J Mol Model (2012) 18:4503–4516
DOI 10.1007/s00894-012-1456-6

ORIGINAL PAPER

Stochastic simulation of structural properties of natively unfolded and denatured proteins

David Curcó · Catherine Michaux · Guillaume Roussel · Emmanuel Tinti · Eric A. Perpète · Carlos Alemán

Received: 14 January 2012 / Accepted: 2 May 2012 / Published online: 29 May 2012

Abstract A new simulation strategy based on a stochastic process has been developed and tested to study the structural properties of the unfolded state of proteins at the atomistic level. The procedure combines a generation algorithm to produce representative uncorrelated atomistic microstructures and an original relaxation method to minimize repulsive non-bonded interactions. Using this methodology, a set of 14 unfolded proteins, including seven natively unfolded proteins as well as seven "classical" proteins experimentally described in denaturation conditions, has been investigated. Comparisons between the calculated and available experimental values of several properties, at hydrodynamic and atomic level, used to describe the unfolded state, such as the radius of gyration, the maximum length, the hydrodynamic radius, the diffusion coefficient, the sedimentation coefficient, and the NMR chemical shifts, reflect a very good agreement. Furthermore, our results indicate that the relationship between the radius of gyration and the hydrodynamic radius deviates from the Zimm's theory of polymer dynamics for random coils, as was recently observed using single-molecule fluorescent methods. Simulations reveal that the interactions between atoms separated by three chemical bonds (1–4 interactions) play a crucial role in the generation process, suggesting that the unfolded state is essentially governed by bonding and short-range non-bonding interactions.

Keywords Denatured state · Hydrodynamic properties · NMR chemical shifts · Stochastic simulation · Uncorrelated atomistic microstructures · Unstructured proteins

D. Curcó
Departament d'Enginyeria Química, Facultat de Química,
Universitat de Barcelona,
Marí i Franques 1,
Barcelona 08028, Spain

C. Michaux (✉) · G. Roussel · E. Tinti · E. A. Perpète
Unité de Chimie Physique Théorique et Structurale,
University of Namur,
rue de Bruxelles, 61,
5000 Namur, Belgium
e-mail: catherine.michaux@fundp.ac.be

C. Alemán
Departament d'Enginyeria Química, E. T. S. d'Enginyeria
Industrial de Barcelona, Universitat Politècnica de Catalunya,
Diagonal 647,
08028 Barcelona, Spain

C. Alemán (✉)
Center for Research in Nano-Engineering,
Universitat Politècnica de Catalunya,
Campus Sud, Edifici C', C/Pasqual i Vila s/n,
Barcelona 08028, Spain
e-mail: carlos.aleman@upc.edu

Introduction

Solving the protein folding problem is one of the most challenging tasks in structural biology [1, 2] as for instance, folding intermediates are inaccessible to X-ray crystallography. The starting point of a folding reaction is the ensemble of structurally disordered molecules in the denatured state [3] where structure can trigger nucleation and initiate the protein folding. In this respect, the unfolded states, studied at structural resolution, can be the «Rosetta Stone» of the protein folding problem [4].

Furthermore, it is now clear that a significant part of eukaryotic genomes encode proteins with substantial regions of disordered structure [5]. Several types of biological functions have been ascribed to these so-called *natively unfolded proteins* (NUPs), such as molecular recognition,

signal transduction or neurodegenerative diseases [6–8]. Unstructured regions are also frequently found in regulatory and cancer associated proteins [9]. For a review of this particular family of proteins, see [10]. Intrinsically and partially disordered under physiological conditions, these particular proteins exist as dynamic ensembles in which the atomic positions significantly vary during time, with no specific equilibrium values. In general, their activity involves the binding to a partner ligand, and these interactions are typically associated with the induction of folding the disordered structures, therefore constituting an unusual structurating refolding process. The existence of these systems questions one of the landmarks in protein biology, claiming that the specific function of a protein is determined by its unique and rigid three-dimensional (3D) structure [11]. The study of this amorphous class of proteins remains therefore out of reach by classical structural biology because of their inherent conformational heterogeneity.

Experimental characterization of denatured states of proteins represents a difficult task, as it is necessary to disfavor the population of native states without adding denaturants [12]. In addition, the relatively flat energy surfaces associated with such a state makes modeling these systems quite problematic [13]. The recent review of Fisher et al. highlights both the strengths and disadvantages of the available strategies established to simulate these structures, *i.e.*, ensemble-restrained molecular dynamics (MD) simulations, ensemble construction using a predefined conformational library, etc. Other methods are based on statistical criteria rather than on energetic ones. For example, Sosnick and coworkers [14] reported a procedure to generate ensembles of unfolded structures using a self-avoiding statistical coil model that is based on backbone conformational frequencies in a coil library (*i.e.*, a subset of the Protein Data Bank). The flexible-meccano method uses a coil-library and a simple volume exclusion term for conformational sampling [15]. The backbone of the unfolded state conformation is generated by using a subset of the database of amino acid-specific backbone dihedral angles $\{\varphi,\psi\}$ propensities, which excludes all residues in α-helices and β-sheets. Residue i is connected to residue i+1 by selecting a random pair of φ- and ψ-angles, taking into account the nature of residue i, and the torsional subset database. The residual dipolar couplings (RDCs) simulated within this model present a reasonable agreement with the experimental couplings measured in both natively unfolded and chemically denatured proteins [16, 17].

Among the current theoretical methods, those based on quantum scale modeling, deserve special attention because of their very accurate predictions. However, as a consequent drawback, the size of the CPU-affordable systems is limited, and their applicability for studying the conformational properties of proteins remains completely impossible. These limitations are partially circumvented by molecular mechanics (MM) approaches, with MD and Monte Carlo (MC) being the most frequent. Although these simulation techniques have allowed investigating a number of properties and phenomena associated with proteins possessing well-defined 3D structures [18–20], they are not efficient enough to examine the conformational behavior of denatured proteins. Indeed, the complexity of the latter arises not only from their conformational disorder but also from the high number of flexible degrees of freedom accompanying their macromolecular nature. Whilst they have successfully described the conformational properties of small peptides [21, 22], neither MD and MC are appropriate to produce large sets of representative uncorrelated microstructures within reasonable computational resources because of the structural correlation imposed by their algorithms.

In this work, we propose a new simulation strategy that has been specifically designed to overcome such problems. Based on a procedure previously developed to study amorphous polymers in the solid state [23–25], it combines an algorithm generating representative atomistic microstructures and a relaxation stage minimizing non-bonded interactions that allows a large number of relaxed and uncorrelated atomistic microstructures to be recurrently obtained at very reasonable computational costs.

As classical proteins under denaturing conditions and NUPs share the same characteristics, we selected two 7-species sets from both families, whose various structural characteristics were available. The reliability and the performance of our generation-relaxation method will be granted if the properties computed for the generated conformations fit the experimental values. Two types of properties in solution substantially changing with the protein conformation, were chosen: at hydrodynamic and atomic-level. The first, like radius of gyration (R_g), hydrodynamic radius (R_H), diffusion (D) and sedimentation (s) coefficients, and the maximum lenght (L_{max}) are key factors to describe the molecular dimensions and compactness of the system. At the atomic-level, we computed NMR chemical shifts which are totally dependent on the conformation and the chemical environment, providing information about local conformational preferences.

The paper is organized as follows. First we present the original strategy developed in this work to study the conformational behavior of disordered proteins. Then, the proteins and their structural properties used to check the reliability of such theoretical strategy are described. After this, the structural parameters calculated using the microstructures produced for 14 unfolded proteins are compared to the available experimental data. Finally, we discuss the main features involved in this work and the conclusions of the work are outlined.

This newly established technique will help to give a detailed characterization of NUPs leading to the development of new drug design strategy. Indeed, the understanding and control of protein–protein interactions based on disorder-to-order transitions of one of the partners and involved in several normal and pathological biological processes should pave the way to several and essential therapeutic targets [26, 27].

Material and methods

Simulation strategy

The strategy, schematically illustrated by Fig. 1, uses the following inputs: (i) the primary structure of the protein; (ii) a library with the data needed for the 20 amino acids found in coded proteins; and (iii) the intervals defined by the user as allowed values for the backbone and side chain dihedral angles in each amino acid. The primary structure is introduced as a list using the conventional three-letters code. On the other hand, the library is formatted as a simple text file and contains: (a) the molecular connectivity, bond lengths and bond angles; (b) the "*atom type*" associated to each atom in each amino acid; and (c) the force-field parameters used to evaluate the energy during the generation-relaxation process. In this work all the parameters were taken from AMBER [28], a well-known force-field that is frequently used to examine the dynamical properties of proteins with 3D order. However, it should be emphasized that this procedure is compatible with many force-fields and our aim is not to investigate the reliability of the AMBER force field in the description of the structural properties of the unfolded state of proteins. Finally, the last input file defines the range of allowed values for the backbone and side chain dihedral angles. This information can be used to improve the efficiency of the generation process by imposing restrictions that exclusively allow low-energy atomic positions. For example, the dihedral angle associated with the peptide bond ω is allowed to take values within the interval $180°\pm10°$, while the dihedral angles associated to the side chain aromatic rings of Phe, Tyr and Trp are $0°\pm0°$. In order to avoid any biasing of the results, no other restriction has been introduced in the remaining backbone (φ, ψ) and side chain dihedral angles (*i.e.*, they are allowed to take any value in the interval $180°\pm180°$) with the exception of Pro, whose values have been defined according to the pseudo-rotational states of the pyrrolidine ring [29, 30].

The input information is used to produce representative microstructures of a given disordered protein using a two-step procedure. In the first one, microstructures are generated with minimum torsional strain, while in the second step such microstructures are relaxed. Specifically, the generation algorithm can be summarized as follows (Fig. 1):

(1) Three bonded atoms are placed in arbitrary positions within the simulation box.

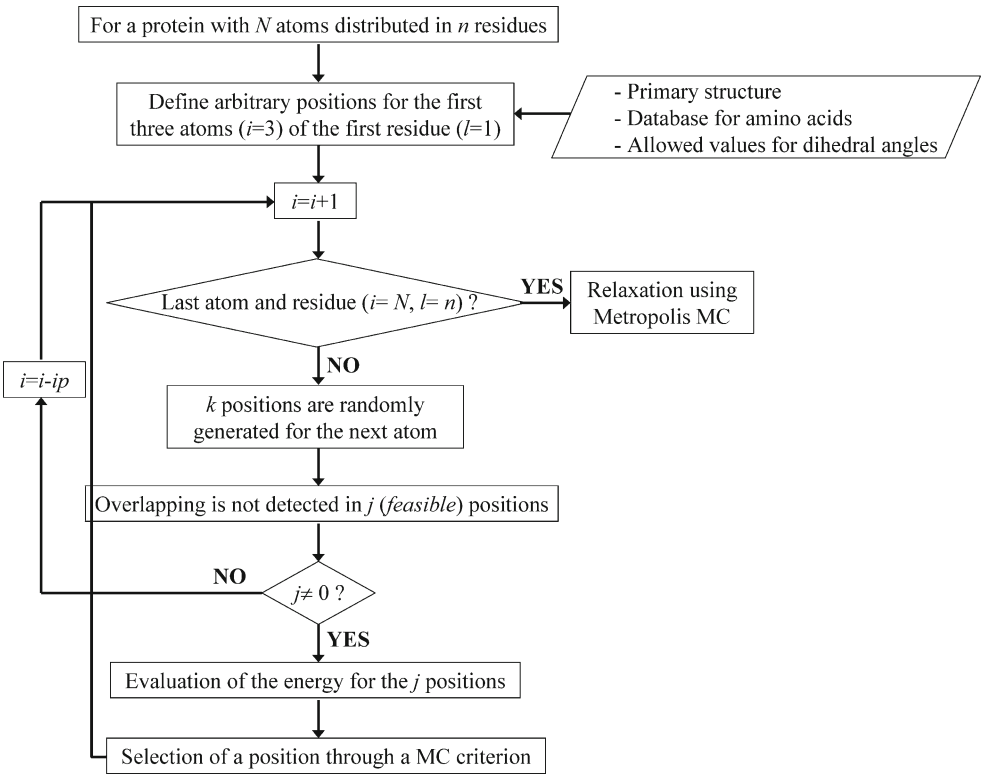

Fig. 1 Flowchart detailing the strategy used to obtain representative microstructures of disordered proteins

(2) A number, k, of positions are randomly generated for the fourth atom, considering the interval of allowed values supplied in the input information for the dihedral angle, the energy associated with every generated position (*i.e.,* E_j where j ranges from 1 to k) being calculated. One of these positions is selected by applying a typical MC acceptance criterion, that is a position is chosen if:

$$\mathrm{Rnd}(1) < \frac{e^{-\tau E_j}}{e^{-\tau E_{j(\min)}}}, \quad (1)$$

where Rnd(1) is a random number between 0 and 1, and τ is a constant taking into account the temperature (298 K) defined by the user.

(3) Again, k positions compatible with the geometry requirements (bond length and bond angle) are randomly generated for the fifth particle. Every position is categorized as *unfeasible* or *feasible* depending on atomic overlaps with atoms separated by more than three bonds. The energy of each *feasible* position is evaluated and one of them is selected by using the MC acceptance criteria mentioned above. This procedure ensures minimum torsional strain.

(4) Step (3) is repeated for the remaining atoms in the protein. If for a given atom i all the generated positions lead to atomic overlaps (*i.e.*, the number of *feasible* positions is zero), the molecule building is restarted from the atom *i-ip* where *ip* increases with the number of failures.

In this work k and *ip* were fixed at 20 and 10 for all the studied proteins. The energies of the generated positions E_j have been calculated as the sum of the torsional and van der Waals contributions associated with the atoms separated by three chemical bonds (1–4 interactions) while all other contributions were neglected. It should be emphasized that, whilst this procedure is not biased to favor microstructures with very low total energies, it guarantees a minimum torsional strain for the generated systems. Moreover, the omission of the electrostatic interactions avoids the participation of charged groups of atoms in the generation procedure (*i.e.*, the electrical neutrality is maintained in the complete residues, once generated, while incomplete residues remain electrically charged during the generation process). Next sections show that the role played by 1–4 interactions in the description of the unfolded proteins is crucial. In order to prove the impact of 1–4 interactions in these systems, the generation algorithm was modified. Specifically, in the "modified" generation algorithm the energetically-based (Eq. 1) MC acceptance criterion used in steps 2 and 3 was replaced by a simple random acceptance criterion. The "modified" algorithm was only used when explicitly indicated, the procedure relying on the 1–4 interactions contribution being switched on otherwise.

In any case the generated positions fulfill the internal geometry restrictions (*i.e.*, molecular connectivity, bond distances and bond angles). The protein is sequentially (*i.e.*, amino acid by amino acid) built using the primary structure provided in the input. Moreover, once a backbone atom has been generated, the procedure continues through the side chain rather than through the next atom of the backbone. An important difference between this method and those developed for the synthetic polymers simulation in the amorphous solid state is that in the latter the radii of the atoms and pseudo-atoms were reduced multiplying by a factor $\lambda < 1$ [24–26]. The scaled radii $\{\lambda \cdot R\}$ were used to categorize each generated position into *feasible* or *unfeasible*. However, in the current study this reduction is unnecessary as no multichain system (*i.e.*, solid state system) was considered.

The generated microstructures were relaxed using a simple Metropolis MC method to eliminate the least favored inter-residue nonbonding interactions as well as to enhance the most attractive ones (*e.g.*, salt bridges). Accordingly, the relaxation was carried out using a complete force-field that includes not only the 1–4 interactions mentioned above but also the van der Waals and electrostatic interactions between atoms separated by more than three chemical bonds. The expressions to describe such energy contributions as well as all the required parameters were taken from AMBER force-field [28]. It should be noted that the sophisticated relaxation methods previously developed for multichain systems (*e.g.*, those based on the geometric aspects of configurational bias and rotation MC methods [31, 32] are not necessary here because of the single-chain nature of the systems under study. Furthermore, disordered proteins are expected to be essentially dominated by bonding and short-range non-bonding interactions, which are implicitly considered in the generation step (*i.e.*, avoiding both torsional strain and steric clashes) rather than by long-range non-bonding interactions. Accordingly, the relaxation of all microstructures generated in this work consists of 1000 MC steps only.

Simulated proteins

As was mentioned above, many biological processes are controlled through disorder-to-order transitions, the understanding of disordered states being essential for the development of therapeutic targets. In this context, seven NUPs whose specific functions are still unknown and which are involved in human diseases or essential biochemical

processes, were selected to check if the stochastic microstructures provided by our computational scheme are representative enough to predict experimental structural properties.

The seven studied NUPs are:

Tau

This is a microtubule-associated protein occurring in the axons of neurons. It is one of the two main proteins involved in the pathology of Alzheimer's disease via the formation of β-sheet aggregates called paired helical filaments [33]. Overall, Tau has a very low content of secondary structure and retains its unfolded state even after binding to microtubules [34].

Nucleocapsid binding domain of Sendai virus phosphoprotein (SeV P)

SeV is the prototype virus of the Paramyxoviridae which infects the respiratory track of mice and causes pneumonia. The RNA polymerase of the virus is constituted by two proteins, the large (L) protein and the phosphoprotein (P). SeV P plays a crucial role in the enzyme by positioning L onto the matrix for transcription and replication formed by the RNA and the nucleoprotein, the N-RNA [35]. The SeV P has a modular structure with two distinct functional domains: a disordered N-terminal domain, and a C-terminal domain which carries the oligomerization region forming a three-helix bundle [36].

Human protein Ki-157 (Ki157)

The expression of the human protein Ki157 in diverse cancer cells and its phosphorylation in blood leukocytes only after mitogenic activation, have suggested its possible role in cell signalling [37]. Moreover, its interaction with several other regulatory proteins possibly plays a role in cellular signaling, transcriptional regulation or RNA metabolism. It was proposed that Ki157 has an extended shape and belongs to the class of NUPs [38].

Prothymosin alpha (ProTα)

ProTα is a small nuclear protein, which was initially isolated from rat thymus. Although the precise function of ProTα is currently unknown, an accumulating line of evidences suggests that it is associated with cell proliferation. This protein is in a random coil conformation at physiological conditions [39], and is therefore a NUP, whereas at acidic pH it is transformed into a premolten globule state [40].

Myelin basic protein (MBP)

MBP is present between the cytoplasmic leaflets of the compact myelin membrane in both peripheral and central nervous systems. Among the existing isoforms, the first species is the most abundant in adult human myelin and is essential for the formation of the latter [41]. MBP is one of the major autoantigens in several scleroses. It was shown to be an extended, intrinsically disordered protein in solution [42].

Mammalian α4 (α4)

This protein, which is also known as immunoglobulin binding protein 1 (IgBP1), is involved in a variety of functions related to differentiation, cell proliferation, apoptosis and vertebrate embryonic development [43]. The structural basis for its regulatory function is still unclear. It contains large unstructured regions [44].

Enterocin EJ97 (EJ97)

Bacteriocins belong to the wide variety of antimicrobial ribosomal peptides synthesized by bacteria and are used in food preservation and in human therapeutics. Particularly, they are a good alternative to conventional antibiotics ineffective against resistant bacterial strains. EJ97 has a broad range of activity against Gram-positive bacteria [45]. The peptide was shown to lack any well-defined tertiary structure and to behave as a disordered protein [46].

In addition to NUPs, the strategy presented in this work has been applied to seven classical proteins, which were experimentally studied in denaturation conditions [47–49]. Once our methodology transferability has been proven for any non-native conformational states, it should be reliable to understand protein unfolding-folding processes. To this purpose, we have selected proteins as diverse as: *apomyoglobin* (APO), an oxygen-binding protein in the muscle tissue; *CheY*, a bacterial chemotaxis response regulator controlling the direction of flagellar rotation; *cytochrome c (Cyto-c)*, a heme protein involved in the electron transport chain of mitochondrion; *OspA*, an antigenic lipoprotein; *protein G*, an immunoglobulin-binding protein expressed in group C and G Streptococcal bacteria; *Snase*, a staphylococcal nuclease; and *Ubiquitin* (*UBI*), a regulatory protein ubiquitously expressed in eukaryotes.

Properties calculated to describe the disordered state

For the hydrodynamic-level properties, we decided to follow the strategy nicely described by Garcia de la Torre et al. [50] who proposed to compute the properties as conformational averages over values obtained for a number of microstructures, each being considered to be rigid. Following this master-plan we produced representative microstructures for each of the seven NUPs and "classical" proteins described above, using the computational strategy presented in the first part of this work.

As the root mean square distance between the constitutive mass elements and the center of mass of the full atomic set, the radius of gyration (R_g) is a structural parameter

characterizing the size and shape of a particle, alongside with the maximal head-to-tail distance, also denoted maximum length (L_{max}). For a rigid system made of i mass elements m_i each separated from the center of mass by a distance r_i, it writes:

$$R_g = \sqrt{\frac{\sum_i m_i r_i^2}{\sum_i m_i}}. \quad (2)$$

Consequently, it can be straightforwardly computed once the atomic positions of any given configuration of the system are known. In the field of proteins, it is often related to the structure compactness, which makes it a key parameter in the folding processes.

Often similar in magnitude, the hydrodynamic (or Stokes) radius (R_H) is directly related to the diffusion coefficient (D) of the macromolecule diffusing in a medium of viscosity η at temperature T, by the Stokes-Einstein relation:

$$D = \frac{kT}{6\pi\eta R_H}. \quad (3)$$

Similarly, the Svedberg relation (Eq. 4) allows obtaining the sedimentation coefficient (s):

$$s = \frac{DM(1-\upsilon\rho)}{RT}, \quad (4)$$

where M is the molecular weight, υ is the specific volume, ρ is the solution density, R is the gas constant and T is the absolute temperature. These properties can be experimentally measured for proteins by several techniques [51].

NMR was recently recognized as a potential method to describe the protein folding mechanisms and the conformational state of NUPs [52, 53]. We therefore predicted chemical shifts for two NUPs and one denatured protein with the SHIFTS program [54] from their atomic coordinates. It was developed based on an additive model of chemical shift contributions, corresponding to various conformational effects found in a database of density functional theory (DFT) calculations on more than 2000 peptides. Some empirical extensions were used for covering additional conformation regions and residue types.

Results

Assessment of our computational scheme performance

The influence of the number of representative microstructures generated and subsequently relaxed was investigated for MBP and UBI, which involve 171 and 76 residues, respectively. Figure 2 represents the accumulated average

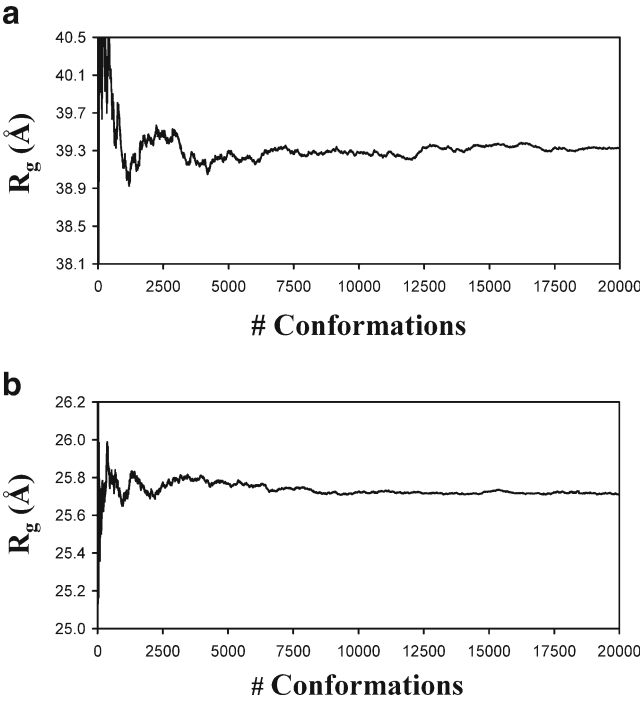

Fig. 2 Accumulated value of the averaged radius of gyration (R_g in Å) against the number of microstructures produced for (**a**) MBP and (**b**) UBI

of R_g against the number of conformations. As it can be seen, this average stabilizes in both cases at 39.3 Å and 25.7 Å for about $6.5 \cdot 10^3$ microstructures. Accordingly, a total of 10^4 microstructures were produced for the set of proteins, independently of their size. That is expected to guarantee the convergence of the calculated properties.

Figure 3a shows the distribution of energies associated to the first 10^4 microstructures obtained for MBP and UBI. The shape of both distributions fits to a Gaussian function, demonstrating the reliability of the method and showing that the conformations can be described in terms of rotational isomers [55]. As expected, the Boltzmann distributions displayed in Fig. 3a depend on the protein size. The relative energy distribution is wider when the number of residues increases, reflecting the enhancement of the conformational flexibility. Similarly, R_g for the same set of conformations of MBP and UBI also follows Gaussian distributions, as depicted in Fig. 3b, the dependence between the width of the functions and the size of the protein being very similar to those found for the relative energies. On the other hand, Fig. 4 represents the number of microstructures compatible with several R_g and relative energy intervals. The remarkable dispersion of microstructures observed for both MBP and UBI reflects the flexibility of the two macromolecules. Indeed, in both cases even low-energy conformations are still compatible with a wide range of R_g values.

Table 1 lists the averaged conformational parameters predicted for the 14 proteins examined in this work (seven

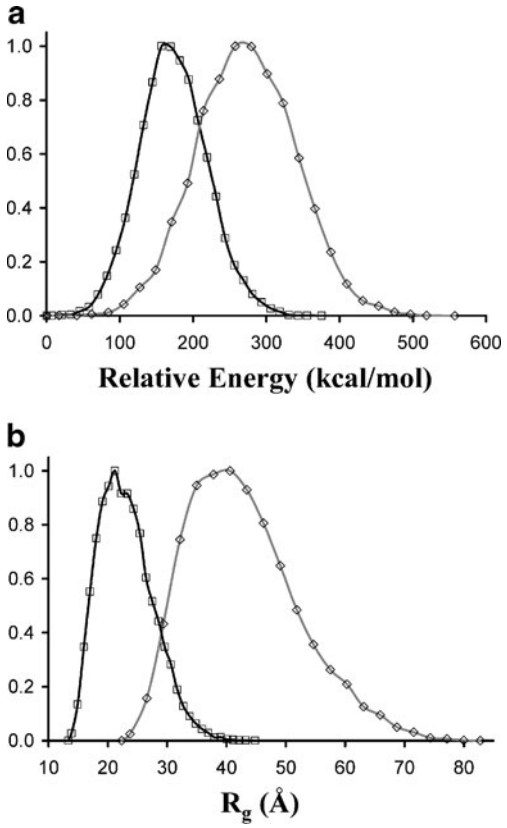

Fig. 3 Normalized distribution of (**a**) relative energies and (**b**) radius of gyration (in Å) obtained for 10^4 microstructures of MBP (gray line and diamonds) and UBI (black line and squares)

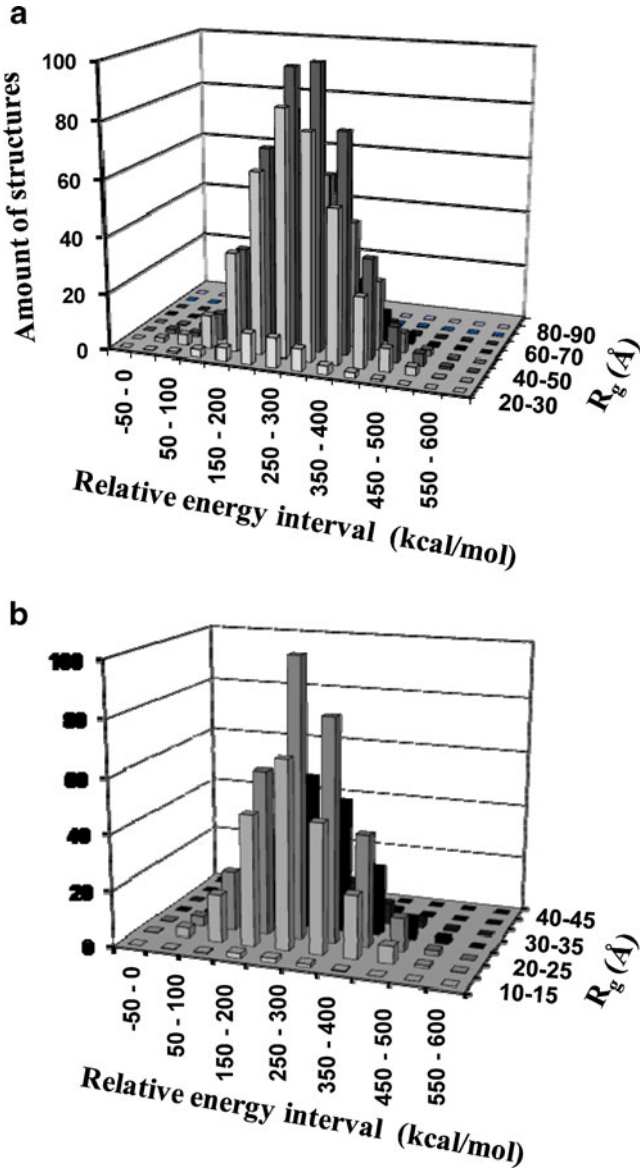

Fig. 4 Three dimensional graphic representing the number of microstructures compatible with defined intervals of radius of gyration (in Å) and relative energy (in kcal mol^{-1}) for (**a**) MBP and (**b**) UBI

NUPs and seven "classical" proteins) and the available experimental values are also included for comparison. In order to illustrate the extensive conformational sampling achieved using this simulation strategy, Fig. 5 represents microstructures with globular, elongated and intermediate shapes produced for Sev P. Figure 6 represents the calculated *versus* the experimental R_g values. An excellent agreement is observed between both values for all the proteins with exception of MBP, α4 and EJ97. In the two former cases the experimental R_g is underestimated and overestimated, respectively, by 27 % and 38 %. This difference is even larger for EJ97, which was attributed to the fact that the experimental estimations of R_g and R_H were roughly derived from the diffusion coefficient D imposing a spherical shape for the NUP [46]. A linear correlation analysis ($R_{g,exp} = c \cdot R_{g,calc}$) between the experimental and calculated on the full set of R_g values leads to a deviation of only 4 % (c= 0.96) with poor correlation (R^2=0.71), while excluding two such proteins (Fig. 6) evidences a very high correlation (R^2=0.95) preserving the 4 % deviation (c=1.04).

Figure 7a represents the calculated R_g against the number of residues (n) for the 14 investigated proteins. It is worth noting that this structural property logarithmically increases with n, the correlation coefficient derived from the fitting being R^2=0.97. In comparison, consideration of the experimental R_g values against n for the whole set led to a logarithmic fitting with a correlation coefficient of only R^2=0.80 (Fig. 7b) confirming that the R_g values experimentally determined for MBP, α4 and EJ97 deviate from the general behavior usually followed by both natively and chemically unfolded proteins. Accordingly to our previous prediction, the regression coefficient increases to R^2=0.90 when these two proteins are eliminated from the fitting (Fig. 7b).

On the other hand, the accuracy in prediction of R_g appears to evolve with the size of the protein. This is reflected by Fig. 8, which represents the square difference between the experimental and calculated R_g against n for 11

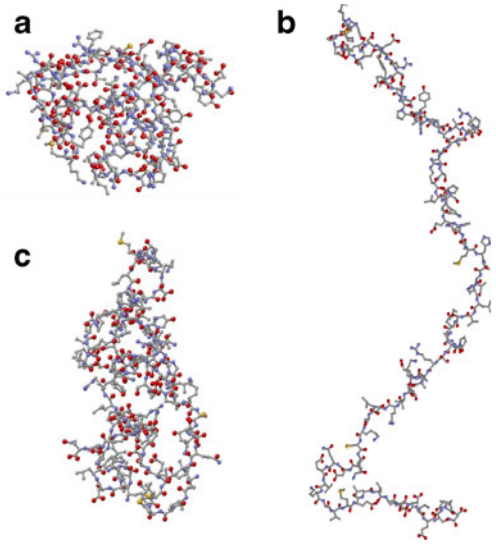

Fig. 5 Representative microstructures obtained for SeV P, which correspond to conformations with (**a**) globular, (**b**) elongated and (**c**) intermediate shapes

proteins (*i.e.*, the MBP, α4 and EJ97 have not been considered in the analysis for above mentioned reasons). Except in the case of Protα, the square difference is larger for the proteins involving more than 250 residues even though the agreement between the experimental and calculated values is excellent for such systems. Moreover, no detailed correlation was found between the square difference and *n*, as reveals the fact that the error is 10 %, 8 % and 5 % for OspA (273 residues), Ki157 (292 residues) and Tau (441 residues), respectively.

The other experimental values (*i.e.*, L_{max}, R_H, D and/or s), are rather sparse in literature. However, very good agreement was obtained between the calculated and the experimental values with the obvious exception of MBP and α4, again. Indeed, their L_{max} is underestimated and overestimated, respectively. Figure 9a represents the logarithm of the calculated R_H against the logarithm of *n*. As expected,

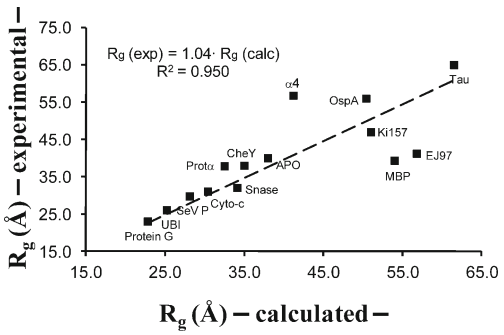

Fig. 6 Graphical representation of calculated radii of gyration (in Å) versus the experimental values for the 14 proteins considered in this work. The linear regression analysis ($y = c{\cdot}x$) was performed excluding the α4 and MBP proteins

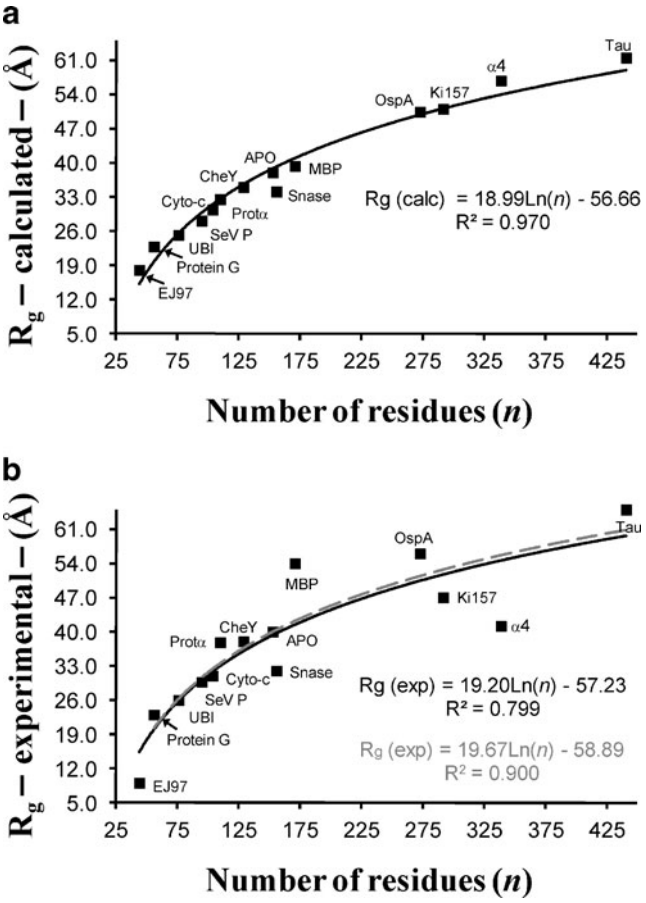

Fig. 7 Graphical representation of (**a**) calculated and (**b**) experimental radii of gyration versus the number of residues for the 14 proteins considered in this work. The equation derived from logarithmic regression analysis ($y = c{\cdot}\ln x + b$) is indicated. In (**b**) the regression obtained considering the proteins is represented with a solid line and black letters, while the fitting performed excluding the values of MBP, α4 and EJ97 proteins are provided in gray and with a dashed line

R_H increases with *n*, the equation resulting from the fitting being very similar to those typically found for random coil macromolecules including globular proteins [51, 56]. Specifically, experimental measures of R_H in different

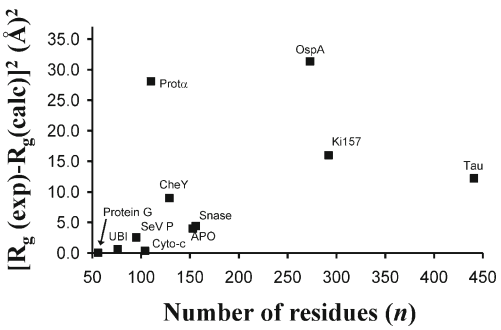

Fig. 8 Graphical representation of square difference between the experimental and calculated radii of gyration versus the number of residues. The results obtained for MBP, α4 and EJ97 are not included

conditions led to:

$$\mathrm{Ln}(R_H) = b + a \cdot \mathrm{Ln}(n), \quad (5)$$

where b≈2.2 and a≈0.5, though both parameters depend on the solvent (water, urea, guanidinium chloride, etc.), the ionic strength, and the method to measure R_H (*i.e.* size exclusion chromatography, dynamic light scattering, pulse field gradient NMR spectroscopy, etc.).

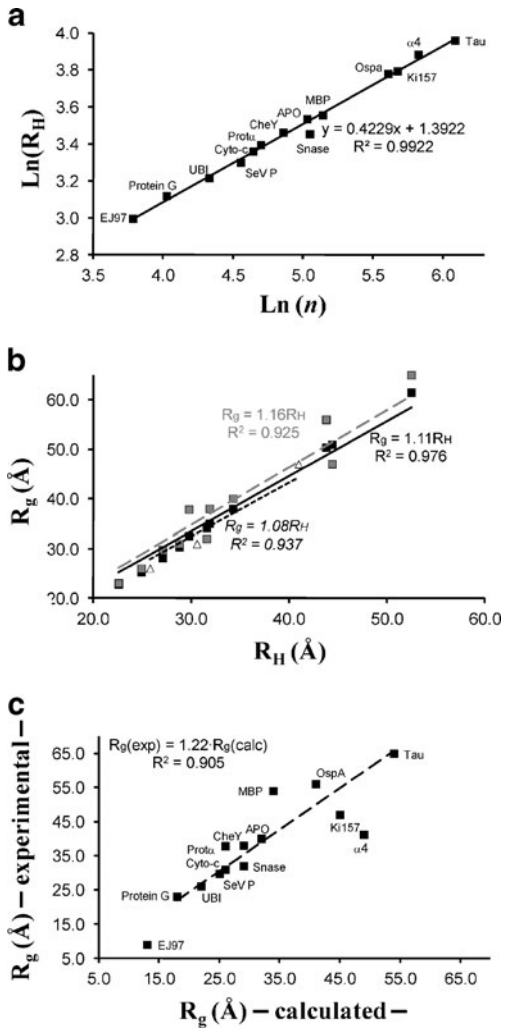

Fig. 9 Graphical representation of: (**a**) the logarithm of the calculated R_H (in Å) versus the logarithm of the number of residues for the proteins considered in this work. The equation derived from a linear regression analysis ($y = a \cdot x + b$) is indicated. (**b**) the R_g versus the R_H : (i) the calculated R_g against the calculated R_H (black squares, solid black line and black plain text); (ii) the experimental R_g against the calculated R_H (gray squares, solid gray line and gray plain text); and (iii) the experimental R_g against the experimental R_H (empty triangles, dashed black line and black italics text). MBP, α4 and EJ97 are not included in the analyses; (**c**) the experimental R_g (in Å) versus the calculated R_g for the studied proteins, which were produced without considering the 1–4 interactions in the generation process. In order to facilitate the comparison with Fig. 5, the linear regression analysis ($y = c \cdot x$) was performed excluding the α4, MBP and EJ97 proteins

The variation of R_g against R_H is displayed in Fig. 9b, where the experimental and calculated values of both properties are considered. Although, R_H experimental values are only available for a few of the 14 investigated proteins, the behavior followed by such parameter and the corresponding R_g have been used as the reference: $R_g(\mathrm{exp}) = 1.08 \cdot R_H(\mathrm{exp})$. Both the variation of the experimental R_g against the calculated R_H and of the calculated R_g against the calculated R_H fit very well to the latter equation [*i.e.*, $R_g(\mathrm{exp}) = 1.16 \cdot R_H(\mathrm{calc})$ and $R_g(\mathrm{calc}) = 1.11 \cdot R_H(\mathrm{calc})$], reflecting again the reliability of

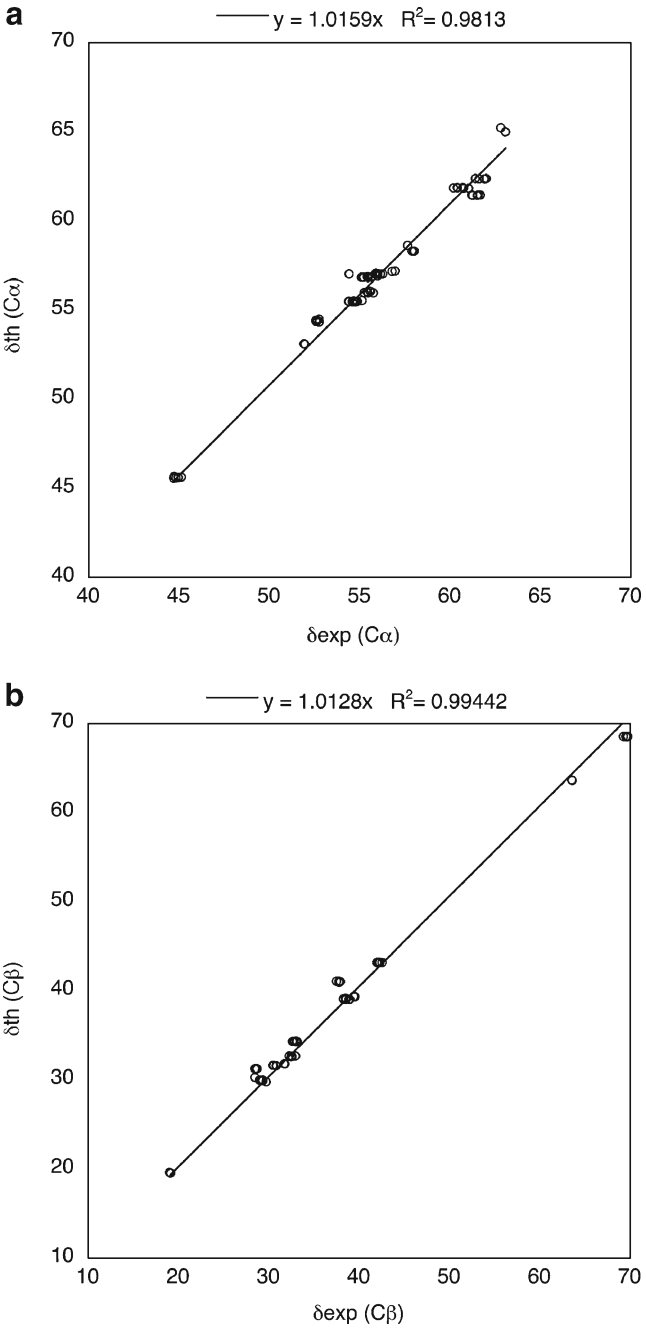

Fig. 10 Experimental (**a**) Cα, (**b**) Cβ chemical shifts (in ppm) of Ubiquitin with respect to calculated values

the microstructures produced by the method. Though the Zimm's theory of polymer dynamics predicts that random coils obey the ratio $R_g \approx 1.5 \cdot R_H$, a recent study of protein L using single-molecule fluorescence methods gave values in the fully denatured state, and the ratio between them was ~1.1 [57].

Finally, chemical shifts of one protein from each family, Ubiquitin and EJ97, were calculated and compared to the available experimental data. First, the Cα and Cβ chemical shifts of Ubiquitin were considered. Figures 10a and b show really good agreement between both values, with R^2=0.98 and 0.99 for Cα and Cβ, respectively. This concordance is maintained all along the amino acid sequence as shown in Fig. 11. Relative errors range from 0.2 to 4.6 and 0.05 to 9.4 % for Cα and Cβ, respectively. The conclusions are the same for the H's chemical shifts (from Hα to Hd) of EJ97 as pictured for Hα and Hd in Fig. 12a and b. The best correlation between experimental and theoretical values is observed for Hb's (R^2=0.94) and Hd's (R^2=0.99). The relative errors are globally higher than for C chemical shifts. For four residues, Ile5, Tyr30, Lys31, Pro38, the relative error is higher than 15 % for two types of H's.

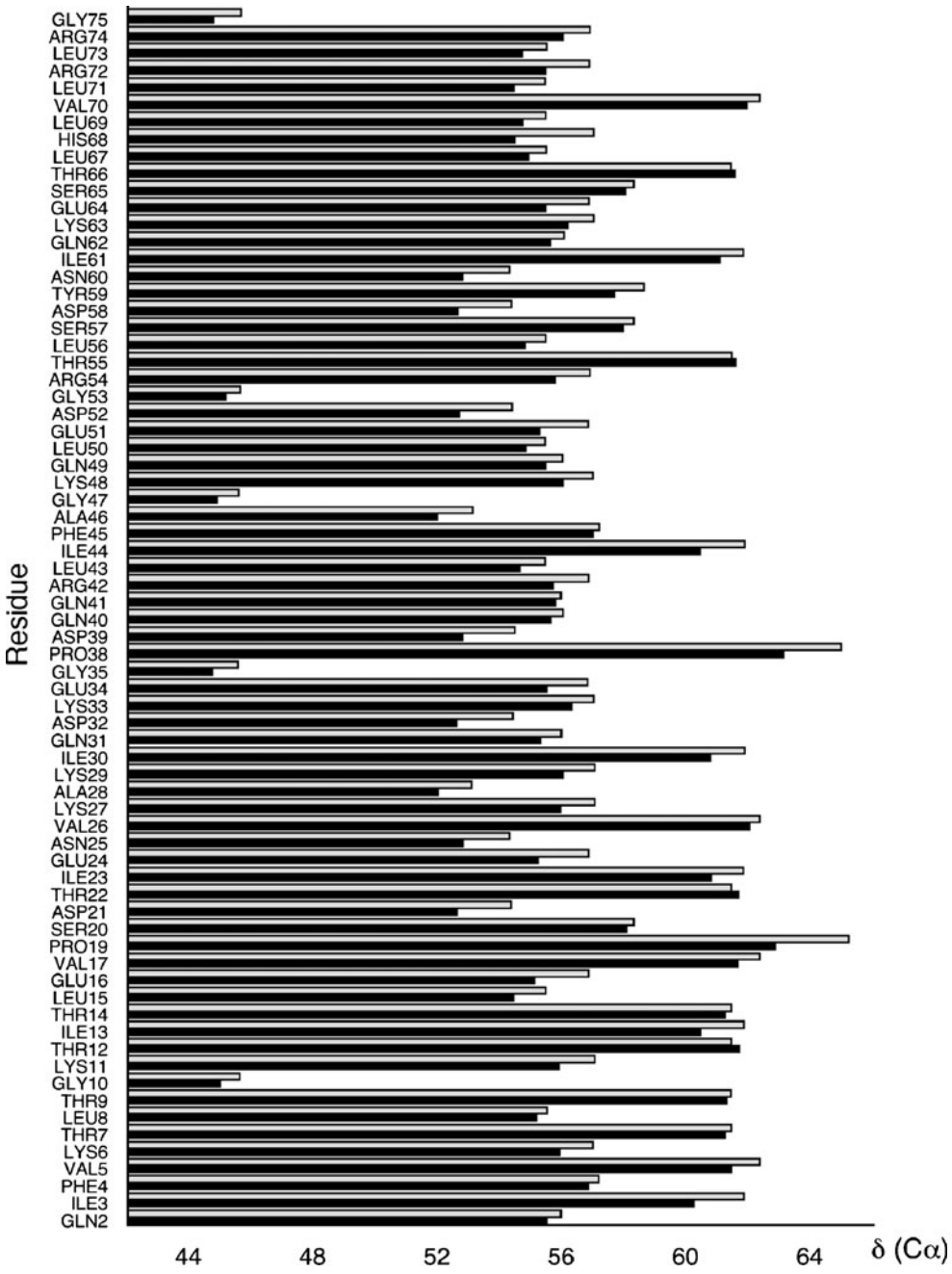

Fig. 11 Experimental (black) and calculated (light gray) Cα chemical shifts (in ppm) of Ubiquitin with respect to amino acid sequence

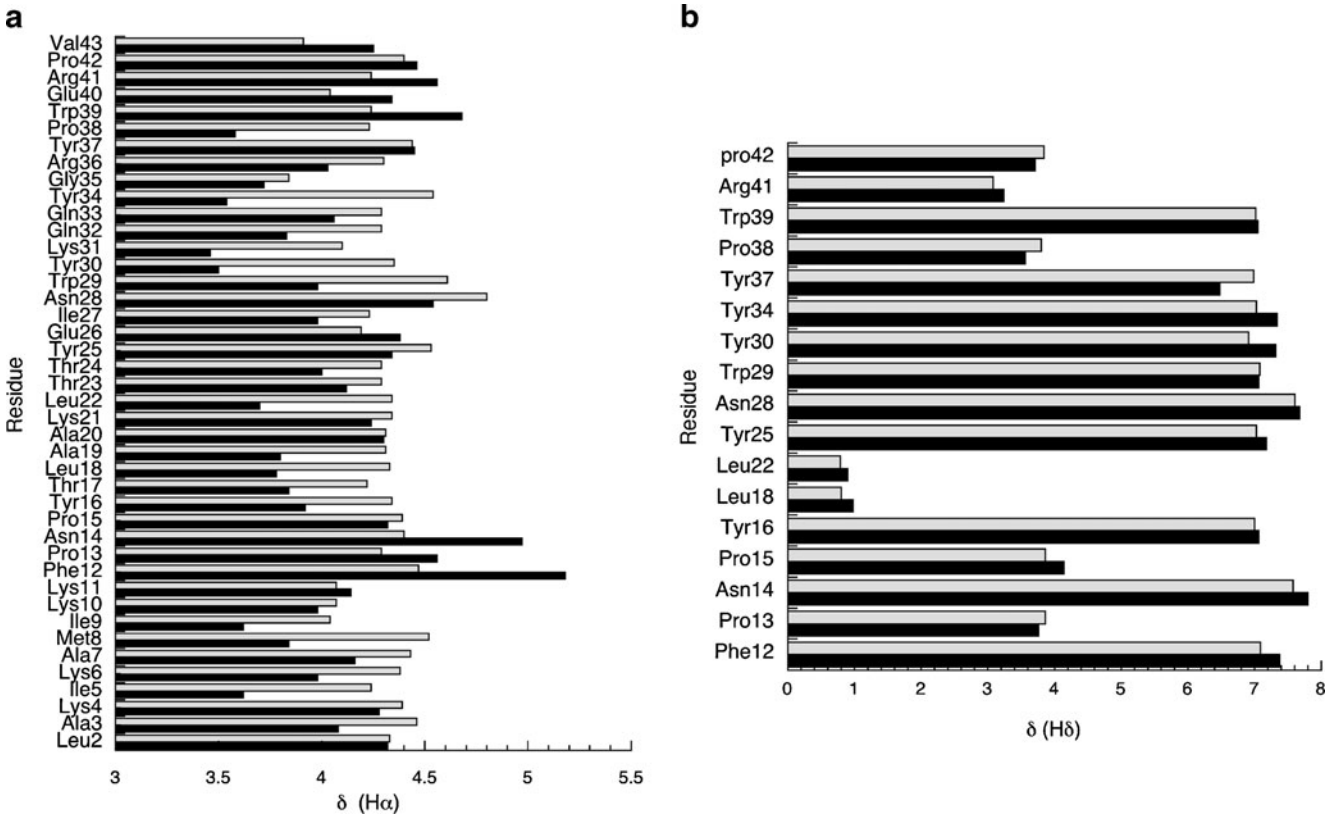

Fig. 12 Experimental (black) and calculated (light gray) (**a**) Hα, (**b**) Hd chemical shifts (in ppm) of EJ97 with respect to amino acid sequence

Discussion

The simulation strategy here presented, which is inspired by those developed for the properties of amorphous polymers (*i.e.*, disordered multichain macromolecules systems) [23–25], has been found to satisfactorily and efficiently describe the conformational behavior of both NUPs and classical proteins in denaturation conditions. Indeed, our generation-relaxation method allows obtain a large number of representative uncorrelated microstructures at the atomistic level using nominal computational resources. For example, the 10^4 microstructures of *Tau* (441 residues) were produced in ~55 h (2.3 days) using a AMD opteron Magny Cours (8 cores 6134) 2.3 GHz double processor, while the requirements needed by conventional methods like MD and MC are significantly higher because of this intrinsically associated structural correlation.

The implemented strategy consists first in the generation of a limited set of characteristic microstructures, corroborated by the relatively low number of conformations required to converge the structural properties. As the present algorithm was designed to produce microstructures with minimal torsional strain and without local and non-local steric clashes, this led to hypothesize that the unfolded state is mainly governed by bonding and short-range non-bonding interactions. Therefore, the 1–4 interactions, which are only to be evaluated during the process, turn out to be essential in disordered proteins. The second stage is a relaxation process based on the application of 10^3 MC steps only, that allows eliminating and enhancing very attractive or repulsive interactions.

This new methodology presents significant advantages with respect to the existing procedures. More specifically, energy evaluation criteria are introduced in the generation algorithm, which allows to significantly reduce the number of generated microstructures with respect to those based on library sampling (*i.e.*, based on the selection of $\{\varphi, \psi\}$ from a torsional subset database). For example, the flexible meccano method needed a huge amount of microstructures (*i.e.*, 50,000) for NUPs α-synuclein and Tau [58], while energy criteria based on 1–4 interactions allows considerable reduction in such number, as shown in this work. From a computational point of view, our method is much more efficient than others also based on energy calculations, as for example MD simulations, as the latter requires a huge amount of CPU resources for the production of uncorrelated microstructures.

Another issue refers to the evaluation of properties using conformational averages rather than conventional Boltzmann

distributions. It is worth noting that the proposed methodology is dedicated to the generation of microstructures without any type of torsional strain, which was thought to be a crucial factor for the description of the unfolded state in proteins. Accordingly, the energies associated to the generated structures follow a Boltzmann distribution by themselves, and the derived properties are consequently directly averaged. The biasing has been introduced through the acceptance MC criterion used to choose among the different generated positions, which is based on the energy evaluated for the 1–4 interactions. Moreover, the latter interactions have been found decisive for the satisfactory description of unfolded "classical" proteins and NUPs. It should be underlined that the effect of 1–4 interactions is intrinsically considered by methods based on statistical criteria, which extract the backbone dihedral angles {φ,ψ} from populations of coil regions found in protein structural databases.

In order to demonstrate the remarkable role of the 1–4 interactions, 10^4 microstructures were produced for the 14 studied unfolded proteins but without considering this class of interactions during the generation process (*i.e.*, using the "modified" generation algorithm described above). Figure 9c, which represents the calculated against the experimental R_g values, evidences that the conformations produced in absence of 1–4 interactions are more compact than those obtained when these interactions are considered in the generation process (Fig. 6). Consequently, the calculated values are underestimated by around 22 % with respect to the experimental ones (R^2=0.753 and 0.905 when the MBP, α4 and EJ97 proteins are included and excluded, respectively, in the analysis). These results confirm the importance of 1–4 interactions in the conformational description of unfolded proteins.

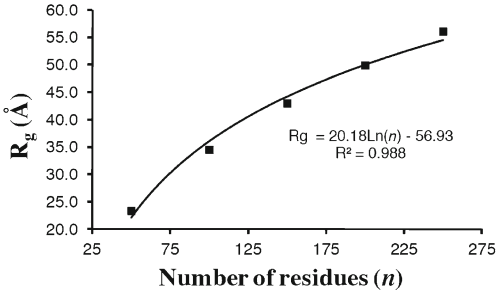

Fig. 13 Graphical representation of the radii of gyration calculated for polyalanine versus the number of residues. The equation derived from logarithmic regression analysis ($y = c \cdot \ln x + b$) is indicated

Another important issue refers to the relationship between the R_g and the number of residues *n*. Results displayed in Fig. 7a and b, which illustrate such dependence for both the calculated and the experimentally determined R_g values, suggests that the influence of the chemical nature on the structure of denatured proteins is relatively small. In order to confirm this feature, the R_g have been calculated considering polypeptides formed by 50, 100, 150, 200 and 250 identical Ala residues. Results, which are displayed in Fig. 13, reflect very similar dependence behavior. Moreover, the R_g calculated for polypeptides of 100 residues obtained by repeating the following sequences –(Ala)–, –(Val)– and –(Ala-Val-Ser)– is 34.5, 29.6 and 32.0 Å. These values are very similar to those calculated and experimentally determined for Cyto-c and SeV P (Table 1), which involve 104 and 95 residues, respectively, and present a significant variability in their primary structure. Accordingly, the role played by the molecular length of the unfolded proteins seems to be more important than that played by the chemical nature.

Table 1 Averaged conformational parameters predicted for the proteins studied in this work: Radius of gyration (R_g; in Å), maximum length (L_{max}; in Å), hydrodynamic radius (R_H; in Å), diffusion coefficient (D; in $cm^2 \cdot s^{-1}$), and sedimentation coefficient (s; in s). Experimental values (if available) are provided in parenthesis

#	R_g	L_{max}	R_H	D (10^{-7})	s
Tau	61.5 (65)	200.0	52.5	4.2	2.1
SeV P	29.7 (28.1)	91.1 (85)	27.1	8.1	0.9
Ki157	51.0 (47)	164.9 (150)	44.4 (41)	5.0 (4.7)	1.8 (2.2)
ProTα	32.5 (37.8)	103.4	29.8	7.4	0.9
MBP	39.3 (54)	125.1 (160)	35.0	6.3	1.2
α4	56.8 (41.2)	183.2 (142)	48.6	4.5	1.9
EJ97	18.4 (9.4)	59.6	20.0 (12.2)	11.0 (14.1)	0.6
APO	38.0 (40)	101.6	34.3	6.4	1.2
CheY	35.0 (38)	91.8	31.9	6.9	1.0
Cyto-c	30.4 (31)	82.1	28.8 (30.6)	7.6	1.0
OspA	50.4 (56.1)	130.7	43.8	5.0	1.5
Protein G	22.8 (23)	71.9	22.6	9.7	0.6
Snase	34.1 (32)	109.8	31.6	6.9	1.1
UBI	25.2 (26)	81.5	24.9 (25.8)	8.8	0.8

Conclusions

To summarize, our computational approach led to an excellent fit between experimental and calculated R_g values indicating that the sets of conformations are statistically able to capture the essential characteristics of real unfolded proteins, though (i) each of such sets involves a relatively low number of conformations with respect to other theoretical methods; and (ii) no adjustable parameter was introduced in our theoretical strategy. The latter is particularly remarkable since no restriction was imposed in the generation process for φ and ψ flexible dihedral angles, avoiding any biasing in the results. However, the allowed values for these angles could be restricted to the sterically accessible conformations of the different amino acids. This strategy will further reduce both the required computational resources and the number of conformations needed to describe the unfolded proteins.

Comparisons between theoretical and experimental values of the several properties considered in this work suggest that MBP and α4 do not behave as could be expected from NUPs. Indeed, the R_g and L_{max} values experimentally determined for MBP are overestimated with respect to those predicted for a disordered protein, while the properties measured for α4 are underestimated. Considering that the experimental and theoretical values found for the other investigated proteins are in excellent agreement, the properties of both MBP and α4 should be re-investigated, or alternatively they should not be considered as completely unstructured systems.

Acknowledgments Financial support from the Ministerio de Ciencia e Innovación (MICINN) and Fondo Europeo de Desarrollo Regional (FEDER) (MAT2009-09138), Generalitat de Catalunya (research group 2009 SGR 925) is gratefully acknowledged. Support for the research of C.A. was received through the prize "Institució Catalana de Recerca i Estudis Avançats (ICREA) Academia" for excellence in research funded by the Generalitat de Catalunya. CM and EAP thank the Belgian National Fund for Scientific Research (FRS-FNRS) for their Research Associate and Senior Research Associate position, respectively.

References

1. Travaglini-Allocatelli C, Ivarsson Y, Jemth P, Gianni S (2009) Folding and stability of globular proteins and implications for function. Curr Opin Struct Biol 19:3–7
2. Dill KA, Ozkan SB, Weikl TR, Chodera JD, Voelz VA (2007) The protein folding problem: when will it be solved? Curr Opin Struct Biol 17:342–346
3. van Gunsteren WF, Bürgi R, Peter C, Daura X (2001) The key to solving the protein-folding problem lies in an accurate description of the denatured state. Angew Chem Int Ed Engl 40:351–355
4. Hammarström P, Carlsson U (2000) Is the unfolded state the Rosetta Stone of the protein folding problem? Biochem Biophys Res Commun 276:393–398
5. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 337:635–645
6. Uversky VN, Dunker AK (2008) Controlled chaos. Science 322:1340–1341
7. Dunker AK, Oldfield CJ, Meng J, Romero P, Yang JY, Chen JW, Vacic V, Obradovic Z, Uversky VN (2008) The unfoldomics decade: an update of intrinsically disordered proteins. BMC Genomics 9:S1–S26
8. Fink AL (2005) Natively unfolded proteins. Curr Opin Struct Biol 15:35–41
9. Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol 323:573–584
10. Uversky VN, Dunker AK (2010) Understanding protein non-folding. Biochim Biophys Acta 1804:1231–1264
11. Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 293:321–331
12. McCarney ER, Kohn JE, Plaxco KW (2005) Is there or isn't there? The case for (and against) residual structure in chemically denatured proteins. Crit Rev Biochem Mol Biol 40:181–189
13. Fisher CK, Stultz CM (2011) Constructing ensembles for intrinsically disordered proteins. Curr Opin Struct Biol 21:426–431
14. Jha AK, Colubri A, Freed KF, Sosnick TR (2005) Statistical coil model of the unfolded state: resolving the reconciliation problem. Proc Natl Acad Sci USA 102:13099–13104
15. Bernadó P, Blanchard L, Timmins P, Marion D, Ruigrok RW, Blackledge M (2005) A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. Proc Natl Acad Sci USA 102:17002–17007
16. Jensen MR, Blackledge M (2008) On the origin of NMR dipolar waves in transient helical elements of partially folded proteins. J Am Chem Soc 130:11266–11267
17. Jensen MR, Houben K, Lescop E, Blanchard L, Ruigrok RW, Blackledge M (2008) Quantitative conformational analysis of partially folded proteins from residual dipolar couplings: application to the molecular recognition element of Sendai virus nucleoprotein. J Am Chem Soc 130:8055–8061
18. Horn AHC, Sticht H (2010) Amyloid-beta 42 oligomer structures from fibrils: a systematic molecular dynamics study. J Phys Chem B 114:2219–2226
19. Zagrovic B, van Gunsteren WF (2007) Computational analysis of the mechanism and thermodynamics of inhibition of phosphodiesterase 5A by synthetic ligands. J Chem Theory Comput 3:301–311
20. Liu J, Nussinov R (2010) Rbx1 flexible linker facilitates Cullin-RING ligase function before neddylation and after deneddylation. Biophys J 99:736–744
21. Steinbach PJ (2004) Exploring peptide energy landscapes: a test of force fields and implicit solvent models. Proteins 57:665–677
22. Kim S, Takeda T, Klimov DK (2010) Mapping conformational ensembles of a beta oligomers in molecular dynamics simulations. Biophys J 99:1949–1958
23. Curcó D, Alemán C (2003) Simulation of dense amorphous polymers by generating representative atomistic models. J Chem Phys 119:2915–2922
24. Curcó D, Alemán C (2004) Performance of SuSi: a method for generating atomistic models of amorphous polymers based on a random search of energy minima. J Comput Chem 25:790–798
25. Curcó D, Laso M, Alemán C (2004) Generation-relaxation algortihms to construct representative atomistic models of amorphous polymers: influence of the relaxation method. J Phys Chem B 108:20337–20339
26. Cheng Y, LeGall T, Oldfield CJ, Mueller JP, Van YY, Romero P, Cortese MS, Uversky VN, Dunker AK (2006) Rational drug design via intrinsically disordered protein. Trends Biotechnol 24:435–442

27. Follis AV, Hammoudeh DI, Wang H, Prochownik EV, Metallo SJ (2008) Structural rationale for the coupled binding and unfolding of the c-Myc oncoprotein by small molecules. Chem Biol 15:1149–1155
28. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA (1995) A second generation force-field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 117:5179–5197
29. Kang YK, Byun BJ (2007) Conformational preferences and cis-trans isomerisation of azaproline residue. J Phys Chem B 111:5377–5385
30. Flores-Ortega A, Jiménez AI, Cativiela C, Nussinov R, Alemán C, Casanovas J (2008) Conformational preferences of α-substituted proline analogues. J Org Chem 73:3418–3427
31. Curcó D, Alemán C (2004) Theoretical strategy to provide atomistic models of comblike polymers: a generation algorithm combined with configurational bias Monte Carlo. J Chem Phys 121:9744–9752
32. Curcó D, Rodríguez-Ropero F, Alemán C (2006) Force-field parametrization of retro-inverso modified residues: development of torsional and electrostatic parameters. J Comput Aided Mol Des 20:13–25
33. Skrabana R, Sevcik J, Novak M (2006) Intrinsically disordered proteins in the neurodegenerative processes: formation of tau protein paired helical filaments and their analysis. Cell Mol Neurobiol 26:1085–1097
34. Jeganathan S, von Bergen M, Mandelkow EM, Mandelkow E (2008) The natively unfolded character of tau and its aggregation to Alzheimer-like paired helical filaments. Biochemistry 47:10526–10539
35. Curran J, Marq JB, Kolakofsky D (1995) An N-terminal domain of the Sendai paramyxovirus P protein acts as a chaperone for the NP protein during the nascent chain assembly step of genome replication. J Virol 69:849–855
36. Blanchard L, Tarbouriech N, Blackledge M, Timmins P, Burmeister WP, Ruigrok RW, Marion D (2004) Structure and dynamics of the nucleocapsid-binding domain of the Sendai virus phosphoprotein in solution. Virology 319:201–211
37. Kobarg J, Schnittger S, Fonatsch C, Lemke H, Bowen MA, Buck F, Hansen HP (1997) Characterization, mapping and partial cDNA sequence of the 57-kD intracellular Ki-1 antigen. Exp Clin Immunogenet 14:273–280
38. Nery FC, Bressan GC, Alborghetti MR, Passos DO, Kuniyoshi TM, Ramos CH, Oyama S Jr, Kobarg J (2006) A spectroscopic analysis of the interaction between the human regulatory proteins RACK1 and Ki-1/57. Biol Chem 387:577–582
39. Gast K, Damaschun H, Eckert K, Schulze-Forster K, Maurer HR, Müller-Frohne M, Zirwer D, Czarnecki J, Damaschun G (1995) Prothymosin alpha: a biologically active protein with random coil conformation. Biochemistry 34:13211–13218
40. Uversky VN, Gillespie JR, Millett IS, Khodyakova AV, Vasiliev AM, Chernovskaya TV, Vasilenko RN, Kozlovskaya GD, Dolgikh DA, Fink AL, Doniach S, Abramov VM (1999) Natively unfolded human prothymosin alpha adopts partially folded collapsed conformation at acidic pH. Biochemistry 38:15009–15016
41. Boggs JM (2006) Myelin basic protein: a multifunctional protein. Cell Mol Life Sci 63:1945–1961
42. Majava V, Wang C, Myllykoski M, Kangas SM, Kang SU, Hayashi N, Baumgärtel P, Heape AM, Lubec G, Kursula P (2010) Structural analysis of the complex between calmodulin and full-length myelin basic protein, an intrinsically disordered molecule. Amino Acids 39:59–71
43. Kong M, Fox CJ, Mu J, Solt L, Xu A, Cinalli RM, Birnbaum MJ, Lindsten T, Thompson CB (2004) The PP2A-associated protein α4 is an essential inhibitor of apoptosis. Science 306:695–698
44. Smetana JH, Oliveira CL, Jablonka W, Aguiar Pertinhez T, Carneiro FR, Montero-Lomeli M, Torriani I, Zanchin NI (2006) Low resolution structure of the human alpha4 protein (IgBP1) and studies on the stability of alpha4 and of its yeast ortholog Tap42. Biochim Biophys Acta 1764:724–734
45. Gálvez A, Valdivia E, Abriouel H, Camafeita E, Méndez E, Martínez-Bueno M, Maqueda M (1998) Isolation and characterization of enterocin EJ97, a bacteriocin produced by Enterococcus faecalis EJ97. Arch Microbiol 171:59–65
46. Neira JL, Contreras LM, de los Paños OR, Sánchez-Hidalgo M, Martínez-Bueno M, Maqueda M, Rico M (2010) Structural characterisation of the natively unfolded enterocin EJ97. Protein Eng Des Sel 23:507–518
47. Goldenberg DP (2003) Computational simulation of the statistical properties of unfolded proteins. J Mol Biol 326:1615–1633
48. Panick G, Malessa R, Winter R, Rapp G, Frye KJ, Royer CA (1998) Structural characterization of the pressure-denatured state and unfolding/refolding kinetics of staphylococcal nuclease by synchrotron small-angle X-ray scattering and Fourier-transform infrared spectroscopy. J Mol Biol 275:389–402
49. Koide S, Bu Z, Risal D, Pham TN, Nakagawa T, Tamura A, Engelman DM (1999) Multistep denaturation of Borrelia burgdorferi OspA, a protein containing a single-layer beta-sheet. Biochemistry 13:4757–4767
50. Carrasco B, Garcia de la Torre J (1999) Hydrodynamic properties of rigid particles. Comparison of different modelling and computational procedures. Biophys J 75:3044–3057
51. Uversky VN (2002) Natively unfolded proteins: a point where biology waits for physics. Protein Sci 11:739–756
52. Bartlett AI, Radford SE (2009) An expanding arsenal of experimental methods yields an explosion of insights into protein folding mechanisms. Nat Struct Mol Biol 16:582–588
53. Mittag T, Forman-Kay JD (2007) Atomic-level characterization of disordered protein ensembles. Curr Opin Struct Biol 17:3–14
54. Xu XP, Case DA (2001) Automated prediction of ^{15}N, ^{13}Ca, ^{13}Cb and $^{13}C'$ chemical shifts in proteins using a density functional database. J Biomol NMR 21:321–333
55. Pérez JJ, Villar HO, Arteca GA (1994) Distribution of conformational energy minima in molecules with multiple torsional degrees of freedom. J Phys Chem 98:2318–2324
56. Wilkins DK, Grimshaw SB, Receveur V, Dobson CM, Jones JA, Smith LJ (1999) Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques. Biochemistry 38:16424–16431
57. Sherman E, Haran G (2006) Coil–globule transition in the denatured state of a small protein. Proc Natl Acad Sci USA 103:11539–11543
58. Cho MK, Kim HY, Bernado P, Fernandez CO, Blackledge M, Zweckstetter M (2007) Amino acid bulkiness defines the local conformations and dynamics of natively unfolded alpha-synuclein and tau. J Am Chem Soc 129:3032–3033