DENPOL: A new program to determine electron densities of polypeptides using extremely localized molecular orbitals

Maurizio Sironi, Michela Ghitti, Alessandro Genoni, Giorgio Saladino, Stefano Pieraccini

DENPOL: a new program to determine electron densities of polypeptides using extremely localized molecular orbitals.

Maurizio Sironi (A), Michela Ghitti (A), Alessandro Genoni (A,B), Giorgio Saladino(A), Stefano Pieraccini(A)

(A) Dipartimento di Chimica Fisica ed Elettrochimica – Università degli Studi di Milano, Via Golgi 19, 20133, Milano, Italy

(B) Quantum Theory Project, University of Florida, P.O. Box 118435, Gainesville, Florida, 32611, USA

E-mail: maurizio.sironi@unimi.it

Tel.: +390250314251

Abstract

A new method to compute high quality electron densities of polypeptides is proposed. The method is based on the transferability properties of extremely localized molecular orbitals, which can be used to describe with great accuracy the different functional groups of a molecule. It is then possible to generate a database of such orbitals, each of them associated with specific amino acids and with the peptide bond. A new program, named DENPOL, has been written aimed at using this database to build up the electron density of a generic polypeptide. Due to both the large number of orbitals required to describe a polypeptide and the non-orthogonal nature of these orbitals, a Divide & Conquer strategy has been used to assemble the final electron density. The application of this strategy is particularly efficient thanks to the extreme localization of the orbitals. The comparison with the corresponding electron densities generated by the Hartree-Fock method, shows the goodness of the proposed approach and indicates that the electron densities generated by DENPOL are very close to those generated by an ab initio approach.

Keywords: extremely localized molecular orbitals, electron densities of polypeptides, divide & conquer

Introduction

The accurate description of the electronic structure of proteins is of fundamental importance for many scientific fields ranging from chemistry to biology, from pharmacy to biotechnology. Accurate force fields developed ad hoc for proteins are largely used in molecular dynamics simulations to obtain information at molecular level about both the chemistry of proteins and their interactions with a variety of substrates. Such information is of course of paramount importance in 'drug design' methodologies.

Despite the relevant improvements of such empirical approaches, there is still the need to develop quantum mechanical strategies in order to accurately describe proteins from ab initio principles. However, the quantum mechanical methods are very demanding from a computational point of view and the scaling of the CPU time with respect to the molecular dimension largely restricts the maximum number of atoms that can be described. Even without considering electronic correlation effects, that can be very important in some situations, e.g. to describe a reaction which takes place in an active site, the scaling behaviour is at least $O(N^2 \log N)$ for the conventional Hartree-Fock method.

The molecular orbitals (MOs) determined by Hartree-Fock calculations are largely spread out on all the atoms of the molecule, and even a slight modification of the molecular structure, e.g. the addition of a methyl group, asks for a new complete determination of all the MOs. Apparently there is 'no chemistry' in the shape of the canonical MOs.

Chemistry is made of local concepts: we learn organic chemistry through the study of functional groups, we understand mechanisms drawing arrows which indicate the movement of electrons to break some bonds and to establish new ones. We feel that a local vision of the molecules is more suitable to get a proper understanding of the molecular world.

The importance of these aspects was immediately recognised by the MO community and different localization strategies have been proposed [1-3]. These strategies, based on different criteria, permit to transform the canonical (delocalised) MOs into a set of localised MOs (LMOs) which can be very useful to get a 'chemical understanding' of the electronic structure. The wavefunction built up using the LMOs or the canonical MOs is the same, due to the double occupancy of the MOs which permits to mix them without changing the wavefunction. So, without loosing in accuracy, we can recover to a great extent the chemical concepts based on bonds, lone pairs etc. However, the LMOs are mainly localised on a few atoms but still preserve some tails on many atoms of the molecule. In other words they permit to identify chemical entities in a intuitive but not rigorous way. This fact reflects in the impossibility to define a group of LMOs able to describe a functional group which can be used in every molecule which contains that functional group. Of course one

possibility could be represented by the tail deletion, i.e. we could just put to zero the coefficients of the atomic functions which are centred on the atoms which do not belong to the functional group we are interested in. In this way we would obtain a transferable group of LMOs but it is evident that we have introduced an arbitrary operation. In addition, it has been clearly evidenced that the small coefficients which describe the tails of the LMOs have non negligible effect on the electronic energy. It should also be pointed out that different LMOs are obtained using different localization schemes, and neither of them can be considered better than others, as they correspond to the same energy value. The definition of a group of LMOs to define a functional group is then also dependent on the adopted localization strategy.

A different approach is based on the use of molecular orbitals which are a priori localised on a small number of atoms but which are variationally determined under this constraint. In this case the orbitals which define the functional group are only biased by the initial choice of the selected atoms to define it, but their determination is a straightforward application of the variational principle. Even if the choice of the atoms to define the functional group can be questionable due to its arbitrariness, it has been demonstrated that in many cases it largely relies on solid chemical ideas. Methods based on this approach permit to obtain extremely localised molecular orbitals (ELMOs) [4]. Their use has been proposed by different research groups using different acronyms, such as strictly localised molecular orbitals (SLMOs) [5] and non-orthogonal localised molecular orbitals (NOLMOs) [6]. All these methods can be considered as an extension of the group function method introduced by McWeeny [7] to extract from a wavefunction functions which describe subsets of electrons.

ELMOs are rigorously defined without the presence of any tails. Once they have been determined on a functional group of a model molecule, they can then be directly transferred to a target molecule which contains the same group without further perturbation. The ELMOs could be considered as an attempt to introduce Valence Bond (VB) ideas into the framework of the MOs theory. Like VB orbitals, ELMOs are non orthogonal and this reflects into a mathematical and algorithmic complexity in the attempt to determine them: non trivial convergence problems arise in their determinations, and particular care must be taken in the development of algorithms to compute ELMOs. Of course, ELMOs, unlike VB orbitals, are doubly occupied and this permits to devise theoretical approaches which are simpler with respect to the VB methods.

The transferability of the ELMOs has already been checked and carefully analyzed on small systems [8].

As the number of atoms increases, also the determination of the electron density based on a large set of non-orthogonal orbitals becomes a demanding task. However, the extremely localised nature

of the ELMOs reflects in a fast decrease of the overlap between them as the distance between the different regions (functional groups) that they describe increases. This permits a direct application of the Divide & Conquer (D&C) strategy [9-12], originally devised to develop a linear scaling approach in the framework of the Density Functional Theory (DFT), to determine the electron density distribution of a large molecule using a set of ELMOs.

The accuracy of this approach and the scaling of the computational resources with respect to the dimension of the system has been already investigated on poly-glycine and poly-leucine [13].

The results obtained suggest that the electron density of a polypeptide can be accurately described using an approach based on the ELMOs determined on the constituent amino acids and transferred to the target polypeptide. For this purpose we have devised a program named DENPOL to compute the electronic density of a general polypeptide based on a database of ELMOs determined on the 20 different amino acids.

In this respect there is an analogy with the work proposed by Walker and Mezey in the approach called Molecular Electron Density Lego Assembler (MEDLA) [14] and further extended in the scheme called Adjustable Density Matrix Assembler (ADMA) [15-17].

In the Theory section after a brief summary of the ELMOs determination and of the D&C strategy to compute electron density, the DENPOL machinery is reported. The applications to some polypeptides and a small protein and the comparison of the results with Hartree-Fock calculations are discussed in the Results section.

Theory

A survey of the DENPOL strategy.

The aim of the DENPOL program is to compute the electron density for an input geometry of a polypeptide using a database of precomputed ELMOs. The requested input is a PDB file which must contain the position of all the atoms (hydrogen atoms included) and that can be downloaded for example from the Protein Data Bank [18] or could be the output of a molecular dynamics program.

The DENPOL program is divided in two principal sections.

In the first one, the structure of the input polypeptide (which will be denoted as the target polypeptide) is sequentially examined in order to recognize the different amino acids and to select the proper ELMOs, stored in the DENPOL database, that must be transferred at the geometry of the

target polypeptide through an opportune rototranslation step. In this way the set of all the ELMOs which constitute the target polypeptide is obtained.

In the second section the electron density must be computed. The DENPOL program needs the overlap matrix between all the atomic functions centred on all the atoms of the target polypeptide, which is easily obtained through an interface with an ab initio package program (GAMESS-US [19] or GAMESS-UK[20]). From this matrix the overlap between the ELMOs can be easily computed. Due to the large number of ELMOs associated with a polypeptide and to their non-orthogonality, the D&C strategy is adopted to compute the electron density, which is evaluated on a grid of points to be successively examined.

We now briefly discuss the algorithm which has been used to determine the ELMOs that constitute the DENPOL database and the method used to compute the electron density. We will then elucidate the definition of the molecular fragments chosen and we will describe some technical details of the DENPOL program.

The determination of the ELMOs.

Let us consider a *2N* electrons closed shell molecule. We subdivide the molecule into n_f molecular fragments (lone pairs, bonds, functional groups, etc). Each of these subsystems is characterized by its own basis set, $\left(\chi_\mu^i\right)_{\mu=1}^{m_i}$, m_i being the number of the atomic orbitals belonging to the *i*-th fragment.

The *i*-th fragment will be described by N_i^o ELMOs, whose number will depend on the chemical nature of the fragment itself (of course $\sum_i^{n_f} N_i^0 = N$). The ELMO wavefunction will be written as:

$$\Psi_{ELMO} = A\left(\Phi\right) \quad \text{with } \Phi = \prod_i^{n_f} \prod_\alpha^{N_i^o} \varphi^i{}_\alpha \alpha \, \varphi^i{}_\alpha \beta$$

where *A* is the antisymmetrizer operator and the α-th ELMO of the *i*-th fragment is described through only the atomic functions which belong to the same fragment:

$$\varphi^i{}_\alpha = \sum_{\mu=1}^{m_i} c_{\mu\alpha}^i \chi_\mu^i$$

In this way each ELMO is rigorously localized on the corresponding fragment. The coefficients $c^i_{\mu\alpha}$ are determined through a minimization of the energy associated with the ELMO wavefunction.

The definition of the molecular fragments is arbitrary and it depends on the application. The choice adopted in the DENPOL program will be discussed later on. Of course, molecular fragments can share atomic functions. For example, if we describe each bond using a "Lewis localization scheme", the atomic functions centred on atoms involved in more than one bond, are common to different molecular fragments.

The ELMOs are determined using the variational principle and in this respect they are the best orbitals which describe the molecular fragments. It can be shown [21-22] that the ELMOs must satisfy the following set of coupled equations:

$$\hat{F}_i \, \varphi^i_\alpha = \varepsilon^i_\alpha \, \varphi^i_\alpha \qquad \left\{ \begin{matrix} i = 1,.....n_f \\ \alpha = 1,....,m_i \end{matrix} \right\} \qquad (1)$$

where $\hat{F}_i$ is a proper hermitian operator for the *i*-th fragment built up using the occupied ELMOs of all the n_f molecular fragments. Each molecular fragment has its own set of occupied and virtual orbitals with the same localization scheme and the orbitals belonging to the same fragments constitute an orthogonal set. ELMOs belonging to different fragments are instead not orthogonal as they are eigenfunctions corresponding to different hermitian operators. When converging, the algorithm based on the resolution of Eq. (1) is quite rapid, but some difficulties in the convergence or some instabilities sometimes arise due to the non orthogonal nature of the orbitals.

In these cases it is necessary to consider a second order algorithm based on a quasi-Newton procedure, where an approximate analytic Hessian computed at the first iteration is updated with a variable metric method in the subsequent iterations. The extremely localized nature of the orbitals permits to introduce approximations in the computation of the full Hessian matrix [23].

The ELMO program has been interfaced with the GAMESS_US [19] and with the GAMESS-UK packages [20]. Both single point calculations (conventional and direct) and geometry optimizations can be carried out.

The determination of the electronic density.

Once the ELMOs have been determined on model molecules to properly define functional groups, they can be transferred to a target molecule.

Let's indicate with $\left(\chi_\mu\right)_{\mu=1}^{m}$ the list of the unique atomic functions centred on all the atoms of the target polypeptide. The transferred ELMOs can be expressed as $\varphi_\alpha^i = \sum_\mu^m c_{\mu\alpha}\chi_\mu^i$, where many coefficients $c_{\mu\alpha}$ will be zero according to the localization pattern.

Recalling that the ELMOs constitute a non orthogonal set, the electron density is given by:

$$\rho(\mathbf{r}) = \sum_{\mu,\nu=1}^{NAO} \chi_\mu(\mathbf{r})\, \mathbf{D}_{\mu\nu}\, \chi_\nu(\mathbf{r})$$

where:

$$\mathbf{D}_{\mu\nu} = 2\sum_{i,j=1}^{N} c_{\mu j}\left[\mathbf{S}^{-1}\right]_{ji} c_{\nu i}^{*} \qquad (2)$$

S is the overlap matrix between the ELMOs, that can be easily computed from the overlap between the atomic basis functions.

For a large system like a polypeptide, the use of the eq. (2) becomes soon prohibitive. To overcome the problem we use a D&C strategy [13] which is made particularly attractive from the extreme localization nature of ELMO which reflects in a sparse nature of the matrix **S**. The target polypeptide is divided into N_s disjoint subunits, each of them corresponding to a different amino acid, so that their union provides the whole system. Owing to this fragmentation, we automatically assign a local basis set β_k to the generic k-th subsystem (k = 1, 2, …, N_s). Furthermore, following the D&C strategy, we introduce the partition matrix $\mathbf{P}^k$ for the k-th subunit:

$$\mathbf{P}_{\mu\nu}^{k} = \begin{cases} 1 & \text{if } \mu\in\beta_k \text{ and } \nu\in\beta_k \\ 1/2 & \text{if } \left(\mu\in\beta_k \text{ and } \nu\notin\beta_k\right) \text{ or } \left(\mu\notin\beta_k \text{ and } \nu\in\beta_k\right) \\ 0 & \text{if } \mu\notin\beta_k \text{ and } \nu\notin\beta_k \end{cases}$$

which obviously satisfies the condition

$$\sum_{k=1}^{N_S} \mathbf{P}_{\mu\nu}^{k} = 1 \quad \forall\, \mu,\nu$$

and allows to assemble the density matrix as a sum of contributions arising from the different subsystems:

$$\mathbf{D}_{\mu\nu} = \sum_{k=1}^{N_S} \mathbf{D}_{\mu\nu}^{k} \quad \textit{with} \quad \mathbf{D}_{\mu\nu}^{k} = \mathbf{P}_{\mu\nu}^{k}\mathbf{D}_{\mu\nu}$$

The $\mathbf{D}^{k}_{\mu\nu}$ elements can be efficiently evaluated considering only n_k ELMOs, namely the ELMOs localized on the k-th subsystem ("Core ELMOs") and the ones that significantly overlap with them ("Buffer ELMOs") i.e. orbitals generally belonging to near subunits in the real space.

Therefore, for the k-th subunit:

$$\mathbf{D}^{k\ (approx)}_{\mu\nu} = 2\,\mathbf{P}^{k}_{\mu\nu} \sum_{i,j=1}^{n_k} C_{\mu j} \left[\mathbf{S}_k^{-1}\right]_{ji} C^{*}_{\nu i}$$

with n_k as the total number of considered ELMOs (i.e., the core and buffer ELMOs) for the k-th subunit and $\mathbf{S}_k$ as their overlap matrix. Of course n_k is generally much lower than N, and the inversion of N_s matrices $\mathbf{S}^k$ is very fast. Using this approach the electron density determined by a large set of ELMOs can be efficiently computed.

The DENPOL program.

The strategy adopted by the DENPOL program is to use a database of ELMOs for each lateral chain of the 20 amino acids, for the amidic bond, and for the N- and C- terminus of a polypeptide. The choice of the localization scheme adopted to define the ELMOs has been done in order to satisfy the following two aspects:

(a) each set of ELMOs used to describe a fragment should have the maximum delocalization in order to give the lowest energy; it is obvious that strict localization schemes are associated with a high reduction in the number of variational coefficients and with a large increase in the energy

(b) each set of ELMOs must be sufficiently localized in order to easily describe the rotations through the different bonds, using a proper rotation of the ELMOs.

In order to clarify these concepts, let's consider as an example the definition of a set of ELMOs for the $-CH_2CH_2CH_3$ moiety. The ELMOs should be obtained through a calculation on a model molecule $X\text{-}CH_2CH_2CH_3$ where an opportune number of ELMOs should describe the X group (i.e. it should use only the atomic functions centred on the X group) and another set of ELMOs should describe the propyl group. The X group should be chosen to properly describe the neighbourhood of the $-CH_2CH_2CH_3$ moiety. To satisfy the condition (a) we should define the ELMOs of the propyl group delocalized on all the atoms of this group. However this choice wouldn't permit to transfer the ELMOs to a target molecule characterized by a different orientation of the CH_3 moiety with

respect to the CH_2 group, unless to perform a new calculation on the model molecule. It is then more appropriate to have a more localized description and to use the following scheme:

- a set of ELMOs localized on the CCH3 atoms, which describe the electrons of the C-C bond and of 3 C-H bonds

- a set of ELMOs localized on the C-CH2 atoms, which describe the electrons of the C-C bond and the 2 C-H bonds of the central CH_2 group

- a set of ELMOs localized on the (first atom of the X group)- CH_2 which describe the electrons of the bond between the first atom of the X group and the C atom of the first CH_2 group, and the 2 C-H bonds.

In this way, for example, a rotation of the terminal methyl group can be described using the appropriate rotation of all the ELMOs describing the CCH3 moiety.

Following the same strategy, the lateral chain of the phenylalanine, C(α)-CH_2-Phe uses a set of ELMOs localized on the C-Phe atoms and a set of ELMOs localized on the C(α)-CH_2 atoms.

Using the above strategy a localization scheme has been introduced for each lateral chain (R-group) of the 20 amino acids. The model molecules used are reported in Fig. 1, and have been chosen in order to reproduce the environment of the target polypeptide. The model molecules permits also to define the ELMOs which describe the bonds C(α)-N, and the fragment C(α)H-C which connect the lateral chain to the two contiguous peptide bonds. A complete description of the localization schemes used for all the amino acids has been reported in the supplementary material (Table 1S), while a pictorial representation of the ELMOs which describe the lateral chain of tyrosine is reported in Fig. 2.

The ELMOs stored in the DENPOL database must be then transferred to the target polypeptide. Of course the geometries of the amino acids of the polypeptide will be not equal to the geometry used in the model molecules. Small differences are observed for bond lengths and bond angles, while significant variations will be observed for the torsion angles (the lateral chain of the amino acids experience large conformational fluctuations in the course of a molecular dynamics simulation, see for example Fig.3). We observe that, due to the above reported strategy to define the ELMOs, a conformational change can be easily taken into account. In our example of the propyl moiety, the rotation around the C-C bonds is easily taken into account, as the ELMOs are defined for the group

of atoms which rotates altogether (for example the methyl group). The ELMOs contained in the database can not take into account a variation for example of the HCH angle in the methyl group, or a variation in the C-CH_3 distance, which however are characterized by small fluctuations. Hence the DENPOL program builds up the geometry of the target polypeptide that will be slightly different from the input geometry; it is constructed taking into full account the conformational variations and discarding small variations in bond lengths. The geometry is assembled using the following strategy. Let's consider the i-th amino acid in the target polypeptide (see Fig.4). DENPOL generates a geometry where the $C^i(\alpha)$ atom is put in the same position of the input target polypeptide. Then the geometry of the the $C^i(\alpha)HC^i$ backbone atoms in the model molecule (where C^i is the carbon atom of the peptide bond connecting the residue i-th with the i+1th one, see Fig.4) is subjected to a roto-translation which minimizes the root mean square deviations (RMSD) with respect to the geometry of the three atoms in the target polypeptide, subject to the condition that the $C^i(\alpha)$ position must remain fixed. In this way the geometry of the H and C^i atoms is determined (and they will not exactly superimposed with the geometry they have in the target polypeptide). Now the geometry of the peptide bond of the corresponding model molecule is roto-translated in order to minimize the RMSD with the corresponding atoms in the target polypeptide, considering that the position of the C^i has been already established. In this way the position of the O, N and H atoms of the peptide bond is determined. The procedure is repeated for the subsequent i+1 th residue. The same strategy is adopted to determine the positions of the atoms of the lateral chains.

In this way the final geometry of the polypeptide built up from DENPOL is very close to the input polypeptide geometry, and the ELMOs are only subjected to the requested roto-translations. Only the ELMOs which describe the N-C(α) bonds must be subjected to a renormalization due to the variations of the bond lengths. The structure of the DENPOL program permits to tune the database structure in order to adopt a larger number of model molecules for the same fragments. For example the peptide bond is associated with different model molecules to take into account the presence of the proline residue, or the less common cis arrangement of the atoms. The same residue can be described by different protonation states according to the input geometry. Some model molecules are considered to describe the N- and C-termini in the zwitterionic form or the most common capped structures. Of course a fine tuning of the database could be further implemented in order to describe some particular situations.

Model calculations

The DENPOL program has been tested on some polypeptides and on a small protein comparing the electron densities obtained with those computed at the Hartree-Fock level.

The database of the ELMOs has been constructed using a standard STO-4G basis set (but different basis set could be otherwise implemented). The geometry of the model molecules have been instead computed using a standard 6-31G basis set at the Hartree-Fock level, as preliminary calculations had shown a much better agreement with experimental geometries. The degree of similarity between two electron densities has been determined using the similarity index $L(a, a')$ introduced by Walker and Mezey [24], which compares point by point two charge distributions evaluated in a grid of points contained in the space bounded by two electron density isosurfaces characterized by the values a and a'.

Table 1 shows the comparison of the electron densities obtained using the ELMOs with those computed at the RHF level for the most common model molecules. We observe that in this case there is no a transfer of the ELMOs. The high degree of agreement is indicative of the goodness of the ELMOs approach in reproducing the electron density despite the great reduction in the number of variational parameters to describe the electronic structure of the molecule, which is evidenced by the large difference in the energy values between the ELMO and the RHF methods. Of course much larger differences would have been observed if we had localized the MOs and subjected them to a tail deletion. The favourable comparison reported in Table 1 supports the accuracy of the DENPOL database, i.e. of the building blocks that will be used to assemble the electronic density of input polypeptides. It is also interesting to observe that the agreement is very good for all the pairs of the isosurfaces chosen to compare the electron densities, which means that regions with high, medium and low electron density are all well described.

Four different polypeptides ranging from 13 to 20 residues have been selected to investigate the goodness of the DENPOL approach in transferring the ELMOs to a target polypeptide (see Table 2). These polypeptides contain small sequences which characterize the most common elements of the secondary structures of proteins (see Fig.5). In addition, in order to analyze the capabilities of DENPOL to treat larger structures, a small protein has been considered, too. The geometries of the studied systems have been obtained minimizing with 10000 steps the starting geometry contained in the reported PDB files (Table 2) with the AMBER03 force field [30] in order to relax the structure and to get a correct position for the hydrogen atoms.

As reported in the Theory section, DENPOL program produces a geometry for the target polypeptide that it is slightly different from the input geometry. However the deviations are very small, as it is evident from the RMSD values between the two geometries reported in Table 3.
To evaluate the accuracy of the DENPOL electron densities generated by the transfer of the ELMOs stored in the database to the target polypeptide, we have performed a RHF calculation to obtain a reference electron density. The comparison of the electron densities is reported in Table 3 and the degree of agreement is striking. With respect to the comparison between the ELMOs and the RHF electron densities obtained on the model molecules, which evidently represents an upper limit to the accuracy, there is only a slight reduction in the similarity index confirming that the ELMOs are really transferable.
The agreement is good for all the polypeptides independently on the secondary structures elements they have. In order to investigate larger systems, we have computed the electron density of a small protein, the immunoglobulin binding domain of streptococcal (2GB1), which contains 56 residues. Once again the electron density computed by DENPOL is very similar to that obtained at the Hartree-Fock level as can be visually checked by an exam of Fig. 6, and as it confirmed more quantitatively from the similarity indexes (Table 3). The result is particularly striking considering the differences in the CPU time requested for the two calculations. The Hartree Fock calculation, which was carried out with the direct strategy due to the large number of basis functions (2603 atomic basis functions), used one week of CPU time on an AMD ATHLON MP 2400+ processor. The DENPOL calculation was carried out in just 1 hour using a code that has not been fully optimised yet.
The results show that the electron density of a protein can be really obtained at ab initio quality from the application of the DENPOL program using very limited computation capabilities.

Conclusions

A new program for the determination of high quality electron densities for polypeptides has been proposed. The quality is comparable to the results obtained through a traditional Hartree-Fock calculation. The novel approach is however much cheaper, as it is based on the use of ELMOs which can be easily transferred from one molecule to another one.
A database of ELMOs to properly describe lateral chains and backbone of polypeptides has been generated using a selected set of model molecules. Particular attention has been devoted to the chosen localization scheme, in order to reduce the approximations associated with any localization

constraint and to preserve the use of molecular fragments that can be freely rotated in order to consider all the conformational freedom degrees of a polypeptide.

The program has been checked on some polypeptides and on two small proteins showing very promising results. The structure of the database is open to further developments. It is possible to define different model molecules for the same residue to better tune the neighbour residues.

The proposed approach could be an interesting starting point to improve the hybrid methods, where a quantum mechanical region (e. g. the active site of an enzyme and the substrate) is surrounded by a molecular mechanics region. The description of the latter region could be improved by a frozen quantum-mechanical region obtained by the methodology introduced in this paper.

Finally, the same approach used in the DENPOL program could be adopted to compute the electronic densities of the nucleic acids.

Acknowledgements.

We thank Dott. Giuseppe Lanzani for technical assistance.

References

1. S.F. Boys, Rev. Mod. Phys. 32 (1960) 296.
2. C. Edmiston and K. Ruedenberg, J. Chem. Phys., 43 (1965) S97.
3. J. Pipek and P.G. Mezey, J. Chem. Phys., 90 (1989) 4916.
4. M. Couty, C.A. Bayse and M.B. Hall, Theor. Chem. Acc., 97 (1997) 96.
5. G. Náray-Szabó, Comput. Chem. 24 (2000) 287.
6. G.F. Smits and C. Altona, Theor. Chim. Acta, 67 (1985) 461.
7. R. Mc Weeny, Methods of Molecular Quantum Mechanics Academic Press, New York (1992).
8. M. Sironi, A. Famulari, M. Raimondi and S. Chiesa, J. Mol. Struct., 47 (2000) 529.
9. W. Yang, Phys. Rev. Lett., 66 (1991) 1438.
10. W.Yang and T.S. Lee, J. Chem. Phys., 103 (1995) 5674.
11. S.L. Dixon and K.M. Merz Jr., J. Chem. Phys., 104 (1996) 6643.
12. S.L. Dixon and K.M. Merz Jr., J. Chem. Phys., 107 (1997) 879.
13. A. Genoni, M. Ghitti, S. Pieraccini and M. Sironi, Chem. Phys. Lett., 415 (2005) 256.
14. P.D. Walker and P.G. Mezey, J. Am. Chem. Soc., 116 (1994) 12022.
15. P.G. Mezey, J. Math. Chem., 18 (1995) 141.
16. T.E. Exner and P.G. Mezey, J. Phys. Chem. A, 106 (2002) 11791.
17. T.E. Exner and P.G. Mezey, J. Comput. Chem., 24 (2003) 1980.
18. H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov and P.E. Bourne, Nucl. Acids Res., 28 (2000) 235.
19. M.W. Schmidt, K.K. Baldridge, J.A. Boatz, S.T. Elbert, M.S. Gordon, J.J. Jensen, S. Koseki, N. Matsunaga, K.A. Nguyen, S. Su, T.L. Windus, M. Dupuis and J.A. Montgomery J. Comput. Chem., 14 (1993) 1347.
20. M.F. Guest, I.J. Bush, H.J.J. van Dam, P. Sherwood, J.M.H. Thomas, J.H. van Lente, R.W.A. Havenith and J. Kendrick, Mol. Phys., 103 (2005) 719.
21. H. Stoll, G. Wagenblast and H. Preuss, Theor. Chim. Acta, 57 (1980) 169.
22. M. Sironi and A. Famulari, Theor. Chem. Acc., 103 (1999) 417.
23. M. Sironi, A. Genoni, M. Civera, S. Pieraccini and M. Ghitti, Theor. Chem. Acc., 117 (2007) 685.
24. P.D. Walzer and P.G. Mezey, J. Am. Chem. Soc., 116 (1994) 12022.
25. G. Wang, P.A. Keifer and A. Peterkofsky, Protein Sci.,12 (2003) 1087.

26. K. Lee, S.Y. Shin, K. Kim, S.S. Lim, K.-S. Hahm and Y. Kim, Biochem. Biophys. Res. Commun., 323 (2004) 712.
27. D. Oh, S.Y. Shin, J.H. Kang, K.-S. Hahm, K.L. Kim and Y. Kim, J. Peptide Res., 53 (1999) 578.
28. F.J. Blanco, G. Rivas and L. Serrano, Nature Struct. Biol., 1 (1994) 584.
29. A.M. Gronenborn, D.R. Filpula, N.Z. Essig, A. Achari, M. Whitlow, P.T. Wingfield and M. Clore, Science, 253 (1991) 657.
30. D.A. Case, T.A. Darden, T.E. Cheatham, III, C.L. Simmerling, J. Wang, R.E. Duke, R. Luo, K.M. Merz, D.A. Pearlman, M. Crowley, R.C. Walker, W. Zhang, B. Wang, S. Hayik, A. Roiteberg, G. Seabra, K.F. Wong, F. Paesani, X. Wu, S. Brozell, V. Tsui, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, P. Beroza, D.H. Mathews, C. Schafmeister, W.S. Ross and P.A. Kollman, AMBER 9, University of California, San Francisco, CA, (2006).

Tab. 1

Energies of the Hartree Fock and of the ELMO calculations using the STO-4G basis set and similarity index L(*a,a'*) for some pairs of isosurfaces values (*a,a'*) obtained for typical model molecules of the DENPOL database. The localization scheme is reported in Fig.1S.

Model Molecules	E_{RHF} (a.u.)	E_{ELMO} (a.u.)	L(*a,a'*)		
			L(*0.001,0.01*)	L(*0.01,0.1*)	L(*0.1,10*)
Ala	-490.062104	-490.035803	99.72	99.72	99.71
Arg	-770.041764	-769.990109	99.55	99.60	99.78
Asn	-656.808079	-656.761244	99.43	99.51	99.73
Asp	-675.759116	-675.700531	99.24	99.32	99.65
Cys	-885.480543	-885.451001	99.71	99.72	99.82
Gly	-451.200790	-451.177725	99.60	99.69	99.81
Glu	-714.578145	-714.545178	99.70	99.67	99.81
Gln	-695.663751	-695.618706	99.44	99.55	99.77
His	-712.515155	-712.471648	99.48	99.55	99.76
Ile	-606.628498	-606.589339	99.66	99.60	99.75
Leu	-606.623790	-606.586696	99.70	99.67	99.81
Lys	-661.810380	-661.727225	99.09	99.18	99.54
Met	-963.203141	-963.168433	99.67	99.70	99.82
Phe	-718.450227	-718.418794	99.76	99.73	99.82
Pro	-566.620850	-566.046552	95.80	97.27	97.69
Ser	-564.415494	-564.381457	99.54	99.63	99.80
Thr	-603.265561	-603.233150	99.65	99.69	99.81
Trp	-848.531740	-848.497610	99.71	99.76	99.85
Tyr	-792.814933	-792.780727	99.74	99.73	99.82
Val1	-567.775393	-567.741388	99.67	99.67	99.81
Nt	-339.588812	-339.581665	99.73	99.85	99.91
Ct	-470.143849	-470.134081	99.69	99.80	99.89
Pep1	-734.508272	-734.490712	99.79	99.85	99.88
Pep2	-811.077065	-811.050975	99.76	99.82	99.87

Tab. 2

List of the systems used to test the DENPOL program with their PDB ID, the number of the residues and the sequence of the amino acids.

Name	PDB ID	Nr. residues	Sequence
Aurein1.2	1VM5	13	GLY LEU PHE ASP ILE ILE LYS LYS ILE ALA GLU SER PHE
[K^7]-IsCT	1T52	13	ILE LEU GLY LYS ILE TRP LYS GLY ILE LYS SER LEU PHE
Cecropin A	1D9J	20	LYS TRP LYS LEU PHE LYS LYS ILE GLY ILE GLY LYS PHE LEU HIS SER ALA LYS LYS PHE
β-hairpin	Sub-sequence from 2GB1	16	GLY GLU TRP THR TYR ASP ASP ALA THR LYS THR PHE THR VAL THR GLU
2GB1	2GB1	56	MET THR TYR LYS LEU ILE LEU ASN GLY LYS THR LEU LYS GLY GLU THR THR THR GLU ALA VAL ASP ALA ALA THR ALA GLU LYS VAL PHE LYS GLN TYR ALA ASN ASP ASN GLY VAL ASP GLY GLU TRP THR TYR ASP ASP ALA THR LYS THR PHE THR VAL THR GLU

Tab. 3

Comparison of the electron densities computed by the DENPOL program with the results obtained at the Hartree-Fock level, using the similarity index (see text) for some pairs of isosurfaces values. The RMSD values (in Angstrom) are the deviations of the geometries generated by DENPOL with respect to the input ones. The number of atomic basis functions (STO-4G basis set) for the target polypeptides is reported as NBF.

Name	NBF	RMSD	L(*a,a'*)		
			L(*0.001,0.01*)	L(*0.01,0.1*)	L(*0.1,10*)
Aurein1.2	638	0.043	98.64	98.82	98.93
[K^7]-IsCT	661	0.053	98.58	98.76	98.87
Cecropin A	1057	0.051	98.44	98.78	98.94
β-hairpin	775	0.045	98.83	98.97	99.11
2GB1	2603	0.044	98.72	98.88	99.05

Fig. 1

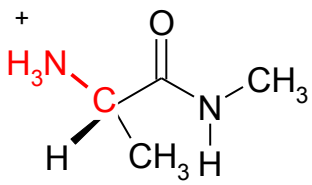
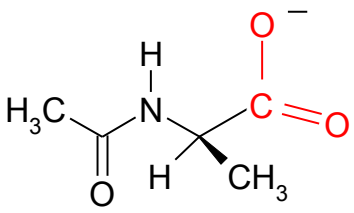
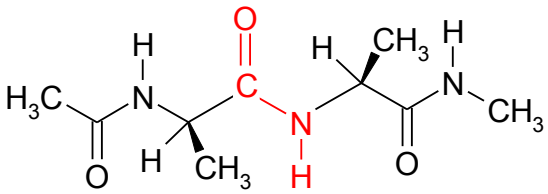
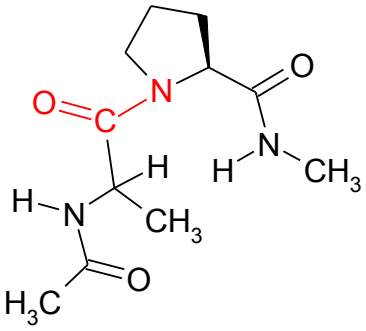
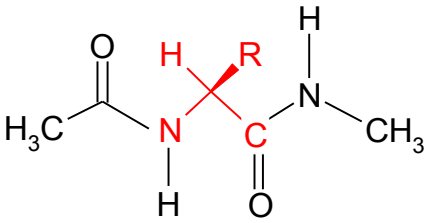
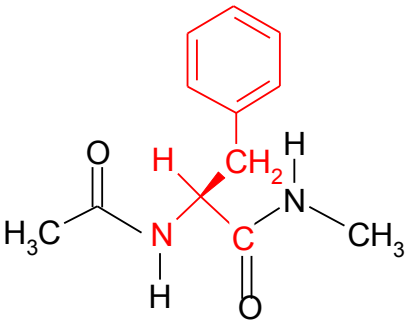
Model Molecule for N-Terminal residues	Model Molecule for C-Terminal residues
	
Model Molecule for peptide bonds	**Model Molecule for peptide bonds**
	
Model Molecule for Central residues	**Example of Model Molecule for the Phenylalanine residue**
	

Fig. 2

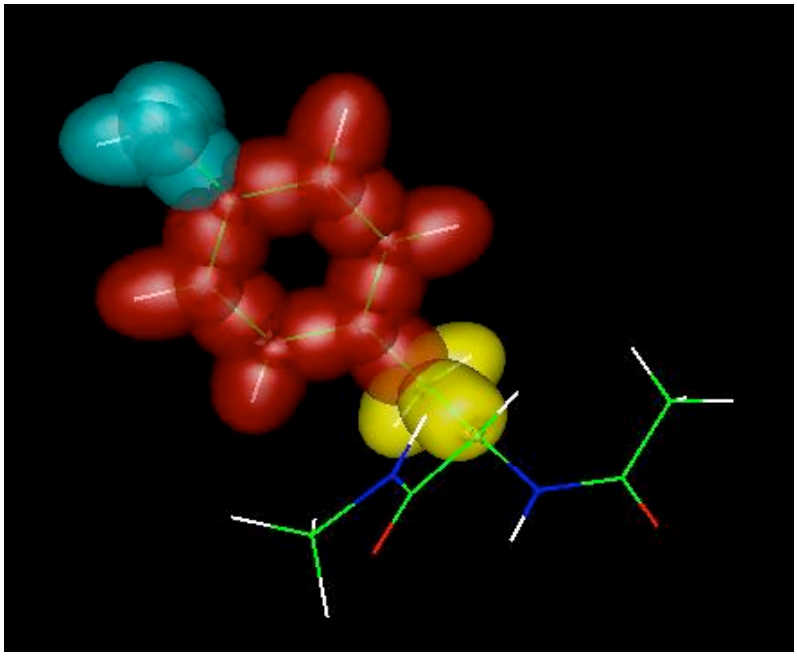


Fig. 3

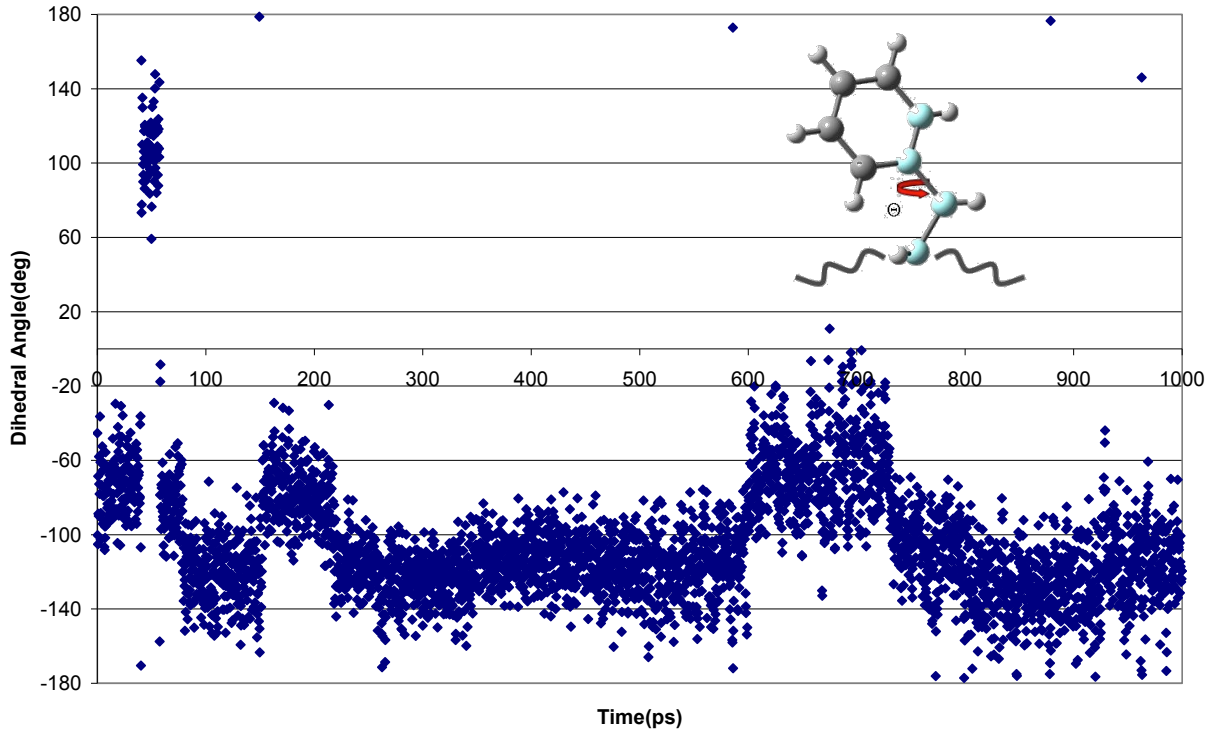

180
140
100
60
20
-20
-60
-100
-140
-180
0
100
200
300
400
500
600
700
800
900
1000
Dihedral Angle(deg)
Time(ps)

Fig. 4

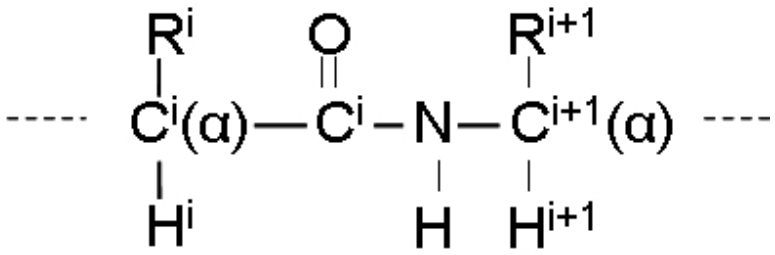


Fig. 5a

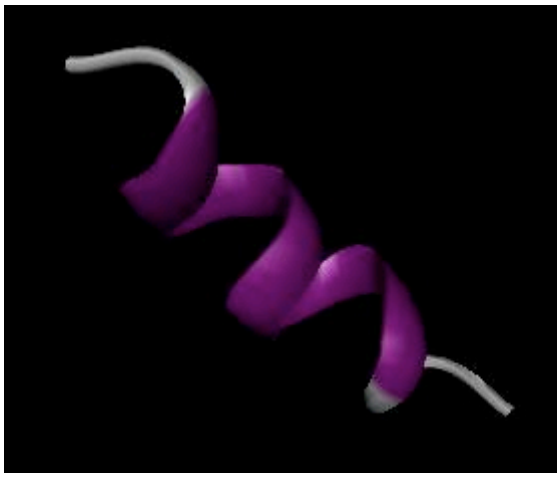


Fig. 5b

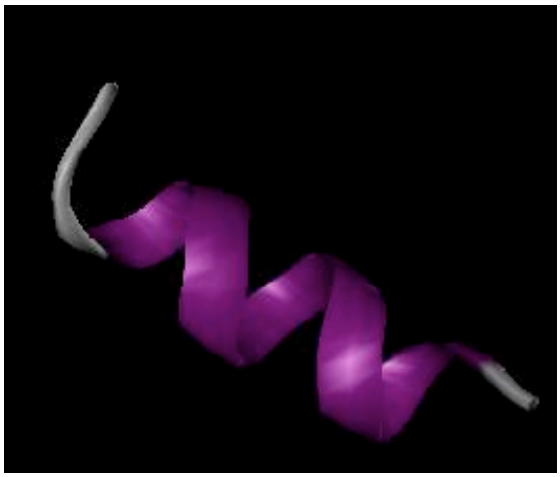


Fig. 5c

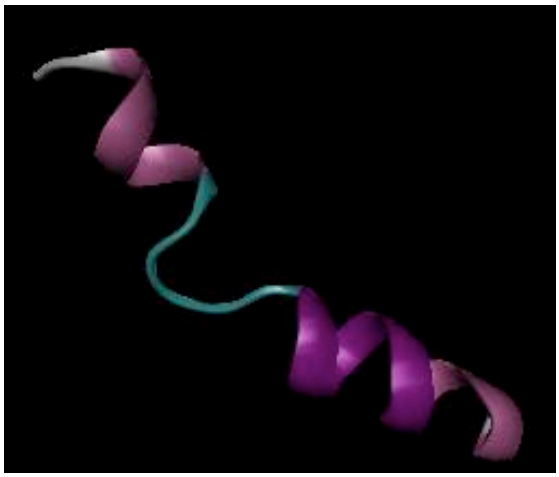


Fig. 5d

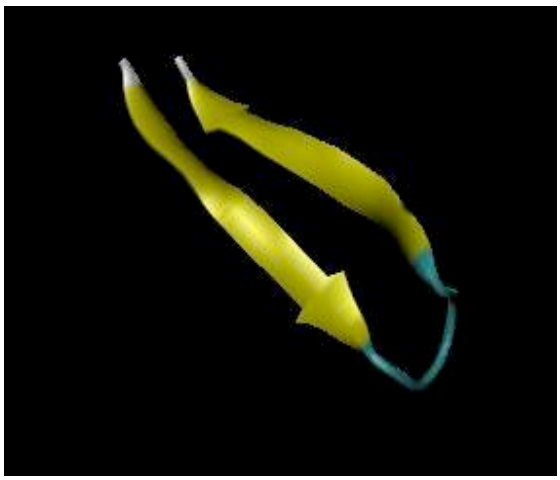


Fig. 5e

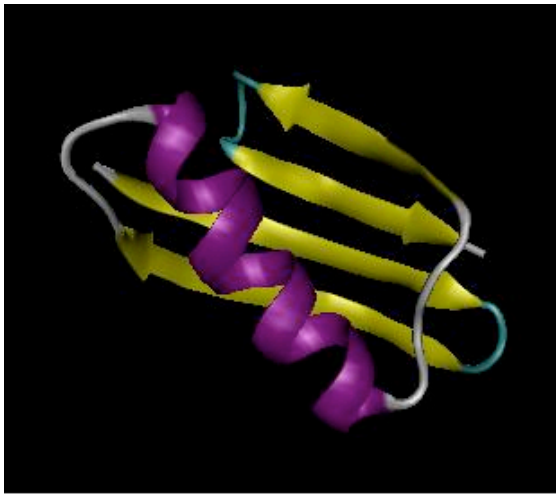
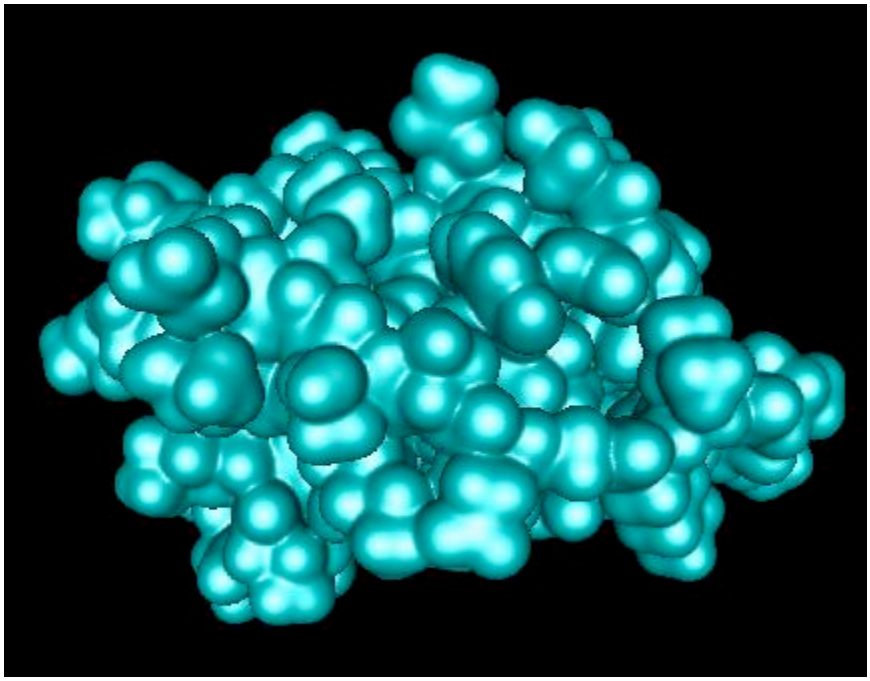


Fig. 6



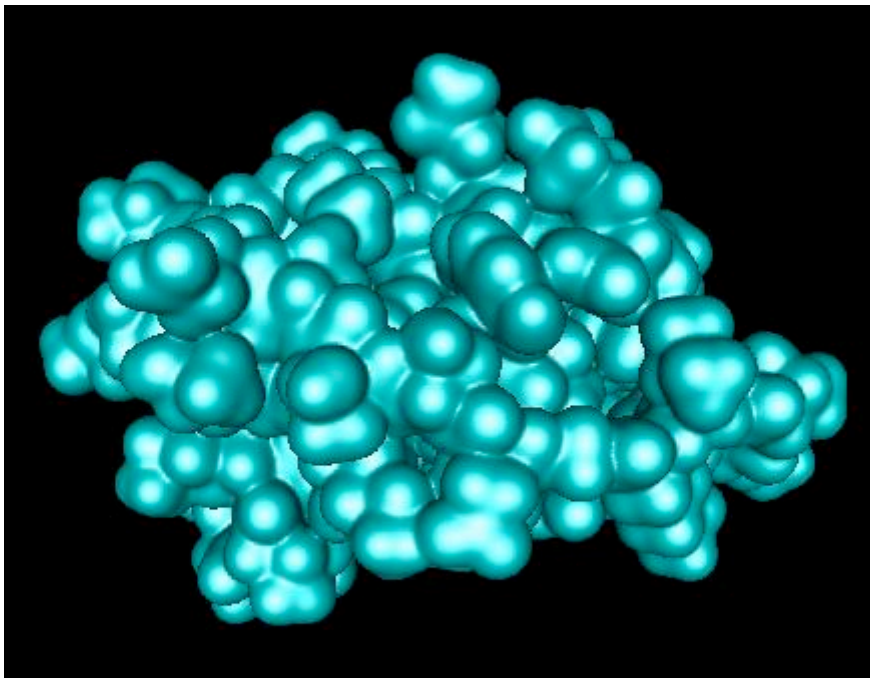


Figure captions

Fig. 1
Model molecules used to build up the ELMOs database used by the DENPOL program.
The atoms used to define the ELMOs are red coloured. The localization scheme is specified in Fig 1S.

Fig. 2

ELMOs used to describe the lateral chain of the Tyrosine amino acid. The ELMOs which describe the -OH, the phenyl ring and the -CH2 groups are respectively blue, red and yellow coloured.

Fig. 3
A typical variation observed for the torsion angle of the Phenylalanine amino acid during a molecular dynamics simulation.
The values of the torsion angle are reported during a 1ns molecular simulation of the β-hairpin polypeptide, one of the systems used in this paper to test the DENPOL program (see Table 2).
The simulation was carried out using the AMBER package with AMBER(2003) force field.

Fig. 4
Peptide bond between the i-th and the i+1 th residues.

Fig. 5
Cartoon diagrams of the polypeptides used to test the DENPOL program: Aurein1.2 (5a), [K^7]-IsCT (5b), Cecropin A (5c), β-hairpin (5d), 2GB1 (5e). The list of the residues for each polypeptide is reported in Table 2.

Fig. 6
Electron density isosurfaces (0.002 a.u) of the 2GB1 protein computed using the Hartree-Fock (top) and the DENPOL (bottom) approaches.

Fig. 1S
Localization schemes used for the model molecules considered to build up the DENPOL database. Each closed curve represents the atoms that are used to define a functional group. Each of them is described by an appropriate number of doubly occupied ELMOs to take into account the number of electrons which constitute the functional group.