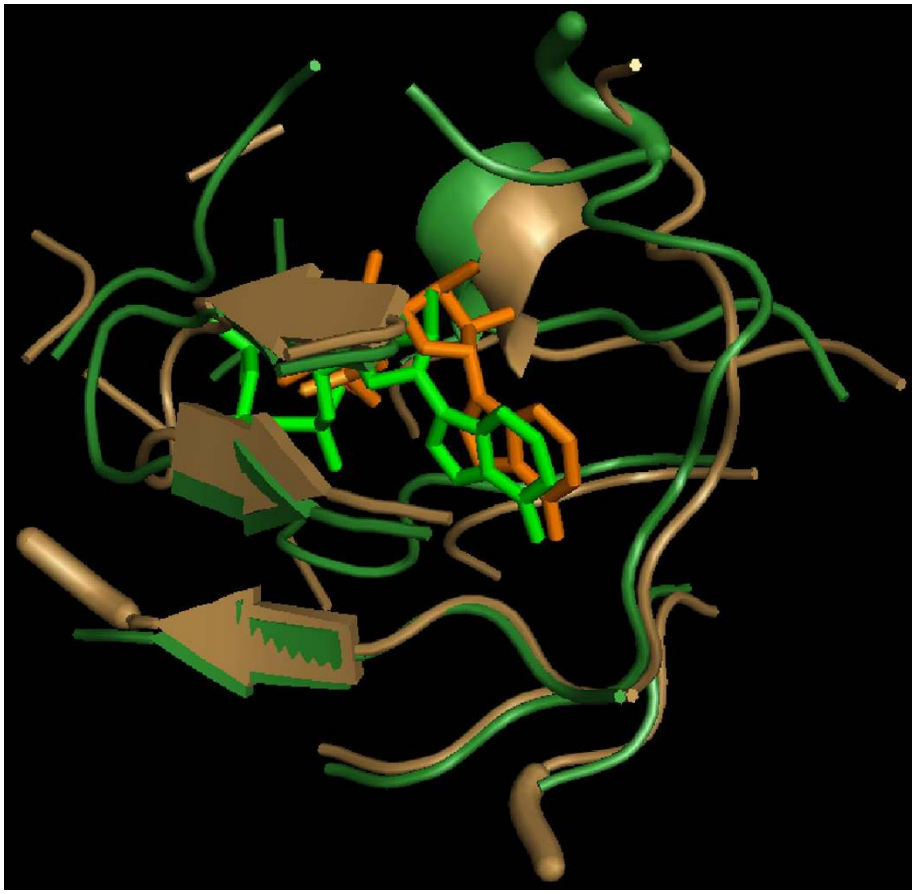

Figure 3 Example of a computed superimposition. Comparison of CO with MolLoc. Pairwise comparisons of the binding site of protein 1atp with other 18 proteins all binding ATP (columns 2). The results of CO (columns 3-5) and MolLoc (columns 6-8). For a definition of SAS see the text. The comparisons are ranked based on the number of corresponding atoms in CO (column 3).

Discussion and Conclusions

The main challenge for a method that compares and classifies binding sites is to be able to cluster the binding sites in groups according to the type of ligands they bind while at the same time allowing some conformational variability within the same group, as is often observed for binding sites of different proteins complexed with the same ligand. The difficulty arises because of the variety of ways in which a ligand can bind proteins. Although we expect a computational method to be able to distinguish among different types of ligands relatively well, there are obviously cases when only experimental methods can determine the binding affinity of two molecules.

Our proposed method, CO, performs well to detect similarity in binding sites when this in fact exists. In the future we plan to do a more comprehensive evaluation of the method by considering large datasets of non-redundant proteins and applying a clustering technique to the results of all comparisons to classify binding sites. A systematic evaluation of CO with other existing methods will be done through the introduction of a common scoring function that will overcome the problem that the available methods use native scoring functions difficult to export to other methods.

Acknowledgements

The authors are indebted to two anonymous Referees whose many helpful comments and suggestions greatly helped improving the paper. The work of C. Guerra was supported by the "Progetto di Ateneo, University of Padova" and by "Fondazione Cariparo".

Author details

[1]Istituto di Analisi dei Sistemi ed Informatica "A. Ruberti", Consiglio Nazionale delle Ricerche, Viale Manzoni, 30, 00185 Rome, Italy. [2]Dipartimento di Ingegneria Informatica, Universitá di Padova, Via Gradenigo, 6a, 35100 Padova, Italy. [3]College of Computing, Georgia Institute of Technology, Atlantic Drive, 801, 30332-0280 Atlanta (GA), USA.

Authors' contributions

All three authors participated in the design of the methodology and in set up of the experiments. GL carried out the implementation. All authors read and approved the final manuscript.

Received: 7 October 2009 Accepted: 29 September 2010
Published: 29 September 2010

References

1. Shatsky M, Shulman-Peleg A, Nussinov R, Wolfson HJ: **The multiple common point set problem and its application to molecule binding pattern detection.** *Journal of Computational Biology* 2006, **13**:407-428.
2. Shulman-Peleg A, Nussinov R, Wolfson HJ: **Recognition of functional sites in protein structures.** *Journal of Molecular Biology* 2004, **339**:607-633.
3. Artymiuk P, Spriggs R, Willett P: **Graph theoretic methods for the analysis of structural relationships in biological macromolecules.** *Journal of the American Society for Information Science and Technology* 2005, **56(5)**:518-528.
4. Chen B, Bryant D, Fofanov V, Kristensen D, Cruess A, Kimmel M, Lichtarge O, Kavraki L: **Cavity-aware motifs reduce false positives in protein function pre-diction.** *Computational System Bioinformatics Conf* 2005, 311-323.
5. Hofbauer C, Lohninger H, Aszodi A: **Surfcomp: A novel graph-based approach to molecular surface comparison.** *Journal of Chemical Information and Computer Sciences* 2004, **44(3)**:837-847.
6. Weskamp N, Kuhn D, Hullermeier E, Klebe G: **Efficient similarity search in protein structure databases by k-clique hashing.** *Bioinformatics* 2004, **20(10)**:1522-1526.
7. Ballester P, Richards W: **Ultrafast shape recognition to search compound databases for similar molecular shapes.** *Journal of Computational Chemistry* 2007, **28(10)**:1711-1723.
8. Sommer I, Müller O, Domingues F, Sander O, Weickert J, Lengauer T: **Moment invariants as shape recognition technique for comparing protein binding sites.** *Bioinformatics* 2007, **23**:3139-3146.
9. Bock M, Garutti C, Guerra C: **Discovery of similar regions on protein surfaces.** *Journal of Computational Biology* 2007, **14(3)**:285-299.
10. Bock M, Garutti C, Guerra C: **Effective labeling of molecular surface points for cavity detection and location of putative binding sites.** *Proc of the VI Int Conf on Computational Systems Bioinformatics; San Diego* 2007, 263-274.
11. Bock M, Garutti C, Guerra C: **Cavity detection and matching for binding site recognition.** *Theoretical Computer Science* 2008, **408(2-3)**:151-162.
12. Angaran S, Bock M, Garutti C, Guerra C: **MolLoc: a web tool for the local structural alignment of molecular surfaces.** *Nucleic Acids Research* 2009, **Web server.**
13. Ausiello G, Gherardini PF, Marcatili P, Tramontano A, Via A, Helmer-Citterich M: **Funclust: a web server for the identification of structural motifs in a set of non-homologous protein structures.** *BMC Bioinformatics* 2008, **9(Suppl 2)**:S2.
14. Jambon M, Olivier A, Combet C, Deleage G, Delfaud F, Geourjon C: **The SuMo server: 3D search for protein functional sites.** *Bioinformatics* 2005, **21(20)**:3929-3930.
15. Kinoshita N, Furui J, Nakamura H: **Identification of protein functions from a molecular surface database, ef-site.** *Journal of Structural and Functional Genomics* 2001, **2**:9-22.
16. Shulman-Peleg A, Shatsky M, Nussinov R, Wolfson HJ: **MultiBind and MAPPIS: webservers for multiple alignment of protein 3D-binding sites and their interactions.** *Nucleic Acids Research* 2008, **36 Web server.**
17. Besl PJ, McKay ND: **A method for registration of 3-D shapes.** *IEEE Trans. on Pattern Analysis and Mach. Intelligence* 1992, **14**:239-255.
18. Goodrich MT, Mitchell JSB, Orletsky MW: **Practical methods for approximate geometric pettern matching under rigid motions.** In *Proc of the 10th Ann Symp on Computational Geometry* Edited by: Press A 1994, 103-112.
19. Efrat A, Itai A, Katz MJ: **Geometry helps in bottle-neck matching and related problems.** *Algorithmica* 2001, **31**:1-28.
20. Connolly ML: **Analytical molecular surface calculation.** *Journal of Applied Crystallography* 1983, **16**:548-558.
21. Liang J, Edelsbrunner H, Woodward C: **Anatomy of protein pockets and caviteis: measurement of binding site geometry and implications for ligand design.** *Protein Science* 1998, **7**:1884-1897.
22. Xu D, Li H, Gu T: **Protein Structure Superposition by Curve Moment Invariants and Iterative Closest Point.** *The 1st Int Conf on Bioinformatics and Biomedical Engineering* 2007, **6**:25-28.
23. Brachetti P, De Felice Ciccoli M, Di Pillo G, Lucidi S: **A new version of the Price's algorithm for global optimization.** *Journal of Global Optimization* 1997, **10**:165-184.
24. Cirio L, Lucidi S, Parasiliti F, Villani M: **A global optimization approach for the synchronous motors design by finite element analysis.** *Journal of Applied Electromagnetics and Mechanics* 2002, **16**:13-27.
25. Price WL: **A controlled random search procedure for global optimization.** In *Towards Global Optimization 2.* Edited by: Dixon L, Szego G. Amsterdam: North-Holland; 1978.
26. Daidone A, Parasiliti F, Villani M, Lucidi S: **A new method for the design optimization of three-phase induction motors.** *IEEE Trans. on Magnetics* 1998, **34**:2932-2935.
27. Liuzzi G, Lucidi S, Piccialli V, Sotgiu A: **A magnetic resonance device designed via global optimization techniques.** *Mathematical Programming* 2004, **101**:339-364.
28. Liuzzi G, Lucidi S, Parasiliti F, Villani M: **Multi-objective Optimization Techniques for the Design of Induction Motors.** *IEEE Transactions on Magnetics* 2003, **39**:1261-1264.
29. Price WL: **Global optimization by controlled random search.** *Journal of Optimization Theory and Applications* 1983, **40**:333-348.
30. Price WL: **Global optimization algorithms for a CAD workstation.** *Journal of Optimization Theory and Applications* 1983, **55**:133-146.

31. Price WL, Woodhams F: **Optimising accelerator for CAD workstations.** *IEEE proceedings* 1988, **135**:214-221.
32. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank.** *Nucleic Acids Research* 2000, **28**:235-242.
33. Horn BKP: **Closed-form solution of absolute orientation using unit quaternions.** *Journal of the Optical Society of America* 1987, **4**:629-642.
34. Goldberg AV, Kennedy R: **An Efficient Cost Scaling Algorithm for the Assignment Problem.** *Mathematical Programming* 1995, **71**:153-178.
35. Schmitt S, Kuhn D, Klebe G: **A New Method to Detect Related Function Among Proteins Independent of Sequence and Fold Homology.** *Journal of Molecular Biology* 2002, **323**:387-406.
36. Kahraman A, Morris RJ, Laskowski RA, Thornton JM: **Shape Variation in Protein Binding Pockets and their Ligands.** *Journal of Molecular Biology* 2007, **368**:283-301.
37. Stockwell GR, Thornton JM: **Conformational diversity of ligands bound to proteins.** *Journal of Molecular Biology* 2006, **356**:928-944.
38. Kolodny R, Koehl P, Levitt M: **Comprehensive Evaluation of Protein Structure Alignment Methods: Scoring by Geometric Measures.** *Journal of Molecular Biology* 2005, **346**:1173-1188.
39. Connolly ML: **Analytical molecular surface calculation.** *Journal of Applied Crystallography* 1983, **16**:548-558.
40. Subbiah S, Laurents D, Levitt M: **Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core.** *Current Biology* 1993, **3(3)**:141-148.
41. Gerstein M, Levitt M: **Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins.** *Protein Science* 1998, **7(2)**:445-456.
42. Comin M, Guerra C, Dellaert F: **Binding Balls: Fast detection of Binding Sites using a property of Spherical Fourier Transform.** *Journal of Computational Biology* 2009, **16(11)**:1577-1591.
43. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *Journal of Molecular Biology* 1995, **247**:536-540.

doi:10.1186/1471-2105-11-488
Cite this article as: Bertolazzi *et al.*: A global optimization algorithm for protein surface alignment. *BMC Bioinformatics* 2010 **11**:488.