Internet Electronic Conference of Molecular Design **2003**, November 23 – December 6
BioChem Press http://www.biochempress.com

Renormalized Protein Folding **of Cancerogenesis Proteins**

Vincenzo Villani[*], Ciro Leonardo Pierri and Alfonso Cascone
Dipartimento di Chimica dell'Università della Basilicata
Via N. Sauro, 85
85100 Potenza (Italy)
villani@unibas.it

Internet Electronic Conference of Molecular Design 2003, November 23 – December 6

Abstract

Motivation. Genomics and proteomics of cancer are well established: more than 100 *proto-oncogenes* can become *oncogenes* due to an alteration acquired by means of mutagen cancerogen agents. The expressed *oncoprotein* is qualitatively different with respect to the original *proto-oncoprotein* and induces the cellular transformation.

This process can be related to inactivated *oncosuppressors*: more than 30 *oncosuppressors* are known. When the genome mutation expresses altered oncosuppressors, the apoptotic process is inhibited, the oncoprotein concentration becomes high and the cellular growth and proliferation uncontrolled.

In general, mutant proteins with functional aberrations show structural and dynamical alterations. In this framework, the study of the protein folding process and of folded proteins of cancerogenesis plays a crucial role.

In this work we have considered the *α-Motif* domain of the *p73* oncosuppressor which presents a 57-aminoacid structure characterized by five *α*-helices and the *Sh3-Sh2* domain of the fusion oncoprotein *Bcr-Abl* with 163 residues in a complex structure.

Method. The protein folding process has been simulated through a well-constructed calculation procedure, which uses, in succession, the software PRIMARIA, PROTEO, HP-PDB, HYPERCHEM 7.5, AMBER 7 and VMD 1.8.2.

The developed *Renormalized Protein Folding* procedure starts with the macromolecule HP model, at a *low resolution* or at an enough large scale to neglect the chain molecular details. For such a simplified system it is possible to perform an efficient *global optimization* through the PROTEO *Monte Carlo-Simulated Annealing*. Then, through a scale transformation by means of HP-PDB, we come back to the *high-resolution* model with the correct molecular details in the found optimal structure. Such a structure is relaxed and refined in the subsequent steps through *local optimization* techniques*.* The goodness of the result is estimated by the *fitting* degree between the simulated structures and the experimental one, minimizing the *rmsd* (*root mean square deviation*) of the Cartesian coordinate sets. The final structure represents an ideal starting point to Molecular Dynamics Simulations of the folded protein.

Results. The folding of the *α-Motif* domain of the p73 oncosuppressor has been successfully simulated. Nevertheless, even in the case of the complex tertiary structure of the *Sh3-Sh2* domain, the procedure has given satisfactory results. The adopted renormalized procedure, which allows the transformation from the high to the small scale, worked satisfactorily: the structures generated by the simplified models have been essentially retained and relaxed during the subsequent step of *Molecular Modeling*. The fitting of the produced structures with the native ones gave *rmsd* < 7 or 15 Å for the *soft* or *hard* target, respectively, improving the performance of similar *ab initio* methods.

Conclusions. The *Renormalized Protein Folding* procedures join successfully the *global optimization* of simplified models, renormalization and *local optimization* and Molecular Dynamics of elaborate models of the resulting structures. Although the protein global conformation is well reproduced (according to the values of the gyration radius) the simulated structures are poor of secondary structures. The final optimization and Molecular dynamics step in aqueous solution of the produced structures, although do not substantially modify the picture, seem to induce the formation of the missing secondary structures.

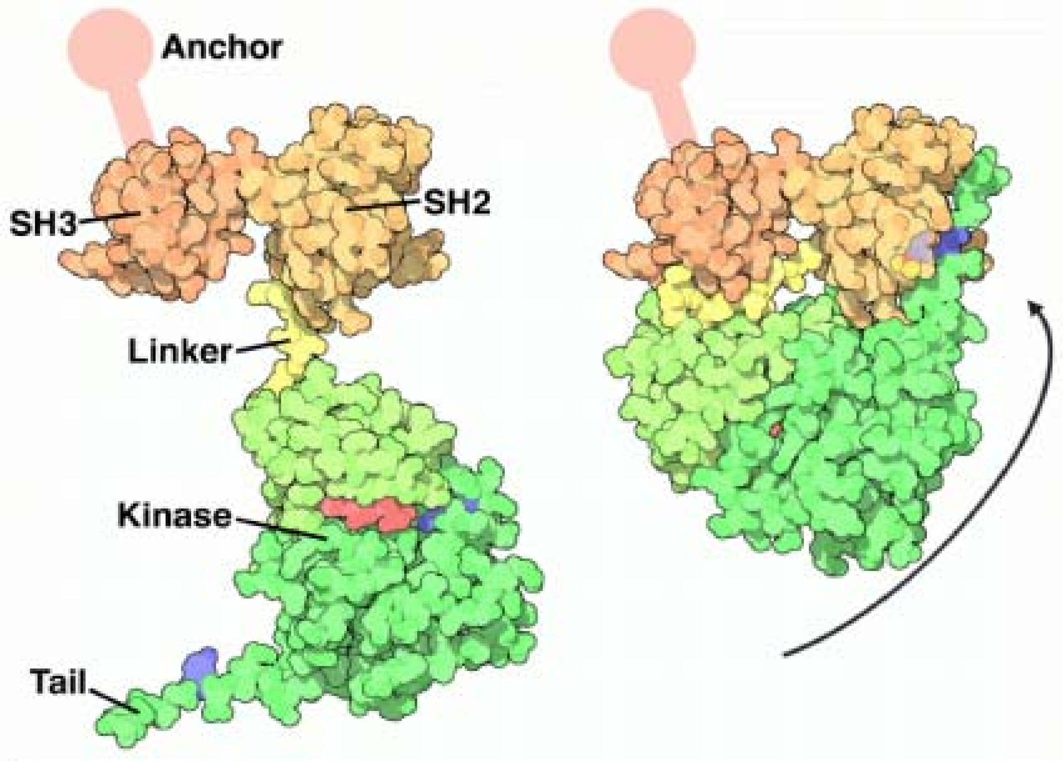
Availability. http://amber.scripps.edu/, http://www.hyper.com/, http://www.ks.uiuc.edu/Research/vmd/

Keywords. Simulated Annealing; Monte Carlo; Molecular Dynamics, HP Models, Self-Avoiding Walk, Protein Folding; Oncoproteins; Oncosuppressor.

Abbreviations and notations	
p73, *p73 α-Motif* Domain	Abl, *Bcr-Abl Sh3-Sh2* Domain
RPF, *Renormalized Protein Folding*	Rmsd, *root mean square deviation*

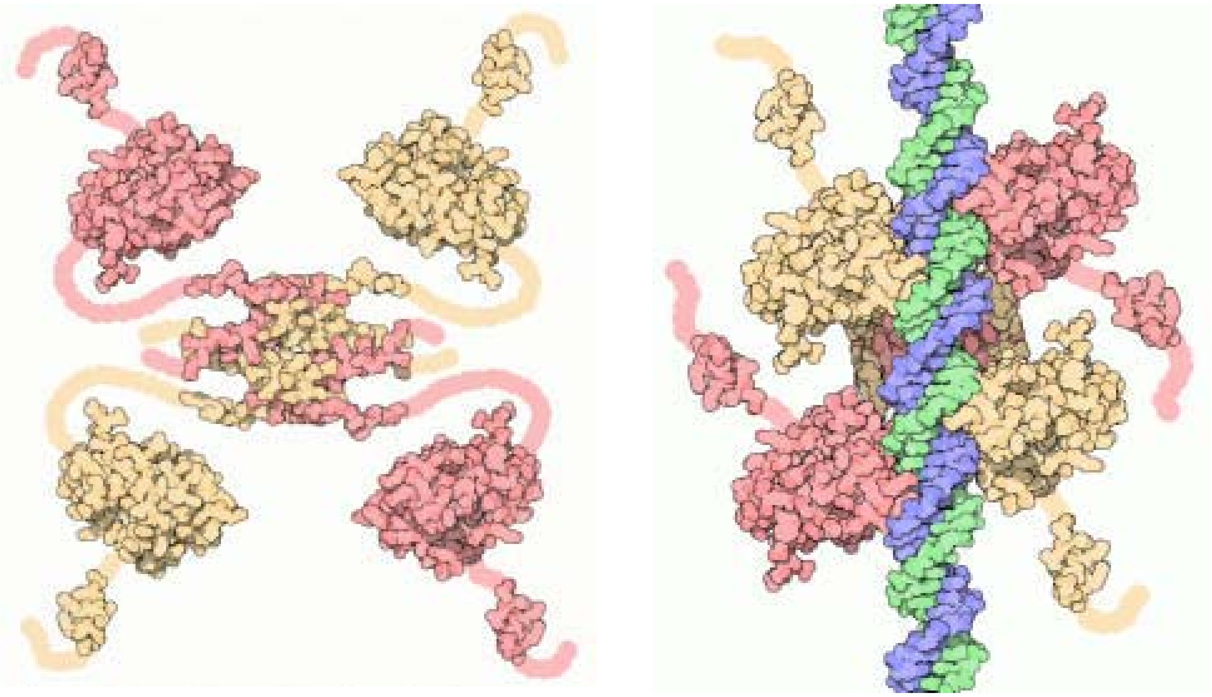
1 INTRODUCTION

Oncoproteins and Oncosuppressors is subject of an extensive Molecular Modeling research. For example, the complex activation-disactivation *Src* mechanism has been described. The inactive form is tightly folded on itself, while, the active form is extended and makes accessible the kinase active site. A large motion, through a flexible sequence, interconverts the two conformations. In the mutant protein the only accessible conformation is the activated one.



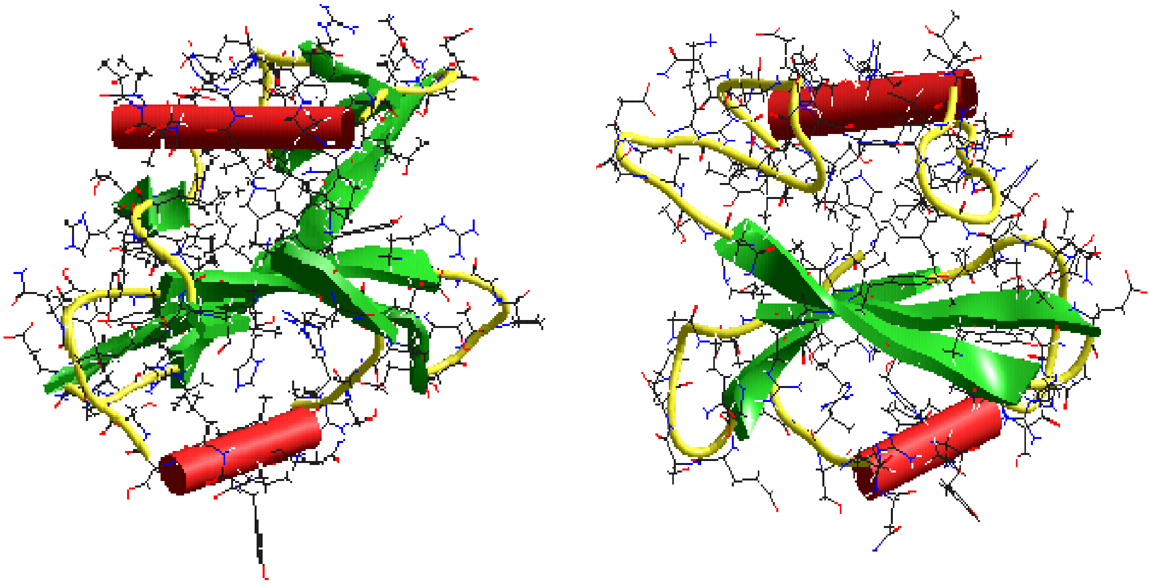
(from http://www.rcsb.org/pdb/molecules/pdb43_1.html)

The *p53* oncosuppressor structure-function relation has been in-depth studied. It is a flexible protein with the central tetramerization domain which holds together the four identical proteic arms with the transactivation domain complexing DNA, inhibiting its expression. In the mutant form the chelating activity is inactivated.



(from http://www.rcsb.org/pdb/molecules/pdb31_1.html)

Schematizing, in the whole the oncoprotein activation and the oncosuppressor inactivation makes the transforming cellule *a car with the accelerator always pushed and the brake broken!*
We have verified that the *Sh2* domain presents significant conformational changes on passing from the *Abl* proto-oncoprotein (pdb id 1AB2, left), to the *Bcr-Abl* fusion oncoprotein (2ABL, right), due to allosteric long-range interactions, although the *Sh2* primary structure is unvaried.



In such a framework the simulations of the *protein folding* for the cancerogenesis proteins plays a crucial role in individuating the system structural and dynamical features, especially in those cases (and they are the most) where the tertiary structure is unknown.

In this work we have considered the *α-Motif* domain of the p73 oncosuppressor which represents a *soft target* with the 57 amonoacid sequence characterized only by five α-helices and the *Sh3-Sh2* domain of the *Bcr-Abl* fusion oncoprotein which represents a *hard target* with 163 residues, organized in a complex structure.

2 METHODS and SOFTWARE

The protein folding process was simulated through an elaborate calculation procedure which uses, in succession, the software PRIMARIA, PROTEO, HP-PDB, HYPERCHEM 7.5, AMBER 7 and VMD 1.8.2. The flow-chart in **Fig. 1** shows the algorithm of the procedure, called *Renormalized Protein Folding, RPF*.

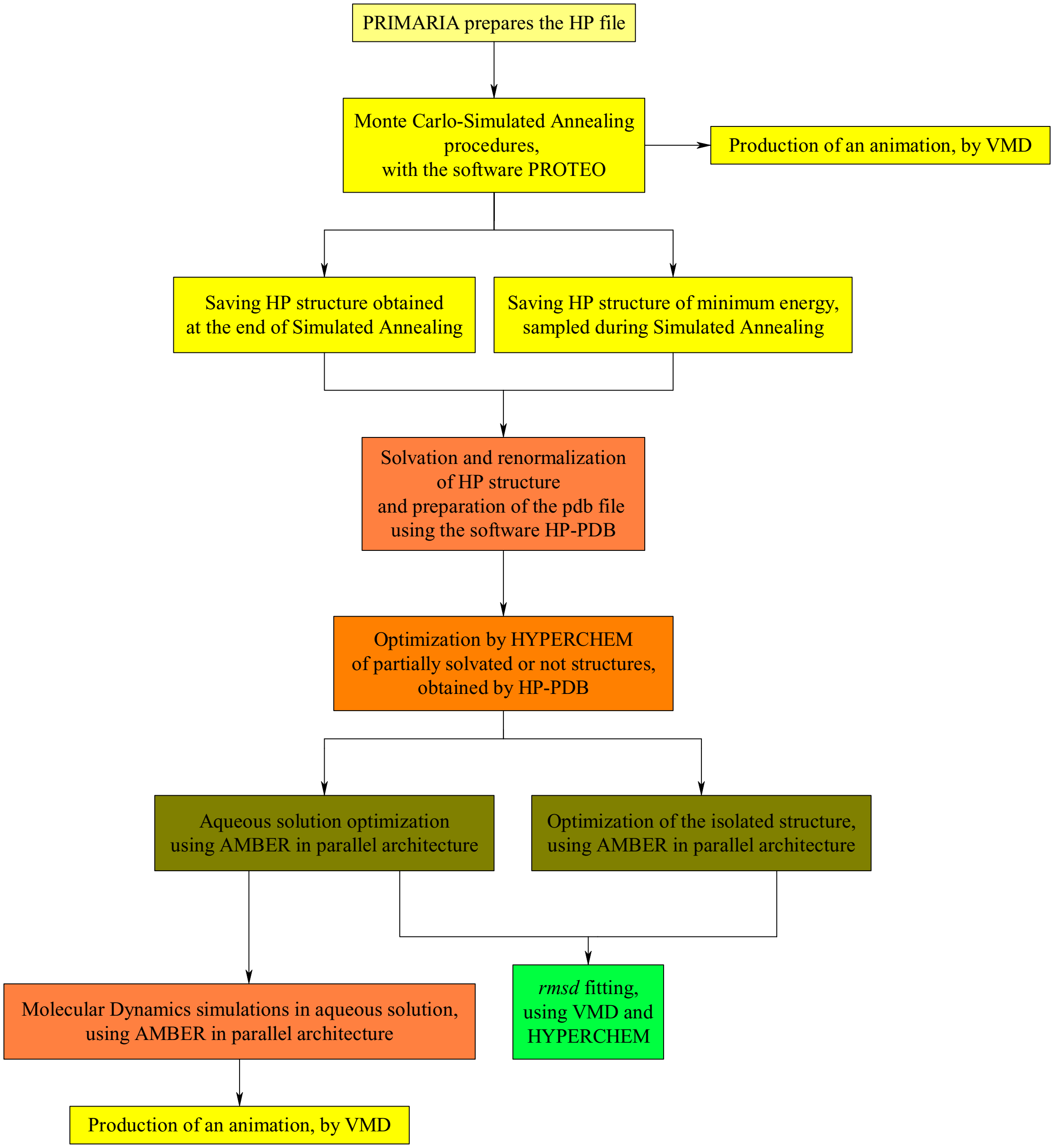

PRIMARIA prepares the HP file
Monte Carlo-Simulated Annealing procedures, with the software PROTEO
Production of an animation, by VMD
Saving HP structure obtained at the end of Simulated Annealing
Saving HP structure of minimum energy, sampled during Simulated Annealing
Solvation and renormalization of HP structure and preparation of the pdb file using the software HP-PDB
Optimization by HYPERCHEM of partially solvated or not structures, obtained by HP-PDB
Aqueous solution optimization using AMBER in parallel architecture
Optimization of the isolated structure, using AMBER in parallel architecture
Molecular Dynamics simulations in aqueous solution, using AMBER in parallel architecture
rmsd fitting, using VMD and HYPERCHEM
Production of an animation, by VMD

Fig. 1 *Renormalized Protein Folding (RPF)* flow chart.

In such a procedure, one starts from the *low-resolution* macromolecular model, i.e. from an enough observation scale to be able to neglect the system molecular details. For such a simplified system it is possible to perform an efficient *global optimization* through the *Monte Carlo-Simulated Annealing* procedure. Then, through a scale transformation, one comes back to the *high-resolution* model, which represents correctly the molecular details and keeps the found optimal structure. Such a structure is refined in the subsequent steps through *local optimization* techniques. Lastly, the goodness of the result is estimated from the degree of *fitting* between the calculated structure and the experimental one.

According to Wolynes, the protein folding is comparable to a phase transition. These problems are characterized by a multiplicity of scale lengths i.e. in correspondence of the critical point a large spectrum of correlations does exist, from short-range (among close residues) to long-range (among far residues). The developed strategy is analogue, in sense, to the transformation, called *renormalization group,* proposed by Wilson and Kadanoff in the study of the phase transitions, therefore *Renormalized Protein Folding*. In this sense, the approach is not reductionistic-like in that one starts from a system global representation and comes back to a local one.

Let us examine in detail the procedure. PRIMARIA transforms the primary structure of the *pdb* file of a protein of *Protein Data Bank* in the HP binary sequence used by PROTEO. The software uses the lipophilicity values of the aminoacid residues tabulated by Kyte and Doolittle and assigns the H (*Hydrophobic*) or the P (*Polar*) code, in correspondence of positive or negative values, respectively. Nevertheless, for close to 0 values (as in the glycine case, 0.4) it is possible to in autonomous way, for the H or P code.

PROTEO, previously developed by Villani and Cascone, is a program which simulates the protein folding according to the Dill HP binary representation, by means of a simulated annealing technique. The protein chain is modeled on a periodic lattice as Self-Avoiding Walk of H or P residues-particles with nearest-neighbor interactions: only nonbonded H residues, which are in adjacent sites, contribute with an energy value (contact energy ε_0) to the system free energy, which considers the solvent effects. The simulated annealing is performed through a succession of Monte Carlo calculations at lower and lower temperatures, starting from an enough high value. In PROTEO, the protein chain is gradually constructed by adding a portion with a residue at a time, as the temperature is lowered. Critical parameters of the procedure are the cooling rate and the chain growth rate, determined by *chain grown temperature* T_c at which the chain construction is terminated. The flow-chart in **Fig. 2** shows the PROTEO algorithm.

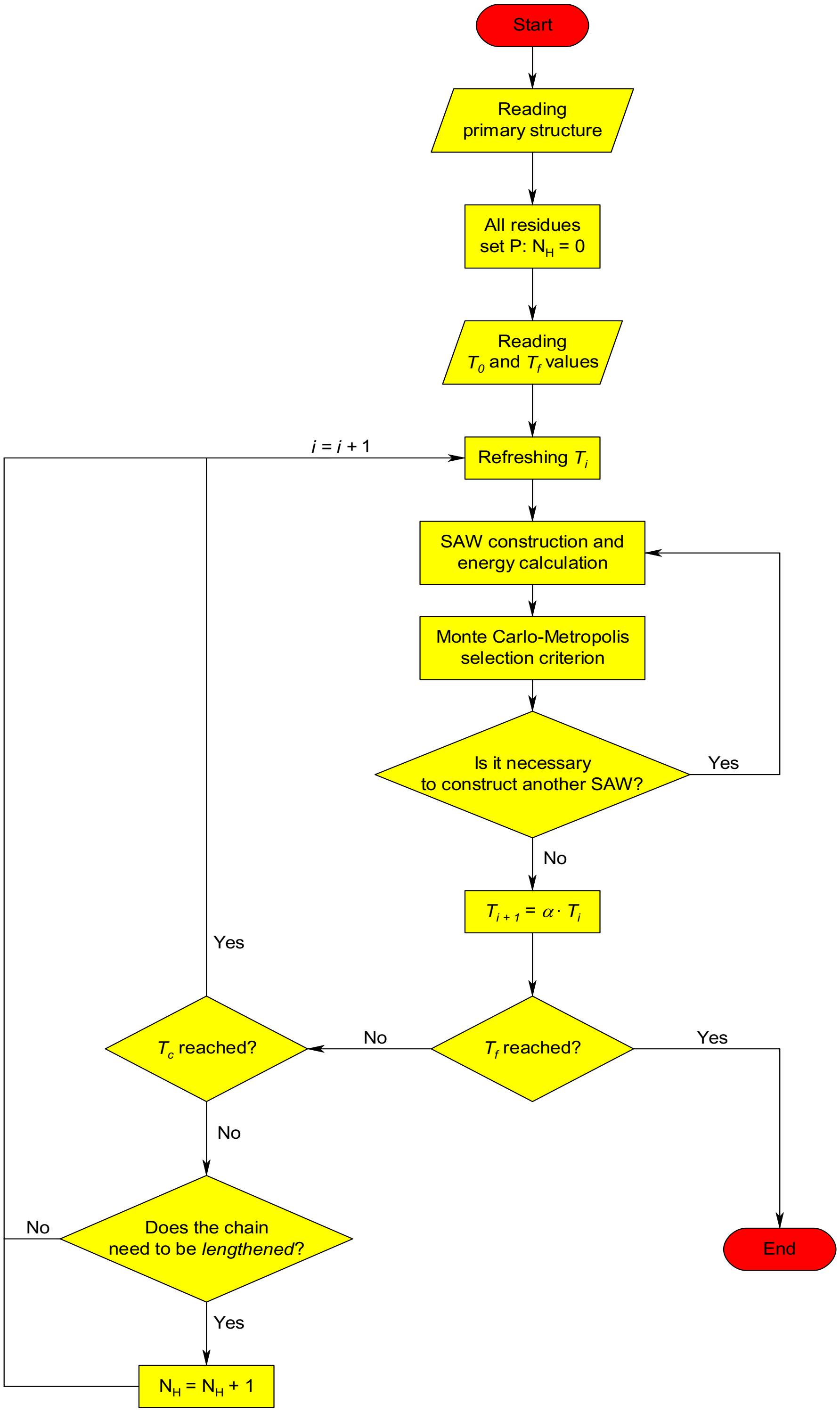

Start
Reading primary structure
All residues set P: N_H = 0
Reading T_0 and T_f values
i = i + 1
Refreshing T_i
SAW construction and energy calculation
Monte Carlo-Metropolis selection criterion
Is it necessary to construct another SAW?
Yes
No
T_{i+1} = α · T_i
T_f reached?
Yes
No
T_c reached?
Yes
No
Does the chain need to be lengthened?
No
Yes
N_H = N_H + 1
End

Fig. 2 PROTEO flow chart. T_0 and T_f are the starting and the final temperatures of the *annealing* procedure, T_c the *chain-grown temperature*, i.e. the temperature at which the chain construction is terminated and N_H the number of the H residues in the growing chain.

At the end of each run, PROTEO individuates an optimal structure, which we will indicate as *hp*, and the sampled one at the lowest energy, that we will indicate as *minhp*, that can be different with respect to the final one.

The HP models, though simplified, represent the protein in aqueous solution. In fact, the lattice sites non-occupied by the macromolecule are implicitly considered as occupied by water. Then, the interface software HP-PDB, adds to the protein structure obtained by PROTEO, solvation water shells in an explicit way, preparing the structure for the subsequent step of molecular modeling. Each hydration shell is constructed occupying the empty sites adjacent to the residues (for the first shell) or to the waters (for the following shells), through water molecules represented at this step by the oxygen atom alone. Moreover, the structure is enlarged by multiplying the system coordinates for a suitable scale factor, in order that the the introduction of the molecular details do not cause prohibited overlaps and the water molecules are not located at too large distances. This is the renormalization phase of the molecular system. In particular, the lattice is dimensioned in A and a scale factor *3x* is resulted as a good compromise in order to the distance between adjacent sites take into account of both the dimension of the peptide unit (two following C_α in a polypeptide chain are distant about 3.85 Å) and the coordination of the hydration shells (oxygens or oxygen-nitrogen bonded by hydrogen-bond are distant about 3 Å).

To substitute the HP particle with the subsequent complete residue specified by the primary structure, HP-PDB uses a reference *pdb* file of the considered protein, and builds a new *pdb* file of the protein in the conformation determined by PROTEO. First of all, the C_α of each aminoacid are re-placed in the corresponding HP chain sites. Then, the remaining residue atoms (excluding hydrogens) are submitted to the same translation, following the corresponding C_α. This procedure is performed calculating the differences Δq between the coordinates q_{PDB} of C_α in the reference *pdb* file and the q_{HP} ones of the corresponding residue in the PROTEO file. Then, the new coordinates Q_{PDB} are obtained subtracting such a difference to the q_{PDB} coordinates of each atom in the considered residue:

$$\Delta q = q_{\mathrm{PDB}} - q_{\mathrm{HP}}$$

$$Q_{\mathrm{PDB}} = q_{\mathrm{PDB}} - \Delta q$$

The obtained structure is hydrogenated and optimized in HYPERCHEM. The HP-PDB algorithm is reported in **Fig. 3**.

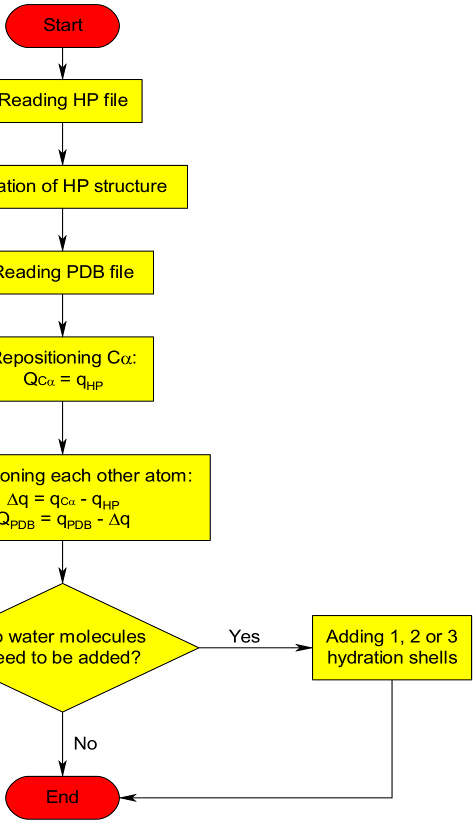

Start
Reading HP file
Dilatation of HP structure
Reading PDB file
Repositioning Cα:
Qcα = q_HP
Repositioning each other atom:
Δq = qcα - q_HP
Q_PDB = q_PDB - Δq
Do water molecules need to be added?
Yes
Adding 1, 2 or 3 hydration shells
No
End

Fig. 3 HP-PDB flow chart

Structures with- or without hydration shells have been optimized through the *Amber99 force field*. The final structures, that we will call as partially hydrated, deprived of the solvation shells, have been compared with the reference crystallographic structure. The match has been quantified in HYPERCHEM and VMD through a *fitting* procedure of the two structures, minimizing the *rmsd* (*root mean square deviation*) of the Cartesian coordinate $\boldsymbol{Q}$ and $\boldsymbol{Q^0}$ sets, either considering all the system atoms

$$rmsd=\left[\sum_{1}^{3N}\left(\mathrm{Q_i}-\mathrm{Q^0{}_i}\right)^2/3N\right]^{1/2}$$

or limiting to those of the main chain.

The structure which showed the best agreement was fully optimized in AMBER 7. A final structure was obtained at a severe convergence level (the energy gradient *rms* lower than 10^{-4} kcalmol^{-1} Å^{-1}, using more than 50,000 optimization cycles). Moreover, the protein in dilute aqueous solution using thousands water molecules. The system was optimized, applying the boundary periodic conditions. Also in this case, the fitting with the crystallographic structure and the optimal *rmsd* calculated.

The procedure was carried out on Pentium 4 PCs running under Windows NT and XP and in serial and parallel architecture on the IBM *Power 3* and *Power 4* CASPUR machines (borneo.caspur.it and man.caspur.it). The NAG or IMSL libraries for the generation of the random numbers were used. Graphics and animations were elaborated by HYPERCHEM and VMD.

3 RESULTS AND DISCUSSION

The *folding* of *p73* oncosuppressor

As a model for our analysis of oncosuppressors protein folding, we have chosen the p73 α-*Motif* domain (which we will indicate as p73 for the sake of brevity) due to the simplicity and peculiarity of its native structure characterized solely by five α–helices. This kind of protein is particularly suitable to the *folding* modeling studies either because it possesses only a kind of secondary structure or because α-helices are those forming faster during the folding. Moreover, p73 has a small length primary structure with 80 residues, of which we will consider a 57 residue partial structure (from VI to LXII residue) according to the known crystallographic structure, codified as 1DXS in the *Protein Data Bank* and reported in **Fig. 4**.

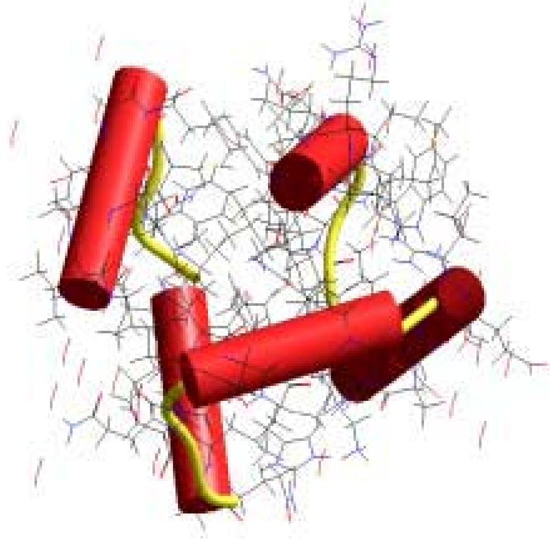


Fig. 4 p73 structure (1DXS) in which the 5 α–helices are evidenced, elaborated by VMD.

According to Dill's HP model used in PROTEO, the chain is codified as a binary sequence H (1) and P (0). Nevertheless, since glycine has a practically amphiphilic, *borderline* between H and P, character, and being the code binary, in our calculations it has been alternatively considered either H or P. In one case the H residues are 33, in the other 26.

As previously, PROTEO produces correctly micellar, compact and globular structures, with a polar surface and a hydrophobic core: these are fundamental requirements of HP models. As said, the p73 known crystallographic structure concerns 57 out of the 80 total residues. Nevertheless, in PROTEO the whole primary structure has been considered and only in the subsequent step the partial structure of interest has been extracted. In such a way possible spurious effects of the chain end portions have been eliminated.

Through PROTEO, fourteen accurate simulations have been carried out, in which the calculation conditions have been suitably varied. In particular, in all the conduced simulated anneal*ing* N_i=3,219 temperature changes have been performed, starting from the initial temperature T_i=1000 K up to the final one T_f=200 K according to the iterative schedule:

$$T_{n+1} = \alpha T_n \qquad \text{with } \alpha = 0.9995$$

As said, a critical parameter of the PROTEO strategy is the temperature T_c at which the protein chain growth is complete. In the performed simulations T_c has been varied from 400 to 850 K that correspond to slow or fast growth, respectively.

At each temperature a long Monte Carlo calculation is carried out, in which a succession of N_{MC} conformations is generated. The final outcome and the computing time depend upon N_{MC} that has been varied from 50,000 to 100,000.

In PROTEO the system free energy (proportional to the contact number N_c), the *end-to-end* square distance and the gyration radius, that have been sampled in an equidistributed way, have been estimated. For the gyration radius R_g only the C_α have been considered to use the theorem of Lagrange referred to particles with equal mass and that requires only the interatomic distances r_{ij}:

$$R_g = \left[\ (n+1)^{-2} \sum_i \sum_{j>i} {r_{ij}}^2 \ \right]^{1/2}$$

Then, either the final structure or the minimum energy structure sampled during the simulations have been used in the renormalization step.

For the optimization step through HYPERCHEM, the solvated structure with two hydration water shells has been prepared. In all cases the fitting of the calculated structures with the experimental one has been performed and the *rmsd* calculated either referred only to the backbone or to the whole protein structure, including the side groups.

Lastly, in AMBER the lowest obtained *rmsd* partially optimized structure has been wholly solvated in a thousand water box, applying the boundary periodic conditions and fully optimized. The *Amber99 force field* has been used. Also in these cases the *rmsd* with respect to the experimental structure has been calculated of the final structures.

The obtained results are reported in **Table I** and **Figs. 5-9**. As it can be inferred from **Table I** more than half simulations that were performed gave reasonably low *rmsd*, lower than 10 Å, and the best result gave the values 6.7-8.0 Å, in agreement with the optimal literature results for p73 comparable sized proteins. It is worthy noting that the size group fitting does not alter significantly the result.

In our *protein folding* method information of the native known structure are not used, as the secondary structures, the disulfur bonds and the gyration radius. Although an improvement of the prediction is in a continuos progress, the obtainment of the correct folding is still very difficult especially using *ab initio* methods. A *rmsd* value of about 6 Å for C_α between the simulated structure and the experimental one is judged as *target* value in the case of small proteins.

As expected, the best fitting is obtained referring only to the backbone and the results are slightly worse including the side groups, which in our strategy are explicited only in the optimization step.

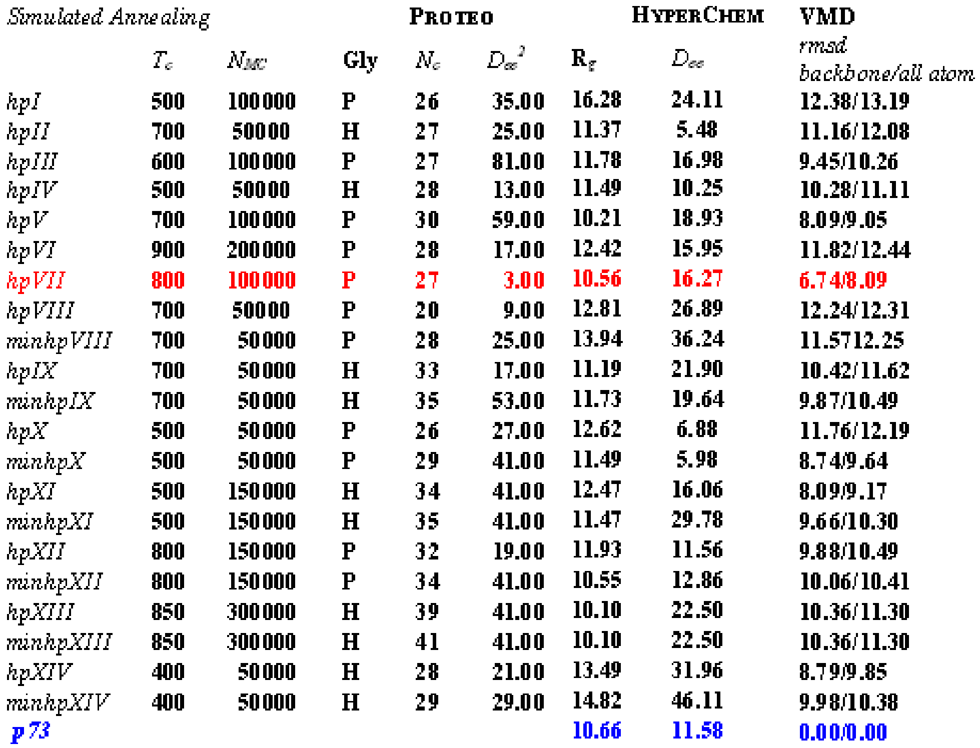
From the predictive viewpoint, the best fittings are associated to the lower values of the gyration radius, as shown **Fig. 10**. The fulfillment of this condition, in the absence of the known native structure, could be used as a necessary requirement in the selection of the more likely tertiary structure.

In all cases the fluctuating, damped trend and the final convergence of D_{ee}^2 and N_c as a function of the annealing iteration N_i confirms the correctness of the simulated annealing.

We observe that the renormalization step is correctly performed: the reticular dilated structure does not show evident overlapping between the side groups, and the solvation at this stage seems efficient. Nevertheless, in the subsequent partial optimization step, while the protein chain efficiently relaxes, often assuming a similar shape with respect to the native one, the solvation shells do not always occupy the space in an optimal way.

The plots in which the calculated structures are overlapped in an optimal way with the native one, confirm, in the best cases, the quality of the obtained result. The global protein conformation is well reproduced. The conflicting aspects are associated to the scarcity of the calculated secondary structures (α-helices) and to the end portion orientation.

It is interesting to note that the final optimization step refines the HP structures, keeping their essential features. This suggests the importance to improve the PROTEO performance including explicitly the driving force of the protein chain to form a-helices.



Simulated Annealing				**PROTEO**		**HYPERCHEM**		**VMD**
	T_c	N_{MC}	**Gly**	N_c	D_{ee}^2	**R$_g$**	D_{ee}	*rmsd* *backbone/all atom*
hpI	**500**	**100000**	**P**	**26**	**35.00**	**16.28**	**24.11**	**12.38/13.19**
hpII	**700**	**50000**	**H**	**27**	**25.00**	**11.37**	**5.48**	**11.16/12.08**
hpIII	**600**	**100000**	**P**	**27**	**81.00**	**11.78**	**16.98**	**9.45/10.26**
hpIV	**500**	**50000**	**H**	**28**	**13.00**	**11.49**	**10.25**	**10.28/11.11**
hpV	**700**	**100000**	**P**	**30**	**59.00**	**10.21**	**18.93**	**8.09/9.05**
hpVI	**900**	**200000**	**P**	**28**	**17.00**	**12.42**	**15.95**	**11.82/12.44**
hpVII	**800**	**100000**	**P**	**27**	**3.00**	**10.56**	**16.27**	**6.74/8.09**
hpVIII	**700**	**50000**	**P**	**20**	**9.00**	**12.81**	**26.89**	**12.24/12.31**
minhpVIII	**700**	**50000**	**P**	**28**	**25.00**	**13.94**	**36.24**	**11.5712.25**
hpIX	**700**	**50000**	**H**	**33**	**17.00**	**11.19**	**21.90**	**10.42/11.62**
minhpIX	**700**	**50000**	**H**	**35**	**53.00**	**11.73**	**19.64**	**9.87/10.49**
hpX	**500**	**50000**	**P**	**26**	**27.00**	**12.62**	**6.88**	**11.76/12.19**
minhpX	**500**	**50000**	**P**	**29**	**41.00**	**11.49**	**5.98**	**8.74/9.64**
hpXI	**500**	**150000**	**H**	**34**	**41.00**	**12.47**	**16.06**	**8.09/9.17**
minhpXI	**500**	**150000**	**H**	**35**	**41.00**	**11.47**	**29.78**	**9.66/10.30**
hpXII	**800**	**150000**	**P**	**32**	**19.00**	**11.93**	**11.56**	**9.88/10.49**
minhpXII	**800**	**150000**	**P**	**34**	**41.00**	**10.55**	**12.86**	**10.06/10.41**
hpXIII	**850**	**300000**	**H**	**39**	**41.00**	**10.10**	**22.50**	**10.36/11.30**
minhpXIII	**850**	**300000**	**H**	**41**	**41.00**	**10.10**	**22.50**	**10.36/11.30**
hpXIV	**400**	**50000**	**H**	**28**	**21.00**	**13.49**	**31.96**	**8.79/9.85**
minhpXIV	**400**	**50000**	**H**	**29**	**29.00**	**14.82**	**46.11**	**9.98/10.38**
p73						**10.66**	**11.58**	**0.00/0.00**

Table I The *Annealing* results are summarized I-XIV for *p73*. The distances and the *rmsd* of the *fitting* with the experimental *p73* structure are expressed in Å. The temperature T_c, the number of Monte Carlo conformations and the glycine hydrophobic or polar character is varied in the calculations. The reported results are referred to the partially optimized structures in HYPERCHEM in the presence of two water molecule solvation shells added by HP-PDB.

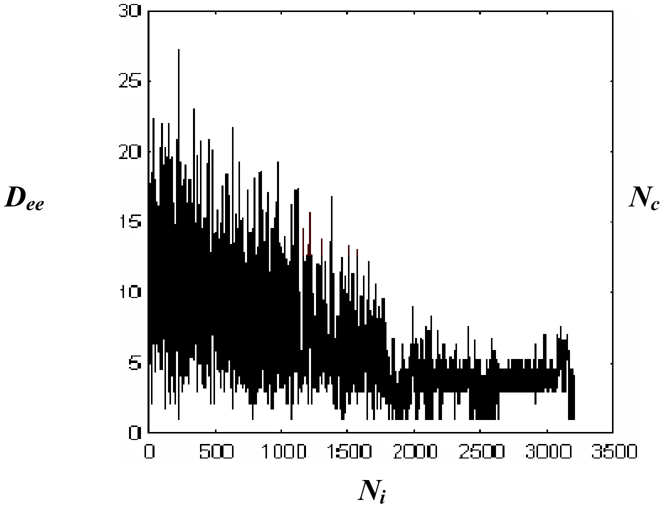


Fig. 5a The *end-to-end* distance *Dee* as function of the *annealing* iteration N_i for *hp VII*

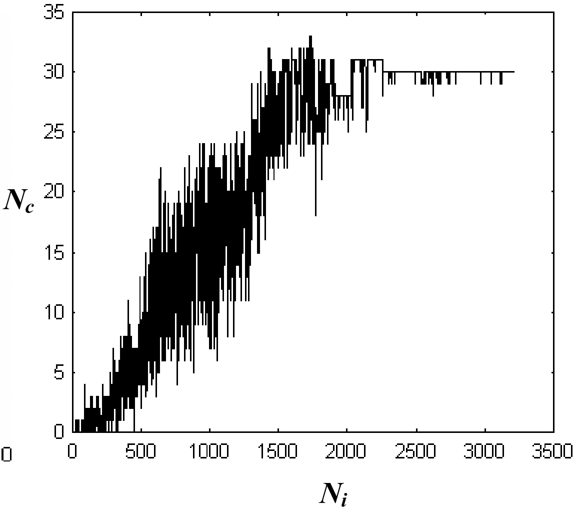


Fig. 5b The contact number N_c as a function of the *annealing* iteration N_i for *hp VII.*

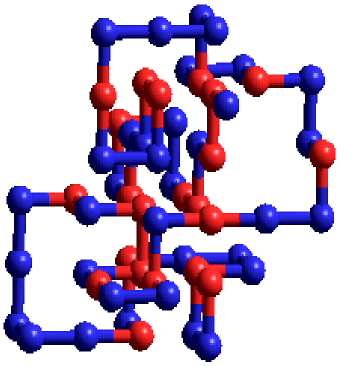


Fig. 6 The P73 *hp VII* structure with glycine as a polar residue (blue).

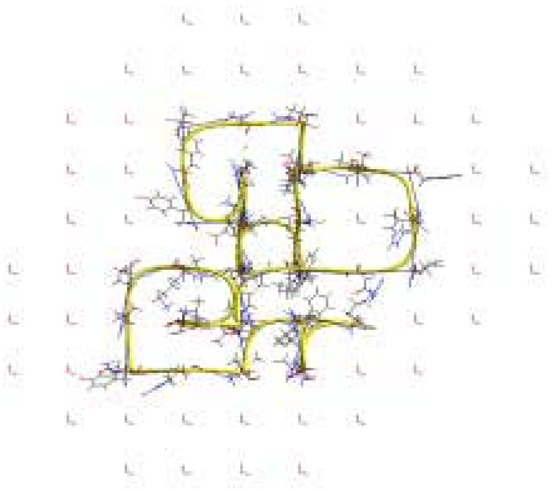


Fig. 7 The *hp VII* renormalized structure of Fig. 6. Two solvation water molecule shells are included.

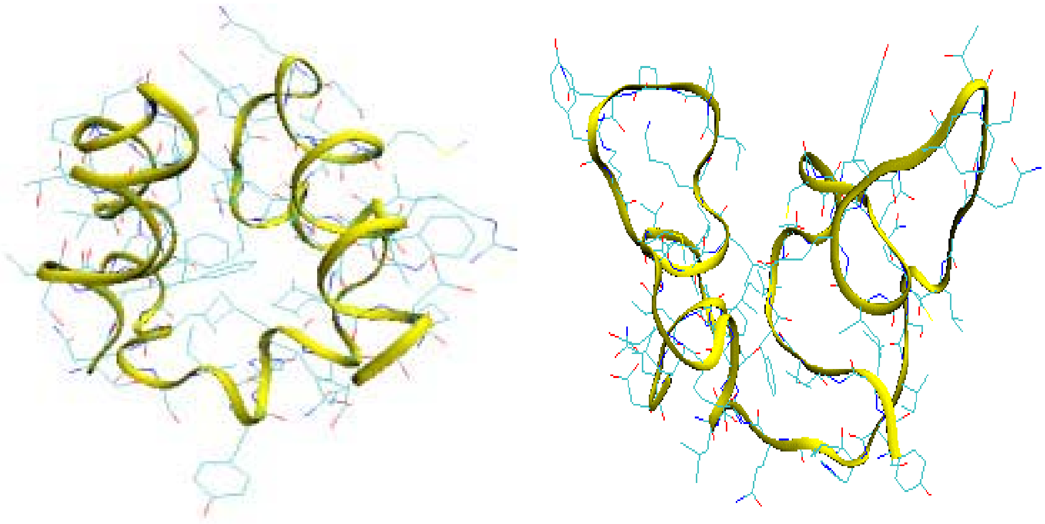


Fig. 8 The optimized *hp VII* structure obtained starting from the renormalized structure of Fig. 7 (right) compared with the corresponding experimental structure (left).

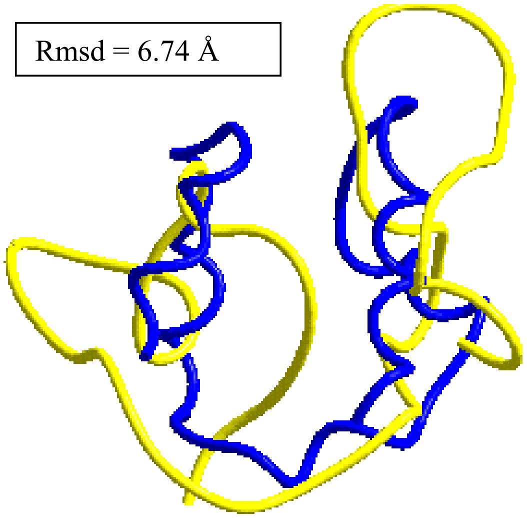


Fig. 9 *Fitting* of the *hp VII* calculated structure (yellow) with the p73 cristallographic structure (blue).

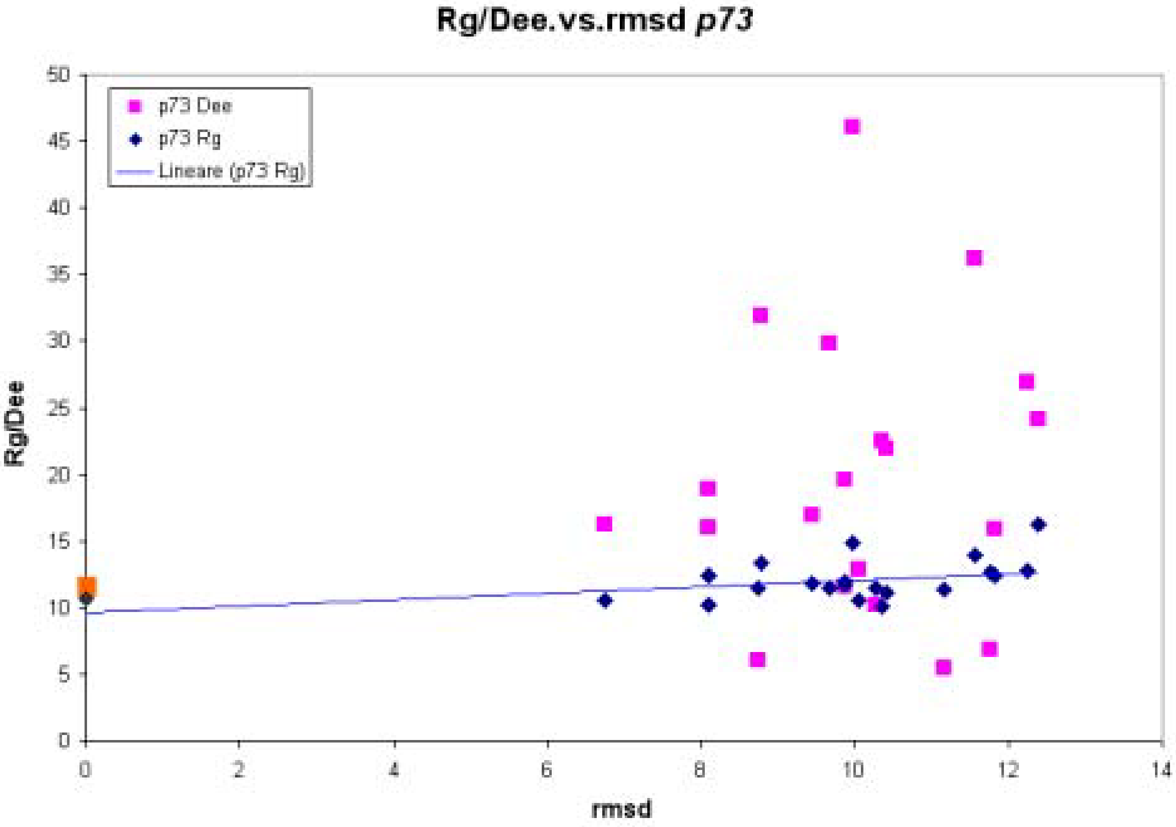


Fig. 10 R_g and D_{ee} vs. *rmsd*, from the data of **Table I**. The fitting is performed for R_g values.

The structure corresponding to the best case, hp VII, has been fully optimized through AMBER in solution (**Fig. 11**) using a box of water molecules and applying the boundary periodic conditions:

In solution optimized structure	*Rmsd (backbone/all atoms)*
In solution p73	6,94/9,07 Å

Then, the heavy optimization in solution does not significantly change the fitting with the experimental structure with respect to the simple partial optimizations (**Fig. 12**). The local optimization, although essential, does not seem to be a crucial step of the procedure. In contrast, the final result is strongly dependent upon the HP structure obtained by PROTEO, renormalized by HP-PDB and lastly relaxed to the corresponding minimum energy conformation.

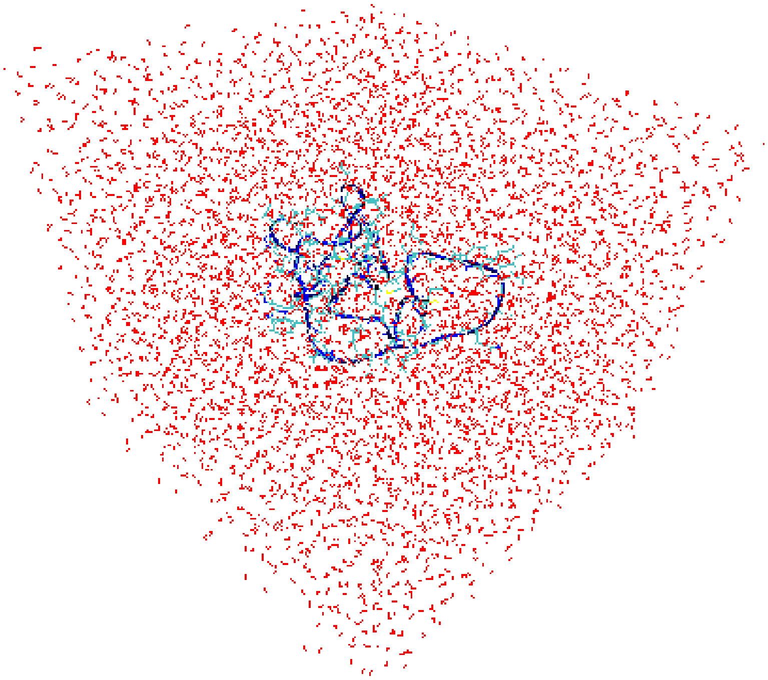


Fig. 11 p73 optimized by AMBER in a box of about 3500 water molecules applying the periodic boundary conditions starting from *hp VII*.

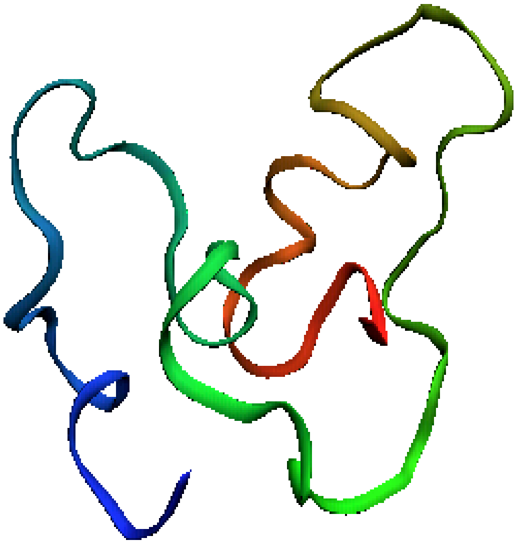


Fig. 12a *p73* structure of Fig. 11 neglecting the water molecules.

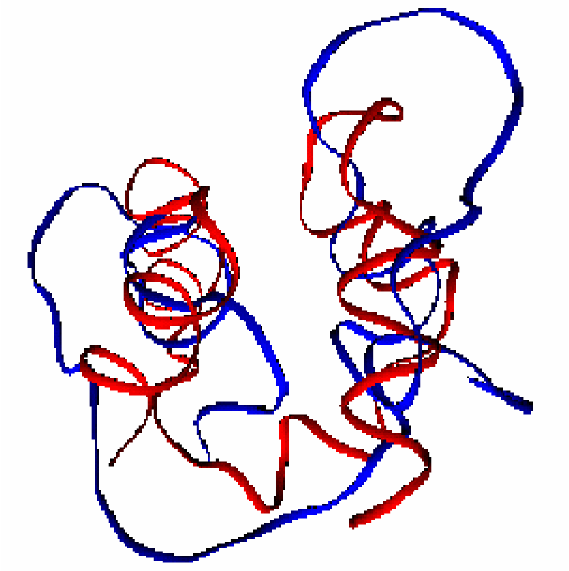


Fig. 12b *p73* fitting of Fig. 12 with the experimental structure (red)

The outcome of the *RFP* strategy can represent an ideal starting point for Molecular Dynamics Simulations of folded proteins. So, a preliminary short but significant aqueous solution simulations of 100 ps was performed starting from the optimized structure of Fig. 11.

Abl oncoprotein *folding*

In the oncoprotein structural analysis, we have chosen the *Sh3-Sh2* domain of the human fusion protein tyrosine-kinase *Bcr-Abl* (2ABL in PDB, which for the sake of brevity we will call as *Abl*) with 163 residues and a complex tertiary structure characterized by α-helices and β-sheets. The CASP meetings indicate that the absolute accuracy in all the *ab initio* methods is still low in comparison to the resolution of the experimental structure, estimating, for the C_α, *rmsd*> 10 A. In such a way the of the strategy of the *Renormalized Protein Folding* can be severely tested. The *Abl* tertiary structure is reported in **Fig. 13**.

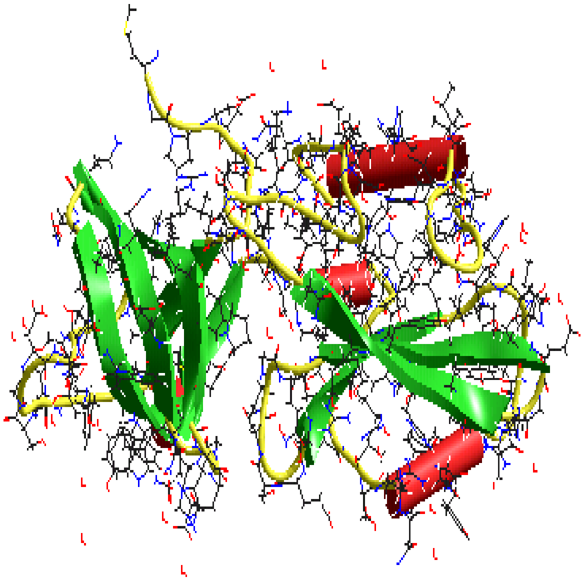


Fig. 13 Abl crystallographic structure (2ABL) where the 3 α–helices, the β-sheets are evidenced.

As for *p73* glycine has been alternatively considered H or P. In the first case the H residues are 47, in the second 61. As previously, PROTEO produces correctly micellar, compact and globular structures. The used calculation conditions are similar to those used previously, although in this case the required computing resources are more severe. 9 accurate simulated annealing have been performed, in which N_i=3,219 Monte Carlo Simulations at lower and lower temperature have been conduced. In one case, the annealing IX, a longer simulations of one order magnitude has been carried out, with N_i = 32,184 iterations and α= 0.99995.

The critical parameters T_c and N_{MC} have been accurately varied. We observe that, higher the temperature T_c, larger the number of generated N_{MC} and more accurate and heavier the simulations. As we will see, it is significant that the better folding are obtained in the more severe calculation conditions.

The contact number N_c, the end-to-end distance D_{ee} and the gyration R_g have been used as the system descriptors. The *hp* and *minhp* structures have been produced and renormalized by HP-PDB.

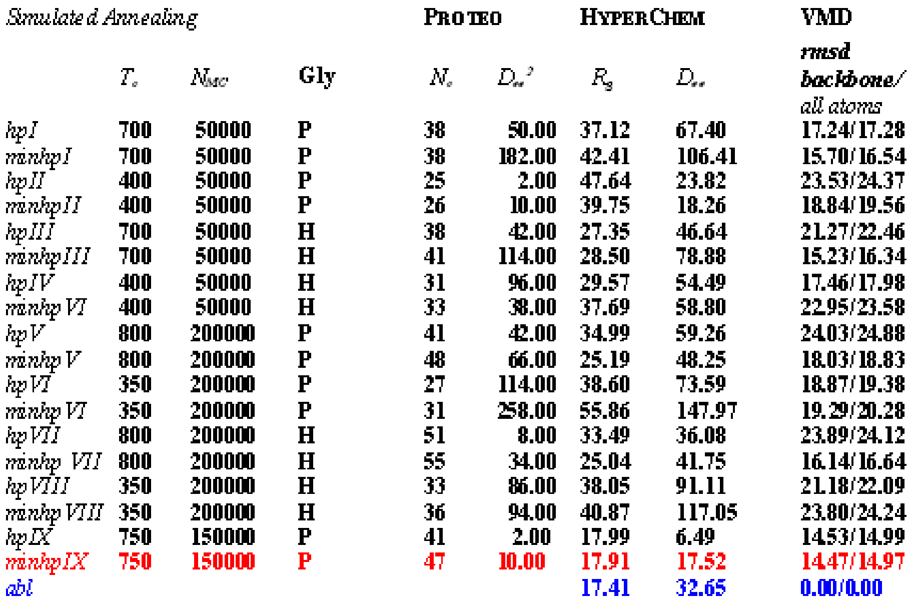
For the optimization step through HYPERCHEM, the structure solvated by a shell of hydration waters has been prepared. The fitting of the calculated structure with the experimental one has been performed and *rmsd* calculated as accuracy indices. The best case has been refined by AMBER.

As it can be inferred from **Table II** and **Figs. 14-18** practically in all cases we obtain reasonably low, lower than 20 Å, *rmsd*, and the best result gave the value 14.5 Å, which is in agreement with the optimal *ab initio* results known in the literature for *hard targets* with comparable size having a *rmsd* for C_α>10 Å.

The D_{ee} and N_c plots as a function of the sampled $N_s = N_i * N_{MC}/2000$ show the correctness and the convergence of the simulated annealing. It is possible to observe the fluctuations towards the minimum energy *minhp* structures, which often show a better agreement than the final annealing structures. This suggests that improving the calculation conditions, as the sampling efficiency towards the low temperatures, it would be possible to obtain even more significant final structures.

It is interesting to observe that the optimal HP structures keep, according to the meaning of the developed strategy, their global shape through the renormalization step and the subsequent relaxation. In fact, the final optimization step is local-like and must improve without altering the previous coarse-grained global optimization result.

The water molecule coordination in the partial optimization step, seems to be open to improvement. Nevertheless, this problem is solved in the final step of optimization in aqueous solution.



Simulated Annealing				PROTEO		HYPERCHEM		VMD
	T_c	N_{MC}	**Gly**	N_c	D_{ee}^2	R_g	D_{ee}	***rmsd backbone/*** *all atoms*
hpI	**700**	**50000**	**P**	**38**	**50.00**	**37.12**	**67.40**	**17.24/17.28**
minhpI	**700**	**50000**	**P**	**38**	**182.00**	**42.41**	**106.41**	**15.70/16.54**
hpII	**400**	**50000**	**P**	**25**	**2.00**	**47.64**	**23.82**	**23.53/24.37**
minhpII	**400**	**50000**	**P**	**26**	**10.00**	**39.75**	**18.26**	**18.84/19.56**
hpIII	**700**	**50000**	**H**	**38**	**42.00**	**27.35**	**46.64**	**21.27/22.46**
minhpIII	**700**	**50000**	**H**	**41**	**114.00**	**28.50**	**78.88**	**15.23/16.34**
hpIV	**400**	**50000**	**H**	**31**	**96.00**	**29.57**	**54.49**	**17.46/17.98**
minhp VI	**400**	**50000**	**H**	**33**	**38.00**	**37.69**	**58.80**	**22.95/23.58**
hpV	**800**	**200000**	**P**	**41**	**42.00**	**34.99**	**59.26**	**24.03/24.88**
minhp V	**800**	**200000**	**P**	**48**	**66.00**	**25.19**	**48.25**	**18.03/18.83**
hpVI	**350**	**200000**	**P**	**27**	**114.00**	**38.60**	**73.59**	**18.87/19.38**
minhp VI	**350**	**200000**	**P**	**31**	**258.00**	**55.86**	**147.97**	**19.29/20.28**
hpVII	**800**	**200000**	**H**	**51**	**8.00**	**33.49**	**36.08**	**23.89/24.12**
minhp VII	**800**	**200000**	**H**	**55**	**34.00**	**25.04**	**41.75**	**16.14/16.64**
hpVIII	**350**	**200000**	**H**	**33**	**86.00**	**38.05**	**91.11**	**21.18/22.09**
minhp VIII	**350**	**200000**	**H**	**36**	**94.00**	**40.87**	**117.05**	**23.80/24.24**
hpIX	**750**	**150000**	**P**	**41**	**2.00**	**17.99**	**6.49**	**14.53/14.99**
minhpIX	**750**	**150000**	**P**	**47**	**10.00**	**17.91**	**17.52**	**14.47/14.97**
abl						**17.41**	**32.65**	**0.00/0.00**

Table II The *Abl Annealing* I-IX results are summarized. The distances and the *rmsd* of the *fitting* with the experimental 2ABL experimental structure, are expressed in Å. The temperature T_c, the Monte Carlo conformation number and the glycine hydrophobic or polar character are varied in the calculations. The reported results are referred to partially optimized structures in HYPERCHEM in the presence of two solvation water molecule shells added by HP-PDB.

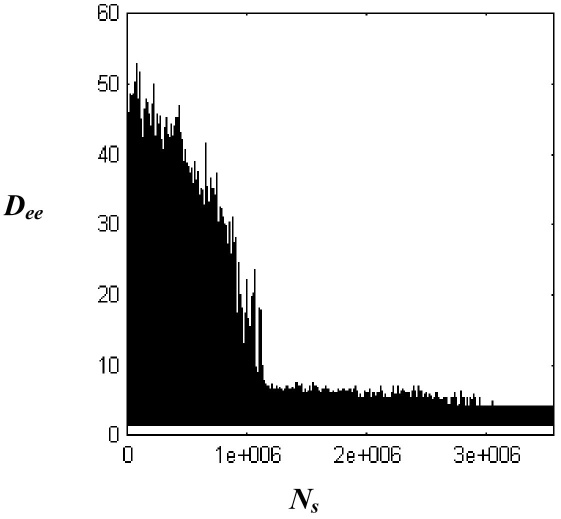


Fig. 14a The end-to-end distance D_{ee} as a function of the sampled conformations $Ns= N_i N_{MC}/2000$ for *hp IX*.

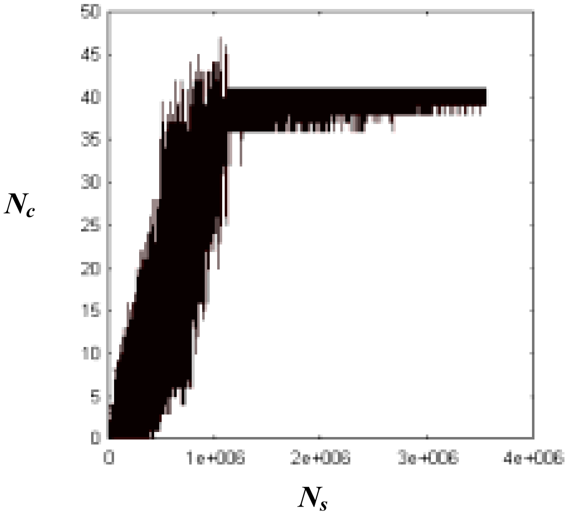


Fig. 14b The contact number N_c as a function of the sampled conformations *Ns* for *hp IX*.

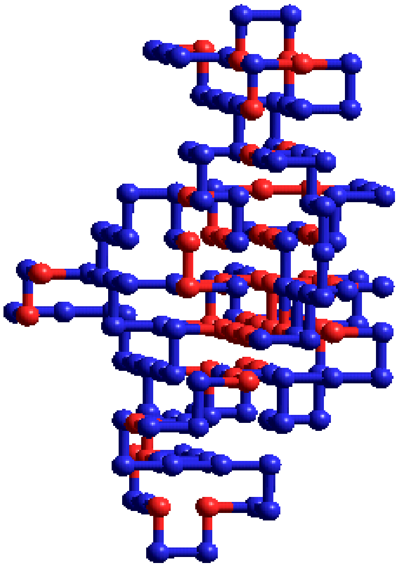


Fig. 15 The *Abl hp IX* structure with 163 residues, with glycine as polar residue (blue).

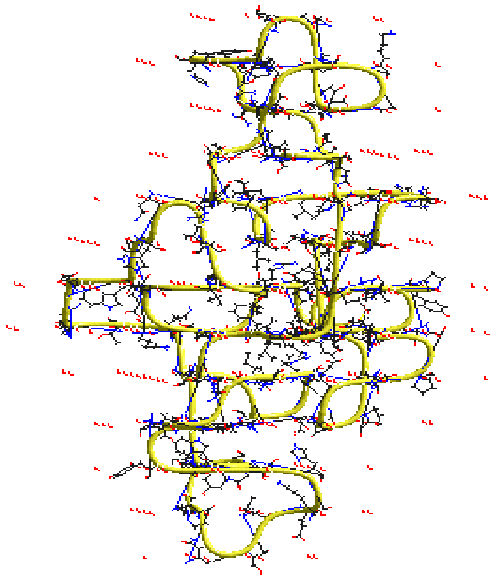


Fig. 16 The *hp IX* renormalized structure obtained starting from the HP structure of Fig. 15. A solvation water molecule shell is included.

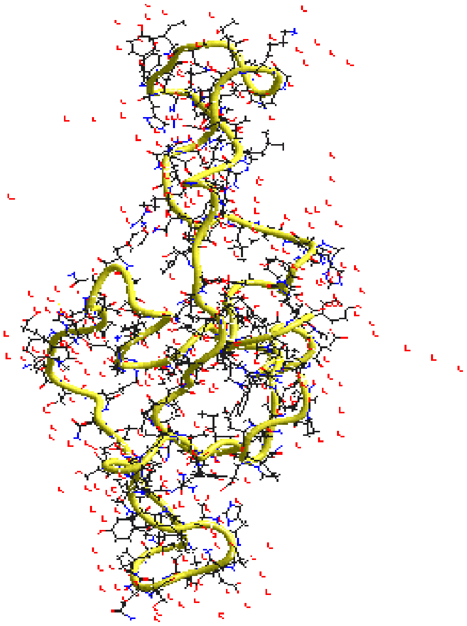


Fig. 17a The *hp IX* partially optimized structure obtained starting from the renormalized structure of Fig. 16.

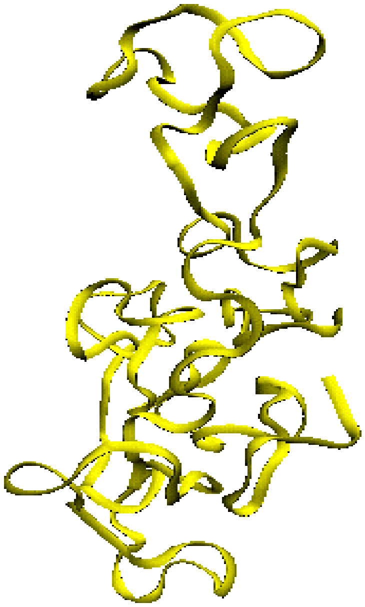


Fig. 17b The partially optimized structure of Fig. 17a in the coil representation, neglecting the water molecules.

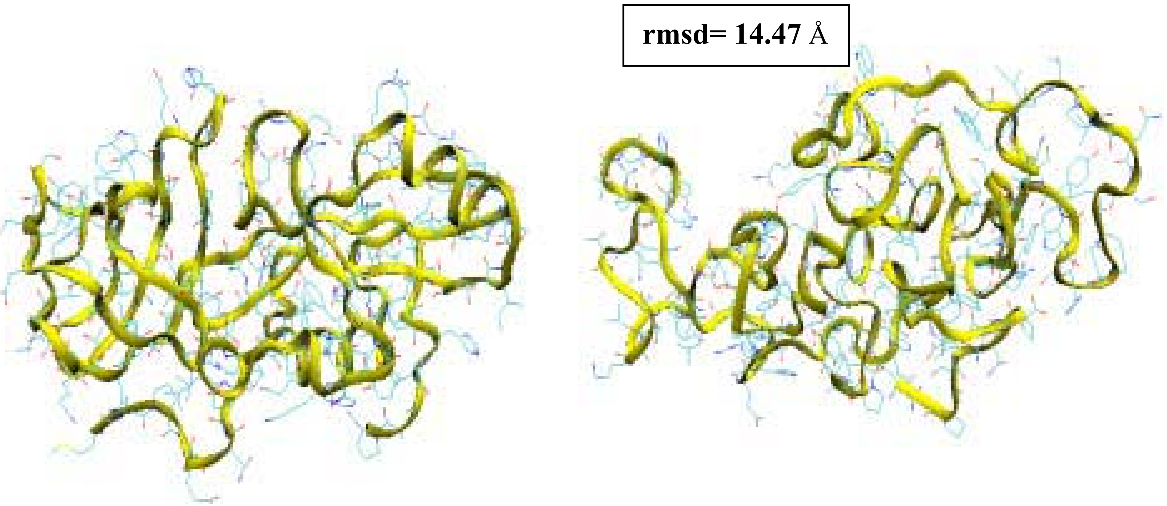


Fig. 18 The *Abl* optimized *hp IX* structure obtained starting from the renormalized structure of Fig. 16 (right) compared with the corresponding experimental structure (left).

As previously, although a good agreement is obtained, the result is affected by the scarcity of secondary structures. This behavior depends upon the difficulty of the HP models of developing secondary structures. In fact, the hydrophobic *driving force*, at the basis of such models, which generates best packing micellar structures, it is not sufficient to make the secondary structures, deriving, in contrast, from the contribution of specific forces of hydrogen bond (and disulfur bond), which are not taken into account by the method. Therefore it needs that a suitable strategy is developed, to induce the formation of the secondary structures in the HP models: embryo structures would be subsequently refined in the renormalization and relaxation steps.

Lastly, the fully solvated structure corresponding to the best case, *hp IX*, has been optimized in aqueous solution through AMBER using a water molecule box and applying the boundary periodic conditions (**Fig. 19**):

Fully solvated structure	*Rmsd (backbone/all atoms)*
Abl *in solution*	15,16/15,61 Å

Then, the heavy full optimizations do not improve the fitting with the experimental structure (**Fig. 20, 21**), with respect to the simple partial optimizations. The relaxation in solution, although determines a globally similar structure with respect to the experimental one, it is not able to evidence the correct secondary structures. Again, the final result is strongly affected by the HP structure obtained by PROTEO and renormalized through HP-PDB and induces to further develop those techniques.

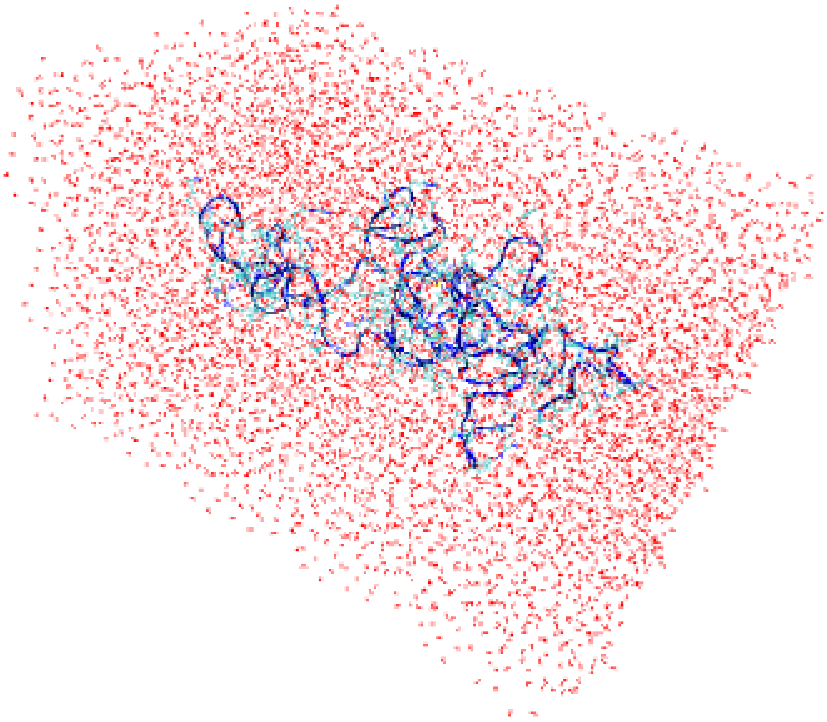


Fig. 19 *Abl* optimized through AMBER in a box of about 7500 water molecules, applying the boundary periodic conditions, starting from *hp VII* in the *ribbon* and *sticks* representation.

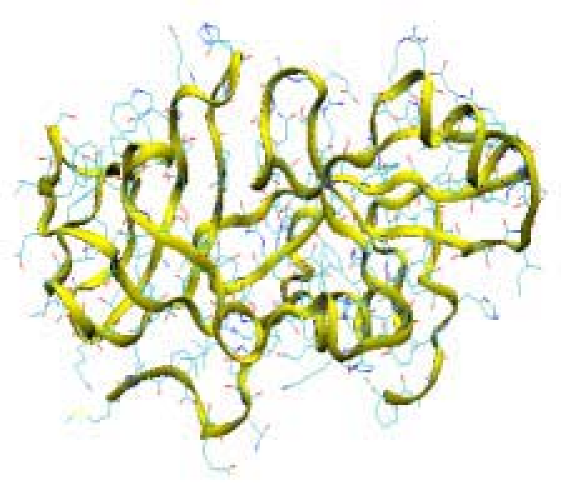


Fig. 20 The *Abl* structure of Fig. 19 neglecting the water molecules.

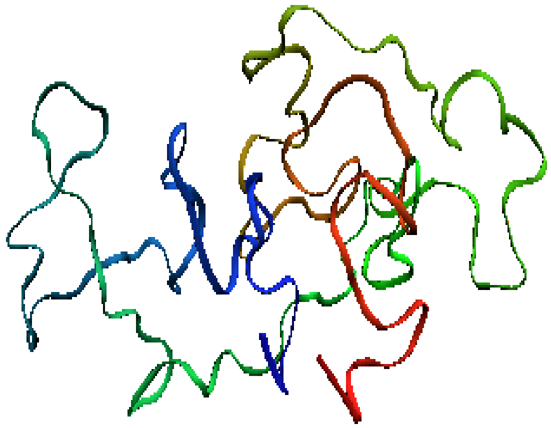


Fig. 21 *Abl fitting* of Fig. 20 with the experimental structure (red).

Also for *Abl*, a preliminary Molecular Dynamics Simulations of 50 ps was performed.

4 CONCLUSIONS

The Nobel Prize for Chemistry has been assigned in 2004 to Ciechanover, Hershko e Rose in understanding the ubiquitin-mediated proteolysis. This small 76 aminoacid protein, complexing to structurally altered enzymes, allows their degradation via the ubiquitin-proteasome pathway, performing a key role in the cellular regulation and, in the opposite case, in the cancerogenesis.

Today the cancer macromolecular basis is well documented and the action mechanism of the involved proteins begins to outline. In general, the mutations in the genes that codify for oncoproteins and oncosuppressors express mutant proteins with structural-dynamic aberrations and functional alterations. In this framework, the study of *protein folding* process and behavior of cancerogenesis folded proteins plays a crucial role.

Unfortunately the *protein folding* problem is still far from having found a definitive solution. A number of approachs do exist and even the biennal CASP meetings (http://predictioncenter.llnl.gov/) in which different predictive methods compete on always new protein sets. The elaborated methods go from the maximum empirism, the tertiary structure is obtained by the comparison with known structure and similar sequence proteins with respect to the protein of interest, to the maximum theory, considering *ab initio* the system physical-chemistry. The second genetic codex problem is not dissimilar from the cryptographic analysis problem: from the coded message (the primary structure) we have to arrive the key to obtain the clear message (the tertiary structure).

The developed procedure, combining the global optimization techniques of simplified systems and the local optimization ones of elaborated models, does not suffer of the typical limits of the Molecular Dynamics Simulations, that can simulate times that are shorter of many orders of magnitude than those typical of the phenomenology (0.1-1000 s), even using techniques of distributed calculation (1 μs) (folding@home).

The adopted renormalized procedure that allows the transformation from the large to the small scale works correctly. The structures generated from the simplified models have been essentially kept and relaxed during the following Molecular Modeling step. The final and heavy step of full optimization of the produced structures did not modify the picture emerged from the preliminary partial optimizations.

The Monte Carlo-Simulated Annealing of HP models produces the expected micellar structures, with polar surface and hydrophobic core, with a compact and globular shape.

The RPF strategy arrives at least similar performance to those obtainable through ab-initio approaches as far as the fitting with the native structure is concerned: *rmsd* of about 6 A for *p73* and about 14 A for *Abl*.

The agreement for the global structural parameters such as the gyration radius, between the simulated structure and the experimental one, is very good.

In perspective, the development is in progress of *modified HP models*, able to produce at a coarse grained level the secondary structures, that can be refined in the subsequent procedure steps, and long Molecular Dynamics Simulations, starting from the generated optimal structures.

Acknowledgments

The authors acknowledge the CASPUR (*La Sapienza*-Rome, Italy) for the use of the cluster of parallel computers and AMBER 7. A MIUR grant (COFIN 2003 *Studi sperimentali e teorico-computazionali di dinamiche reattive e non reattive in sistemi molecolari*) is gratefully acknowledged.

5 REFERENCES

V. Villani e A. Cascone, 'From Random Walks to Protein Folding Simulations' *Recent Research Developments in Polymer Science* **8**, 21 (2004)

National Cancer Institute, *http://www.nci.nih.gov/*

National Center for Biotechnology Information, *http://www.ncbi.nlm.nih. gov/*

D. W. Ross, *Introduction to Oncogenes and Molecular Cancer Medicine*, Springer-Verlag, 1998

B.Z. Lu, B. H. Wang, W. Z. Chen and C. X. Wang, 'A new computational approch for real protein folding prediction', *Protein Engieneering* **16**, 659-663 (2003)

T. Hunter, 'Oncoprotein Networks', *Cell* **88**, 333-346, 1997

G.A. Chassea, A.M. Rodriguezb, M.L. Maka, E. Dereteya, A. Perczelc, C.P. Sosad,
R.D. Enrizb, I.G. Csizmadiaa, 'Peptide and protein folding', *Journal of Molecular Structure (Theochem)* **537**, 319-361 (2001)

C. I. Branden, J. Tooze, *Introduction to Protein Folding*, Garland Publishing, 1999

A. M. Lesk, *Introduction to Protein Architecture: The structural Biology of Proteins*, Oxford University Press, 2001

R. Goldstain, Z. Luthey-Schulten and P.G. Wolynes *PNAS* **87**, 4918 (1992); **89**, 9029 (1992); **90**, 9949 (1993)

H. Frauenfelder e P. G. Wolynes, ‘Biomolecules: where the physics of complexity and simplicity meet’, *Physics Today*, 47-58 (1994)

R.B. Laughlin and D. Pines, ‘The Theory of Everything’, *PNAS* **97**, 28 (2000)

R.B. Laughlin, D. Pines, J. Schmalian, B. P. Stojkovic and P. Wolynes ‘The middle way’, *PNAS* **97**, 32 (2000)

Dill Research Group University of California, San Francisco, *http://www.dillgroup.ucsf.edu/*

H. S. Chan & K. A. Dill, ‘The protein folding problem’, *Physics Today*, 24 (1993)

K. F. Lau & K. A. Dill, ‘A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins’ *Macromolecules* **22**, 3986 (1989)

RCSB Protein Data Bank, *http://www.rcsb.org/pdb/*

Protein Structure Prediction Center, *http://predictioncenter.llnl.gov/*

Folding@home, http://folding.stanford.edu/

J. Kyte & R. F. Doolittle, ‘A simple method for displaying the hydropathic character of a protein’, *J. Mol. Biol.* **157**, 105 (1982)

E. H. M. Van Laarhoven & E. H. L. Aarts, *Simulated Annealing: Theory and applications*, Reidel Publishing, 1987

S. Kirkpatrick, C. D. Gelatt, Jr. & M. P. Vecchi, ‘Optimization by Simulated Annealing’ *Science* **220**, 671 (1983)

Biography

Vincenzo Villani is researcher in Macromolecular Chemistry at the University of Basilicata (Italy). He is professor of *Theory of Macromolecules* and group leader. Graduate at University of Naples, he is author of about 60 scientific publications. He has developed models of catalytic sites in the Ziegler-Natta isospecific polymerization. He has investigated of drug design at a semiempiric and *ab-initio* level. He has developed methods of conformational search and of simulated annealing of complex molecular systems. He has studied the dynamics of biomolecules in aqueous solution using the time series nonlinear analysis. He has proposed an entropic mechanism and a model morphology for the Protein Rubbers within the theory of complex systems. He works in protein folding and folded protein behavior in the framework of self-organized structures and cancerogenesis. He is involved in epistemology and in science diffusion.