IL NUOVO CIMENTO Vol. 35 C, N. 5 Suppl. 1 Settembre-Ottobre 2012
DOI 10.1393/ncc/i2012-11335-1

Colloquia: PR PS BB 2011

Continuous global optimization for protein structure analysis

P. Bertolazzi([1]), C. Guerra([2])([3]), F. Lampariello([1]) and G. Liuzzi([1])

([1]) *Istituto di Analisi dei Sistemi ed Informatica (IASI) "A. Ruberti", CNR Viale Manzoni 30, 00815 Roma, Italy*
([2]) *Dipartimento di Ingegneria Informatica, Università di Padova - Via Gradenigo 6a 35100 Padova, Italy*
([3]) *College of Computing, Georgia Institute of Technology - Atlantic Drive 801 30332-0280 Atlanta, GA, USA*

ricevuto il 30 Settembre 2011; approvato l' 1 Dicembre 2011

Summary. — Optimization methods are a powerful tool in protein structure analysis. In this paper we show that they can be profitably used to solve relevant problems in drug design such as the comparison and recognition of protein binding sites and the protein-peptide docking. Binding sites recognition is generally based on geometry often combined with physico-chemical properties of the site whereas the search for correct protein-peptide docking is often based on the minimization of an interaction energy model. We show that continuous global optimization methods can be used to solve the above problems and show some computational results.

PACS 02.70.-c – Computational techniques; simulations.
PACS 02.60.Pn – Numerical optimization.
PACS 87.15.bg – Tertiary structure.

1. – Introduction

Two relevant problems in drug design are, among others, the comparison and recognition of protein binding sites and the protein-peptide docking. Indeed, the identification of protein binding sites, their classification and analysis is of much interest for treatment of diseases. Moreover, when designing a new drug or protein, the interaction with a particular peptide is typically required since biochemical specificity relies on the selective binding of molecules to a given protein in a well-defined orientation.

As concerns the function of a protein, it typically depends on the structure of specific binding sites located at the surface of the protein where the interaction with a ligand takes place. 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.

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].

As regards the docking problem, classical approaches are based on discrete optimization algorithms which perform either a deterministic or stochastic search of the conformational space of the ligand in the binding pocket. Such methods are at the basis of many well-known software packages such as: FlexX [17], GOLD [18], Dock [19], AutoDock [20] and DynaDock [21].

We present two methods for continuous global optimization of multivariate functions that can be profitably used in protein structure analysis problems like those mentioned above. We present also some preliminary numerical results that appear to confirm the usefulness of the proposed approach.

2. – Preliminaries on global optimization methods

In this section we provide some basic notions and definition about continuous global optimization problems and methods. To this aim, let $f(x) : \Re^n \to \Re$ be a real-valued function of n unknowns and consider the problem of finding the global minimum points of $f(x)$ onto a feasible set $\Omega \subseteq \Re^n$, that is

$$\text{glob}\min\ f(x), \qquad \text{subject to } x \in \Omega. \tag{1}$$

We denote by Ω^* the set of global minimum points of f onto Ω. We assume, as usual, that the function $f(x)$ is at least Lipschitz-continuous [22] on Ω, that is, for every $x, y \in \Omega$, a constant $L > 0$ exists such that

$$f(x) - f(y) \leq L\|x - y\|.$$

The methods that allow to find a point $x^* \in \Omega^*$ can be roughly classified into two main classes: a) probabilistic methods and b) deterministic ones. Probabilistic methods are typically quite fast in locating a good approximation of a global minimum point but have weak convergence properties, that is, they can only be proved to converge to x^* with probability one. On the contrary, deterministic methods may be very slow in approximating a global minimum of $f(x)$, but the convergence to a solution can be proved. Typically, deterministic methods exhibit the so-called everywhere dense convergence, *i.e.* they generate a set of points in Ω (or in a relaxation of it) that tends to become dense in the limit as the iteration number increases.

To solve the binding site comparison problem, we use of a probabilistic method which is a modification of the controlled random search algorithm proposed in [23-26]. For the protein-peptide docking problem we use a deterministic method [27] based on the introduction of an additive Gaussian-filling function thus allowing the exploration of a large search region.

3. – Binding site comparison

In this section we present the modified controlled random search algorithm used to solve the binding site comparison problem and we present some numerical results.

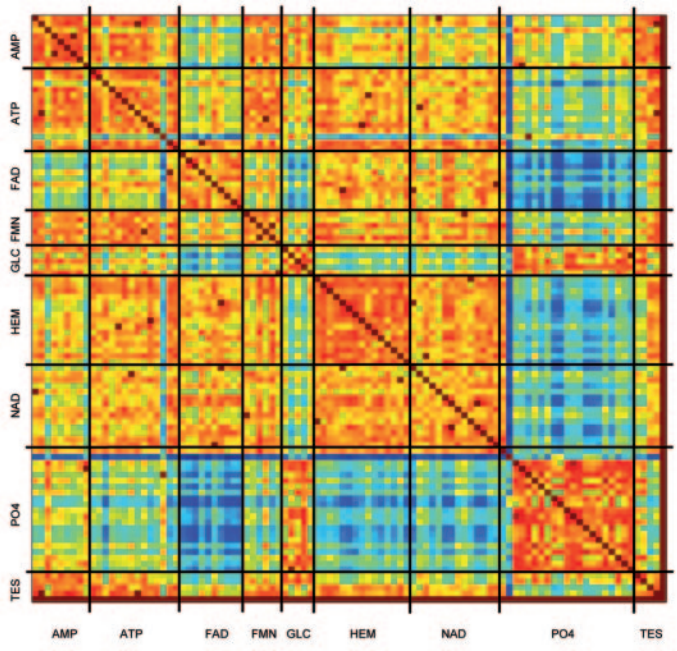

Fig. 1. – Distance matrix for all-to-all comparison.

The global optimization algorithm we use which is a modification of the method proposed in [28]. It is a population based algorithm in the sense that, throughout 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.

The objective function that we want to globally minimize depends on the six parameters describing an isometric transformation of a binding site with respect to the other, that is, three traslations and three rotation angles. The objective function, given an isometric transformation, measures the dissimilarity between the two binding sites. Hence, the lower the objective function value the more similar the two binding sites.

The proposed continuous global optimization (CO) method has been benchmarked on a dataset of 100 proteins in complex with 9 ligands: AMP, ATP, FAD, FMN, GLC, HEME, NAD, PO4, and Steroid; the ligands differ in chemical composition as well as in size and shape. This dataset was used in [29] 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 pairwise comparisons are visualized by means of a distance matrix. 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.

The results of all-to-all comparisons are illustrated by means of the distance matrix of fig. 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

$$2\left(\frac{\text{no. aligned atoms}}{n+m}\right),$$

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 is known [29], 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 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 [30], 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.

4. – Protein-peptide docking

This section concerns the use of a global optimization method for the solution of protein-peptide docking problems. Molecular docking programs play a crucial role in drug design and development. In recent years, much attention has been devoted to this problem where docking of a flexible peptide with a known protein is sought. We consider a docking algorithm based on the use of a filling function method for continuos unconstrained global optimization [27]. The correct protein-peptide docking position is obtained by minimizing the function representing the total potential energy according to a specific mathematical model. In order to preserve the primary sequence of the given peptide it is necessary to take into account some constraints on the problem variables, and then we construct the Lagrangian of the original problem.

The resulting optimization problem has two main features; it is a large-scale one in constrained global optimization, and the total potential energy function has many local minima. Once a local minimum has been found, the method modifies the original objective function by adding to it a filling term. This allows the algorithm to escape from the local minimum so that it may explore large regions in the search space.

As regards the mathematical model employed, we assume that the protein tertiary structure and the peptide primary residue sequence are known and we denote by N and M the number of peptide and protein residues, respectively.

In order to find the docking pocket and position of the peptide onto the protein, we consider that peptide is flexible in such a way that all its atoms have three degrees of freedom. Moreover, the given peptide primary residue sequence is not modified while calculating the docking position by means of suitable simple nonlinear constraints. Let

$$E_{LJ}^{ij} = 4\epsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right]$$

and

$$E_{C}^{ij} = \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}}$$

be the Lennard-Jones and Coulomb potentials, respectively. They represent the interaction between protein atom i and peptide atom j and r_{ij} denotes their distance. Hence, we consider the following total potential energy function [21],

$$E(r) = \sum_{i,j}\left(E_{LJ}^{ij} + E_{C}^{ij}\right), \tag{2}$$

where the summation is calculated over all pairs $(N \times M)$ of atoms.

The approaches proposed in the literature are based on the minimization of successive approximations of (2), obtained by introducing some suitable parameters (see, *e.g.* [31, 21]) in order to simplify the single minimization step. Moreover, these methods determine the docking between the peptide and a prefixed small part of the protein, namely the prefixed receptor or binding site. Most of these methods allow for receptor flexibility [32, 21, 31]. The minimization process is based on a multi-start strategy that uses the steepest descent or conjugate gradient methods as local minimization tools.

Here we consider the problem of finding the docking position of a given peptide by taking into account all the given protein atoms, so that the binding site is not prefixed, but directly determined by the algorithm. In order to avoid that the peptide primary sequence is modified, we consider the following constraints between the carbon alpha atoms of the peptide residues:

$$r_{i,i+1}^{c} = 3.8, \qquad \forall\, i = 1, \ldots, N-1, \tag{3}$$

and

$$r_{i,k}^{c} \geq 2, \qquad \forall\, i = 1, \ldots, N-2, \quad k = i+2, \ldots, N. \tag{4}$$

Therefore, our algorithm computes the final docking position by solving the following problem:

$$\text{glob}\min\ E(r), \qquad \text{subject to (3) and (4)}. \tag{5}$$

Finally, once the binding site has been identified as that corresponding to the lowest potential energy function value, it is possible to adjust the final solution allowing for receptor flexibility, by applying, for instance, a proviously proposed tool (*e.g.*, DynaDOCK, AutoDOCK, FDS).

REFERENCES

[1] Shatsky M., Shulman-Peleg A., Nussinov R. and Wolfson H. J., *J. Comput. Biol.*, **13** (2006) 407.
[2] Shulman-Peleg A., Nussinov R. and Wolfson H. J., *J. Mol. Biol.*, **339** (2004) 607.
[3] Artymiuk P., Spriggs R. and Willett P., *J. Am. Soc. Inf. Sci. Technol.*, **56** (2005) 518.
[4] Chen B., Bryant D., Fofanov V., Kristensen D., Cruess A., Kimmel M., Lichtarge O. and Kavraki L., *Cavity-aware motifs reduce false positives in protein function prediction*, in *Proceedings of Computational System Bioinformatics Conference*, 2005, pp. 311–323.
[5] Hofbauer C., Lohninger H. and Aszodi A., *J. Chem. Inf. Comp. Sci.*, **44** (2004) 837.
[6] Weskamp N., Kuhn D., Hullermeier E. and Klebe G., *Bioinformatics*, **20** (2004) 1522.
[7] Ballester P. and Richards W., *J. Comput. Chem.*, **28** (2007) 1711.
[8] Sommer I., Müller O., Domingues F., Sander O., Weickert J. and Lengauer T., *Bioinformatics*, **23** (2007) 3139.
[9] Bock M., Garutti C. and Guerra C., *J. Comput. Biol.*, **14** (2007) 285.
[10] Bock M., Garutti C. and Guerra C., *Effective labeling of molecular surface points for cavity detection and location of putative binding sites*, in *Proceedings of the VI International Conference on Computational Systems Bioinformatics; San Diego*, 2007, pp. 263–274.
[11] Bock M., Garutti C. and Guerra C., *Theor. Comput. Sci.*, **408** (2008) 151.
[12] Angaran S., Bock M., Garutti C. and Guerra C., *Nucl. Acids Res.*, Web server issue (2009).
[13] Ausiello G., Gherardini P. F., Marcatili P., Tramontano A., Via A. and Helmer-Citterich M., *BMC Bioinformatics*, **9** (2008) S2.
[14] Jambon M., Olivier A., Combet C., Deleage G., Delfaud F. and Geourjon C., *Bioinformatics*, **21** (2005) 3929.
[15] Kinoshita N., Furui J. and Nakamura H., *J. Struct. Funct. Genom.*, **2** (2001) 9.
[16] Shulman-Peleg A., Shatsky M., Nussinov R. and Wolfson H. J., *Nucl. Acids Res.*, **36** (Web server issue) (2008).
[17] Rarey M., Kramer B., Lengauer T. and Klebe G., *J. Mol. Biol.*, **261** (1996) 470.
[18] Jones G., Willett P., Glen R. C., Leach A. R. and Taylor R., *J. Mol. Biol.*, **267** (1997) 727.
[19] Shoichet B. K. and Kuntz I. D., *Protein Engin.*, **6** (1993) 723.
[20] Morris G. M., Goodsell D. S., Halliday R. S., Huey R., Hart W. E., Belew R. K. and Olson A. J., *J. Comput. Chem.*, **19** (1998) 1639.
[21] Antes I., *Proteins: Struct. Funct. Bioinf.*, **78** (2010) 1084.
[22] Ortega J. M. and Rheinboldt W. C., *Iterative Solution of Nonlinear Equations in Several Variables*, in *Classics in Applied Mathematics*, **30** (SIAM, Philadelphia) 2000.
[23] Price W. L., *A controlled random search procedure for global optimization*, in *Towards Global Optimization*, **2**, edited by Dixon L. and Szego G. (North-Holland, Amsterdam) 1978.
[24] Brachetti P., De Felice Ciccoli M., Di Pillo G. and Lucidi S., *J. Global Optim.*, **10** (1997) 165.
[25] Liuzzi G., Lucidi S., Parasiliti F. and Villani M., *IEEE Trans. Magn.*, **39** (2003) 1261.
[26] Liuzzi G., Lucidi S., Piccialli V. and Sotgiu A., *Math. Progr.*, **101** (2004) 339.
[27] Lampariello F., Tech. Rep. IASI-CNR R.615, IASI-CNR (2004).
[28] Cirio L., Lucidi S., Parasiliti F. and Villani M., *J. Appl. Electromagn. Mech.*, **16** (2002) 13.
[29] Kahraman A., Morris R. J., Laskowski R. A. and Thornton J. M., *J. Mol. Biol.*, **368** (2007) 283.
[30] Stockwell G. R. and Thornton J. M., *J. Mol. Biol.*, **356** (2006) 928.
[31] Taylor R. D., Jewsbury P. J. and Essex J. W., *J. Comput. Chem.*, **24** (2003) 1637.
[32] Apostolakis J., Plückthun A. and Caflisch A., *J. Comput. Chem.*, **19** (1998) 21.