Discrete analyses of protein dynamics

Akhila Melarkode Vattekatte, Tarun Jairaj Narwani, Pierrick Craveur, Nicolas K Shinada, Aline Floch, Hubert Santuz, Akhila Melarkode Vattekatte, Narayanaswamy Srinivasan, Joseph Rebehmed, Jean-Christophe Gelly, et al.

Preprint – accepted for publication in *Journal of Biomolecular Structure and Dynamics*

Discrete analyses of protein dynamics

Tarun Jairaj Narwani[1,2,3,+], Pierrick Craveur[1,2,3,4,+], Nicolas K. Shinada[1,2,3,5,+], Aline Floch[2,6,7,8], Hubert Santuz[1,2,3], Akhila Melarkode Vattekatte[1,2,3,9], Narayanaswamy Srinivasan[10], Joseph Rebehmed[1,2,3,11], Jean-Christophe Gelly[1,2,3,9,12], Catherine Etchebest[1,2,3,9] & Alexandre G. de Brevern[1,2,3,9,12,*]

[1] Biologie Intégrée du Globule Rouge UMR_S1134, Inserm, Univ. Paris, Univ. de la Réunion, Univ. des Antilles, F-75739 Paris, France.
[2] Laboratoire d'Excellence GR-Ex, F-75739 Paris, France.
[3] Institut National de la Transfusion Sanguine (INTS), F-75739 Paris, France.
[4] Department of Integrative Structural and Computational Biology, The Scripps Research Institute, La Jolla, California, USA.
[5] Discngine, SAS, 75012, Paris, France.
[6] Etablissement Français du Sang Ile de France, Créteil, France.
[7] IMRB - INSERM U955 Team 2 « Transfusion et maladies du globule rouge », Paris Est- Créteil Univ., Créteil, France.
[8] UPEC, Université Paris Est-Créteil, Créteil, France.
[9] Faculté des Sciences et Technologies, Saint Denis Messag, F-97715 La Réunion, France.
[10] Molecular Biophysics Unit, IISc, Bangalore, India.
[11] Department of Computer Science and Mathematics, Lebanese American University, 1h401 2010 Byblos, Lebanon.
[12] IBL, F-75015 Paris, France.

[+] These authors must be considered as first authors.

Short title: Protein flexibility.

* Corresponding author:
Mailing address: Dr. Alexandre G. de Brevern, INSERM UMR_S 1134, DSIMB, Université Paris, Institut National de Transfusion Sanguine (INTS), 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France
e-mail: alexandre.debrevern@univ-paris-diderot.fr

Abstract

Protein structures are highly dynamic macromolecules. This dynamics is often analysed through experimental and/or computational methods only for an isolated or a limited number of proteins. Here, we explore large-scale protein dynamics simulation to observe dynamics of local protein conformations using different perspectives. We analysed molecular dynamics to investigate protein flexibility locally, using classical approaches such as RMSf, solvent accessibility, but also innovative approaches such as local entropy.

Firstly, we focussed on classical secondary structures and analysed specifically how β-strand, β–turns, and bends evolve during molecular simulations. We underlined interesting specific bias between β–turns and bends, which are considered as same category, while their dynamics show differences.

Secondly, we used a structural alphabet that is able to approximate every part of the protein structures conformations, namely Protein Blocks (PBs) to analyse (i) how each initial local protein conformations evolve during dynamics and (ii) if some exchange can exist among these PBs. Interestingly, the results are largely complex than simple regular/rigid and coil/flexible exchange.

Key words: local protein conformations, structural alphabet, molecular dynamics, disorder, flexibility, secondary structure, Protein DataBank, solvent accessibility.

List of abbreviations: PB: Protein Blocks, RMSf: Root Mean Square fluctuations, PDB: Protein DataBank, N_{eq}: Number of equivalent.

Introduction

More than 65 years ago, Pauling and Corey presented a series of local protein conformations stabilized by intramolecular hydrogen bonds called, shortly after, secondary structures (Pauling & Corey 1951; Pauling, Corey & Branson 1951; Eisenberg 2003). The α-helix and the β-sheet have since been extensively analysed (Richardson 1981; Richardson & Richardson 1988; Aurora & Rose 1998; Pal, Chakrabarti & Basu 2003; Craveur, Joseph, Rebehmed & de Brevern 2013). The secondary structure description of protein structures had led to the development of more than 20 assignment methodologies (Levitt & Greer 1977; Sklenar, Etchebest & Lavery 1989; Colloc'h, Etchebest, Thoreau, Henrissat & Mornon 1993; Frishman & Argos 1995; Labesse, Colloc'h, Pothier & Mornon 1997; Dupuis, Sadoc & Mornon 2004; Martin et al. 2005; Offmann, Tyagi & de Brevern 2007; Hosseini, Sadeghi, Pezeshk, Eslahchi & Habibi 2008; Tyagi, Bornot, Offmann & de Brevern 2009; Park, Yoo, Shin & Cho 2011), the most widely used being DSSP (Kabsch & Sander 1983). Analysis of local protein conformations had also been focussed on other types of secondary structures such as β-turns (Venkatachalam 1968; Hutchinson & Thornton 1996; Bornot & de Brevern 2006; de Brevern 2016), PolyProline II (Adzhubei & Sternberg 1993; Creamer 1998; Mansiaux, Joseph, Gelly & de Brevern 2011; Chebrek, Leonard, de Brevern & Gelly 2014; Narwani et al. 2017) and loops categorization (Fernandez-Fuentes, Querol, Aviles, Sternberg & Oliva 2005; Hermoso et al. 2009). Nonetheless, secondary structure analyses have also some limitations such as the *non*-definition of coil state, the characterization of some conformations or not, some known issues with short repetitive structures or the discrepancies between the different assignment algorithms (Colloc'h et al. 1993; Fourrier, Benros & de Brevern 2004; Martin et al. 2005; Tyagi et al. 2009; Tyagi, Bornot, Offmann & de Brevern 2009).

Hence, alternative views have been proposed using systematic analysis of all local protein conformations. Several scientific teams have developed local protein structure libraries able to approximate all (or almost all) local protein structures and without relying on classical secondary structures. These libraries categorize 3D structures with no preconceptions into small prototypes that are specific to local conformations found in proteins. The complete set of local structure prototypes defines a structural alphabet (de Brevern 2001; Karchin, Cline, Mandel-Gutfreund & Karplus 2003; Offmann et al. 2007). After the precursor research of Unger and co-workers (Unger, Harel, Wherland & Sussman 1989), numerous applications of structural alphabets, from the analysis of sequence-structure relationship (Rooman, Rodriguez & Wodak 1990) to the prediction of short loops (Fourrier et al. 2004) have been conducted. Protein Blocks (PBs) is actually the most used structural alphabet (de Brevern, Etchebest & Hazout 2000; Joseph et al. 2010) including several studies such as 3D protein backbone description (de Brevern 2005), local structure prediction (de Brevern et al. 2000; de Brevern, Valadie, Hazout & Etchebest 2002; Etchebest, Benros, Hazout & de Brevern 2005; de Brevern, Etchebest, Benros & Hazout 2007), description and prediction of long fragments (de Brevern, Camproux, Hazout, Etchebest & Tuffery 2001; Benros, Hazout & de Brevern 2002; de Brevern & Hazout 2003; Benros, de Brevern, Etchebest & Hazout 2006; de Brevern et al. 2007; Zimmermann & Hansmann 2008; Benros, de Brevern & Hazout 2009; Bornot, Etchebest & de Brevern 2009; Li, Zhou & Liu 2009; Rangwala, Kauffman & Karypis 2009) or short loops (Fourrier et al. 2004; Tyagi et al. 2009), analysis of protein contacts (Faure, Bornot & de Brevern 2008), building of transmembrane proteins (de Brevern et al. 2005; de Brevern, Autin, Colin, Bertrand & Etchebest 2009), definition of a reduced amino acid alphabet dedicated to mutation design (Etchebest, Benros, Bornot, Camproux & de Brevern 2007; Zuo & Li 2009; Zuo & Li 2009), protein structure superimposition and comparison

(Tyagi, Gowri, Srinivasan, de Brevern & Offmann 2006; Tyagi et al. 2006; Tyagi, de Brevern, Srinivasan & Offmann 2008; Leonard, Joseph, Srinivasan, Gelly & de Brevern 2014), reconstruction of globular protein structures (Dong, Wang & Lin 2007), design of peptides (Thomas et al. 2006), and definition of binding site signatures (Dudev & Lim 2007). The most recent impressive developments concern the inclusion of PBs in threading approaches (Mahajan, de Brevern, Sanejouand, Srinivasan & Offmann 2015; Vetrivel et al. 2017) and fold recognition especially with ORION (Ghouzam, Postic, de Brevern & Gelly 2015; Ghouzam, Postic, Guerin, de Brevern & Gelly 2016).

More specifically, PBs are also useful for analysis of protein flexibility (Craveur et al. 2015), for instance, using molecular dynamics in the specific cases of integrins (Jallu, Poulain, Fuchs, Kaplan & de Brevern 2012; Jallu, Poulain, Fuchs, Kaplan & de Brevern 2014; Goguet, Narwani, Petermann, Jallu & de Brevern 2017), Duffy Antigen Chemokine Receptor (DARC) protein (de Brevern et al. 2005), KiSS1-derived peptide receptor (KISS1R) (Chevrier et al. 2013), HIV-1 capsid protein (Craveur et al. 2019), α-1,4-glycosidic hydrolase (Pandurangan et al. 2019), NMDA Receptor Channel Gate (Ladislav et al. 2018), and analysis of local dynamics of proteins and DNA estimated from crystallographic B-factors (Schneider et al. 2014; Schneider, Gelly, de Brevern & Cerny 2014). These researches motivated us to go further to design the analysis of large-scale molecular dynamics simulation from a large number of folds, in order to understand the structural evolution of the local conformation of each region (see Figure 1). One of our previous studies (Narwani et al. 2018), revealed that there is dynamical conversion between α-helices and, the less frequent 3_{10}- and π-helices (Donohue 1953; Pal & Basu 1999; Pal, Basu & Chakrabarti 2002; Pancsa, Raimondi, Cilia & Vranken 2016). More than three quarter of α-helix residues remain in this helical conformation while it drops to 40.5% for 3_{10}-helices. We also underline the fact that using the

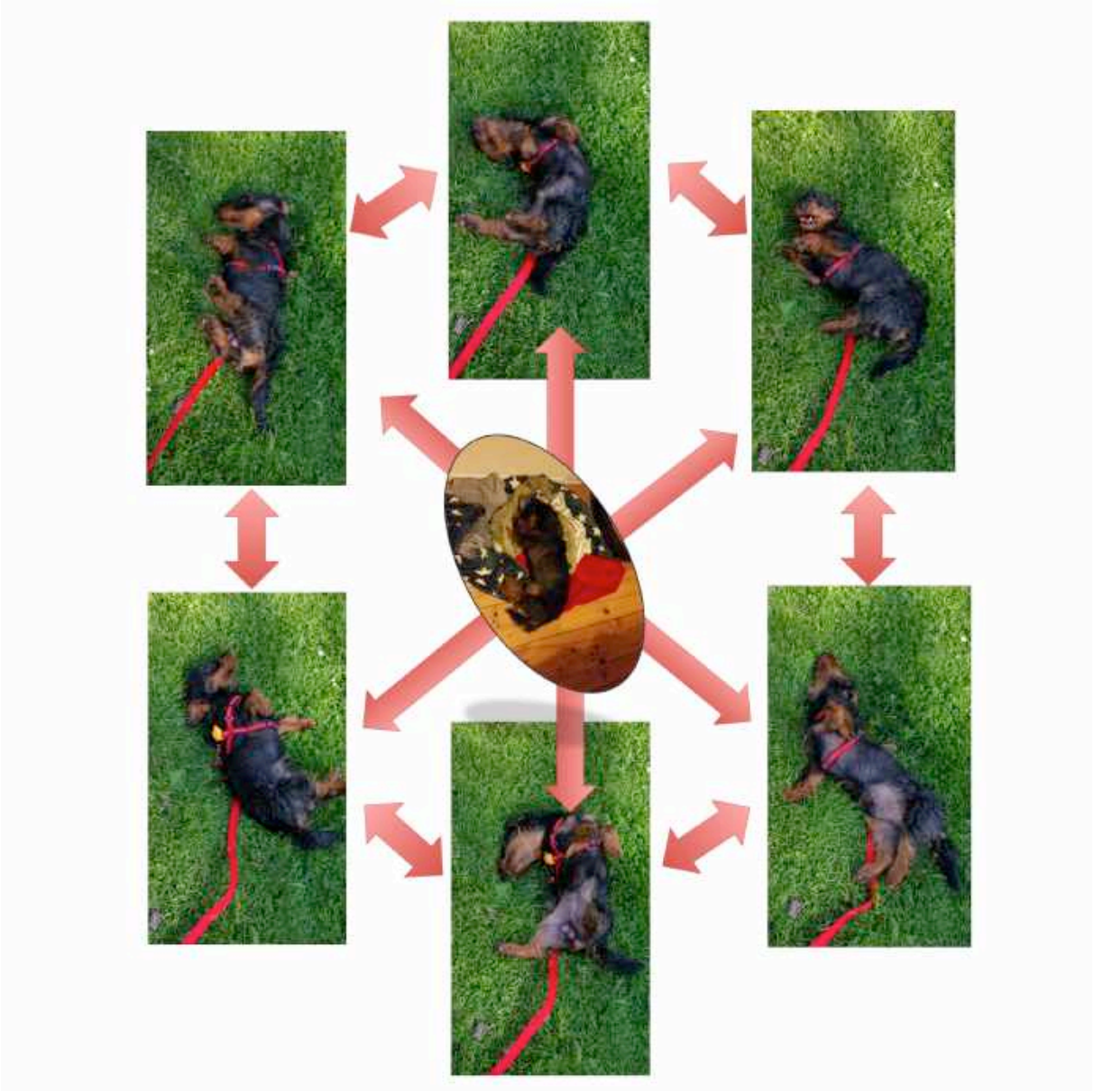

Figure 1. *A dachshund-analogy to illustrate the analysis of protein dynamics at the light of protein local backbone conformation.* This example used a one-year wirehaired dachshund named Snoopy playing on the grass. In the centre, Snoopy is sleeping representing the protein structures as obtained from X-ray crystals. It seems that Snoopy is highly rigid, while when Snoopy is playing on the grass, a large spectra of local backbone dynamics is observed. The analysis here will focus precisely on local conformations, namely protein backbone as it can be seen with Snoopy. Please notice that Snoopy was not harmed to take these pictures and received an optimized dose of dry fishes that he likes a lot, but can also make him sick and they stink.

classical DSSP (namely CMBI 2000) version, the π–helices cannot be described as stable, but with the most recent DSSP version (namely v2.2.1), these results are totally scrambled, the π–helices showed behaviours equivalent to 3_{10}-helix.

The number of large-scale molecular dynamics simulation analyses is relatively limited. We can notice the Dynameomics project (Jonsson, Scott & Daggett 2009), a representative sample of all globular protein metafolds was used to perform simulations under both native and unfolding condition. A related database is available and provides visual results (van der Kamp et al. 2010), but no analyses of protein flexibility have been conducted. Dynasome (Hensen et al. 2012) from Grubmüller's group, comes closest to our questioning. With the simulation of 112 proteins, they showed that collective Dynasome descriptors defined to characterize each simulation, describe most of the movements. These Dynasome descriptors are defined from only a few of the 34 different descriptors. They focus on global level, while the design of this study is to analyse local protein conformations.

In this research, we continue the examination of large-scale protein dynamics simulation to see how β-strands, β-turns, and bends behave during dynamics. Following our previous work (Narwani et al. 2018), PBs are used to analyse (i) how each initial local protein conformation, i.e. PBs, evolves during dynamics and (ii) if transitions are possible between PBs. These analyses have been done using protein flexibility from X-ray data and from molecular dynamic simulations (MD), as well as solvent accessibility, and entropy computed from PB distributions.

Materials and Methods

Data sets. A databank of 169 X-ray structures, taken from Protein DataBank (PDB), (Berman et al. 2000) was extracted using ASTRAL 2.03 at 40% sequence identity (Chandonia

et al. 2004; Fox, Brenner & Chandonia 2014) (PDB ids and corresponding chain provided in Supplementary Data 1). The databank was filtered out based on structure resolution better than 1.5 Å, and without presence of heteroatoms (other than water), without alternate, without missing or modified residues in the chain. Only globular proteins, with chain length ranging between 50 and 250 residues, were selected. In-house parser was used to filter out and to fetch the information (Craveur, Rebehmed & de Brevern 2014). The 169 domains represent a rather equilibrated repartition among the different SCOP classes: all-α represents 18.9% of the chains, all-β 29.6%, α/β 24.8% and α+β 26.7%.

Molecular dynamics simulations. Three molecular dynamic (MD) simulations were performed for each protein structure with GROMACS 4.5.7 software (Pronk et al. 2013), using AMBER99sb force field (van Gunsteren et al. 1996). Each protein structure was put in a periodic dodecahedron box, using TIP3P water molecules (Jorgensen & Madura 1983), and neutralised with Na^+ or Cl^- counter ions. The system was then energetically minimised with a steepest-descent algorithm for 2000 steps. The MD simulations were performed in isotherm-isobar thermodynamics ensemble (NPT), with temperature fixed at 300 K and pressure at 1 bar. A short run of 1ns was performed to equilibrate the system, using Berendsen algorithm for temperature and pressure control (Berendsen, Postma, van Gunsteren, DiNola & Haak 1984). The coupling time constants were equal to 0.1 ps for each physical parameter. Then, a production step of 50 ns was done using Parrinello-Rahman algorithm (Parrinello & Rahman 1981) for temperature and pressure control, with coupling constants of T=0.1 ps and P=4 ps. All bond lengths were constrained with LINCS algorithm (Hess, Bekker, Berendsen & Fraaije 1997), which allowed an integration step of 2 fs. The PME algorithm (Darden, Perera, Li & Pedersen 1999) was used for long-range electrostatic interactions using a cut-off of 1 nm for

non-bonded interactions.

This protocol was applied on each of the 169 protein chains. Conformations were saved every ps. For each MD simulation, the secondary structures were analysed and the structural deviation of each snapshot from the initial structure was measured. Trajectory analyses were done with the GROMACS software, in-house Python and R scripts. Root mean square deviations (RMSD) and root mean square fluctuations (RMSf) were computed on Cα atoms. Normalized RMSfs and normalized B-factors were computed as in Bornot et al. study (Bornot, Etchebest & de Brevern 2011).

Local protein conformation analyses. Secondary structure assignment was performed using DSSP (Kabsch & Sander 1983) (with classic CMBI version 2000 and with a recent DSSP 2015 version 2.2.1) with default parameters (Touw et al. 2015), the latest DSSP distribution is available at GitHub (https://github.com/cmbi/xssp). Protein Blocks (PBs), which are a structural alphabet composed of 16 local prototypes (de Brevern et al. 2000), were also employed to analyse local conformations. Each specific PB is characterized by the φ, ψ dihedral angles of five consecutive residues. The PBs *m* and *d* can be roughly described as prototypes for central α-helix and central β-strand, respectively. PBs *a* through *c* primarily represent the N-cap region of β-strand while PBs *e* and *f* correspond to the C-caps; PBs *g* through *j* are specific to coils, *k* and *l* correspond to the N cap region of α-helix, and PBs *n* through *p* to that of C-caps (de Brevern 2005; Joseph et al. 2010; Joseph, Bornot & de Brevern 2010). PB assignment was carried out for every residue from every snapshot extracted from MD simulations using our PBxplore tool (Barnoud et al. 2017) available at GitHub (https://github.com/pierrepo/PBxplore). We also developed a useful measure to quantify the flexibility of each amino acid, the so-called N_{eq} (for equivalent number of PBs)

(de Brevern et al. 2000). N_{eq} is a statistical measurement similar to entropy; it represents the average number of PBs a residue may adopt at a given position. N_{eq} is calculated as follows (de Brevern et al. 2000):

$$Neq = exp(-\sum_{x=1}^{16} f_x \, ln \, f_x) \qquad (1)$$

Where, f_x is the frequency of PB x in the position of interest. An N_{eq} value of 1 indicates that only one type of PB is observed, while a value of 16 is equivalent to an equal probability for each of the 16 states, *i.e.* random distribution. We have also computed average N_{eq} values.

Finally, to analyse the difference between two series of PBs, a distance named ΔPB was used. It corresponds to:

$$\Delta PB = \sum_{x=1}^{16} (f^1_x - f^2_x) \qquad (2)$$

Where, f^1 and f^2 are two distributions of PB and x is the frequency of PB x for these two distributions.

To trace the evolution of a local protein conformation represented by a given PB in regard to its original assignment, we compute a simple PB ratio, named C^{PB}. It is calculated as the fraction of PBs staying in the initial conformations to that of altered conformations expressed as a percentage of time.

For analyses purpose, we define 8 different classes: 100%, 99-90%, 89-75%, 74-50%, 49-25%, 24-10%, 9-1% and finally less than 1%. For instance, the 100% class is defined by the position that stayed 100% of time in their initial PB assignment during simulations.

Relative solvent accessibility (rASA) was obtained using DSSP output that provides accessible area. The accessible area was then normalized using the maximum surface area of each amino acid type to obtain the rASA.

Clustering approach. A *k-means* clustering approach was then used (Hartigan & Wong 1979) to analyse protein local conformation behaviours during simulations. The general principle is (i) to fix *k* as the number of clusters, they have the same dimension *S* that the data. (ii) The clusters are randomly initialized by taking some examples in the dataset *D*. Then, (iii) one by one all the observations of the dataset *D* is presented to each cluster with a proper distance, for instance Euclidean distance. (iv) Each observation is so associated to the cluster that has the smallest distance. (v) After the first presentation of the dataset, each cluster is readjusted as the average value of all the observations associated to it. And then, step (iii) is done again and the cluster values evolve. Often a large number of learning of the dataset is done, here 50.

Here, a residue is initially associated to a local protein conformation state, namely (i) one of the 8 defined secondary states assigned by DSSP (Kabsch & Sander 1983) (i.e. α-, 3_{10}- and π-helices, β-strand, turns, bends, β-bridges and coil) and (ii) one of the 16 PBs (de Brevern et al. 2000). During MDs, DSSP and PBxplore are used to assign the protein chain local protein structure conformations. Hence, each residue is associated to a vector of size $S = 8$, representing the 8 defined secondary states and more specifically the occurrence of each observed state, and $S' = 16$ when PBs are analysed.

A first series of analyses look at the evolution of each local protein conformation (secondary structure or PB). For a specific state, a subset is created. It represents all the residues that were associated to this state before MD, e.g. PB *d*. Then a fixed number of clusters *k* is determined. The *k* clusters have length of size *S* for secondary structures and *S'* for PBs. All the data of the subset is compared to each of the *k* cluster, and the one with the minimal Euclidean distance is considered as the winner. After one cycle (all the subset had been used), the values of the *k* clusters are modified to correspond the associated

observations, *i.e*. each cluster is the barycentre of the associated observations. After few cycles, the *k* clusters are stable and can be analysed, *e.g*. how the residues associated to PB *d* assignment have evolved.

The final analysis is slightly different as the whole dataset encoded in terms of PBs is used. The number of cluster *k* was higher (k = 30) and showed less stability than before, the approach had begun with a quite larger number of clusters (k = 100), then a hierarchical clustering was performed to select a lower number of *k* clusters and repeated until k = 30. The hierarchical clustering allows finding clusters that have some similarity and so can be associated. We so reduce the number accordingly, but also used these results to initialize the next k-means clustering (with a more limited number of clusters).

Results & Discussions

Analyses of protein structures and simulations. As presented in (Narwani et al. 2018), the percentage of helical structures is impacted by the version of DSSP (Kabsch & Sander 1983) used. With DSSP CMBI 2000, the α-helix represents 31.5%, the 3_{10}-helix 3.99% and 0.02% for the π-helix; similar to the distribution observed by (Tyagi et al. 2009). Protein Blocks frequencies is strictly comparable to previous study (de Brevern 2005) (see Supplementary Data 2). With the new DSSP (version 2.2.1) (Touw et al. 2015), no change is observed at all for 3_{10}-helices (still 3.5%) and near no impact on α-helix, but the frequencies of π–helices increase by 16 times going from 0.02 to 0.32%, i.e. 2/3 of previously assigned α-helices and 1/3 of turns (Narwani et al. 2018).

Distributions of normalized B-factor, normalized RMSf, and relative solvent accessibility are also following distributions close to previous study (see Figures 2A to 2C). The correlation between normalized B-factor and normalized RMSf is of 0.43, while it is

quite low with the N_{eq}, i.e. 0.24 and 0.14. This last result was expected since N_{eq} does a local conformation analyses (see Supplementary Data 3) with more than 60% of residues having a N_{eq} of 1.0, *i.e*. does not change during all simulations, while only 0.8% having a N_{eq} higher than 4 (see Figure 2D).

As we have already analysed the helical structures (Narwani et al. 2018), we have analysed β-sheets and turns (β-turns and bends) evolution during dynamics in the next sections. Figure 3A summarizes the dynamical evolution of all non-helical states. Near 95% of the residues assigned, as β-sheet remains unchanged during the dynamics, while 4.2% goes to coil state, very rarely to bends (0.7%), turns (0.1%) and β-bridge (0.6%). Interestingly the rare β-bridge exchanges to coil state in 12.8% and β-sheet in 10.3%.

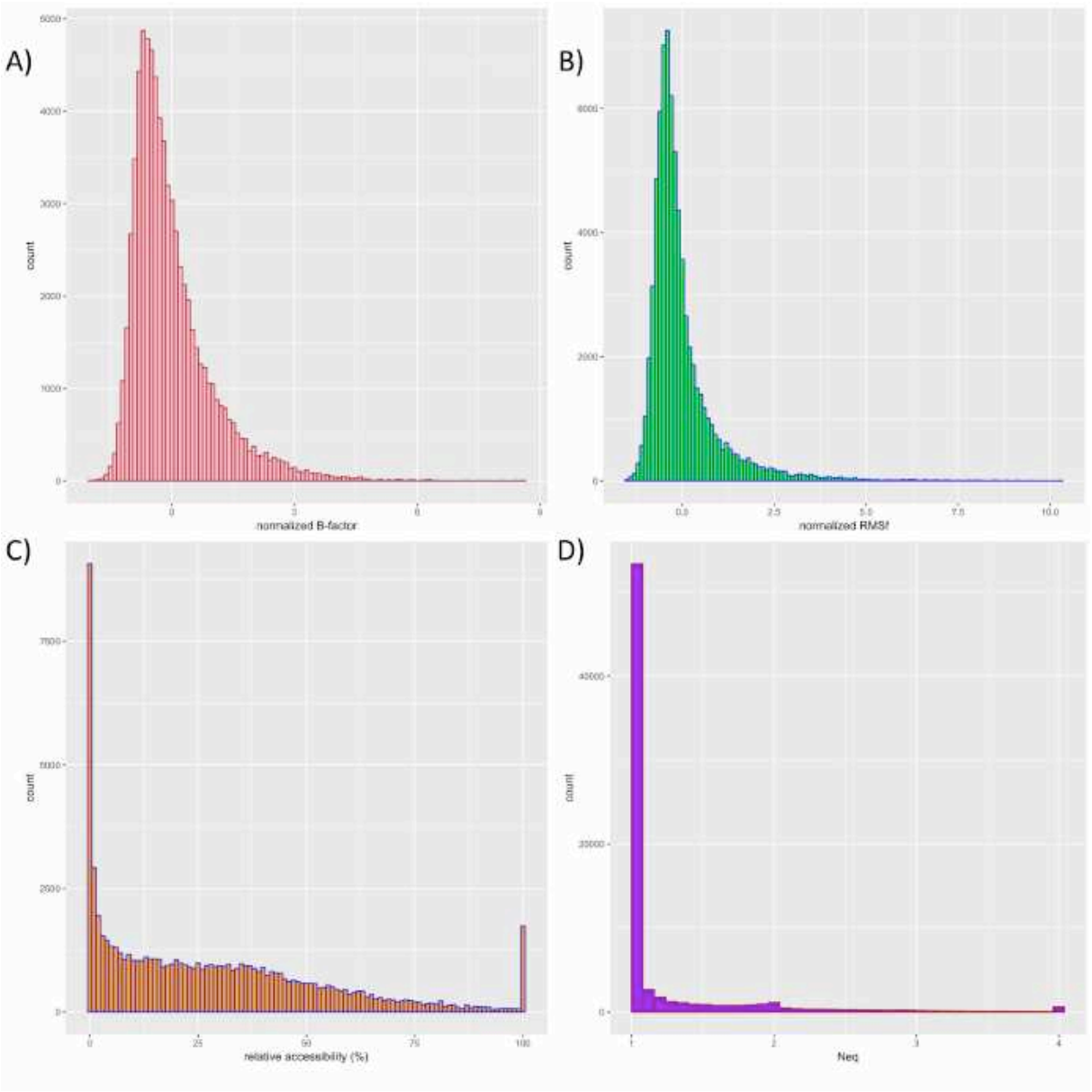

Figure 2. *Analyses of the complete dataset.* Are shown the A) Normalized B-factor, B) normalized RMSf, C) relative solvent accessibility (%) and D) N_{eq} with all N_{eq} higher than 4.0 being grouped in '4.0' and representing 0.82% of the residues (0.26 have $N_{eq} > 5$ and 0.08 higher than 6).

The β-sheets. As presented in M&M section, clusters were generated using *k*-means algorithm with *k*=5 clusters. Clusters were named with the most frequent state observed during simulations (see Figure 3B and Supplementary Data 4A).

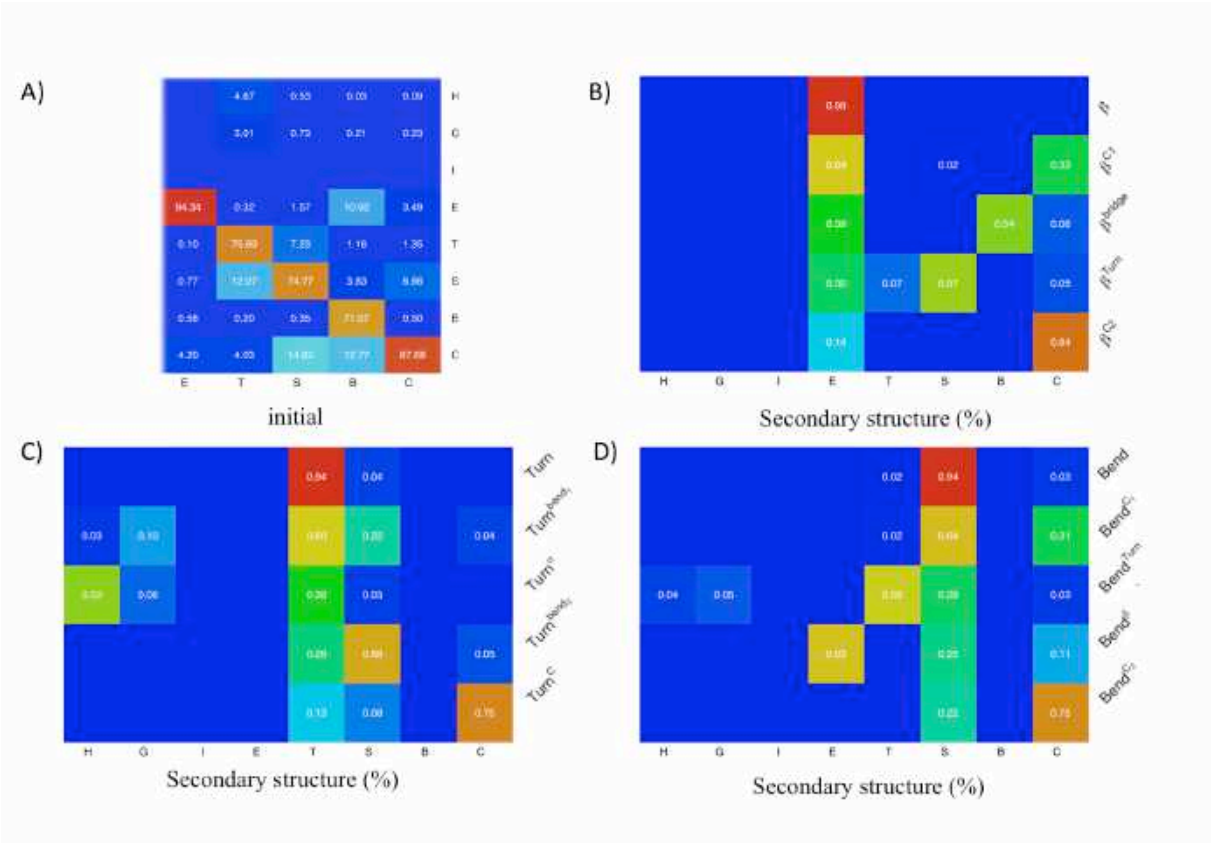

Figure 3. *Secondary structure analyses.* A) Occurrences of non-helical states (see (Narwani *et al*. 2018) for the analysis of helical states), followed by the results of *k-means* clustering (with 5 clusters) for B) β-sheets, C) (hydrogen bond) turns and D) (non-hydrogen bond) bends. Colour gradient goes from red (1.0) to dark blue (0.0), as described previously (Narwani *et al*. 2018). Secondary structures are abbreviated in 1-letter thus: α-helix (H), 3_{10}-helix (G), π-helix (I) β-sheet (E), turn (T), bend (S), β-bridge (B) and coil (C).

The first cluster can be considered as the cluster β, as it remains in β–sheet conformations (> 99%) during all the dynamics. It represents 92.2% of the occurrences with extremely low normalized B-factor (-0.48) and normalized RMSf (-0.53). It corresponds to the most buried part of the dataset (mean relative accessibility of 14.4).

The last four clusters have all higher relative accessibility ranging between 21.6 and 28.9. They are named β^{C1}, β^{bridge}, β^{Turn}, and β^{C2}, according to the state they transition into. As expected, clusters β^{C1} and β^{C2} (transition to coil state) are the most occurring with 3.9% and

2.4% respectively. The first one is the most rigid (nBfact of -0.20 and nRMSf of -0.22 vs -0.10 and -0.03 for the second one) and also has the most β-sheet content (near 2/3rd vs only 14%). A β^{Turn} cluster representing 0.84% also appeared; it is quite flexible in regards to the other clusters and had a higher accessibility (with highest N_{eq} when looking at PB contents). The surprising cluster with only 0.6% of occurrence is the β^{bridge} cluster; it is the most rigid one after cluster β. It is composed of 56% of β-bridge, 6% of coil and 39% of β-sheet.

The Turns. Venkatachalam described and classified the first β-turns as hydrogen bond turns (Venkatachalam 1968). Later, the definition of turns evolved from an energetic to a distance criterion between Cα (Richardson 1981). DSSP differentiates between hydrogen bond turns (namely turns, 'T') and non-hydrogen bond turns (namely bends, 'b') (Kabsch & Sander 1983). We have so analysed them separately.

A turn stays as a turn with a frequency of 75.7% (see Figure 3A) and changes to bends in 12.1%, then to the α-helix (4.7%), coil (4.0%), 3_{10}-helix (3.0%) and extremely rarely to extended state. The percentage of bends and helical state is expected, it must not be forgotten that 3_{10}-helix was often confused with β-turn type III (erased since) and α-helix can be considered as a succession of specific turns (de Brevern 2016).

The clustering (see Figure 3C and Supplementary Data 4B) reflects these results with a cluster *Turn* representing 63.8% of the initial turns. The following four clusters are *Turn*^{bend1}, *Turn*$^{\alpha}$, *Turn*^{bend2}, and *Turn*^{c}, with 18.7%, 6.3%, 7.6% and 3.5%, respectively. They are characterized by higher normalized B-factor (0.42 to 0.88) and normalized RMSf (0.32 to 0.74). All clusters have high relative solvent accessibility (52%) with the exception of *Turn*$^{\alpha}$, with only 36.9% which is the counterpart of a helical cluster previously described (Narwani et al. 2018) also with low accessibility state and with exchange between α-helix (52%) and turns

(38%). In conclusion, (hydrogen bond) turns are not so rigid, the extreme being the *Turn*^{c}, with nBfact of 0.88 and nRMSf of 0.74.

The bends. Slightly less frequent than turns, they change less to helical states than turns. Indeed, 71% stays as bends, 14.8% goes to coils, 7.2% to turns and 1.6% to β-sheet. These obtained clusters reflected this result. The canonical *bend* cluster represents 63.8% of the occurrences, followed by *bend*^{C1}, *bend*^{turn}, *bend*$^{\beta}$ and *bend*^{C2} with 16.1, 9.2, 1.7 and 10.3%, respectively (see Figure 3D and Supplementary Data 4C). The *bend* cluster, as *bend*^{C1} and *bend*^{C2}, is more rigid than *Turns* clusters with lower nBfact, nRMSf and rASA. The *bend*^{turn} is an equivalent of the two *Turn*^{bend}. The unexpected one is the cluster *bend*$^{\beta}$ that is 63% β-sheet, 25% bends and 11% coil with a lower rASA of 30.1% and nBfact of 0.21 (lowest of 5 bend clusters) but nRMSF of 0.98 (the highest).

Hence, even if turns and bends are highly comparable, they have unexpected specificities. The lack of hydrogen bonds at short range allows for a limited number of them to participate dynamically in β-sheet and forms a specific recurrent cluster. For the turns, a specific cluster exchanges with α-helix state, but no cluster reflects an exchange with 3_{10}-helix, that may have been expected.

PB analysis: I. General analysis of PBs. As seen in Figure 2D, more than 60% of the residues have a N_{eq} equal to 1.0, i.e. they did not change of PB assignment during the whole simulation. This rigid-constituency is highly dependent of the type of PBs. Indeed, PB geometrically related to core of repetitive structures namely PB *m* (for α-helix) and PB *d* (for β-sheet) stick to their original assignment with respective frequency of 86.2% and 75.4%. It decreases very rapidly with a strong gradient with PBs *n* (66.6%), *l* (63.2%), *i* (60.7%), *a*

(60.0%), *f* (59.9%), *k* (56.5%), *h* (53.7%), *o* (51.2%), *c* (50.1%), *b* (46.5%), *e* (43.1%), and *p* (39.9%); it became extremely low for PBs *j* (19.1%) and *g* (16.2%).

This gradient is not only the reflection of an ultra-stable assignment, but truly of the dynamics behaviours. Hence, by analysing PBs that stay less than 50% of the times associated to their initial assignment, i.e. the C^{PB} less than 50%, the highest C^{PB} values are associated to PBs *g* (40.3%) then *j* (23.4%) while PBs *d* (4.8%) and *m* (3.9%, see Supplementary Data 5). Hence a strong correlation exists between the original assignment and the conservation of the local protein conformations.

A simple question arise, can it be not simply due to the accessibility of the residues? Indeed, if a local protein conformation is accessible, will it not simply enhance its probability to be changed more easily? In fact, the tendency exists but is not a simple binary case for every PBs (see Supplementary Data 6). For PBs *m* and *d*, the percentage of rigid position (C^{PB}> 75%) is largely higher than the deformable one (C^{PB} <25%) and is directly linked to the solvent accessibility. However, PB *n* does not show this simple (and expected) tendency, the difference between rigid and deformable position is not significant. Depending on the type of PBs, it goes from a slight tendency to no tendency at all. A surprising result is the PB *j* that is one of the two less constraint PBs, it is more exposed than the others but have same distribution of relative accessibility. Hence, no specific rules can be observed here.

Figure 4 shows the distribution of PBs accordingly to the initial assigned PB. It reflects the previous results underlying the high frequencies of PBs *m* and *d* (96% and 94%), and the low of PBs *g* and *j* (59% and 72%). It shows also were the PBs go to another local conformation. Hence 7 PBs go to PB *m* with a frequency higher than 2% (a threshold used in all representations) and 8 to PB *d*. In fact, 22% of these transitions are observed, the highest one being PB *g* to PB *p* (9%), to PB *c* (8%) and to PB *e* (6%), and from PB *p* to PB *m* (6%).

While most of the transitions are logical as geometrically they stay in similar neighbourhood, i.e. between PBs *a*, *b*, *c* and *d*, these most frequent ones are not (de Brevern 2005). Another way to look at the evolution of PBs is the computation of N_{eq}. Figure 4 shows the N_{eq} computed from the PB distribution (named here PB N_{eq}), it is also possible to compute an average N_{eq} on all the observations associated to each PBs (named here average N_{eq}, see Supplementary Data 7). These last is quite smaller as expected but show an impressive Pearson correlation coefficient of 0.94 with PB N_{eq} underlying the possibility to use both for the analyses.

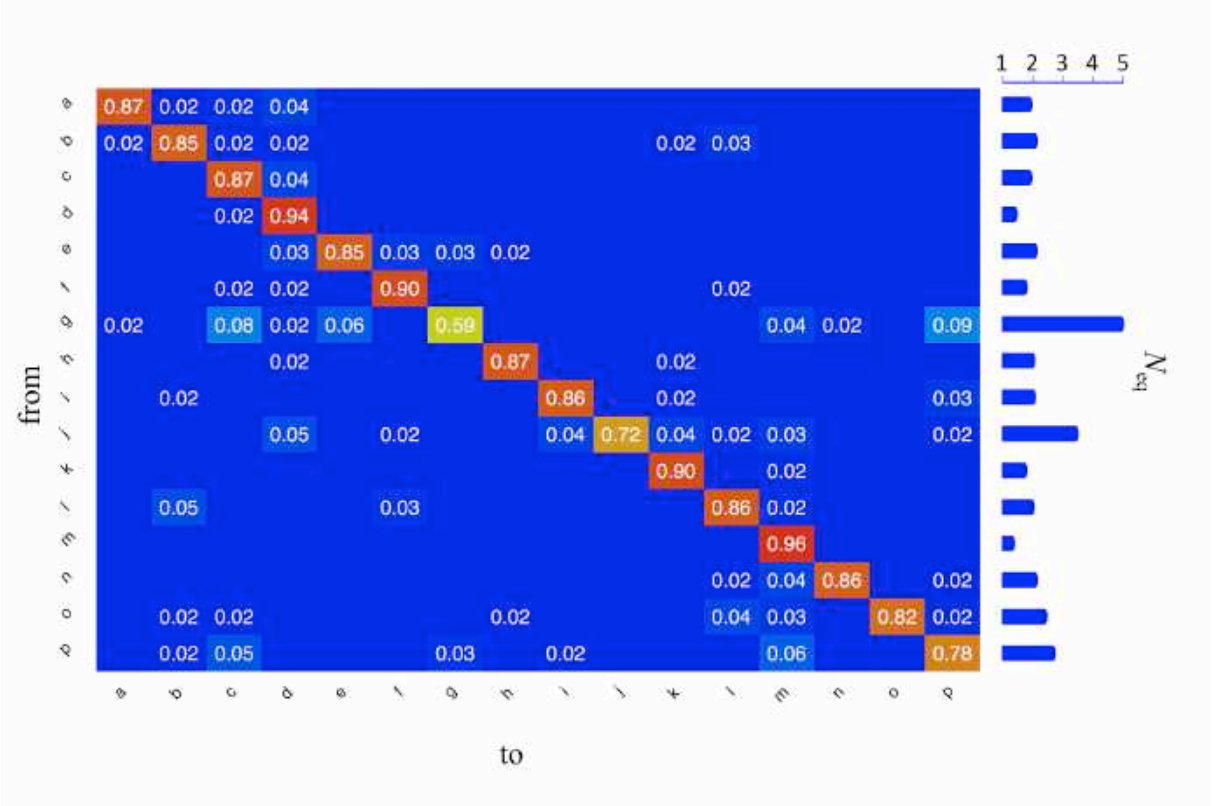

Figure 4. *Dynamical evolution of PBs.* The initial PBs are shown on the y-axis; the frequencies of each PB observed during dynamics are provided following color gradients (frequencies under 0.02 are not written for clarity); The N_{eq} observed for each PB are on the right.

PB analysis: II. PB analysis. Using *k-means* approach, we clustered the behaviour of

each PB. Hence, 5 clusters were generated for each of the 16 PBs and analysed (see a complete analysis of all clusters in Supplementary Data 8 and for detailed values Supplementary Data 9). Figure 5 shows a summary of recurrent behaviour which highlight PBs *a*, *b*, *g* and *f*.

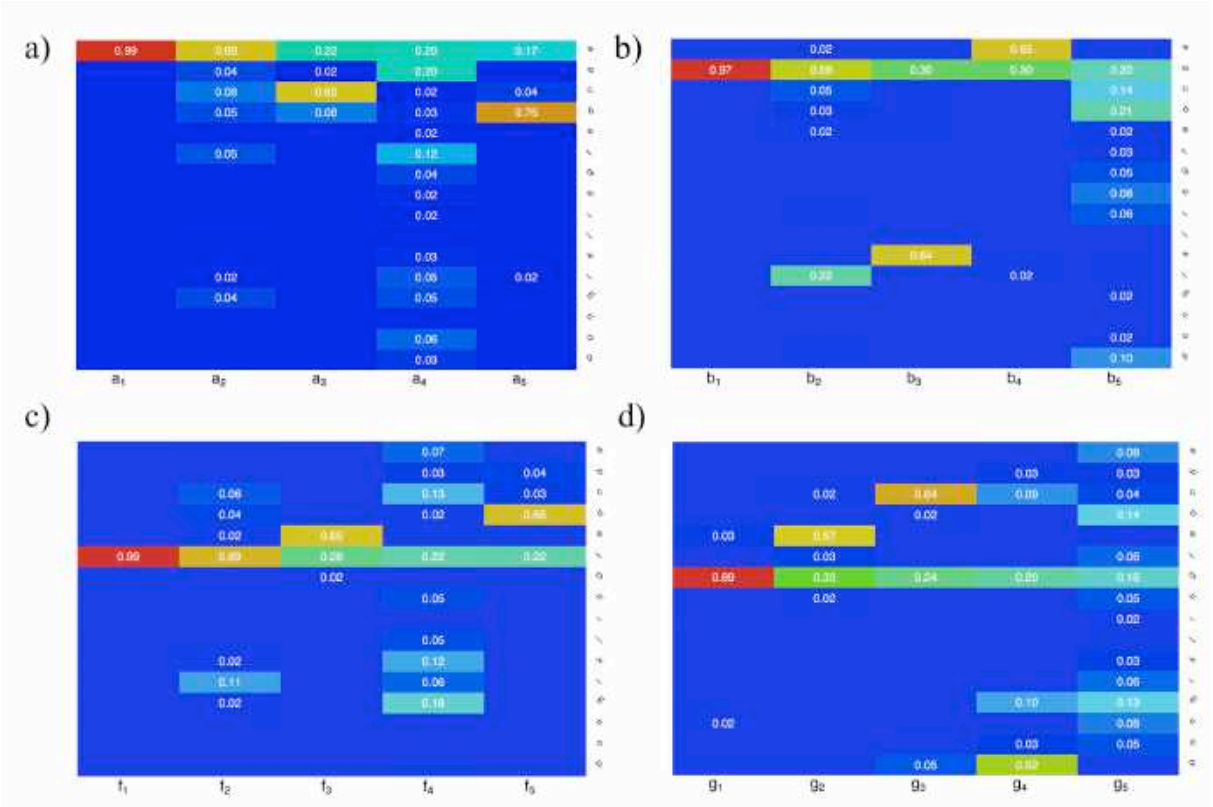

Figure 5. *Example of PB clusters.* The clusters for PBs *a*, *b*, *f* and *g* are shown on the x-axis (see Sup data 8 for all the different PBs) with the distribution of associated PBs on the y-axis.

An important point is that in a previous study, we have analysed the geometrical compatibilities between PBs by looking at the second best PB for every local conformation, i.e. defining major geometrical transition for a PB to another (de Brevern 2005). For instance, for PB *a*, the major geometrical transitions were PB *c* (51%), PB *f* (17%) and PB *d* (9.4%).

The five clusters of PB *a* give interesting results (see Figure 5A). Cluster a_1 is the PB *a* cluster (with >98% of PB *a*) and represents $4^{th}/5$ of positions initially associated to PB *a*.

Unexpectedly, it has not the lowest normalized B-factor values, but the fourth (0.10 vs. -0.03 and -0.02 for clusters a_3 and a_4). It is associated to the lowest normalized RMSf and one of the lowest relative solvent accessibility values. It so represents a first typical case, the more stable cluster (i.e. PB *x* that stay PB *x*) which is not always associated to lowest nBf and rSA.

Clusters a_3 and a_5 followed this idea of geometrical resemblance, i.e. the major geometrical transitions, as they have respectively high frequency of PB *c* (65%) and PB *d* (76%). No clusters with strong evolution to PB *f* can be seen.

Cluster a_2 represents another behaviour; it is still high content by the original PB, but also goes to a large number of other local conformations. In cluster a_2, PB *a* still represents 67% of the occurrences, but with 6 PBs at more than 2%.

Finally, cluster a_4 represented the fuzzy cluster, represents 5.3% of initial PB and is the more deformable, i.e. average N_{eq} of 2.36, and cluster N_{eq} of 8.43. It is also associated to highest accessibility, highest nBf and highest nRMSf.

Figure 5B shows the PB *b*. The cluster b_1 (>97% of PB *b*) represents 79% of original PB *b* positions had lowest rSA and lowest nRMSf but the second lowest nBf (0.00 *vs*. -0.06 for cluster b_3). Interestingly none of the three following clusters has adopted the expected major geometrical transitions (i.e., PBs *d*, *c* and *f*), except PB *l* for cluster b_2 (22%), PB *k* for cluster b_3 (64%) and PB *a* for cluster b_4 (65%). Only cluster b_3 can be considered as comparable with cluster b_1 in terms of nRMSf and nBf and the closest rSA.

Figure 5C shows the PB *f*. As often seen the majority cluster, namely cluster f_1 (>98% of PB *f*) represents 83% of original PB *b* positions had lowest nBf, nRMSf and rSA. The expected geometrical transitions (PBs *b* and *k*) have not been adopted during dynamics to substitute PB *f*. They have been replaced by PB *e* for cluster f_3 (65%) and PB *d* for cluster f_5 (66%). Interestingly, this last is associated to high nBf, high nRMSf and high rSA that is quite

uncommon for PB *d*. Cluster f_2 is less stable than cluster b_1 with a lower PB content of PB *f* (69%).

Finally Figure 5D shows the PB *g*. Cluster g_1, the canonical cluster *g* comprised more than 89% of PB *g*; it represents 55% of original PB *g* positions and had lowest rSA and lowest nRMSf, the second lowest nBf but with a slight difference (0.04 vs. 0.03 for cluster g_2). As seen in the previous section, PB *g* does not stay as PB *g* as often that other PBs. It is seen again here, the following clusters are only composed of 33%, 24%, 20% and 16% of PB *g*. The first surprising cluster is cluster g_2 dominated by PB *e* (57%) that is quite comparable to cluster g_1 in terms of protein flexibility characteristics (similar nBf and nRMSf). Cluster g_3 was more expected as PB *c* is an expected geometrical transition; it represents 64% of the cluster. The second surprise cluster is cluster g_4 controlled by PB *p* (52%), which is not a major geometrical transition. Cluster g_5 is the second most occurring cluster (15.1% of original PB), it encompass the more deformable regions with high average N_{eq} (2.61) and highest cluster N_{eq} (11.69) associated to high accessibility, highest nBf and highest nRMSf.

To summarize this PB clustering, we have each time:

- One cluster represents the initial PB with high frequency with more than 95% of initial PB, except for PB *g* with only 89.2%. This 'stable' PB is not always associated to lowest normalized B-factor (not for 11 PBs *a*, *b*, *e*, *g*, *h*, *i*, *j*, *l*, *n*, *o*, and *p*) and lowest mean relative solvent accessibility (not for 6 PBs *a*, *e*, *h*, *l*, *n* and *p*). Even for PB *e*, the cluster e_1 (>95% of PB *e*) represents 80% of original PB *e* positions, and is not associated with the lowest values of nBf, nRMSf and rSA, in this case it is the cluster e_4. This last cluster is mainly directed by PB *d* (63%) the second best geometrically compatible PB.
- One cluster is a fuzzy cluster with the highest average N_{eq} and especially highest cluster N_{eq} with only three bellow 8.0, 4 between 8 and 9, 3 between 9 and 10 and 6 higher

than 10. They often represent 5 to 12% of the occurrences with exception of PBs *m* with cluster m_2 (2.7%) and PB *d* with cluster d_5 (2.2%). They are often associated with highest nBf, highest nRMSf and highest rSA, but not always. There are 6 cases for the B-factor (PBs *b*, *e*, *f*, *k*, *o* and *p*), 4 with nRMSf (PBs *f*, *i*, *k* and *p*) and 9 cases with relative solvent accessibility (PBs *e*, *f*, *g*, *i*, *j*, *m*, *n*, *o* and *p*).

- The three remaining clusters are divided into (a) a cluster that is a degenerated version of the PB cluster, often with more than 60% of initial PB and a lot of others (12 on 16), i.e. clusters a_2, b_2, c_2, d_2, f_2, h_2, i_2, j_2, k_2, n_2, o_2, and p_2. They are not always with structural features close to the majority clusters, e.g. cluster o_2 has a very high nBf (0.58 *vs*. 0.34), higher nRMSf (0.81 *vs*. 0.25), and higher rSA (44.6 *vs* 32.0) than cluster o_1. (b) Clusters that are directed by unexpected PB, i.e. not from the major geometrical transitions, namely 21 clusters b_2, b_3, b_4, c_3, e_2, f_3, f_5, g_2, g_4, i_3, i_4, j_3, j_4, k_3, l_3, l_5, m_3, n_3, n_4, o_3, and p_3. (c) The remaining clusters were in fact expected, as they followed major geometrical transitions, namely 13 clusters a_3, a_4, c_4, d_3, d_4, e_4, g_3, k_4, l_2, m_4, m_5, o_4, and p_4. Interestingly, no simple correlation between one or another category can be found to predict simply the behaviours of a local conformation.

Hence, some common behaviour can be found as the conservation of local protein conformation with one cluster, a slightly degenerated one after, and a largely the fuzzy ones. This is also entirely unexpected type of change that is 1.5 times more frequent than the expected ones. Comparison of obtained clusters showed that most of the fuzzy clusters are highly similar and most of the others do not cluster with their associated PB clusters (see Supplementary Data 10) highlighting that the initial local conformations can go to very different local confirmations behaviours. Interestingly near no cluster coming from a given PB is associated to one of its related generated cluster, i.e. cluster f_1 is closest to cluster e_3,

cluster f_2 is closest to cluster d_4, cluster f_3 is closest to cluster e_1, cluster f_4 is a fuzzy cluster, and cluster f_5 is closest to cluster a_5. The dynamics had so a strong local protein conformation impact that can be clustered and properly described and it is not a simple task.

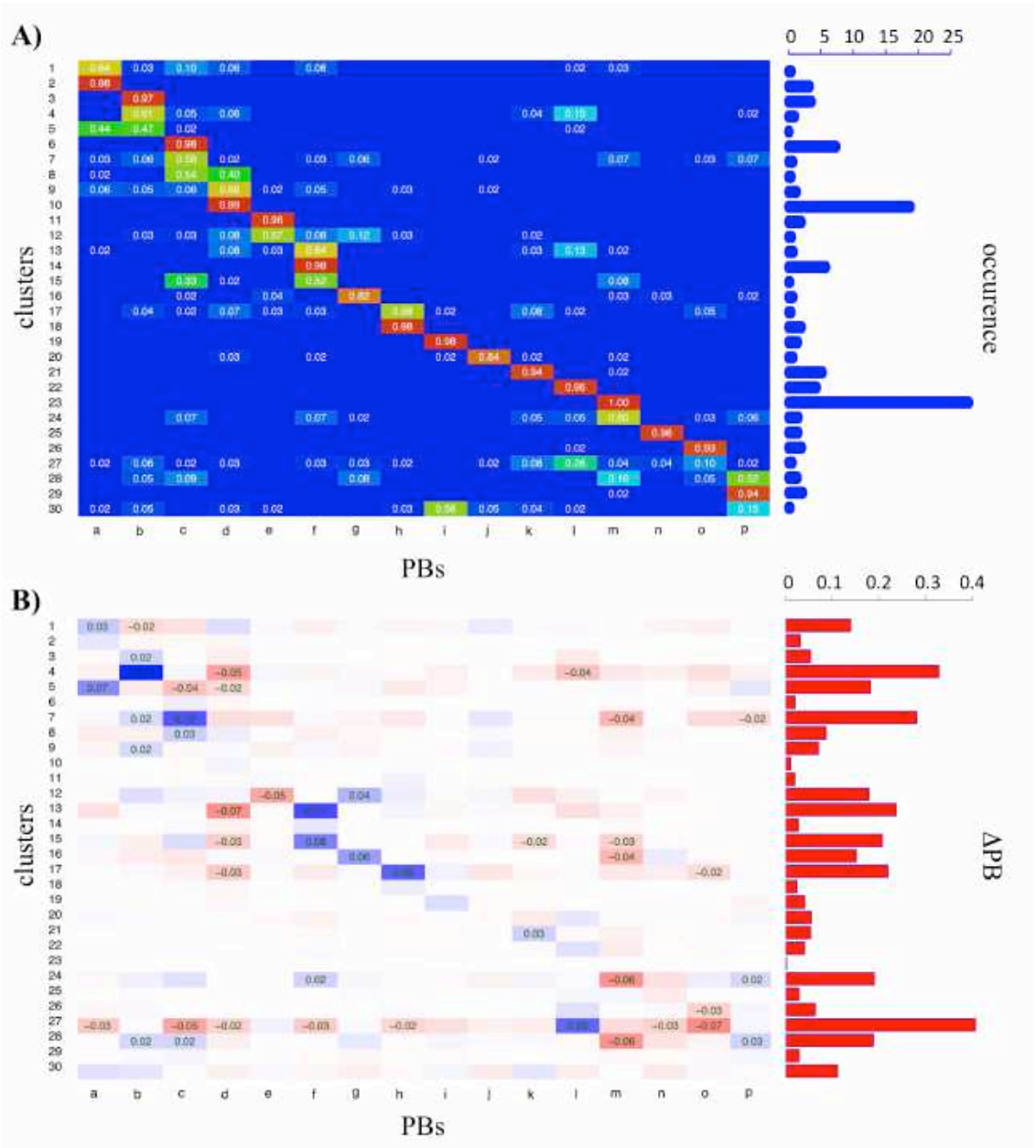

Figure 6. *Clustering of the protein dynamics observed at the light of PBs, namely PB(r)evolution.* A) Is represented with the same rules the 30 ordered clusters obtained through *k-means* approach (y-axis the clusters, x-axis the PB), on the right is provided the occurrence of each cluster. B) Difference between the obtained clusters (A) and the distribution of initial PB (see Sup Data 12), white being no significant differences, positive difference are in blue, negative in red, only difference higher than 0.02 are written. On the left is provided the ΔPB.

PB analysis: III. PB(r)evolution. In a last step, we have clustered directly all the observations without any *a priori* in regards to the initial PBs. The main question here is to see if some similar dynamical behaviour can be observed coming potentially from different initial PB conformation. We have chosen at first a large number of clusters (i.e. 100 clusters), this number was gradually reduced to 30. The final choice is based mainly to assure a minimal number of occurrences associated to each cluster. It must be noticed that the clusters have been rearrange for clarity using the majority PB (see Figure 6A), i.e. the first clusters are associated to PB *a*, while the last ones are mainly with PB *p*.

The 30 clusters can be considered as associated to a majority PB in 29 of 30 cases (see Figure 6A and Supplementary Data 11C). In complete agreement with previous results PB *g* and PB *j* are associated to only one cluster with the lowest major frequency (cluster 16 at 82% and cluster 20 at 84%, resp.) while all other PBs have a major cluster at more than 93%. The most directed one is cluster 23 with cluster N_{eq} of 1.04 and so near exclusively PB *m*, the second one is the PB *d* core, namely cluster 10 with a cluster N_{eq} of 1.07. Only cluster 27 with a cluster N_{eq} of 7.5 is highly fuzzy. Three others have high N_{eq} higher than 5, i.e. cluster 7 (PB *c* frequency of 58% and 9 other PBs with frequency higher than 2%), cluster 17 (PB *h* frequency of 56% and 9 other PBs with frequency higher than 2%), and cluster 30 (PB *i* frequency of 56% and 9 other PBs with frequency higher than 2%).

Hence, with 30 clusters, only 6 PBs have only 1 cluster (PBs *g*, *j*, *k*, *l*, *n*, *o* and *p*) while PBs *b*, *c,* and *f* have 3 clusters. There is a direct correlation between mean normalized B-factor and mean normalized RMSf of the clusters (see Supplementary Data 11A and 11B), and it is also linked to mean rSA and N_{eq} values (no effect to analyse average or cluster N_{eq}).

These analyses differ from the previous ones, as the initial PB is not associated with

any *a priori*. It underlines that some local protein conformations remain highly constant (and it is expected), but also that some have tendencies to oscillate between two local protein conformations. For instance, cluster 5 is PB *a* (44%) and PB *b* (47%) with cluster N_{eq} of 3.1, cluster 15 is PB *f* (52%) and PB *c* (33%) with cluster N_{eq} of 3.3. Others have the tendencies to stay with a majority PB, but goes to many different others, i.e. cluster 1 that is PB *a* (64%) and goes to 6, other PBs for a cluster N_{eq} of 4.3.

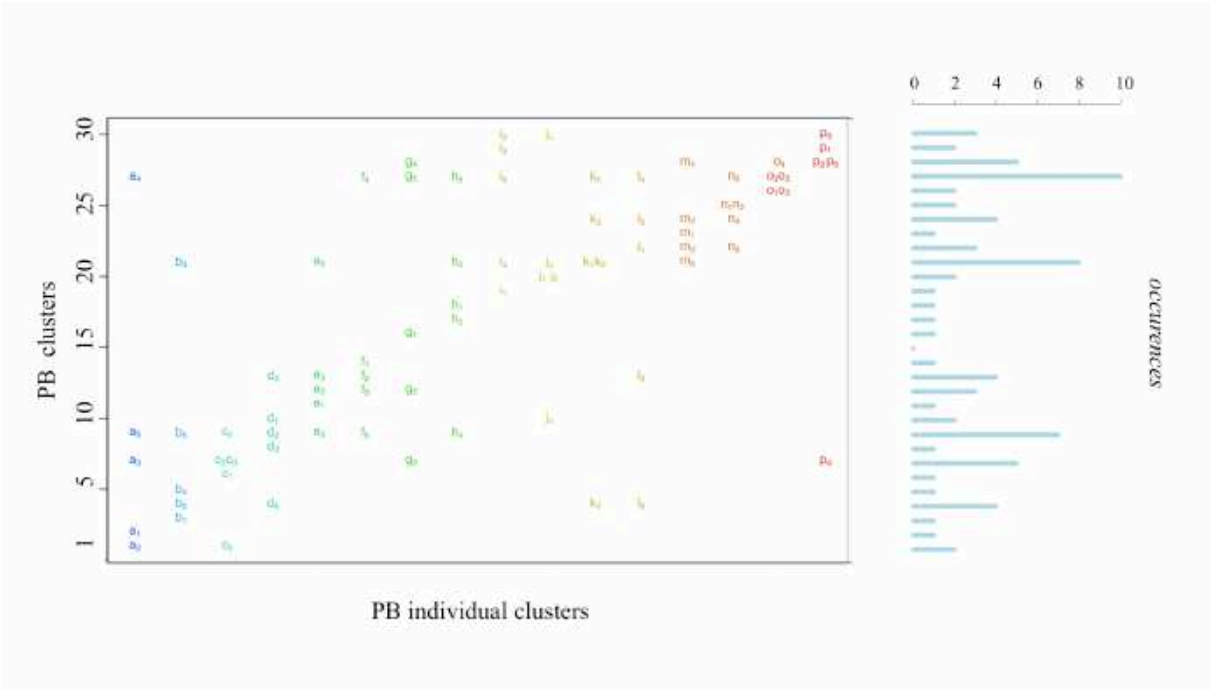

Figure 7. *Comparison of the clustering of the protein dynamics.* On the left is represented the repartition of the 5 clusters obtained for each PB accordingly to the 30 clusters (see Figure 6). On the left is the occurrence of these PB clusters in regards to the 30 clusters, i.e. the number of times a PB cluster is found associated to one of the 30 clusters.

The analysis of the initial distribution of PB and their evolution during dynamics underline other features. Of course the high frequency PB cluster are associated to very similar initial PB distribution (see Figure 7 and Supplementary Data 12), i.e. low ΔPB. Highest ΔPB is found with cluster 27 (ΔPB of 0.41, see Figure 6B), so this cluster is highly fuzzy and can be associated to the most flexible or deformable and without constraints

regions. Nonetheless, it is not associated to the highest rSA observed (cluster 27 has a rSA of 55.3, while cluster 27 has a rSA of 71.5). A surprising observation is cluster 4 (PB *b* 61% and PB *l* 15% for a cluster N_{eq} of 4.3), it has the second highest ΔPB of 0.32. It is displaced from PB *l* at the beginning to PB *d* during the dynamics. Eight of the 30 clusters have a ΔPB higher than 0.15 showing interesting changes in regards to original PB distribution.

Figure 7 shows the confusion between the 5 clusters associated to each PB and the (unsupervised) 30 clusters. These last ones have shown some unexpected tendencies such as the entirely fuzzy cluster 27, or the clusters that do not like to conserve a major PB but also sample a large number of conformations. The final analysis shows that 10 clusters obtained with the 5-clusters (clusters a_4, f_4, g_5, h_5, i_5, k_5, l_4, n_5, o_3 and o_5) approach are in fact the cluster number 27. Similarly 8 clusters (clusters b_3, e_5, h_3, i_4, j_3, k_1, k_2, and m_5) are in fact the cluster number 21 directed by PB *k* (94% and a cluster N_{eq} of 1.40). Then the fuzzy cluster 7 corresponds to 5 different clusters.

Conclusion

Protein structures are often viewed as rigid macromolecules merely composed of helices, sheets and coils. In our study, MDs simulations were performed on a large set of 169 representative protein domains. Previously, we have shown that only 76.4% of the residues associated to α-helices retain the conformation, while this tendency drops to 40.5% for 3_{10}-helices and is never seen for π-helices. In fact, we mainly noticed that the view on π-helices drastically changes with the change in DSSP assignment approach, leading to behaviour similar to 3_{10}-helices underlining the importance of secondary structure assignment methods (Narwani et al. 2017; Narwani et al. 2018).

Here, we have gone further. Firstly, it confirms the rigidity of β-sheet, but also

underline its capacity to transform into turns. Secondly, while the dynamics between turns (with hydrogen bond) and bends (without hydrogen bond) have some strong similarities, they also showed that turns convert easily to helical structures while bends prefer the extended conformations.

Then, for the first time, we perform an entire analysis of a large set of protein dynamics simulations using a structural alphabet. It was done on three levels: (i) a global view in terms of PBs, (ii) performing clustering for each types of PBs and (iii) analysing without any *a priori*.

As expected a large part of the buried positions remain highly stable, but it is not a fixed rule. In fact, for at least half of the PBs, the fact to be buried or exposed does not change at all its dynamics. The majority of PBs remain as their original PB, or at least with a high frequency. Some PBs have a higher tendency to be not as rigid as others and it is particularly true for PB *g* and PB *i*. The intriguing fact is that the change from a PB to another one is not an obvious geometrical change. It is more frequent to go to an unexpected PB than an expected one (based on its geometrical compatibility). Analysis without any *a priori*, i.e. the classification of the 30 clusters, underlined some similar dynamics coming from initial distinct PBs. The choice of K-means clustering, an unsupervised approach, seemed appropriate as the clustering needed was very close from Sequence Families approach we developed few years ago (de Brevern et al. 2000; Etchebest et al. 2005).

The use of these two types of classification shows the difficulty to cluster properly these dynamical properties but indeed it is a first step able to improve (i) our knowledge of protein dynamics and (ii) the relationship between sequence – structure and dynamics. In this field the use of Protein Blocks tends to be an interesting asset.

Acknowledgments

We would like to thanks Nenad Mitić, organizers and participants of Belbi'2016 and Belbi'2018 for fruitful discussions, and Snoopy for the pictures. This work was supported by grants from the Ministry of Research (France), University Paris Diderot, Sorbonne, Paris Cité (France), National Institute for Blood Transfusion (INTS, France), National Institute for Health and Medical Research (INSERM, France) and labex GR-Ex. The labex GR-Ex, reference ANR-11-LABX-0051 is funded by the program "Investissements d'avenir" of the French National Research Agency, reference ANR-11-IDEX-0005-02. TJN, NS and AdB acknowledge to Indo-French Centre for the Promotion of Advanced Research / CEFIPRA for collaborative grant (number 5302-2). This work is supported by a grant from the French National Research Agency (ANR): NaturaDyRe (ANR-2010-CD2I-014-04) to JR and AdB. NSh acknowledges support from ANRT. AMV is supported by Allocation de Recherche Réunion granted by the Conseil Régional de la Réunion and the European Social Fund EU (ESF).

The authors were granted access to high performance computing (HPC) resources at the French National Computing Centre CINES under grant no. c2013037147, no. A0010707621 and A0040710426 funded by the GENCI (Grand Equipement National de Calcul Intensif). Calculations were also performed on an SGI cluster granted by Conseil Régional Ile de France and INTS (SESAME Grant).

Authors' contributions

AdB designed the study. PC did all the molecular dynamics simulations under the guidance of JR and AdB, with the technical help of HS. TJN, AF, AMV and NKS performed all analyses of the results under the guidance of AdB, NS, JCG and CE. AdB and CE prepared the manuscript. All authors discussed the results and implications and commented on the manuscript at all stages.

Conflict of interest

The authors have no conflict of interest to declare. JCG and ADB are associated with IBL, Paris, France.

References

Adzhubei, A. A. & Sternberg, M. J. (1993). Left-handed polyproline II helices commonly occur in globular proteins. *J Mol Biol 229*: 472-493.

Aurora, R. & Rose, G. D. (1998). Helix capping. *Protein Sci 7*: 21-38.

Barnoud, J., Santuz, H., Craveur, P., Joseph, A. P., Jallu, V., de Brevern, A. G. & Poulain, P. (2017). PBxplore: A Tool To Analyze Local Protein Structure And Deformability With Protein Blocks. *PeerJ 5*: e4013.

Benros, C., de Brevern, A. G., Etchebest, C. & Hazout, S. (2006). Assessing a novel approach for predicting local 3D protein structures from sequence. *Proteins 62*: 865-880.

Benros, C., de Brevern, A. G. & Hazout, S. (2009). Analyzing the sequence-structure relationship of a library of local structural prototypes. *J Theor Biol 256*: 215-226.

Benros, C., Hazout, S. & de Brevern, A. G. (2002). *Extension of a local backbone description using a structural alphabet. "Hybrid Protein Model": a new clustering approach for 3D local structures*. International Workshop on Bioinformatics ISMIS, Lyon, France.

Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics 81*: 3684-3690.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res 28*: 235-242.

Bornot, A. & de Brevern, A. G. (2006). Protein beta-turn assignments. *Bioinformation 1*: 153-155.

Bornot, A., Etchebest, C. & de Brevern, A. G. (2009). A new prediction strategy for long local protein structures using an original description. *Proteins 76*: 570-587.

Bornot, A., Etchebest, C. & de Brevern, A. G. (2011). Predicting protein flexibility through the prediction of local structures. *Proteins 79*: 839-852.

Chandonia, J. M., Hon, G., Walker, N. S., Lo Conte, L., Koehl, P., Levitt, M. & Brenner, S. E. (2004). The ASTRAL Compendium in 2004. *Nucleic Acids Res 32*: D189-192.

Chebrek, R., Leonard, S., de Brevern, A. G. & Gelly, J. C. (2014). PolyprOnline: polyproline helix II and secondary structure assignment database. *Database (Oxford) 2014*.

Chevrier, L., de Brevern, A., Hernandez, E., Leprince, J., Vaudry, H., Guedj, A. M. & de Roux, N. (2013). PRR repeats in the intracellular domain of KISS1R are important for its export to cell membrane. *Mol Endocrinol 27*: 1004-1014.

Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B. & Mornon, J. P. (1993). Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein Eng 6*: 377-382.

Craveur, P., Gres, A. T., Kirby, K. A., Liu, D., Hammond, J. A., Deng, Y., Forli, S., Goodsell, D. S., Williamson, J. R., Sarafianos, S. G. & Olson, A. J. (2019). Novel Intersubunit Interaction Critical for HIV-1 Core Assembly Defines a Potentially Targetable Inhibitor Binding Pocket. *MBio 10*.

Craveur, P., Joseph, A. P., Esque, J., Narwani, T. J., Noel, F., Shinada, N., Goguet, M., Leonard, S., Poulain, P., Bertrand, O., Faure, G., Rebehmed, J., Ghozlane, A., Swapna, L. S., Bhaskara, R. M., Barnoud, J., Teletchea, S., Jallu, V., Cerny, J., Schneider, B., Etchebest, C., Srinivasan, N., Gelly, J. C. & de Brevern, A. G. (2015). Protein flexibility in the light of structural alphabets. *Front Mol Biosci 2*: 20.

Craveur, P., Joseph, A. P., Rebehmed, J. & de Brevern, A. G. (2013). beta-Bulges: extensive structural analyses of beta-sheets irregularities. *Protein Sci 22*: 1366-1378.

Craveur, P., Rebehmed, J. & de Brevern, A. G. (2014). PTM-SD: a database of structurally resolved and annotated posttranslational modifications in proteins. *Database (Oxford) 2014*.

Creamer, T. P. (1998). Left-handed polyproline II helix formation is (very) locally driven. *Proteins 33*: 218-226.

Darden, T., Perera, L., Li, L. & Pedersen, L. (1999). New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations. *Structure 7*: R55-60.

de Brevern, A. G. (2001). Nouvelles stratégies d'analyses et de prédiction des structures tridimensionnelles des protéines. Biology (Analyses de Génomes et Modélisation Moléculaire). Paris, University Paris 7. *PhD***:** 208.

de Brevern, A. G. (2005). New assessment of a structural alphabet. *In Silico Biol 5*: 283-289.

de Brevern, A. G. (2016). Extension of the classical classification of beta-turns. *Sci Rep 6*: 33191.

de Brevern, A. G., Autin, L., Colin, Y., Bertrand, O. & Etchebest, C. (2009). In silico studies on DARC. *Infect Disord Drug Targets 9*: 289-303.

de Brevern, A. G., Camproux, A.-C., Hazout, S., Etchebest, C. & Tuffery, P. (2001). Protein structural alphabets: beyond the secondary structure description. *Recent Research Developments in Protein Engineering*. S. Sangadai. Trivandrum Research Signpost. *1***:** 319-331.

de Brevern, A. G., Etchebest, C., Benros, C. & Hazout, S. (2007). "Pinning strategy": a novel approach for predicting the backbone structure in terms of protein blocks from sequence. *J Biosci 32*: 51-70.

de Brevern, A. G., Etchebest, C. & Hazout, S. (2000). Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins 41*: 271-287.

de Brevern, A. G. & Hazout, S. (2003). 'Hybrid protein model' for optimally defining 3D protein structure fragments. *Bioinformatics 19*: 345-353.

de Brevern, A. G., Valadie, H., Hazout, S. & Etchebest, C. (2002). Extension of a local backbone description using a structural alphabet: a new approach to the sequence-structure relationship. *Protein Sci 11*: 2871-2886.

de Brevern, A. G., Wong, H., Tournamille, C., Colin, Y., Le Van Kim, C. & Etchebest, C. (2005). A structural model of a seven-transmembrane helix receptor: the Duffy antigen/receptor for chemokine (DARC). *Biochim Biophys Acta 1724*: 288-306.

Dong, Q. W., Wang, X. L. & Lin, L. (2007). Methods for optimizing the structure alphabet sequences of proteins. *Comput Biol Med 37*: 1610-1616.

Donohue, J. (1953). Hydrogen bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci USA 39*: 470– 478.

Dudev, M. & Lim, C. (2007). Discovering structural motifs using a structural alphabet: application to magnesium-binding sites. *BMC Bioinformatics 8*: 106.

Dupuis, F., Sadoc, J. F. & Mornon, J. P. (2004). Protein secondary structure assignment through Voronoi tessellation. *Proteins 55*: 519-528.

Eisenberg, D. (2003). The discovery of the alpha-helix and beta-sheet, the principal structural features of proteins. *Proc Natl Acad Sci U S A 100*: 11207-11210.

Etchebest, C., Benros, C., Bornot, A., Camproux, A. C. & de Brevern, A. G. (2007). A reduced amino acid alphabet for understanding and designing protein adaptation to mutation. *Eur Biophys J 36*: 1059-1069.

Etchebest, C., Benros, C., Hazout, S. & de Brevern, A. G. (2005). A structural alphabet for

local protein structures: improved prediction methods. *Proteins 59*: 810-827.
Faure, G., Bornot, A. & de Brevern, A. G. (2008). Protein contacts, inter-residue interactions and side-chain modelling. *Biochimie 90*: 626-639.
Fernandez-Fuentes, N., Querol, E., Aviles, F. X., Sternberg, M. J. & Oliva, B. (2005). Prediction of the conformation and geometry of loops in globular proteins: testing ArchDB, a structural classification of loops. *Proteins 60*: 746-757.
Fourrier, L., Benros, C. & de Brevern, A. G. (2004). Use of a structural alphabet for analysis of short loops connecting repetitive structures. *BMC Bioinformatics 5*: 58.
Fox, N. K., Brenner, S. E. & Chandonia, J. M. (2014). SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res 42*: D304-309.
Frishman, D. & Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins 23*: 566-579.
Ghouzam, Y., Postic, G., de Brevern, A. G. & Gelly, J. C. (2015). Improving protein fold recognition with hybrid profiles combining sequence and structure evolution. *Bioinformatics 31*: 3782-3789.
Ghouzam, Y., Postic, G., Guerin, P. E., de Brevern, A. G. & Gelly, J. C. (2016). ORION: a web server for protein fold recognition and structure prediction using evolutionary hybrid profiles. *Sci Rep 6*: 28268.
Goguet, M., Narwani, T. J., Petermann, R., Jallu, V. & de Brevern, A. G. (2017). In silico analysis of Glanzmann variants of Calf-1 domain of alphaIIbbeta3 integrin revealed dynamic allosteric effect. *Sci Rep 7*: 8001.
Hartigan, J. A. & Wong, M. A. (1979). A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 28*: 100-108.
Hensen, U., Meyer, T., Haas, J., Rex, R., Vriend, G. & Grubmuller, H. (2012). Exploring protein dynamics space: the dynasome as the missing link between protein structure and function. *PLoS One 7*: e33931.
Hermoso, A., Espadaler, J., Enrique Querol, E., Aviles, F. X., Sternberg, M. J., Oliva, B. & Fernandez-Fuentes, N. (2009). Including Functional Annotations and Extending the Collection of Structural Classifications of Protein Loops (ArchDB). *Bioinform Biol Insights 1*: 77-90.
Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. M. (1997). LINCS: a linear constraint solver for molecular simulations. *J. Comp. Chem. 18*: 1463-1472.
Hosseini, S. R., Sadeghi, M., Pezeshk, H., Eslahchi, C. & Habibi, M. (2008). PROSIGN: a method for protein secondary structure assignment based on three-dimensional coordinates of consecutive C(alpha) atoms. *Comput Biol Chem 32*: 406-411.
Hutchinson, E. G. & Thornton, J. M. (1996). PROMOTIF--a program to identify and analyze structural motifs in proteins. *Protein Sci 5*: 212-220.
Jallu, V., Poulain, P., Fuchs, P. F., Kaplan, C. & de Brevern, A. G. (2012). Modeling and molecular dynamics of HPA-1a and -1b polymorphisms: effects on the structure of the beta3 subunit of the alphaIIbbeta3 integrin. *PLoS One 7*: e47304.
Jallu, V., Poulain, P., Fuchs, P. F., Kaplan, C. & de Brevern, A. G. (2014). Modeling and molecular dynamics simulations of the V33 variant of the integrin subunit beta3: Structural comparison with the L33 (HPA-1a) and P33 (HPA-1b) variants. *Biochimie 105*: 84-90.
Jonsson, A. L., Scott, K. A. & Daggett, V. (2009). Dynameomics: a consensus view of the protein unfolding/folding transition state ensemble across a diverse set of protein

folds. *Biophys J 97*: 2958-2966.
Jorgensen, W. L. & Madura, J. D. (1983). Quantum and statistical mechanical studies of liquids. 25. Solvation and conformation of methanol in water. *J. Am. Chem. Soc., 105*: 1407-1413.
Joseph, A. P., Agarwal, G., Mahajan, S., Gelly, J. C., Swapna, L. S., Offmann, B., Cadet, F., Bornot, A., Tyagi, M., Valadie, H., Schneider, B., Etchebest, C., Srinivasan, N. & De Brevern, A. G. (2010). A short survey on protein blocks. *Biophys Rev 2*: 137-147.
Joseph, A. P., Bornot, A. & de Brevern, A. G. (2010). Local Structure Alphabets. *Protein Structure Prediction* H. Rangwala and G. Karypis, wiley**:** in press.
Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*: 2577-2637.
Karchin, R., Cline, M., Mandel-Gutfreund, Y. & Karplus, K. (2003). Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. *Proteins 51*: 504-514.
Labesse, G., Colloc'h, N., Pothier, J. & Mornon, J. P. (1997). P-SEA: a new efficient assignment of secondary structure from C alpha trace of proteins. *Comput Appl Biosci 13*: 291-295.
Ladislav, M., Cerny, J., Krusek, J., Horak, M., Balik, A. & Vyklicky, L. (2018). The LILI Motif of M3-S2 Linkers Is a Component of the NMDA Receptor Channel Gate. *Front Mol Neurosci 11*: 113.
Leonard, S., Joseph, A. P., Srinivasan, N., Gelly, J. C. & de Brevern, A. G. (2014). mulPBA: an efficient multiple protein structure alignment method based on a structural alphabet. *J Biomol Struct Dyn 32*: 661-668.
Levitt, M. & Greer, J. (1977). Automatic identification of secondary structure in globular proteins. *J Mol Biol 114*: 181-239.
Li, Q., Zhou, C. & Liu, H. (2009). Fragment-based local statistical potentials derived by combining an alphabet of protein local structures with secondary structures and solvent accessibilities. *Proteins 74*: 820-836.
Mahajan, S., de Brevern, A. G., Sanejouand, Y. H., Srinivasan, N. & Offmann, B. (2015). Use of a structural alphabet to find compatible folds for amino acid sequences. *Protein Sci 24*: 145-153.
Mansiaux, Y., Joseph, A. P., Gelly, J. C. & de Brevern, A. G. (2011). Assignment of PolyProline II conformation and analysis of sequence--structure relationship. *PLoS One 6*: e18401.
Martin, J., Letellier, G., Marin, A., Taly, J. F., de Brevern, A. G. & Gibrat, J. F. (2005). Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. *BMC Struct Biol 5*: 17.
Narwani, T. J., Craveur, P., Shinada, N. K., Santuz, H., Rebehmed, J., Etchebest, C. & de Brevern, A. G. (2018). Dynamics and deformability of α-, 310- and π-helices. *Archives of Biological Sciences 70*: 21-31.
Narwani, T. J., Santuz, H., Shinada, N., Vattekatte, A. M., Ghouzam, Y., Srinivasan, N., Gelly, J. C. & de Brevern, A. G. (2017). Recent advances on PolyProline II. *Amino Acids 49*: 705-713.
Offmann, B., Tyagi, M. & de Brevern, A. G. (2007). Local Protein Structures. *Current Bioinformatics 3*: 165-202.
Pal, L. & Basu, G. (1999). Novel protein structural motifs containing two-turn and longer

3(10)-helices. *Protein Eng 12*: 811-814.
Pal, L., Basu, G. & Chakrabarti, P. (2002). Variants of 3(10)-helices in proteins. *Proteins 48*: 571-579.
Pal, L., Chakrabarti, P. & Basu, G. (2003). Sequence and structure patterns in proteins from an analysis of the shortest helices: implications for helix nucleation. *J Mol Biol 326*: 273-291.
Pancsa, R., Raimondi, D., Cilia, E. & Vranken, W. F. (2016). Early Folding Events, Local Interactions, and Conservation of Protein Backbone Rigidity. *Biophys J 110*: 572-583.
Pandurangan, S., Meganathan, I., Ragavan, S., Ramudu, K. N., Shanmugam, E., Shanmugam, G. & Niraikulam, A. (2019). Engineering of a skin-fiber-opening enzyme for sulfide-free leather beam house operation through xenobiology. *Green Chemistry 21*: 2070-2081.
Park, S. Y., Yoo, M. J., Shin, J. & Cho, K. H. (2011). SABA (secondary structure assignment program based on only alpha carbons): a novel pseudo center geometrical criterion for accurate assignment of protein secondary structures. *BMB Rep 44*: 118-122.
Parrinello, M. & Rahman, A. (1981). Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics 52*: 7182-7190.
Pauling, L. & Corey, R. B. (1951). The pleated sheet, a new layer configuration of polypeptide chains. *Proc Natl Acad Sci U S A 37*: 251-256.
Pauling, L., Corey, R. B. & Branson, H. R. (1951). The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A 37*: 205-211.
Pronk, S., Pall, S., Schulz, R., Larsson, P., Bjelkmar, P., Apostolov, R., Shirts, M. R., Smith, J. C., Kasson, P. M., van der Spoel, D., Hess, B. & Lindahl, E. (2013). GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics 29*: 845-854.
Rangwala, H., Kauffman, C. & Karypis, G. (2009). svmPRAT: SVM-based protein residue annotation toolkit. *BMC Bioinformatics 10*: 439.
Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Adv Protein Chem 34*: 167-339.
Richardson, J. S. & Richardson, D. C. (1988). Amino acid preferences for specific locations at the ends of alpha helices. *Science 240*: 1648-1652.
Rooman, M. J., Rodriguez, J. & Wodak, S. J. (1990). Relations between protein sequence and structure and their significance. *J Mol Biol 213*: 337-350.
Schneider, B., Cerny, J., Svozil, D., Cech, P., Gelly, J. C. & de Brevern, A. G. (2014). Bioinformatic analysis of the protein/DNA interface. *Nucleic Acids Res 42*: 3381-3394.
Schneider, B., Gelly, J. C., de Brevern, A. G. & Cerny, J. (2014). Local dynamics of proteins and DNA evaluated from crystallographic B factors. *Acta Crystallogr D Biol Crystallogr 70*: 2413-2419.
Sklenar, H., Etchebest, C. & Lavery, R. (1989). Describing protein structure: a general algorithm yielding complete helicoidal parameters and a unique overall axis. *Proteins 6*: 46-60.
Thomas, A., Deshayes, S., Decaffmeyer, M., Van Eyck, M. H., Charloteaux, B. & Brasseur, R. (2006). Prediction of peptide structure: how far are we? *Proteins 65*: 889-897.
Touw, W. G., Baakman, C., Black, J., te Beek, T. A., Krieger, E., Joosten, R. P. & Vriend, G. (2015). A series of PDB-related databanks for everyday needs. *Nucleic Acids Res 43*:

D364-368.
Tyagi, M., Bornot, A., Offmann, B. & de Brevern, A. G. (2009). Analysis of loop boundaries using different local structure assignment methods. *Protein Sci 18*: 1869-1881.
Tyagi, M., Bornot, A., Offmann, B. & de Brevern, A. G. (2009). Protein short loop prediction in terms of a structural alphabet. *Comput Biol Chem 33*: 329-333.
Tyagi, M., de Brevern, A. G., Srinivasan, N. & Offmann, B. (2008). Protein structure mining using a structural alphabet. *Proteins 71*: 920-937.
Tyagi, M., Gowri, V. S., Srinivasan, N., de Brevern, A. G. & Offmann, B. (2006). A substitution matrix for structural alphabet based on structural alignment of homologous proteins and its applications. *Proteins 65*: 32-39.
Tyagi, M., Sharma, P., Swamy, C. S., Cadet, F., Srinivasan, N., de Brevern, A. G. & Offmann, B. (2006). Protein Block Expert (PBE): a web-based protein structure analysis server using a structural alphabet. *Nucleic Acids Res 34*: W119-123.
Unger, R., Harel, D., Wherland, S. & Sussman, J. L. (1989). A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins 5*: 355-373.
van der Kamp, M. W., Schaeffer, R. D., Jonsson, A. L., Scouras, A. D., Simms, A. M., Toofanny, R. D., Benson, N. C., Anderson, P. C., Merkley, E. D., Rysavy, S., Bromley, D., Beck, D. A. & Daggett, V. (2010). Dynameomics: a comprehensive database of protein dynamics. *Structure 18*: 423-435.
van Gunsteren, W. F., Billeter, S. R., Eising, A. A., Hünenberger, P. H., Krüger, P., Mark, A. E., Scott, W. R. P. & Tironi, I. G. (1996). Biomolecular Simulation: The GROMOS96 Manual and User Guide. 1042.
Venkatachalam, C. M. (1968). Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units. *Biopolymers 6*: 1425-1436.
Vetrivel, I., Mahajan, S., Tyagi, M., Hoffmann, L., Sanejouand, Y. H., Srinivasan, N., de Brevern, A. G., Cadet, F. & Offmann, B. (2017). Knowledge-based prediction of protein backbone conformation using a structural alphabet. *PLoS One 12*: e0186215.
Zimmermann, O. & Hansmann, U. H. (2008). LOCUSTRA: accurate prediction of local protein structure using a two-layer support vector machine approach. *J Chem Inf Model 48*: 1903-1908.
Zuo, Y. C. & Li, Q. Z. (2009). Using K-minimum increment of diversity to predict secretory proteins of malaria parasite based on groupings of amino acids. *Amino Acids*.
Zuo, Y. C. & Li, Q. Z. (2009). Using reduced amino acid composition to predict defensin family and subfamily: Integrating similarity measure and structural alphabet. *Peptides 30*: 1788-1793.