BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 5 2014, pages 668–674
doi:10.1093/bioinformatics/btt586

Structural bioinformatics

Advance Access publication October 15, 2013

SecStAnT: secondary structure analysis tool for data selection, statistics and models building

Giuseppe Maccari[1,*], Giulia L.B. Spampinato[2,3] and Valentina Tozzini[2]

[1]Center for Nanotechnology and Innovation @NEST, Istituto Italiano di Tecnologia, [2]NEST, Istituto Nanoscienze – CNR and Scuola Normale Superiore, Piazza San Silvestro 12-56127 Pisa and [3]Dipartimento di Fisica 'E. Fermi', Università di Pisa Largo B. Pontecorvo 3-56127 Pisa, Italy

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: Atomistic or coarse grained (CG) potentials derived from statistical distributions of internal variables have recently become popular due to the need of simplified interactions for reaching larger scales in simulations or more efficient conformational space sampling. However, the process of parameterization of accurate and predictive statistics-based force fields requires a huge amount of work and is prone to the introduction of bias and errors.

Results: This article introduces SecStAnT, a software for the creation and analysis of protein structural datasets with user-defined primary/secondary structure composition, with a particular focus on the CG representation. In addition, the possibility of managing different resolutions and the primary/secondary structure selectivity allow addressing the mapping-backmapping of atomistic to CG representation and study the secondary to primary structure relations. Sample datasets and distributions are reported, including interpretation of structural features.

Availability and implementation: SecStAnT is available free of charge at secstant.sourceforge.net/. Source code is freely available on request, implemented in Java and supported on Linux, MS Windows and OSX.

Contact: giuseppe.maccari@iit.it

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 22, 2013; revised on September 17, 2013; accepted on October 5, 2013

1 INTRODUCTION

Molecular dynamics (MD) computer simulations, and more specifically force field (FF)-based atomistic MD (Adcock and McCammon, 2006), are considered invaluable tools to get insight in the structure and function of biological matter. Within this approach, the inter-atomic interactions are represented by means of a sum of analytical terms, whose parameters were optimized in the course of the past decades, based on quantum chemistry calculations or experimental data. This approach is implemented in a number of widely used software packages (van Gunsteren, 1996; Jorgensen *et al*., 1996; Vanommeslaeghe *et al*., 2010; Wang *et al*., 2004).

Despite its undoubted utility, the atomistic MD presents some weaknesses. Single proteins simulations can now reach the sub-μs scale with ordinary computational resources. However, most biologically interesting phenomena occur on wider time and space scales, needing large parallelism. This problem is not likely to be simply resolved by the increase of the processors power and of parallelism, becoming increasingly harder as the system complexity grows. Recent efforts have focused on the development of dedicated hardware. An example is the supercomputer Anton (Shaw *et al*., 2008), which implements specialized hardware for protein dynamics, leading to simulation time scales into the range of hundreds of micro seconds to milliseconds. However, such systems are not broadly available to the scientific community. A second problem of the atomistic approach is related to the model itself. As longer time scales are explored in the simulations, the traditional FFs show inaccuracies especially in the evaluation of the relative energies of different secondary structures (Freddolino *et al*., 2008; Lindorff-Larsen *et al*., 2012). A great effort is currently in the course to produce a new generation of FFs to fix these problems, although this task appears hard without the introduction of more complex interactions with larger number of parameters (Chaudret *et al*., 2013; Zhao *et al*., 2010). This, in turn, worsens a problem already existing in the traditional FFs, i.e. the complexity of the optimization procedure.

Apparently paradoxically, the reductionist approach has recently been considered as a possible alternative. Minimizing the number of parameters of the model allows applying more efficient parameters optimization strategies to accurately reproduce given properties. Direct emanations of this approach are the coarse grained (CG) models, representing group of atoms with single interacting centers (beads) (Tozzini, 2005) and the so called 'knowledge based' or statistical potentials (SP) (Tadmor *et al*., 2011), i.e. potentials with a relatively small number of parameters, derived by the statistical analysis of the increasingly larger experimental structures databases. CG models solve directly the first class of problems, as they immediately reduce the computational cost of orders of magnitude. On the other hand, SPs, though bearing many limitations specifically residing in the difficulty of combining transferability and predictive power with structural accuracy (Vendruscolo and Domany, 1998), have shown better performance than traditional atomistic FFs for docking or homology modeling applications (Poole and Ranganathan, 2006).

The combination of CG with SPs has been used in models for MD simulations, such as MARTINI (Marrink *et al*., 2007),

*To whom correspondence should be addressed.

representing the amino acid at an intermediate resolution level (with 2–6 beads), embedded in CG explicit water, or the one by Bahar and Jernigan, with a similar resolution but with implicit water(Jernigan and Bahar, 1996), or the one developed by us (Tozzini *et al*., 2006), based on a single bead per amino acid (placed on Cα) in implicit water. This resolution level can be considered the minimal where internal variables can still explicitly describe secondary structures, and therefore called 'minimalist'.

Specifically referring to minimalist models (Tozzini, 2010a), different algorithms were considered to produce SPs, such as direct or iterative Boltzmann inversion (BI) (Reith *et al*., 2003), relative entropy minimization (Chaimovich and Shell, 2011) and reverse Monte Carlo (Lyubartsev and Laaksonen, 1995). They all rely on the statistical distributions of the internal variables, either used as direct input or as target quantity for the potential optimization. This implies that the quality of statistical distribution determines the accuracy of the model (Trovato and Tozzini, 2012). In turn, the quality of the statistical distribution is determined by the statistical relevance of the dataset (i.e. number and diversity of included structures) and its composition in terms of sequence or secondary structures. The latter in particular is relevant for the parameterization of potentials capable of accurately reproducing the secondary structure tendency of different amino acids.

The RSCB Protein Data Bank (PDB) (Rose *et al*., 2013), the most comprehensive database of biomolecular—and specifically proteins—structures, is the natural source of data for building statistical sets and corresponding probability distributions. Biomolecules coordinates are stored in a format that is a widely used, internationally referred representation for macromolecular data, including experimental and structural information. These, however, are integrated within the coordinates file, and not of immediate use to the aim of building, e.g. primary or secondary structure-dependent dataset.

In this article, we describe and validate SecStAnT, a tool with an intuitive and sleek interface able to automatically create from PDB user-defined datasets of protein structural composition or primary sequence motives at different levels of resolution (atomistic or CG). Furthermore, a large number of internal variable distributions can be evaluated together with two and three body correlation functions. The latter point is particularly innovative and useful for the parameterization and validation of the CG models. In fact, the correlation map between the internal variables describing the backbone conformation within the CG representation has the same role of the well-known Ramachandran plot (Tozzini *et al*., 2006). Although there are a number of tools to evaluate the latter [for instance (Gopalakrishnan *et al*., 2007)], to our knowledge, none are freely available to evaluate the corresponding correlation maps within the minimalist representations. The ability to evaluate the SP by means of BI facilitates the parameterization process of CG models. In addition, the possibility to consider both atomistic and CG resolutions allows in principle to directly make connections between the two levels. This is particularly important in view of generating CG models fully compatible with atomistic FFs to be included in a coherent multiscale representation, which are often considered as possible solutions to combine the advantages of CG and atomistic representation and eliminate their disadvantages (Colombo and Micheletti, 2005; Tozzini, 2010b).

We illustrate SecStAnT and its potentialities reporting sample datasets distributions and correlations. Interesting novel features emerge from this statistical analysis to which we give a physical-chemical interpretation.

2 METHODS

2.1 Model definition

Three typical resolution levels used to represent a protein model are reported in Figure 1: (i) the fully atomistic; (ii) the 'backbone-only' and (iii) the 'minimalist' (Cα only). In the latter case, the internal variables defining the backbone conformation are the bond angle θ between three subsequent Cαs and the dihedral φ between four subsequent Cαs (see Fig. 1C). These are the homologues of the Φ and Ψ dihedrals defined within the atomistic representation of the backbone. Consequently, the (θ, φ) correlation map can be considered the homolog of the Ramachandran plot, which is, in fact, the (Φ, Ψ) correlation map (Tozzini *et al*., 2006). Other important internal variables, which are relevant to the CG representation and specifically to the minimalist one, are the distances between Cα in general, and specifically those separated by a given number of amino acids (i.e. r_{14}, r_{15}, ...). SecStAnT is able to treat every custom subset of atoms, including atomistic (A), minimalist (C) and also intermediate representation, such as the (B) and others. Because the main focus is on the minimalist, a large number of distribution calculations for the minimalist representation are implemented.

2.2 Statistical analysis and normalization

Three classes of statistics calculations are available: single variable distributions, two- and three-variables correlation maps (hereafter 2D and 3D maps). A list of distributions and correlations is represented in Table 1. Single variables distributions can be done of any of the defined internal variables, and in some cases of their complementary (e.g. the distribution of 'non-bonded' beads, which are the complementary of the $r_{i,i+1}$, ..., $r_{i,i+3}$ with respect to the set of all $r_{i,j}$s). The primary output data are the non-normalized occurrences, i.e. the ΔNs:

$$\Delta N_i = N(x_i)\Delta x_i \tag{1}$$

$$\sum_i N(x_i)\Delta x_i = N_{tot} \tag{2}$$

with N_{tot} the total number of occurrences of a given variable within the dataset and Δx_i is the width of the histogram intervals (bins). Besides the raw data ΔN_i, additional differently normalized are of interest. One is the normalized relative occurrence, tending to the probability distribution in the $\Delta x_i \rightarrow 0$, $\Delta N_i \rightarrow \infty$ limit

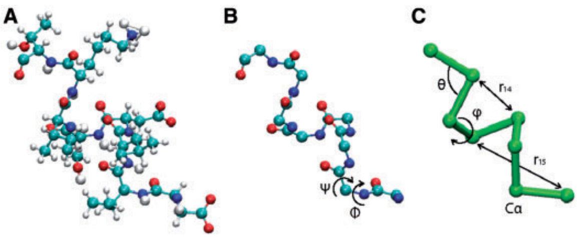

Fig. 1. An illustration of proteins representations available in SecStAnT: (**A**) atomistic representation, (**B**) backbone-only representation and (**C**) minimalist (Cα-only) representation. Relevant internal variables are indicated in B and C

Table 1. List of available distributions and correlations

	Name	Description
1D	$r_{1,1+n}$ ($1 \le n \le 6$)	Cα distance distributions
	θ	Bond angle distribution
	φ	Dihedral angle distribution
	g(r) total	g(r) distribution
	g(r) non-bonded	Distribution of non-bonded g(r)
2D	(θ_-, θ_+)	Theta$^+$, theta$^-$ angles correlation
	(φ, θ_-)	Phi, Theta$^-$ angles correlation
	(φ, θ_+)	Phi, Theta$^+$ angles correlation
	($r_{1\text{–}3}$, θ)	$r_{1\text{–}3}$, Theta- angles correlation
	(Φ,Ψ)	Ramachandran's correlation plot
3D	($r_{1\text{–}4}$, φ, θ_-)	$r_{1\text{–}4}$, Phi, Tetha$^-$ angles correlation
	($r_{1\text{–}4}$, φ, θ_+)	$r_{1\text{–}4}$, Phi, Tetha$^+$ angles correlation

$$\frac{N(x_i)}{N_{tot}} \approx P(x_i) \tag{3}$$

In addition, especially to the aim of using the distribution for the FFs parameterization, it is often useful to separate the purely geometric effect defining

$$\overline{P}(x_i) = \frac{P(x_i)}{P_0(x_i)} = \frac{N(x_i)}{N_0(x_i)} \tag{4}$$

where N_0 (P_0) is the probability distribution in the non-interacting particles system (ideal gas), which can be evaluated analytically in some cases. For instance, if x is $r_{i,j}$, then N_0 is the number of uniformly distributed particles at distance r from a given one, and one has

$$N_0(r) \propto \rho 4\pi r^2 \qquad \overline{P}(r) \propto \frac{N(r)}{\rho 4\pi r^2} = g(r) \tag{5}$$

where g(r) is called the pair distribution function and contains the same information as P or N but, having the ideal part extracted, reflects specifically the effect of the interactions. The g(r) here defined with $N_0 = 4\pi r^2$ is the one corresponding to the ideal infinite gas, with standard normalization [$g(r) \to 1$ for $r \to \infty$]. However, when applied to finite size systems, it brings uneven behavior of g(r), making it vanishing at large r values. Besides the standard normalization, two additional are implemented accounting for the finite size of the proteins:

$$N_0(r) = Cr^{\gamma} \tag{6}$$

$$N_0(r) = \rho 4\pi r^2 \left[1 - \left(\frac{3}{4}\right)\left(\frac{r}{R}\right) + \left(\frac{1}{16}\right)\left(\frac{r}{R}\right)^3\right] \tag{7}$$

where in Equation (6), $\gamma \sim 1.5$, fitted on a statistical dataset of large proteins (Zhou and Zhou, 2002). The normalization in Equation (7) corresponds to the distribution of the ideal gas confined in a sphere of radius R, analytically evaluated (reducing to $N_0 = \rho 4\pi r^2$ for small r). This normalization works well if the proteins in the dataset are not excessively dispersed, R being their average gyration radius (Tozzini, 2010a).

Another noticeable case of non-trivial N_0 is the bond angle, for which $N_0(\theta) \propto \sin(\theta)$. In any case, the ΔN_i, the *P*s and the $\overline{P}$s are given in the output (see the Supplementary Material for additional details).

Two variables (2D) and three variables (3D) maps are implemented for combinations of variables particularly relevant to the secondary structure analysis. Included in this list, there are the (θ, φ) map for the Cα representation, and the (Φ, Ψ) (Ramachandran plot) for the backbone-only representation. In this case, the output is only the ΔN_i, and the same normalized to its maximum value, which is a convenient normalization for visualization of 2D and 3D maps.

2.3 Potentials generation and other applications

Statistical distributions can be used for the generation of the FFs, specifically in the case of the minimalist model. A commonly used representation of the FFs for these models is (Tozzini, 2010a; Trovato and Tozzini, 2012):

$$U = U_b + U_\theta + U_\phi + U_{loc} + U_{nloc} \tag{8}$$

where U_{loc} and $U_{non\text{-}loc}$ are the local and non-local parts of the non-bonded interactions. Depending on the model, the separation between U_{loc} and $U_{non\text{-}loc}$ can be based either on physical-chemical criteria, or on geometrical-structural criteria or a combination of them (Tozzini *et al.*, 2007). In any case, at least the $r_{i,i+n}$ distances with i = 4,5,6 are usually included in the local part, thus an expansion of the FFs terms reads:

$$\begin{aligned} U_b &= \sum_i u_b(r_{i,i+1}) \\ U_\theta &= \sum_i u_\theta(\theta_i) \\ U_\phi &= \sum_i u_\phi(\phi_i) \\ U_{loc} &= \sum_{i,n=4,5,6} u_{loc,n}(r_{i,i+n}) \\ U_{nloc} &= \sum_{i,j>i+6} u_{nloc}(r_{i,j}) \end{aligned} \tag{9}$$

A rough first approximation to evaluate FF terms of a given internal variable is to consider the potential of mean force, namely:

$$w(x) = -kT \ln \overline{P}(x) \approx -kT \ln \left(\frac{N(x)}{N_0(x)}\right) \tag{10}$$

where x is any of the variables on which the FF terms depend, kT the Boltzmann factor. This is defined the BI and gives an operative way to evaluate numerically the interactions in a system on statistical basis (the numerical *w* can subsequently be fitted with an analytical form). $w(x)$ can be obtained with a single passage from Equation (4) and is delivered in the output files.

3 RESULTS

3.1 Workflow

The program is roughly composed of two modules, as described in Figure 2. The first one, the parsing module, performs the dataset building, extracting structures from PDB and fragmenting them in elements with defined secondary structures. The input selection is performed through a graphical interface by combining secondary structure information with any other selection criterion available on the Research Collaboratory for Structural Bioinformatics (RCSB) advanced search interface as, for instance, the experimental method for the structure determination, the release year and so forth. The downloading process is performed through RCSB FTP interface, according to the server guidelines; a cache mechanism is implemented to avoid multiple downloads of a single entry. As anticipated, the queries consist of secondary structures composition and sequence motifs. Secondary structures can be selected either based on the information included into the PDB file itself (provided by the PDB file author) (Berman *et al.*, 2003) or on the DSSP file [provided by RSCB and based on the DSSP algorithm (Andersen *et al.*, 2002)]. Secondary structures identified by the two algorithms are listed in Supplementary Table S1 in the Supplementary Material. Primary structure is defined by standard regular expression search. During the extraction process, information on primary, secondary and super-secondary structures (when available) is mined and stored. Either the whole proteins or only the structure

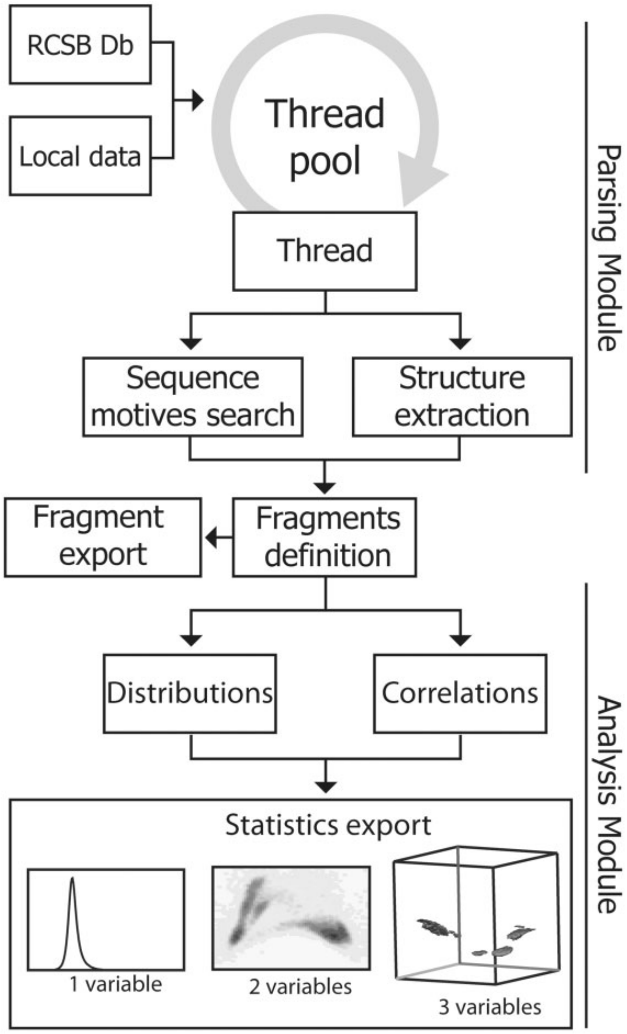

Fig. 2. Schematic illustration of the SecStAnT workflow. The process is separated in two modules. In the parsing module, input data are downloaded and processed by a thread pool. Each entry is fragmented by user-defined primary and secondary structure criterions. In the analysis module, each fragment is then saved separately and a series of statistics is calculated

fragments with the selected secondary structure can be stored in hierarchical organized folders for future consultation (see Supplementary Material for the detailed description of the output dataset organization). In the analysis module, the fragments dataset is used to build different kinds of distributions of internal variables and their correlations. A description of the statistical analysis process as well as sample distributions and correlations are discussed below. The output format (described in detail in the Supplementary Material) is given in numerical form, conveniently readable by a large number of commonly used graphics software packages.

3.2 Structural dataset

Two sample datasets were built, one for α helical fragments and one for unstructured fragments. Each dataset was separated in two subsets of X-ray crystallographic and nuclear magnetic resonance (NMR)-derived structures, although only distributions of X-ray data are shown (the others are given in the Supplementary Material). For each database query, search results were filtered by the RCSB server on the basis of a similarity threshold of 30%. To further limit the unstructured dataset size, we additionally selected those entries released after the year 2001. The data were saved at the ‘minimalist’ resolution level. The search resulted in a total of 16 400 structures for the X-ray α helical structures, 3550 for NMR α helical, 11 672 for X-ray unstructured and 3839 for NMR unstructured. Detailed RCSB queries are reported in Supplementary Material.

3.3 Internal variables distribution

Sample distributions and maps evaluated on the datasets described in the previous section are here reported. Figure 3 reports the single variable distributions (evaluated on X-ray datasets).

The difference between red and black lines reflects the secondary structure difference, the black lines representing the α-helices dataset and the red lines the unstructured proteins dataset. For the helices, the local variables distributions of (θ, φ, $r_{1\text{–}3}$ and $r_{1\text{–}4}$) are single peaked and little dispersed, the signature of the local order. In particular, the α-helix is characterized by $\theta \approx 91$ deg, $\varphi \approx 50$ deg. Conversely, those of the unstructured proteins are multimodal and more disordered.

The distribution of the non-bonded distances (Fig. 3D) shows similar differences concerning the comparison between helical and unstructured datasets, although the distribution itself in both cases is more complex.

It can be observed that the θ distributions in panel A show a striking similarity with the $r_{1\text{-}3}$ distributions, the solid lines in panel C, which is obvious considering that the two variables are related by $r_{1\text{–}3} = 2l\sin(\theta/2)$ (l is the Cα-Cα distance). In fact, this relationship is also directly measured by the ($r_{1\text{–}3}$, θ) correlation map reported in the Supplementary Material.

3.4 Internal variables correlation

A comparison between the correlation maps of the conformational variables of the all-atom and CG models is reported in Figure 4, where the (θ, φ) correlation map of α helices and unstructured proteins (Panels A and B, respectively) is compared with the Ramachandran plots (Φ, Ψ correlation map) of the same datasets (Panel C and D).

The (θ, φ) correlation map can be considered the equivalent of the Ramachandran plot for the minimalist representation (Tozzini, 2010a). In fact, the helices plot shows a peaked concentration in a specific area both in Panel A and C (green area), defining the ‘helical region’. The helical peak is also present in the unstructured proteins plot (Panel B and D), which, however, also shows occurrence in other areas, corresponding to extended structures (around $\theta = 130$, $\varphi \approx -175$ and $\Phi = -135$, $\Psi = 135$, blue area), turns and coils, as well as out of the region corresponding to defined secondary structures, as expected. The ‘unstructured’ proteins Ramachandran map (Fig. 4D) shows a more broad population of all the allowed areas, with concentration in all the secondary structures areas (delimited by green, red and blue lines, representing the right-handed, and left-handed helices

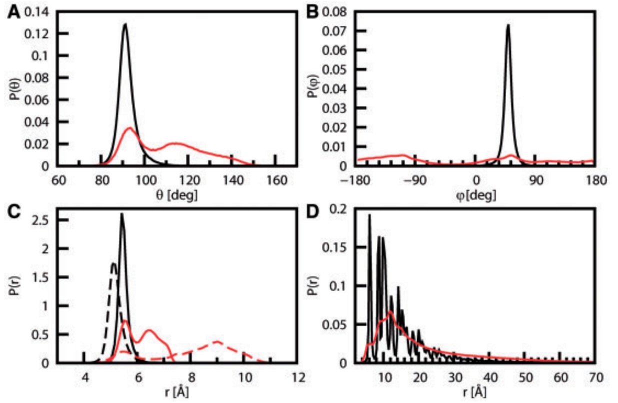

Fig. 3. X-ray distributions of internal variables. (**A**) θ; (**B**) φ; (**C**) $r_{1\text{–}3}$ (solid lines) and $r_{1\text{–}4}$ (dashed lines); and (**D**) r 'non–non-bonded' (r_{ij}, with $j>i+3$). Black lines: α helices dataset, red lines: unstructured proteins

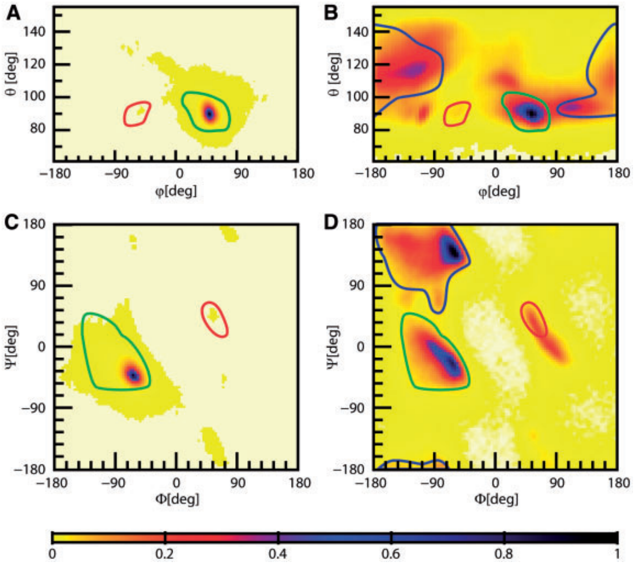

Fig. 4. Two variables correlation maps. (**A**) $\theta-$, φ map of the X-ray PDB α helices; (**B**) $\theta-$, φ map of the X-ray PDB unstructured proteins; (**C**) Ramachandran plot (Φ, Ψ) of the X-ray PDB α helices; and (**D**) Ramachandran plot (Φ, Ψ) of the X-ray PDB unstructured proteins. The color bar is reported at the bottom

and extended structures), because of the residual presence of folded secondary structures even in the unstructured proteins. The corresponding areas are well visible and separated also in the (θ, φ) correlation map, confirming its usability in the analysis of the secondary structures as the Ramachandran map. Additional concentrations are visible in the two maps out of the secondary structures regions, related to unstructured transition or random conformations. It must be noted that in the (θ, φ) correlation map, two subsequent θ are involved with any given φ. As a consequence, two possible maps can be evaluated: N($\theta-$, φ) between a dihedral and the preceding bond angle and N($\theta+$, φ) between a dihedral and the following bond angle. These are different due to the directionality of the polypeptide (see a comparison in Supplementary Material). A detailed discussion about this point is beyond the scope of this article and is addressed elsewhere (Spampinato *et al.*, in preparation).

It is interesting to note that, with respect to the Ramachandran map, the (θ, φ) map has a more direct interpretation: the φ dihedral directly represents the helicity, thus $\varphi = \pm$ 180 corresponds to flat structures, $\varphi = 0$ to rings, whereas positive and negative φ correspond to right- or left-handed structures with different degrees of helicity. For instance, the presence of a populated area at $\varphi \sim -170$ and $\theta \sim 120$ indicates that the extended structures tend to have a weak left-handed torsionality. This representation allows more immediately to identify different kind of helices.

More detailed information for the unstructured proteins dataset is given in the ($r_{1\text{-}4}$, θ, φ) 3D map represented in Figure 5. The relations between $r_{1\text{–}4}$ distribution and the previously defined correlation are:

$$P(\theta,\phi) \propto \int P(\theta,\phi,r_{1-4})dr_{1-4} \quad (11)$$

$$P(r_{1-4}) \propto \int P(\theta,\varphi,r_{1-4})d\theta d\varphi \quad (12)$$

In the 3D map, the highly populated regions distributed in the volume can be visualized with iso-values surfaces (in gray, in the upper part of Fig. 5), making the separation between secondary structures even more immediate than in the 2D map. Again this is a consequence of choosing immediately physically interpretable variables for the 3D map building. In fact, the $r_{1\text{–}4}$ is a particularly important variable especially in certain kind of helices, being associated to the formation of local hydrogen bonds stabilizing certain kind of helices and turns.

For this reason, an alternative visualization of the 3D map by means of the iso-variable sections, e.g. the iso-$r_{1\text{–}4}$ (lower part of Fig. 5) is also particularly interesting. Figure 5 reports the sections corresponding to three relevant values of the single variable $r_{1\text{–}4}$ distribution (red dots A, B, C in the top right plot). 2D maps of these slices are also reported in the tree bottom plots (corresponding letters) each with its colors bar. By definition, the single variable $r_{1\text{–}4}$ distribution (right upper plot) is the (renormalized) integral over θ and φ of the 3D map. The 3D representation is generated with Visual Molecular Dynamics (VMD) software from the CUBE file. The surface corresponding to the helical region (residually populated also in the 'unstructured' dataset) is a roughly ellipsoid shape located at $r_{1\text{–}4} \sim 5.75$ Å. This is also confirmed by the $r_{1\text{–}4} = 5.75$ section (plot A Fig. 5), in which a high concentration in the helical area is observed. In this plot, one can also observe an upside-down parabolic shape is populated (red-blue shades). This kind of correlation is, in fact, induced among the variables θ and φ by keeping constant the $r_{1\text{–}4}$ (see Trovato and Tozzini, 2012).

At higher levels of $r_{1\text{–}4}$ other structures appear, first a transition region (plot B) and then the extended structures region (plot C). We defer a discussion of the structural meaning of the information present in the 2D and 3D map to a forthcoming work. Other sample 2D and 3D maps involving different variables are reported in the Supplementary Material.

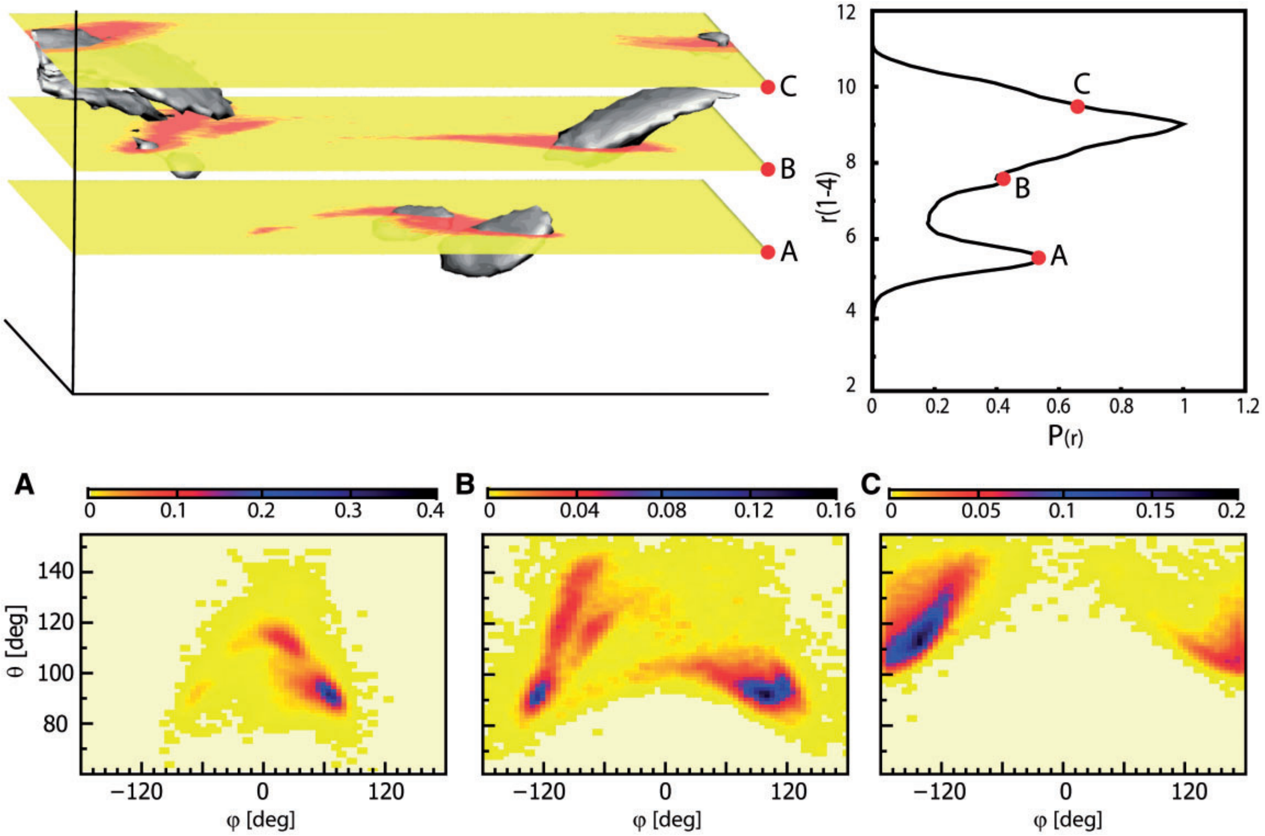

Fig. 5. r_{1-4}, θ, φ map for the X-ray PDB unstructured proteins. An iso-surface (level = 120) is represented in gray and three r_{1-4} = const sections are in color. The three r_{1-4} values are chosen corresponding to three relevant values of the single variable r_{1-4} distribution (red dots A, B, C in the top right plot). 2D maps of these slices are also reported in the tree bottom plots (corresponding letters) each with its colors bar. By definition, the single variable r_{1-4} distribution (right upper plot) is the (renormalized) integral over θ and φ of the 3D map. The 3D representation is generated with VMD from the CUBE file

4 CONCLUSIONS

SecStAnT is an efficient and flexible software tool to create selective databases of structures extracted from the PDB and to calculate statistical distribution of internal variables. The focus of selection criteria is on the secondary and primary structure, for which accurate algorithms are considered. Additionally, all the selection criteria for the initial input data, implemented in the RCSB 'advanced query' form, are available. Despite the focus on the minimalist CG representation, SecStAnT can be as well used for different CG models with a compatible Cα-based backbone representation, like the popular MARTINI (Marrink *et al.*, 2007). Statistical distributions and correlations of a number of internal variables can be performed with different normalization, allowing the generation of SP by BI.

Moreover, this software was thought as a part of a larger package for proteins (and other biomolecules) multiscale modeling. Consequently, it was designed to be extended to include advanced techniques for SP generation, such as iterative BI, multivariable potential generation. The ability to the directly evaluate correlations among variables gives the possibility to obtain more accurate potentials correcting the potential of mean forces by subtracting the correlation effects. In addition, given the capability of SecStAnT to produce highly selective distributions, the primary and secondary structure selectivity can be easily introduced into these potentials.

Apart from SP parameterization, another important class of SecStAnT applications concerns the secondary structure tendency evaluation of amino acids or sequence motifs, i.e. the problem of prediction of secondary structure from primary sequence. SecStAnT currently offers the possibility to address it at the minimalist level, with the advantage of simplifying the problem, and possibly including additional information by means of 3D maps. The Ramachandran plot indicates a limited and confined set of conformations, governed by steric overlap between atoms in the side chains of adjacent residues (Betancourt and Skolnick, 2004). Using this information, structure prediction methods allow a limited search in areas of the conformations space where the correct conformation is most likely to be (Keskin *et al.*, 2004). The possibility to calculate Ramachandran plots of extremely well-defined dataset of secondary structure fragments permit to have informative statistics. Furthermore, SecStAnT allows to calculate the probability distribution of (θ, φ) and (Φ, Ψ) for Cα and backbone-only representations, respectively, giving the ability to address the mapping-backmapping of atomistic to CG representation. In conclusion, SecStAnT is designed with the aim to facilitate the extraction of information from proteins structures datasets, aid the parameterization of statistics-based potentials and investigate the sequence–structure relationship. Furthermore, the expansion of atomistic-level statistics, as well as the introduction of new minimalistic representations will allow in principle to realize accurate, transferable and predictive potentials for multiscale models. More generally, it can be considered as tool to find useful directions navigating the continuously expanding ocean of protein structures.

ACKNOWLEDGEMENTS

The authors thank Dr Fabio Trovato and Dr Paolo Mereghetti for useful discussions and for help in the software design and programing. A special thanks to the Platform ‘Computation’ of IIT (Italian Institute of Technology) for the allocation of computer resources.

Conflict of Interest: None declared.

REFERENCES

Adcock,S.A. and McCammon,J.A. (2006) Molecular dynamics: survey of methods for simulating the activity of proteins. *Chem. Rev.*, **106**, 1589–1615.

Andersen,C.A.F. *et al.* (2002) Continuum secondary structure captures protein flexibility. *Structure*, **10**, 175–184.

Berman,H. *et al.* (2003) Announcing the worldwide Protein Data Bank. *Na. Struct. Biol.*, **10**, 980.

Betancourt,M.R. and Skolnick,J. (2004) Local propensities and statistical potentials of backbone dihedral angles in proteins. *J. Mol. Biol.*, **342**, 635–649.

Chaimovich,A. and Shell,M.S. (2011) Coarse-graining errors and numerical optimization using a relative entropy framework. *J. Chem. Phys.*, **134**, 094112.

Chaudret,R. *et al.* (2013) Further refinements of next-generation force fields — Nonempirical localization of off-centered points in molecules. *Can. J. Chem.*, 1–7.

Colombo,G. and Micheletti,C. (2005) Protein folding simulations: combining coarse-grained models and all-atom molecular dynamics. *Theor. Chem. Acc.*, **116**, 75–86.

Freddolino,P.L. *et al.* (2008) Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. *Biophys. J.*, **94**, L75–L77.

Gopalakrishnan,K. *et al.* (2007) Ramachandran plot on the web (2.0). *Protein Pept. Lett.*, **14**, 669–671.

Jernigan,R.L. and Bahar,I. (1996) Structure-derived potentials and protein simulations. *Curr. Opin. Struct. Biol.*, **6**, 195–209.

Jorgensen,W.L. *et al.* (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, **118**, 11225–11236.

Keskin,O. *et al.* (2004) Relationships between amino acid sequence and backbone torsion angle preferences. *Proteins*, **55**, 992–998.

Lindorff-Larsen,K. *et al.* (2012) Systematic validation of protein force fields against experimental data. *PLoS One*, **7**, e32131.

Lyubartsev,A. and Laaksonen,A. (1995) Calculation of effective interaction potentials from radial distribution functions: A reverse Monte Carlo approach. *Phys. Rev. E Stat. Phy. Plasmas Fluids Relat. Interdiscip. Topics*, **52**, 3730–3737.

Marrink,S.J. *et al.* (2007) The MARTINI force field: coarse grained model for biomolecular simulations. *J. Phys Chem. B*, **111**, 7812–7824.

Poole,A.M. and Ranganathan,R. (2006) Knowledge-based potentials in protein design. *Curr. Opin. Struct. Biol.*, **16**, 508–513.

Reith,D. *et al.* (2003) Deriving effective mesoscale potentials from atomistic simulations. *J. Comput. Chem.*, **24**, 1624–1636.

Rose,P.W. *et al.* (2013) The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res.*, **41**, D475–D482.

Shaw,D.E. *et al.* (2008) Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM*, **51**, 91.

Tadmor,E.B. *et al.* (2011) The potential of atomistic simulations and the knowledgebase of interatomic models. *JOM*, **63**, 17.

Tozzini,V. (2005) Coarse-grained models for proteins. *Curr. Opin. Struct. Biol.*, **15**, 144–150.

Tozzini,V. *et al.* (2007) Flap opening dynamics in HIV-1 protease explored with a coarse-grained model. *J. Struct. Biol.*, **157**, 606–615.

Tozzini,V. *et al.* (2006) Mapping all-atom models onto one-bead Coarse Grained Models: general properties and applications to a minimal polypeptide model. *J. Chem. Theory Comput.*, **2**, 667–673.

Tozzini,V. (2010a) Minimalist models for proteins: a comparative analysis. *Q. Rev. Biophys.*, **43**, 333–371.

Tozzini,V. (2010b) Multiscale modeling of proteins. *Acc. Chem. Res.*, **43**, 220–30.

Trovato,F. and Tozzini,V. (2012) Minimalist models for biopolymers: Open problems, latest advances and perspectives. In: *AIP Conference Proceedings*. American Institute of Physics, Pavia, Italy, pp. 187–200.

van Gunsteren,W.F. (1996) *Biomolecular Simulation: The GROMOS96 Manual und User Guide vdf*. Hochschulverlag an der ETH; BIOMOS, Zürich, Switzerland.

Vanommeslaeghe,K. *et al.* (2010) CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J. Comput. Chem.*, **31**, 671–690.

Vendruscolo,M. and Domany,E. (1998) Elusive unfoldability: learning a contact potential to fold crambin. *Fold. Des.*, **3**, 13.

Wang,J. *et al.* (2004) Development and testing of a general amber force field. *J. Comput. Chem.*, **25**, 1157–1174.

Zhao,D.-X. *et al.* (2010) Development of a polarizable force field using multiple fluctuating charges per atom. *J. Chem. Theory Comput.*, **6**, 795–804.

ZhouA,H. and Zhou,Y. (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.*, **11**, 2714–2726.