Impact of protein dynamics on secondary structure prediction

Alexandre G. de Brevern

Version of Record: https://www.sciencedirect.com/science/article/pii/S0300908420302133
Manuscript_771206f49f911f0aace51d4a955ba0cd

Impact of protein dynamics on secondary structure prediction.

Alexandre G. de Brevern[1,2,3,4,*]

[1] Biologie Intégrée du Globule Rouge UMR_S1134, Inserm, Université de 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] IBL, F-75015 Paris, France.

Short title: Predictions and dynamics

* Corresponding author:
Mailing address: Dr. de Alexandre G. de Brevern, INSERM UMR_S 1134, DSIMB, Université de Paris, Institut National de Transfusion Sanguine (INTS), 6, rue Alexandre Cabanel, 75739 Paris cedex 15, France
E-mail : alexandre.debrevern@univ-paris-diderot.fr
Orcid ID: https://orcid.org/0000-0001-7112-5626

Abstract

Protein 3D structures support their biological functions. As the number of protein structures is negligible in regards to the number of available protein sequences, prediction methodologies relying only on protein sequences are essential tools. In this field, protein secondary structure prediction (PSSPs) is a mature area, and is considered to have reached a plateau.

Nonetheless, proteins are highly dynamical macromolecules, a property that could impact the PSSP methods. Indeed, in a previous study, the stability of local protein conformations was evaluated demonstrating that some regions easily changed to another type of secondary structure.

The protein sequences of this dataset were used by PSSPs and their results compared to molecular dynamics to investigate their potential impact on the quality of the secondary structure prediction. Interestingly, a direct link is observed between the quality of the prediction and the stability of the assignment to the secondary structure state. The more stable a local protein conformation is, the better the prediction will be. The secondary structure assignment not taken from the crystallized structures but from the conformations observed during the dynamics slightly increase the quality of the secondary structure prediction. These results show that evaluation of PSSPs can be done differently, but also that the notion of dynamics can be included in development of PSSPs and other approaches such as *de novo* approaches.

Key words: secondary structure prediction, molecular dynamics, protein structures, B-factors, RMSf, solvent accessibility, DSSP, PSIPRED, structural alphabet, helix, sheet, loop.

1. Introduction

The protein 3D structure directly supports their function(s). They are often analysed through the prism of classical repetitive secondary structures, namely α-helices and β-strands connected by loops (i.e. coil state) [1-4]. Nonetheless, the secondary structures are more complex than only a 3-state description. They include two other types of helices, namely 3_{10}- and π-helices [5-7], β-turns [8-10], polyproline type II helices [11, 12] and β-bridges [13]. Designed in 1983, Dictionary of Secondary Structure of Proteins (DSSP) remains the most widely used methodology to assign the protein secondary structures [13]. In fact, assigning secondary structure is not a so simple task and is associated to large number of known issues [14, 15], e.g. differences in the delimitation of the helix ends. Nowadays DSSP can be considered as a gold standard; it had been recently modified to change the assignment priority in the helical regions, which slightly increased the occurrence of the π-helix [16].

The number of protein sequences is incredibly higher than the number of avalaible protein structures (millions against >169.000 in the Protein DataBank [17]). Protein secondary structure predictions (PSSPs) remain essential, as it is not conceivable to have experimental structures associated to each known sequence. PSSPs are useful for instance for predicting protein folds, for building structural models of proteins, and in functional prediction methods such as binding site prediction.

The first generation of PSSPs was based on simple analyses [18] and classical statics, e.g. Chou-Fasman and GOR approaches [19, 20], leading to a prediction rate around two thirds. The introduction of neural networks coupled with evolutionary information made a great impact increasing to 80% the prediction success rate [21, 22], while the deep learning approaches did not drastically improve it [23-25]. PSSP was thereby considered as a mature field [26], the prediction rates have reached a plateau of 82 to 84% [25, 27-30]. It was also commonly believed that a large fraction of the mistakes was due to insufficient precision in

the identification of the borders of the secondary structural elements. Different studies have estimated around 88% the theoretical limits of PSSPs [25, 31, 32], by taken into account the variations of secondary structure assignment between SSAMs, the divergence between homologous sequences, and the prediction errors at boundaries of repetitive helical and sheet structures [30].

Nonetheless, another factor can (or even should) be taken into account, namely the inner dynamics of the proteins [33]. Obviously, protein structures are not rigid macromolecules, and the flexibility / mobility of a large part of the protein structures is essential for their biological function [34]. From the X-ray structures, experimental B-factors can be used to capture the protein flexibility [35, 36]. Molecular dynamics simulations are also valuable to apprehend this flexibility [37]. Sometimes, a simple minimisation made a change in a secondary structure assignment, as seen in [38], underlying the impact of crystallisation on the local protein conformations.

Analysis of the stability of protein local conformations was considered through molecular dynamics (MD) simulations [6]. In a first study focussed on helical regions, it was showed that only 3/4 of the residues associated to α-helices retain the conformation while this tendency drops to 40.5% for 3_{10}-helices. Similarly, the infrequent π–helices go to β-turn, bend and coil conformations, but not to α- or 3_{10}-helices. Only 3_{10}-helices goes to α-helices [6]. Additional analyses underlined that hydrogen bonded turn went more frequently to helical conformation while the non-bonded turn preferred to go to coil conformations [39].

These finding raised a simple question: *'Can the dynamics of the protein explain (partially) erroneous protein secondary structure prediction?'* To answer to this question the protein MD dataset was used [6, 39-41]. Their sequences were predicted by PSSPs and compared with the results of simulations to see if the dynamics can have some influence the prediction.

2. Materials and Methods

2.1 Data sets. A databank of 169 X-ray structures, taken from Protein Data Bank (PDB) [17] was extracted using ASTRAL 2.03 at 40% sequence identity [42, 43] (PDB ids and corresponding chain provided in Supplementary Data 1 of [39]). 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 complete chain length ranging between 50 and 250 residues, were selected. In-house parser was used to filter out and to fetch the information [44]. 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%. This distribution is well distributed among all types of protein folds.

2.2 Molecular dynamics simulations. Three molecular dynamic (MD) simulations were performed for each protein structure with GROMACS 4.5.7 software [45], using AMBER99sb force field [46]. Each protein structure was put in a periodic dodecahedron box, using TIP3P water molecules [47], 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 [48]. 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 [49] 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 [50], which allowed an integration step of 2 fs. The PME algorithm [51] 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 [52].

2.3 Local protein conformation analyses. Secondary structure assignment was performed using DSSP [13] (with DSSP 2015 version 2.2.1) with default parameters [53]. DSSP provides an 8-state assignment with 'H' (α-helix), 'G' (3_{10}-helix), 'I' (π-helix), 'T' (hydrogen-bond turn), 'S' (bends, i.e. non-hydrogen-bond turn), 'B' (β-bridge), 'E' (β-strand) and a blank (coil).

Protein Blocks (PBs), a structural alphabet composed of 16 local prototypes [54], were also employed to analyse local conformations. Each specific PB is characterized by the φ, ψ dihedral angles of five consecutive residues [55-57]. PB assignment was carried out for every residue from every snapshot extracted from MD simulations using PBxplore tool [58]. Flexibility of each amino acid was quantified with the N_{eq} (for equivalent number of PBs) [54], 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 [54]:

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

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. N_{eq} was used for MD analyses [59] and even for disorder proteins [60].

2.4 Secondary structure prediction. Two different approaches were used to predict the secondary structure, namely PSIPRED (for PSI-blast based secondary structure PREDiction) [61, 62] and SSpro [63, 64]. The first one was based on the use of Artificial Neural Networks and evolutionary information, while the second one used more complex machine learning approach (Bidirectional Recursive Neural Networks) also with evolutionary information. PSIPRED and SSpro3 predicted three states (helical, extended and coil), while SSpro8 predicted the 8 states proposed by DSSP [13]. SSpro8 results were simplified to generate a 3 state prediction named SSpro8to3.

It must be noticed that PSIPRED and SSpro3 did not reduce the 8 states to 3 states in a similar way. For PSIPRED [62, 65, 66], helical state was defined as 'H', 'G' and 'I', extended state was 'E' and rest is the coil state ('T', 'S', 'B', blank). It is slightly different for SSpro3 [63, 64], helical state was only 'H', and 'G', extended state was 'E' and 'B', and coil state was ('T', 'S', 'B', 'I', blank).

Two distinct prediction modes exist in SSpro, namely *with* homologues or *without* homologues. The *with* homologues mode added a specific search for homologous sequence in the PDB. If one (or more) sequence (s) was (were) found, it was used as supplementary information. The weights depended on the degree of detected homology (i.e. sequence identity). It would therefore improve largely the prediction, especially if highly related sequences are in the PDB. It was consequently an issue for this study as all the proteins included our dataset came from PDB. It biased the analysis, so the mode *without* homologues was the only one that had been taken into account.

To assess the quality of the prediction, classical accuracy measure Q_3 and Q_8 were used, both simply corresponded to the number of cases correctly predicted:

$$Q = \mathrm{TP} / \mathrm{N}$$

With *TP* the number of true positive and *N* the number of residues used in the

prediction.

Another criterion was also used, it focussed on the highly inadequate prediction (*inad.*), i.e. a residue predicted as extended when it was assigned as helical or inversely. To compare two distinct secondary structure prediction methods, an agreement rate was used. It was the proportion of residues associated with the same state (α-helix, β-strand and coil). Noted C_3 [14], it showed some similarity to Q_3, it monitored if the two clustering were superimposable. PSIPRED also proposed an useful confidence index (CIndex) that ranged between 0 (poor predictability) and 9 (high predictability) as it was done previously by Rost and Sander [67].

2.5 Data analysis. Most of the analyses were done with R software [68] with some C [69] and Python [70] codes.

3. Results

3.1 Principle. The principle of the study was to look at the PSSPs and compare their results with results from MDs. Firstly, PSSPs results were independently analysed, namely, PSIPRED, SSPro3 and SSPro8to3 (SSPro8 reduced to 3 states). It was essential to evaluate the quality of prediction on this specific dataset to know their pertinences. Secondly, all PSSPs were directly compared to apprehend their specificities. Thirdly, the predictions have been analysed in the light of MDs results, underlying potential correlations between the quality of the prediction and the dynamics behaviours. Finally, the prediction have been re-analysed in the light of the regions that have changed of secondary assignment during the simulations to see if incorrect predictions were linked to potential dynamical properties.

3.2 PSIPRED prediction analysis. The occurrence of assigned α-helix, β-sheet and coil state were of 35.8%, 24,5% and 39,7% respectively, an expected distribution seen in most

of the studies [15, 71, 72]. The prediction rate of secondary structure Q_3 was excellent with 83.6% on average, and more specifically 80.3% for the α-helix, 80.9% for the β-sheet and 88.3% for the coil state. The results were particularly good in regards to β-sheet that can be considered as the most difficult secondary structure to predict [73]. Moreover, the number of inadequate problematic issue was highly limited, i.e. the α-helix predicted as β-sheet or inversely, it represented only 0.64% of the residues (see Table 1a).

The prediction quality was directly linked to the PSIPRED's CIndex (see Figure 1 and SI Table 1). For instance, 36.9% of the residues were predicted with a CIndex of 9, associated to a Q_3 of 96.2%, dispatched between 98.7% for the coil, 96.1% for the β-sheets and 93.9% for the α-helices; corresponding *inad* was only equals to 0.03%.

The Q_3 values decreased with the CIndex values, from 96.2% (CIndex = 9) to 90.1% (CIndex = 8), 86.5% (CIndex = 7), 80.9% (CIndex = 6), 77.4% (CIndex = 5), 73.3% (CIndex = 4), 70.3% (CIndex = 3), 63.8% (CIndex = 2), 57.8% (CIndex = 1) and 50.7% (CIndex = 0, see SI Table 1). This drop was observed for the three-states with the same tendency. The confidence index was quite efficient, for instance for a CIndex value of 5 or higher, it corresponded to 73.9% of the residues with a Q_3 of 90.7%, dispatched between 95.1% for the coil, 89.9% for the β-sheets and 86.7% for the α-helices; *inad* was only of 0.20%. Hence, the relationship can be considered very robust.

3.3 SSpro3 prediction analysis. SSpro was an interesting tool as it can predict both 3-states (SSpro3) and 8-states (SSpro8). Another option of SSpro was to mine PDB and used existing structures to enhance the prediction. In our case, it strongly biased the prediction rate, as all the dataset structures were present in the PDB. The SSpro3 (i.e. *with* homologues) accuracy Q_3 reached 96.6%, with 97.9% for the β-sheets, 97.1% for the α-helix and 95.3% for the coil (see SI Table 2a).

Hence, the real accuracy prediction of SSpro3 (i.e. *without* homologues) was of 81.2%, with 84.5% for the α-helices, 82.5% for the coil and 74.4% for the β-sheets (see Table 1b). It was slightly less effective than PSIPRED (minus 2.4%) with an increase of the α-helices prediction rate of 4.4%, and a decrease of 5.8% for the coil by and of 6.5% for the β-sheets. Its *inad* was also higher (1.4%).

It must be noticed that the definition of the three-states was slightly different between PSIPRED and SSpro3, the helical state comprised for the first one the α-helices, π-helices and 3_{10}-helices, while π-helices were excluded for the second. At the contrary, extended state for SSpro3 encompassed β-strands and β-bridges, while these last were considered as coil for PSIPRED. The assignment difference affected less than 2% of the residues.

3.4 SSpro8to3 prediction analysis. Without homologous structure mining, the Q_8 accuracy of SSpro8 values reached 69.4% (see SI Table 2b for the results with homologous proteins). The α-helix state was well predicted at 92.2%, while β-sheet dropped to 83.3%, coil was only of 66.6% and (hydrogen-bond) turn of 55.5%. The remaining was poorly predicted with accuracy of 8.7% for the bends and of 2.3% for the 3_{10}-helix. β-bridge and π-helix were never predicted. Some clear disequilibrium were observed such as 72.5% of π-helix and 36.0% for the 3_{10}-helix were predicted as α-helix, and 55.1% of β-bridge and 49.1% of bends were predicted as coil (see SI Table 2c).

SSpro8to3 corresponded to use the 8 states prediction to a classical 3 states. The Q_3 of SSpro8to3 without homologous proteins equalled to 80.8% (see Table 1c) with prediction rate of 86.2% for α-helix, 81.2% for β-sheet, and 75.8% for coil; *inad* was of 1.7% (see SI Table 2d for the results with homologous proteins).

3.5 Comparison of the two prediction approaches. As the Q_3 values of PSIPRED,

SSpro3 and SSpro8to3 may seem very close (i.e. 83.6, 81.2 and 80.8%, respectively), one can wonder if they did not make similar predictions.

The predictions were so compared. The C_3 values of PSIPRED *vs* SSpro3 equalled to 83.1% (see SI Table 3a) and of PSIPRED *vs* SSpro8to3 to 82.8% (see SI Table 3b). Hence, nearly 1 residue on five was predicted differently. 81.1% of α-helix predicted by SSpro3 were predicted as α-helix and 17.8% as coil by PSIPRED. 77.9% of β-sheet predicted by SSpro3 were predicted as β-sheet and 21.3% as coil by PSIPRED. 0.65% of the residues were antithetically predicted (α-helix for one, β-sheet for the other). The predictions of PSIPRED and SSpro were so significantly different (p-values of 3.10^{-2}).

A major interest of PSIPRED was the robust CIndex values. Surprisingly, this index could be also used for SSpro results (see SI Figure 1). The differences were limited and the trends highly similar (see SI Table 4).

3.6 Secondary structure prediction in the light of protein dynamics. The MD simulations of the 169 proteins were assigned in terms of secondary structure by DSSP [13, 53], and the 8-states description was reduced to the 3-state description. 41.2% of the residues remained always associated to the same initial secondary structure assignment. To analyse the potential impact of protein dynamics, 6 classes, named MD classes, were defined. They were based on the initial assigned state: (i) the residue always assigned to the initial secondary structure assignment, i.e. 100% (41.2% of the residues), (ii) between >100 and 90% (39.6%), (iii) between > 90 and 70% (9.4%), (iv) between > 70 and 50% (4.4%), (v) between > 50 and 10% (4.47%) and (vi) less than 10% (0.95%).

Figure 2 showed the decline of the Q_3 prediction rates in regards to the diminution of the assignment stability. Q_3 decreased (black line in Figure 2) from (i) 93.8% to (ii) 85.6%, (iii) 66.3%, (iv) 57.2% and finished when the stability was less than 50% at (v) only 46.2%

and (vi) 34.1%. This plot highlighted a direct relationship between the quality of the prediction and the local conformation stability.

Interestingly the α-helix (red line in Figure 2) was the most properly predicted for MD classes (i) and (ii) (Q_3 of 98.5% and 89.4%). This feature was slightly different from the information provided by CIndex analysis that emphasized more the coil state (see Figure 1). A similar tendency was seen for SSpro (see SI Figure 2), α-helices being always better predicted than β-sheet, especially when the stability was better than 50%, i.e. MD classes (i) to (iv).

Prediction analyses of dynamical behaviours both in terms of normalized B-factors (see Figure 3a) and normalized RMSf (see Figure 3b) can be seen as *counter intuitive* at the first sight. Indeed, as the normalized B-factors and normalized RMSf mean (black lines) and the median values (red lines) decreased with the decrease of the confidence index, i.e. higher was the certainty of the prediction, higher was the flexibility. CIndex of 9 was a perfect example, the mean and the median normalized B-factor values equalled to 0.13 and -0.22, i.e. higher than for CIndex of 0 with values of -0.04 and -0.25, respectively. This unexpected tendency came from the wrong predictions that were in limited number for high CIndex, but still present. Hence, for CIndex of 9, erroneous mean prediction (brown line) equalled to 0.23 when the correct mean prediction (blue line) equalled to 0.12. Similar tendencies can be observed also with RMSFs (see Figure 3b). Occurrences of the first CIndex values were the most populated while the last were associated with fewer occurrences (see also Figure 1). Hence, the prediction errors were more associated to the 'flexibility' than the 'rigidity' of the proteins.

The relative surface accessibility (see Figure 3c) also provided a similar view with mean and median values that slightly decreased with the quality of the predictions, but again, the erroneous predictions were associated to more accessible regions (mean and median

relative surface accessibility of 30-35% for the CIndex value of 9, while their false prediction had relative surface accessibility higher than 45%).

Number of equivalent (N_{eq}, see Figure 3d) provided a slightly different view as the mean values increased with confidence index. It was logical as N_{eq} is a local conformation measure (on 5 residues) while B-factors and RMSf depended of global measures (on the whole protein structure). It underlined the interest of a local measure and that more it changed, more difficult was it to predict.

The analyses of SSpro (see SI Figure 3) showed similar tendencies for normalized B-factors, relative accessibility, normalized RMSf and N_{eq}. Thus, these four criteria indicated that the accuracy of the prediction correlated with the dynamic properties of the proteins. It was indispensable to look both experimental B-factor and computational RMSf as they encompassed similar notions (rigidity to flexibility), but shared a limited correlation (around 0.45 as observed previously [52]); they therefore have different specificities. To see that these two distinct criteria give results going in similar directions was very positive in order to be able to draw conclusions from these analyzes. [33].

3.7 Secondary structure prediction assessed in the light of protein dynamics. For more than 5% of the residues, the most frequently observed secondary structure state was not the one initially assigned (see SI Table 5). On these 5.4% of the residues, the coil state represented 77.4% of the residues; they mostly went to helical state for 59.3% and extended state for 18.2%. The helical state corresponded to 10.4% of these residues and went only to coil state. The extended state represented 12.2% of these residues and mainly went to coil state (12.0%), and very rarely to helical state (0.2%).

The secondary structure prediction was evaluated again using this *new assignment*. Table 2 provided the Q_3 accuracy showing an increase from 83.6% to 84.1%; the inadequacy

rate decreased from 0.64% to 0.54%. Prediction of helical state increased from 80.3% to 84.4%, extended state remained stable from 80.9% to 80.8%, and coil state decreased from 88.3% to 86.1%. Interestingly, the increase of the prediction mainly concerned the best CIndex values (see SI Table 6 and SI Figure 4). Helical state won for every CIndex value while coil state always decreased.

For SSpro3, the interest was more limited with a limited improvement of Q_3 from 81.2% to 81.4%. It was also mainly directed by the helical state that increased from 84.5% to 87.5%, and the extended state from 74.4% to 74.9%; the coil state decreased from to 82.5% to 80.5% (see SI Table 7a). For SSpro8to3, the Q_3 did not change too much from 80.8% to 80.7%. Again, it was the helical structure that gained from 86.2% to 89.9%, the extended structures increased lightly from 81.2% to 81.6%, while the coil greatly decreased again from 75.8% to 73.7% (see Table 7b). For both SSpro methods, the number of problematic predictions mixing helical and extended structures increased by 0.2% to 1.4% and 1.7%. Hence, the idea to use dynamical properties suited more PSIPRED approach than SSpro.

4. Discussion & Conclusion

PSSP was supposed to reach a theoretical plateau close to 88% [25, 74], i.e. 12% less than a perfect prediction. Different factors explained this difference. Some were linked to biophysical constrains, methodological limitations or evolutionary variations, e.g. protein secondary structure formation are controlled by long-range interactions [75], and by the environment [76], exact sequence fragments can be found in helices as in sheets (namely chameleon sequences [77], a property that could make the prediction more complex), and protein structure can be constrained in crystals [78]. Another point was that proteins are dynamical flexible macromolecules, the local protein conformation changes during proteins' lifetime [6, 39, 79], often with links to their functions [80-83].

Prediction and dynamics

PSSPs were assessed at Critical Assessment of Methods of Protein Structure Prediction (CASP), but were dropped after CASP 4. As noted by Moult and colleagues [27]: "It is likely that these methods are as close to accurate as they will ever be, given that secondary structure is partially determined by tertiary factors. Very small improvements continue to be made but probably only as a consequence of increased sequence database size." Nonetheless, PSSPs are are still particularly used today. For instance, *de novo* protein structure prediction, protein structural classification, or fold recognition approaches often used PSSP [22, 24, 25, 30, 84-86]. Following our analysis of protein dynamics behaviours [6, 39, 41], a logical and intuitive question rose: '*Is the prediction of a given secondary state easier when the residue is in a rigid region, while a flexible region has less constraints and so information to predict it?*' Two points were important for this study: (i) the choices of PSSPs and (ii) the information for protein dynamics.

Concerning the first point, at least dozens of methods have been published [30] and the choices could be easily criticized. It was dictated by simplicity of installation and usage, a good number of papers' citations, and also author's personal experience. PSIPRED was quite intuitive as it was one of the most famous approaches, while SSpro was easy to use and also highly recognized and efficient (Q_3 of 83.6% and 81.2%, respectively, see Table 1). Moreover, PSIPRED proposed a prediction confidence index that was highly pertinent and robust (see Figure 1).

Concerning the second point, different options could be considered, experimental ones such as Nuclear Magnetic Resonance (NMR) [87], and/or computational ones with Normal Mode Analysis [88] or the usage of server for fast modelling of protein structure flexibility, e.g. CABS-flex [89, 90]. As for the previous criteria, the choice of protein X-ray structures and molecular dynamics was partly personal. It was classical and recognized, but also had shown interesting correlations with previously presented approaches [91]. It seemed logical to

use this protein structure dataset [6, 39, 41].

This study underlined that more stable was the assignment of secondary structure better was the prediction (see Figure 2). For instance, the PSIPRED Q_3 accuracy of the residues always assigned to the same secondary structure during all the simulations (100%) equalled to 93.8% while it drastically dropped to 34.1% when the initial assignment was conserved less than 10% of the times. This strong relationship was found in the light of MD classes, but also with CIndex classes, using *experimental* B-factors, and *computational* RMSf, relative accessibility and N_{eq}. For instance, analysis of true and false prediction showed that the highest normalized B-factors and normalized RMSf values have poorest prediction rates.

The link between protein dynamics and protein secondary structure prediction was direct. Hence, using the most observed secondary structure assigned state as new assignment (corresponding to a change for 5.4% of the residues) provides an increase of Q_3 value from 83.6% to 84.1% (see Table 2). It mainly benefited to residues originally associated to coil regions that are now considered as helical regions (more than half of the cases). This last result underlined the interest to take into account protein flexibility [39, 92, 93] both in the evaluation, but also the assignment [94]. It must be noticed that these tendencies were limited to the 3-states prediction. Indeed, the 8-states prediction was characterized (i) by two non-predicted (and rare) states, namely π-helix [7] and β-bridge [13], (ii) the 3_{10} helix, turns and bends [9] prediction did not gain by using MD results, being biased to prediction of α-helix for the two first and coil for the last one (*data not shown*). It could also underline the potentiality of improvement for these more complex predictions.

Interestingly, the only case of a change in the assignment from extended to helical regions was a small region of a *Plasmodium falciparum* lactate dehydrogenase. The β-strands analysed with DSSP were short and highly twisted. Using STRIDE [95] software to assign, a methodology highly related to DSSP (i.e. 95% of identical assignment [72]), the region was

considered as coil not β-strands. Indeed, secondary structure assignment methodologies were not always optimal, and divergences between different approaches were common [7, 14, 71, 96-98].

This study showed that the use of flexibility and/or dynamics could be an interesting addition to the analyses of proteins as (i) it had direct implication in the quality of the predictions, and (ii) that further improvements could be made. It provided interesting insight for potential use of secondary structure predictors to spot flexible regions in the protein. These analyses would need to distinguish between wrong prediction and flexible properties, probably using confidence index values in a first step.

It also underlined the possibility to evaluate, taking into account the flexibility / dynamics of proteins, the prediction rates of PSSPs. It also provided the idea of including this notion of dynamics directly in predictive methods, whether for PSSPs, but also other methodologies based on them, such as *de novo* approaches. Most the approaches have tried to improve the methodologies, but the information used was often similar to the one used in PHD near 30 years ago [21]. It also underlined that evolutionary information was able to provide information about the local protein dynamics, and partially explained why no substantial improvements in the field of protein secondary structure prediction have been reported during the last two decades.

Acknowledgments

AdB thanks a lot all the reviewers who spent time evaluating this work and asked particularly relevant questions; their comments have greatly improved the manuscript. AdB would also like to thank Professor Nenad Mitic, Professor Catherine Etchebest, Dr. Jean-Christophe Gelly, Dr. Tatiana Galochkina, Dr. Julien Diharce, Dr Aline Floch, Mr Gabriel Cretin, his former students, Pierrick Craveur, Agnel Praveen Joseph, Tarun J. Narwani, Snoopy de Brevern, Nicolas K. Shinada, and Akhila Melarkode Vattekatte, and the organizers and participants of Belbi'2016 (Belgrade, Serbia), International Conference on IDP'2017

(IISER Mohali, Chandigarh, India) and Belbi'2018 (Belgrade, Serbia) for fruitful discussions.

This work was supported by grants from the Ministry of Research, France, University of Paris (formerly 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. AdB acknowledges the Indo-French Centre for the Promotion of Advanced Research / CEFIPRA for collaborative grants (numbers 5302-2). AdB acknowledges the French National Research Agency with grant ANR-19-CE17-0021 (BASIN). This study was supported by grants from the Laboratory of Excellence GR-Ex, reference ANR-11-LABX-0051. The labex GR-Ex is funded by the programme "Investissements d'avenir" of the French National Research Agency, reference ANR-11-IDEX-0005-02. Calculations were performed on an SGI cluster granted by Conseil Régional Ile de France and INTS (SESAME Grant). The author was granted access to high performance computing (HPC) resources at the French National Computing Centre CINES under grants no. c2013037147, no. A0010707621, no. A0040710426 and no. A0070710961 funded by the GENCI (Grand Equipement National de Calcul Intensif).

The funding bodies have no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Conflict of interest

The author has no conflict of interest to declare. AdB is associated with IBL, Paris, France.

Legends

Figure 1. *Prediction rate per confidence index values for PSIPRED.* Is given per confidence index value, the Q_3 in black, the prediction rate of α-helix in red, of β-sheet in brown and coil in blue. The occurrence of each confidence index value is provided as histogram.

Figure 2. *Prediction rate per MD class for PSIPRED.* Is given per MD class, the Q_3 in black, the prediction rate of α-helix in red, of β-sheet in brown and coil in blue. The occurrence of each MD class is provided as histogram.

Figure 3. *Analyses of PSIPRED prediction.* For each classes of confidence index (CI) is given the mean and median values of (A) normalized B-factors (nBfactors), (B) normalized RMSf (nRMSf), (C) relative accessibility (rSA), and (D) N_{eq}. Black lines are the mean values, red lines are the median values, is also provided the corresponding values for true and false prediction, in grey mean true prediction, in brown mean false prediction, in orange median true prediction, and in pink median false prediction.

Reference

[1] L. Pauling, R.B. Corey, The pleated sheet, a new layer configuration of polypeptide chains, Proc Natl Acad Sci U S A, 37 (1951) 251-256.

[2] L. Pauling, R.B. Corey, Atomic coordinates and structure factors for two helical configurations of polypeptide chains, Proc Natl Acad Sci U S A, 37 (1951) 235-240.

[3] L. Pauling, R.B. Corey, H.R. Branson, The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain, Proc Natl Acad Sci U S A, 37 (1951) 205-211.

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

[5] M.N. Fodje, S. Al-Karadaghi, Occurrence, conformational features and amino acid propensities for the pi-helix, Protein Eng, 15 (2002) 353-358.

[6] T.J. Narwani, P. Craveur, N.K. Shinada, H. Santuz, J. Rebehmed, C. Etchebest, A.G. de Brevern, Dynamics and deformability of α-, 310- and π-helices, Archives of Biological Sciences, 70 (2018) 21-31.

[7] P. Kumar, M. Bansal, Dissecting pi-helices: sequence, structure and function, FEBS J, 282 (2015) 4415-4432.

[8] C.M. Venkatachalam, Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units, Biopolymers, 6 (1968) 1425-1436.

[9] E.G. Hutchinson, J.M. Thornton, PROMOTIF--a program to identify and analyze structural motifs in proteins, Protein Sci, 5 (1996) 212-220.

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

[11] Y. Mansiaux, A.P. Joseph, J.C. Gelly, A.G. de Brevern, Assignment of PolyProline II conformation and analysis of sequence--structure relationship, PLoS One, 6 (2011) e18401.

[12] T.J. Narwani, H. Santuz, N. Shinada, A.M. Vattekatte, Y. Ghouzam, N. Srinivasan, J.C. Gelly, A.G. de Brevern, Recent advances on PolyProline II, Amino Acids, 49 (2017) 705-713.

[13] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers, 22 (1983) 2577-2637.

[14] L. Fourrier, C. Benros, A.G. de Brevern, Use of a structural alphabet for analysis of short loops connecting repetitive structures, BMC Bioinformatics, 5 (2004) 58.

[15] J. Martin, G. Letellier, A. Marin, J.F. Taly, A.G. de Brevern, J.F. Gibrat, Protein secondary structure assignment revisited: a detailed analysis of different assignment methods, BMC Struct Biol, 5 (2005) 17.

[16] R.P. Joosten, T.A. te Beek, E. Krieger, M.L. Hekkelman, R.W. Hooft, R. Schneider, C. Sander, G. Vriend, A series of PDB related databases for everyday needs, Nucleic Acids Res, 39 (2011) D411-419.

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

[18] A.V. Guzzo, The influence of amino-acid sequence on protein structure, Biophys J, 5 (1965) 809-822.

[19] P.Y. Chou, G.D. Fasman, Prediction of protein conformation, Biochemistry, 13 (1974) 222-245.

[20] J. Garnier, D.J. Osguthorpe, B. Robson, Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins, J Mol Biol, 120 (1978) 97-120.

[21] B. Rost, C. Sander, Improved prediction of protein secondary structure by use of sequence profiles and neural networks, Proc Natl Acad Sci U S A, 90 (1993) 7558-7562.

[22] F. Meng, L. Kurgan, Computational Prediction of Protein Secondary Structure from Sequence, Curr Protoc Protein Sci, 86 (2016) 2 3 1-2 3 10.

[23] S. Wang, J. Peng, J. Ma, J. Xu, Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields, Sci Rep, 6 (2016) 18962.

[24] M. Torrisi, G. Pollastri, Q. Le, Deep learning methods in protein structure prediction, Computational and Structural Biotechnology Journal, (2020) in press.

[25] W. Wardah, M.G.M. Khan, A. Sharma, M.A. Rashid, Protein secondary structure prediction using neural networks and deep learning: A review, Comput Biol Chem, 81 (2019) 1-8.

[26] W. Pirovano, J. Heringa, Protein secondary structure prediction, Methods Mol Biol, 609 (2010) 327-348.

[27] J. Moult, K. Fidelis, A. Zemla, T. Hubbard, Critical assessment of methods of protein structure prediction (CASP)-round V, Proteins, 53 Suppl 6 (2003) 334-339.

[28] A. Drozdetskiy, C. Cole, J. Procter, G.J. Barton, JPred4: a protein secondary structure prediction server, Nucleic Acids Res, 43 (2015) W389-394.

[29] M. Torrisi, M. Kaleel, G. Pollastri, Porter 5: fast, state-of-the-art ab initio prediction of protein secondary structure in 3 and 8 classes, bioRxiv, (2018).

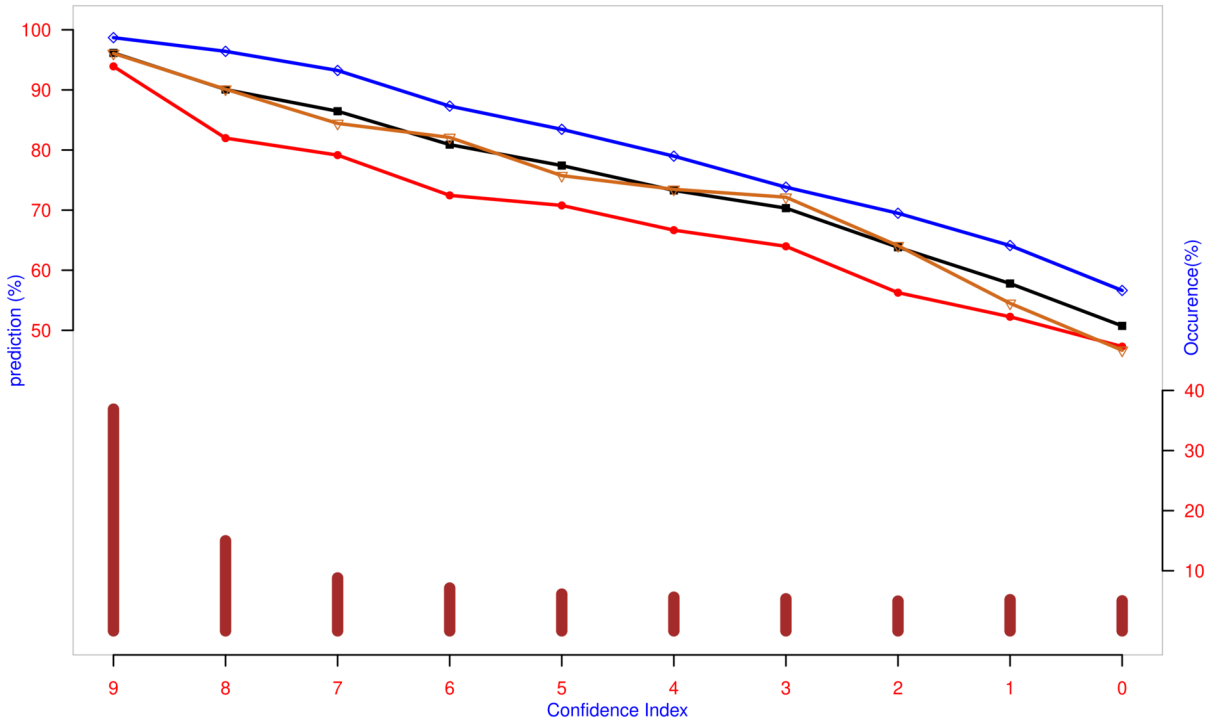
[30] Y. Yang, J. Gao, J. Wang, R. Heffernan, J. Hanson, K. Paliwal, Y. Zhou, Sixty-five years of the long march in protein secondary structure prediction: the final stretch?, Brief Bioinform, 19 (2018) 482-494.

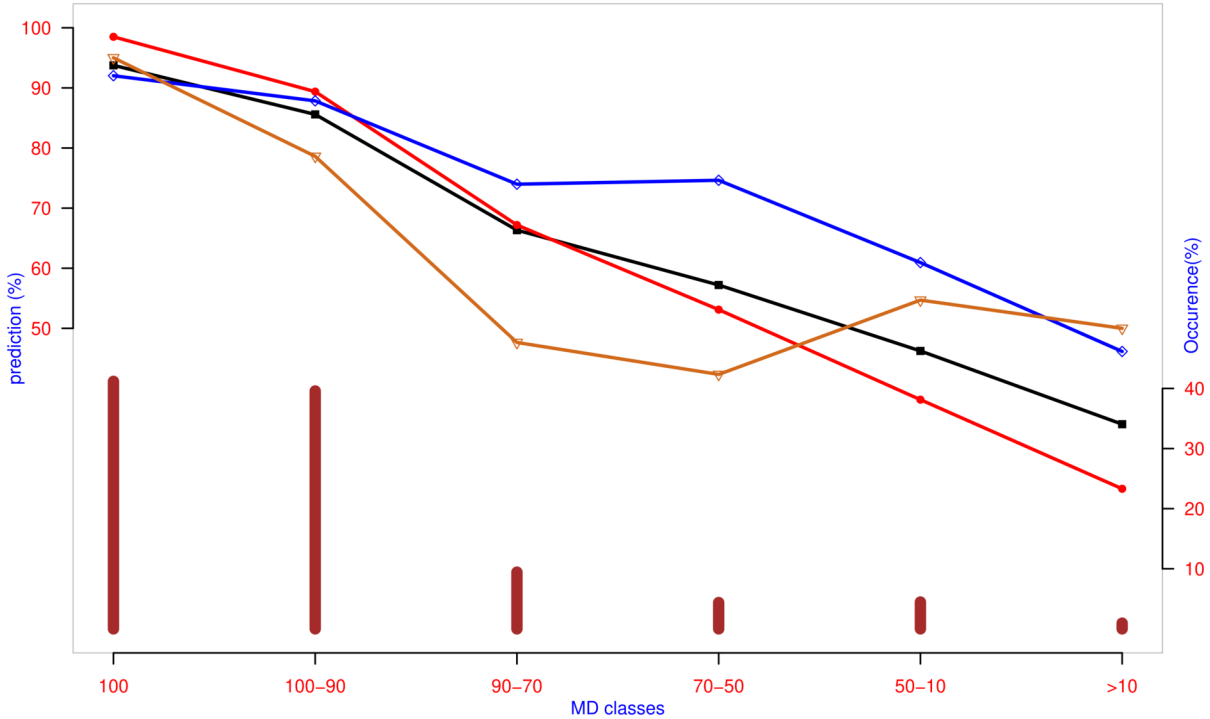
[31] B. Rost, C. Sander, R. Schneider, Redefining the goals of protein secondary structure prediction, J Mol Biol, 235 (1994) 13-26.

[32] B. Rost, Review: protein secondary structure prediction continues to rise, J Struct Biol, 134 (2001) 204-218.
[33] M. Orozco, A theoretical view of protein dynamics, Chem Soc Rev, 43 (2014) 5051-5066.
[34] P. Craveur, A.P. Joseph, J. Esque, T.J. Narwani, F. Noel, N. Shinada, M. Goguet, S. Leonard, P. Poulain, O. Bertrand, G. Faure, J. Rebehmed, A. Ghozlane, L.S. Swapna, R.M. Bhaskara, J. Barnoud, S. Teletchea, V. Jallu, J. Cerny, B. Schneider, C. Etchebest, N. Srinivasan, J.C. Gelly, A.G. de Brevern, Protein flexibility in the light of structural alphabets, Front Mol Biosci, 2 (2015) 20.
[35] B. Erman, Universal features of fluctuations in globular proteins, Proteins, 84 (2016) 721-725.
[36] O. Carugo, How large B-factors can be in protein crystal structures, BMC Bioinformatics, 19 (2018) 61.
[37] D.A. Beck, V. Daggett, Methods for molecular dynamics simulations of protein folding/unfolding in solution, Methods, 34 (2004) 112-120.
[38] P. Craveur, A.P. Joseph, J. Rebehmed, A.G. de Brevern, beta-Bulges: extensive structural analyses of beta-sheets irregularities, Protein Sci, 22 (2013) 1366-1378.
[39] T.J. Narwani, P. Craveur, N.K. Shinada, A. Floch, H. Santuz, A.M. Vattekatte, N. Srinivasan, J. Rebehmed, J.C. Gelly, C. Etchebest, A.G. de Brevern, Discrete analyses of protein dynamics, J Biomol Struct Dyn, (2019) 1-15.
[40] A. Melarkode Vattekatte, T.J. Narwani, A. Floch, M. Maljkovic, S. Bisoo, N.K. Shinada, A. Kranjc, J.C. Gelly, N. Srinivasan, N. Mitic, A.G. de Brevern, Data set of intrinsically disordered proteins analysed at a local protein conformation level, Data Brief, 29 (2020) 105383.
[41] A. Melarkode Vattekatte, T.J. Narwani, A. Floch, M. Maljkovic, S. Bisoo, N.K. Shinada, A. Kranjc, J.C. Gelly, N. Srinivasan, N. Mitic, A.G. de Brevern, A structural entropy index to analyse local conformations in intrinsically disordered proteins, J Struct Biol, 210 (2020) 107464.
[42] N.K. Fox, S.E. Brenner, J.M. Chandonia, SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures, Nucleic Acids Res, 42 (2014) D304-309.
[43] J.M. Chandonia, G. Hon, N.S. Walker, L. Lo Conte, P. Koehl, M. Levitt, S.E. Brenner, The ASTRAL Compendium in 2004, Nucleic Acids Res, 32 (2004) D189-192.
[44] P. Craveur, J. Rebehmed, A.G. de Brevern, PTM-SD: a database of structurally resolved and annotated posttranslational modifications in proteins, Database (Oxford), 2014 (2014).
[45] S. Pronk, S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M.R. Shirts, J.C. Smith, P.M. Kasson, D. van der Spoel, B. Hess, E. Lindahl, GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit, Bioinformatics, 29 (2013) 845-854.
[46] W.F. van Gunsteren, S.R. Billeter, A.A. Eising, P.H. Hünenberger, P. Krüger, A.E. Mark, W.R.P. Scott, I.G. Tironi, Biomolecular Simulation: The GROMOS96 Manual and User Guide, (1996) 1042.
[47] W.L. Jorgensen, J.D. Madura, Quantum and statistical mechanical studies of liquids. 25. Solvation and conformation of methanol in water, J. Am. Chem. Soc., , 105 (1983) 1407-1413.
[48] H.J.C. Berendsen, J.P.M. Postma, W.F. van Gunsteren, A. DiNola, J.R. Haak, Molecular dynamics with coupling to an external bath, The Journal of Chemical Physics, 81 (1984) 3684-3690.
[49] M. Parrinello, A. Rahman, Polymorphic transitions in single crystals: A new molecular dynamics method, Journal of Applied Physics, 52 (1981) 7182-7190.
[50] B. Hess, H. Bekker, H.J.C. Berendsen, J.G.E.M. Fraaije, LINCS: a linear constraint solver for molecular simulations, J. Comp. Chem. , 18 (1997) 1463-1472.
[51] T. Darden, L. Perera, L. Li, L. Pedersen, New tricks for modelers from the crystallography toolkit: the particle mesh Ewald algorithm and its use in nucleic acid simulations, Structure, 7 (1999) R55-60.
[52] A. Bornot, C. Etchebest, A.G. de Brevern, Predicting protein flexibility through the prediction of local structures, Proteins, 79 (2011) 839-852.
[53] W.G. Touw, C. Baakman, J. Black, T.A. te Beek, E. Krieger, R.P. Joosten, G. Vriend, A series of PDB-related databanks for everyday needs, Nucleic Acids Res, 43 (2015) D364-368.
[54] A.G. de Brevern, C. Etchebest, S. Hazout, Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks, Proteins, 41 (2000) 271-287.
[55] A.G. de Brevern, New assessment of a structural alphabet, In Silico Biol, 5 (2005) 283-289.
[56] A.P. Joseph, G. Agarwal, S. Mahajan, J.C. Gelly, L.S. Swapna, B. Offmann, F. Cadet, A. Bornot, M. Tyagi, H. Valadie, B. Schneider, C. Etchebest, N. Srinivasan, A.G. De Brevern, A short survey on protein blocks, Biophys Rev, 2 (2010) 137-147.
[57] A.P. Joseph, A. Bornot, A.G. de Brevern, Local Structure Alphabets, in: H. Rangwala, G. Karypis (Eds.) Protein Structure Prediction wiley2010, pp. in press.
[58] J. Barnoud, H. Santuz, P. Craveur, A.P. Joseph, V. Jallu, A.G. de Brevern, P. Poulain, PBxplore: A Tool To Analyze Local Protein Structure And Deformability With Protein Blocks, PeerJ, 5 (2017) e4013.
[59] J. Cerny, P. Bozikova, A. Balik, S.M. Marques, L. Vyklicky, NMDA Receptor Opening and Closing-Transitions of a Molecular Machine Revealed by Molecular Dynamics, Biomolecules, 9 (2019).
[60] A.G. de Brevern, Analysis of Protein Disorder Predictions in the Light of a Protein Structural Alphabet, Biomolecules, 10 (2020).
[61] D.T. Jones, Protein secondary structure prediction based on position-specific scoring matrices, J Mol Biol, 292

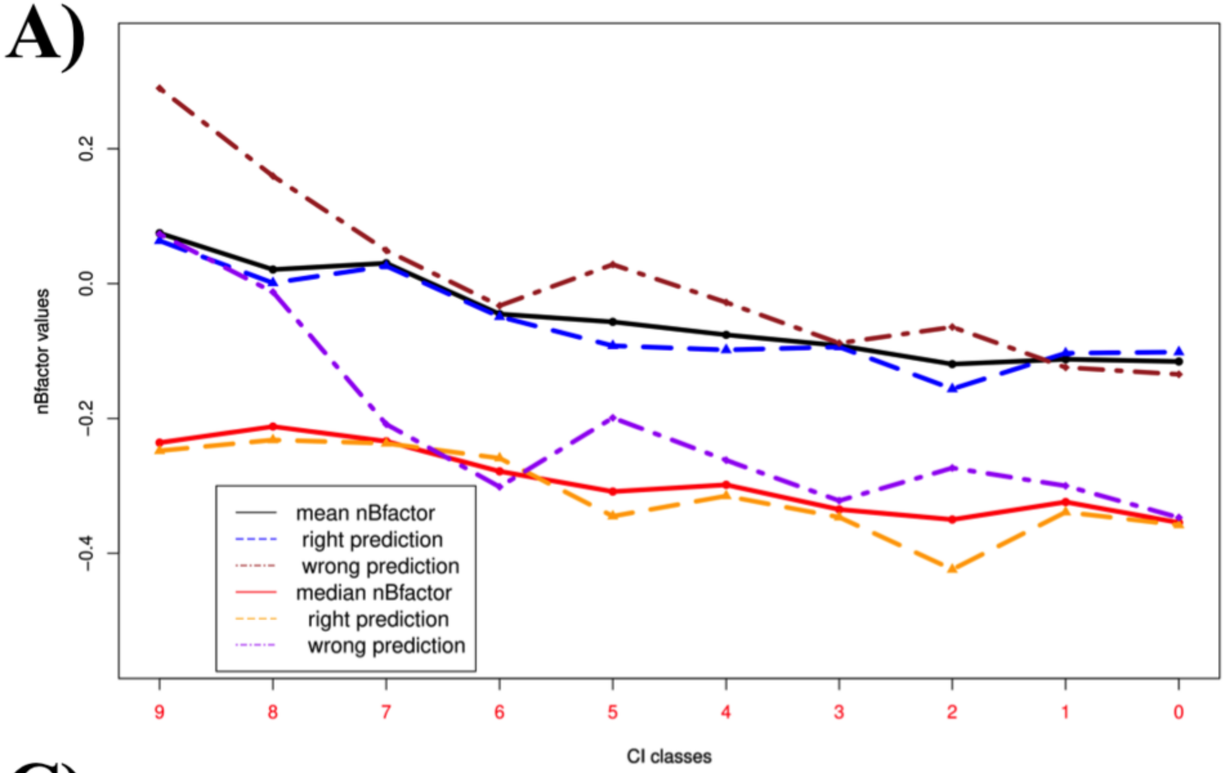
(1999) 195-202.
[62] D.W.A. Buchan, D.T. Jones, The PSIPRED Protein Analysis Workbench: 20 years on, Nucleic Acids Res, 47 (2019) W402-W407.
[63] J. Cheng, A.Z. Randall, M.J. Sweredoski, P. Baldi, SCRATCH: a protein structure and structural feature prediction server, Nucleic Acids Res, 33 (2005) W72-76.
[64] C.N. Magnan, P. Baldi, SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity, Bioinformatics, 30 (2014) 2592-2597.
[65] D.W. Buchan, F. Minneci, T.C. Nugent, K. Bryson, D.T. Jones, Scalable web services for the PSIPRED Protein Analysis Workbench, Nucleic Acids Res, 41 (2013) W349-357.
[66] L.J. McGuffin, K. Bryson, D.T. Jones, The PSIPRED protein structure prediction server, Bioinformatics, 16 (2000) 404-405.
[67] B. Rost, C. Sander, Prediction of protein secondary structure at better than 70% accuracy, J Mol Biol, 232 (1993) 584-599.
[68] R.C. Team, R: A Language and Environment for Statistical Computing, in: R.F.f.S. Computing (Ed.)Vienna, Austria, 2017.
[69] B.W. Kernighan, D.M. Ritchie, The C Programming Language (2nd edition), Prentice Hall Software Series1988.
[70] G. Van Rossum, F.L. Drake Jr, Python tutorial, Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands1995.
[71] M. Tyagi, A. Bornot, B. Offmann, A.G. de Brevern, Protein short loop prediction in terms of a structural alphabet, Comput Biol Chem, 33 (2009) 329-333.
[72] M. Tyagi, A. Bornot, B. Offmann, A.G. de Brevern, Analysis of loop boundaries using different local structure assignment methods, Protein Sci, 18 (2009) 1869-1881.
[73] J. Andreani, J. Soding, bbcontacts: prediction of beta-strand pairing from direct coupling patterns, Bioinformatics, 31 (2015) 1729-1737.
[74] B. Rost, Rising accuracy of protein secondary structure prediction, in: D. Chasman (Ed.) Protein structure determination, analysis, and modeling for drug discovery, Dekker, New York, 2003, pp. 207-249.
[75] X.-M. Pan, W.-D. Niu, Z.-X. Wang, What Is the Minimum Number of Residues to Determine the Secondary Structural State?, J Prot Chem, 18 (1999) 579-584.
[76] I. Jacoboni, P.L. Martelli, P. Fariselli, M. Compiani, R. Casadio, Predictions of protein segments with the same aminoacid sequence and different secondary structure: a benchmark for predictive methods, Proteins, 41 (2000) 535-544.
[77] A. Ghozlane, A.P. Joseph, A. Bornot, A.G. de Brevern, Analysis of protein chameleon sequence characteristics, Bioinformation, 3 (2009) 367-369.
[78] O. Carugo, P. Argos, Protein-protein crystal-packing contacts, Protein Sci, 6 (1997) 2261-2263.
[79] T.J. Narwani, C. Etchebest, P. Craveur, S. Leonard, J. Rebehmed, N. Srinivasan, A. Bornot, J.C. Gelly, A.G. de Brevern, In silico prediction of protein flexibility with local structure approach, Biochimie, 165 (2019) 150-155.
[80] H. Shukla, V. Kumar, A.K. Singh, N. Singh, M. Kashif, M.I. Siddiqi, M. Yasoda Krishnan, M. Sohail Akhtar, Insight into the structural flexibility and function of Mycobacterium tuberculosis isocitrate lyase, Biochimie, 110 (2015) 73-80.
[81] T. Masuda, K. Okubo, K. Murata, B. Mikami, M. Sugahara, M. Suzuki, P.A. Temussi, F. Tani, Subatomic structure of hyper-sweet thaumatin D21N mutant reveals the importance of flexible conformations for enhanced sweetness, Biochimie, 157 (2019) 57-63.
[82] M.C. Nonato, R.A.P. de Padua, J.S. David, R.A.G. Reis, G.P. Tomaleri, H. D'Muniz Pereira, F.A. Calil, Structural basis for the design of selective inhibitors for Schistosoma mansoni dihydroorotate dehydrogenase, Biochimie, 158 (2019) 180-190.
[83] P. Goettig, H. Brandstetter, V. Magdolen, Surface loops of trypsin-like serine proteases as determinants of function, Biochimie, 166 (2019) 52-76.
[84] M. Kaleel, M. Torrisi, C. Mooney, G. Pollastri, PaleAle 5.0: prediction of protein relative solvent accessibility by deep learning, Amino Acids, 51 (2019) 1289-1296.
[85] J. Zhou, H. Wang, Z. Zhao, R. Xu, Q. Lu, CNNH_PSS: protein 8-class secondary structure prediction by convolutional neural network with highway, BMC Bioinformatics, 19 (2018) 60.
[86] T. Sidi, C. Keasar, Redundancy-Weighting the PDB for Detailed Secondary Structure Prediction Using Deep-Learning Models, Bioinformatics, (2020).
[87] G. Bouvignies, P. Markwick, R. Bruschweiler, M. Blackledge, Simultaneous determination of protein backbone structure and dynamics from residual dipolar couplings, J Am Chem Soc, 128 (2006) 15100-15101.
[88] E. Frezza, R. Lavery, Internal Normal Mode Analysis (iNMA) Applied to Protein Conformational Flexibility, J Chem Theory Comput, 11 (2015) 5503-5512.
[89] M. Jamroz, A. Kolinski, S. Kmiecik, CABS-flex: Server for fast simulation of protein structure fluctuations, Nucleic Acids Res, 41 (2013) W427-431.

[90] M. Kurcinski, T. Oleniecki, M.P. Ciemny, A. Kuriata, A. Kolinski, S. Kmiecik, CABS-flex standalone: a simulation environment for fast modeling of protein flexibility, Bioinformatics, 35 (2019) 694-695.
[91] L. Salmon, G. Bouvignies, P. Markwick, M. Blackledge, Nuclear magnetic resonance provides a quantitative description of protein conformational flexibility on physiologically important time scales, Biochemistry, 50 (2011) 2735-2747.
[92] A.G. de Brevern, A. Bornot, P. Craveur, C. Etchebest, J.C. Gelly, PredyFlexy: flexibility and local structure prediction from sequence, Nucleic Acids Res, 40 (2012) W317-322.
[93] A. Schlessinger, G. Yachdav, B. Rost, PROFbval: predict flexible and rigid residues in proteins, Bioinformatics, 22 (2006) 891-893.
[94] I. Majumdar, S.S. Krishna, N.V. Grishin, PALSSE: a program to delineate linear secondary structural elements from protein structures, BMC Bioinformatics, 6 (2005) 202.
[95] D. Frishman, P. Argos, Knowledge-based protein secondary structure assignment, Proteins, 23 (1995) 566-579.
[96] N. Colloc'h, C. Etchebest, E. Thoreau, B. Henrissat, J.P. Mornon, Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment, Protein Eng, 6 (1993) 377-382.
[97] P. Kumar, M. Bansal, Structural and functional analyses of PolyProline-II helices in globular proteins, J Struct Biol, 196 (2016) 414-425.
[98] S. Kumar, M. Bansal, Geometrical and sequence characteristics of alpha-helices in globular proteins, Biophys J, 75 (1998) 1935-1944.

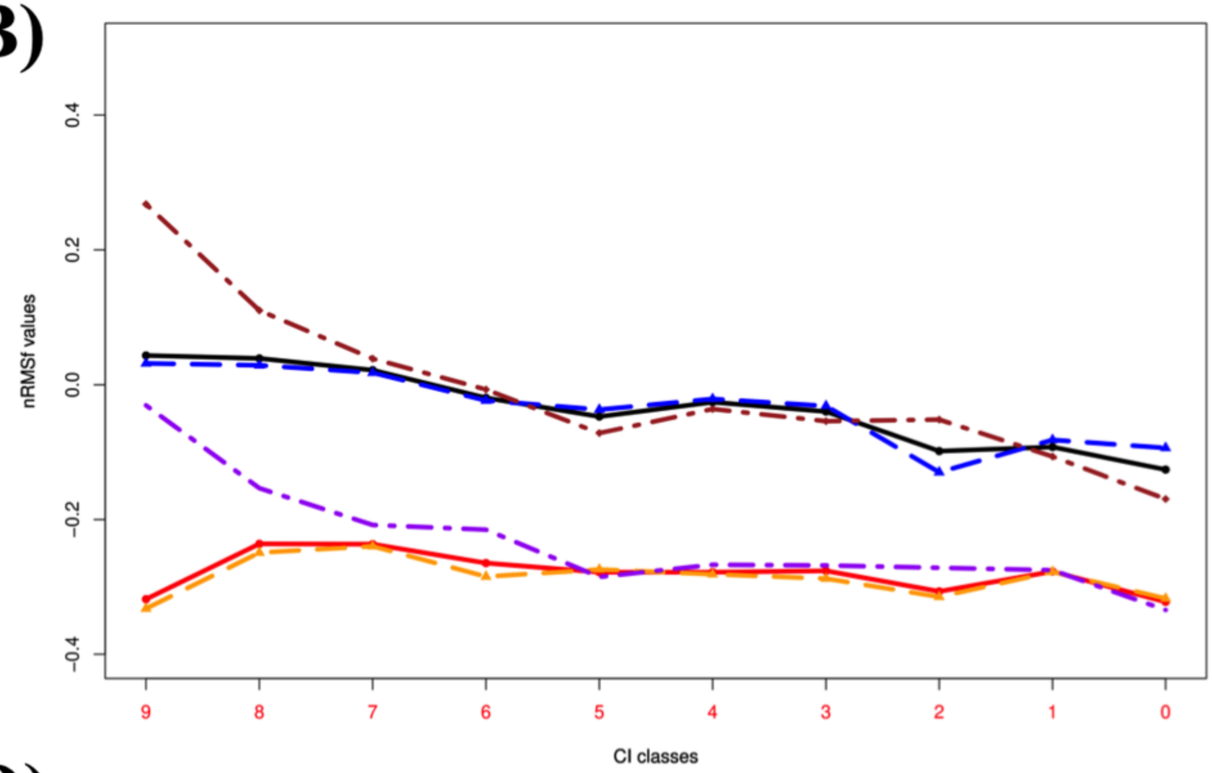

100
90
80
70
60
50
prediction (%)
40
30
20
10
Occurence(%)
9
8
7
6
5
4
3
2
1
0
Confidence Index


100
90
80
70
60
50
prediction (%)
Occurence(%)
40
30
20
10
100
100–90
90–70
70–50
50–10
>10
MD classes

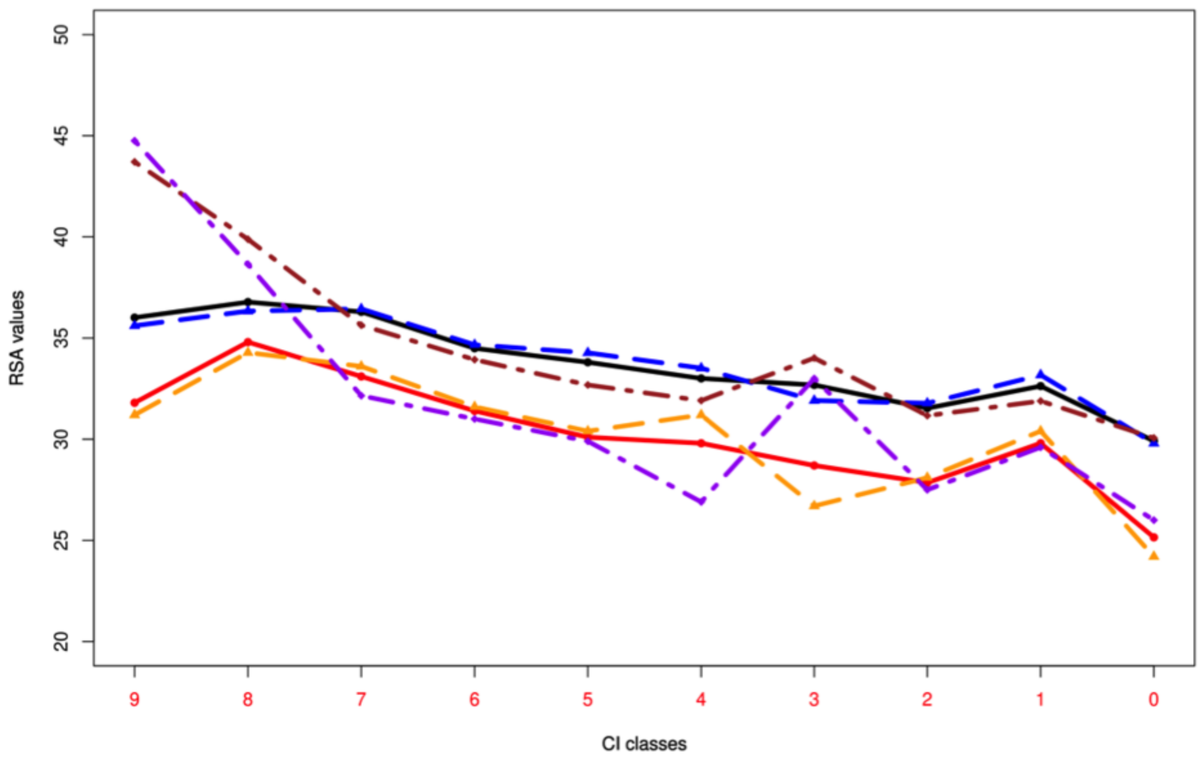
A)


nBfactor values
0.2
0.0
−0.2
−0.4
mean nBfactor
right prediction
wrong prediction
median nBfactor
right prediction
wrong prediction
9
8
7
6
5
4
3
2
1
0
CI classes

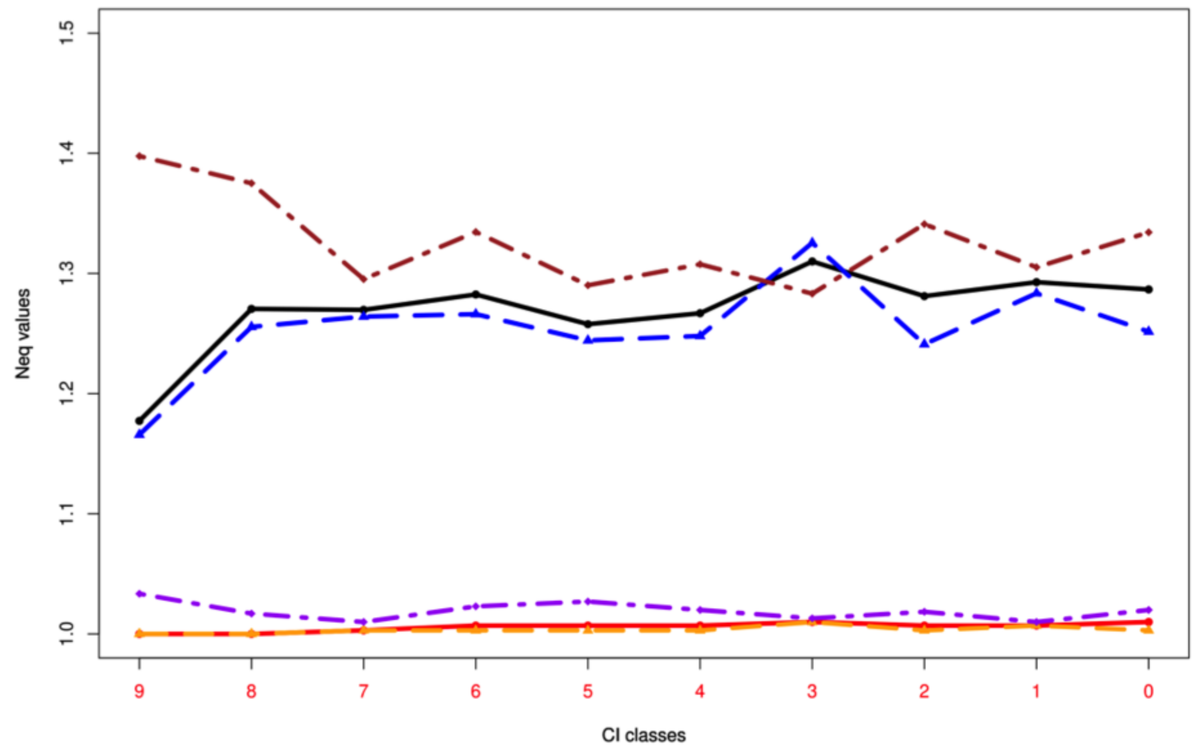
B)


nRMSf values
0.4
0.2
0.0
−0.2
−0.4
9
8
7
6
5
4
3
2
1
0
CI classes

C)


RSA values
50
45
40
35
30
25
20
9
8
7
6
5
4
3
2
1
0
CI classes

D)


Neq values
1.5
1.4
1.3
1.2
1.1
1.0
9
8
7
6
5
4
3
2
1
0
CI classes

(a)

PSIPRED	DSSP	H	C	E	
H	35.8	**80.3**	18.4	1.3	
C	39.7	4.5	**88.3**	7.2	
E	24.5	0.77	18.29	**80.9**	
Q_3					**83.6**
inad.					0.64

(b)

SSpro3	DSSP	H	C	E	
H	35.5	**84.5**	13.6	1.8	
C	39.0	8.3	**82.5**	9.2	
E	25.5	2.9	22.6	**74.4**	
Q_3					**81.2**
inad.					1.4

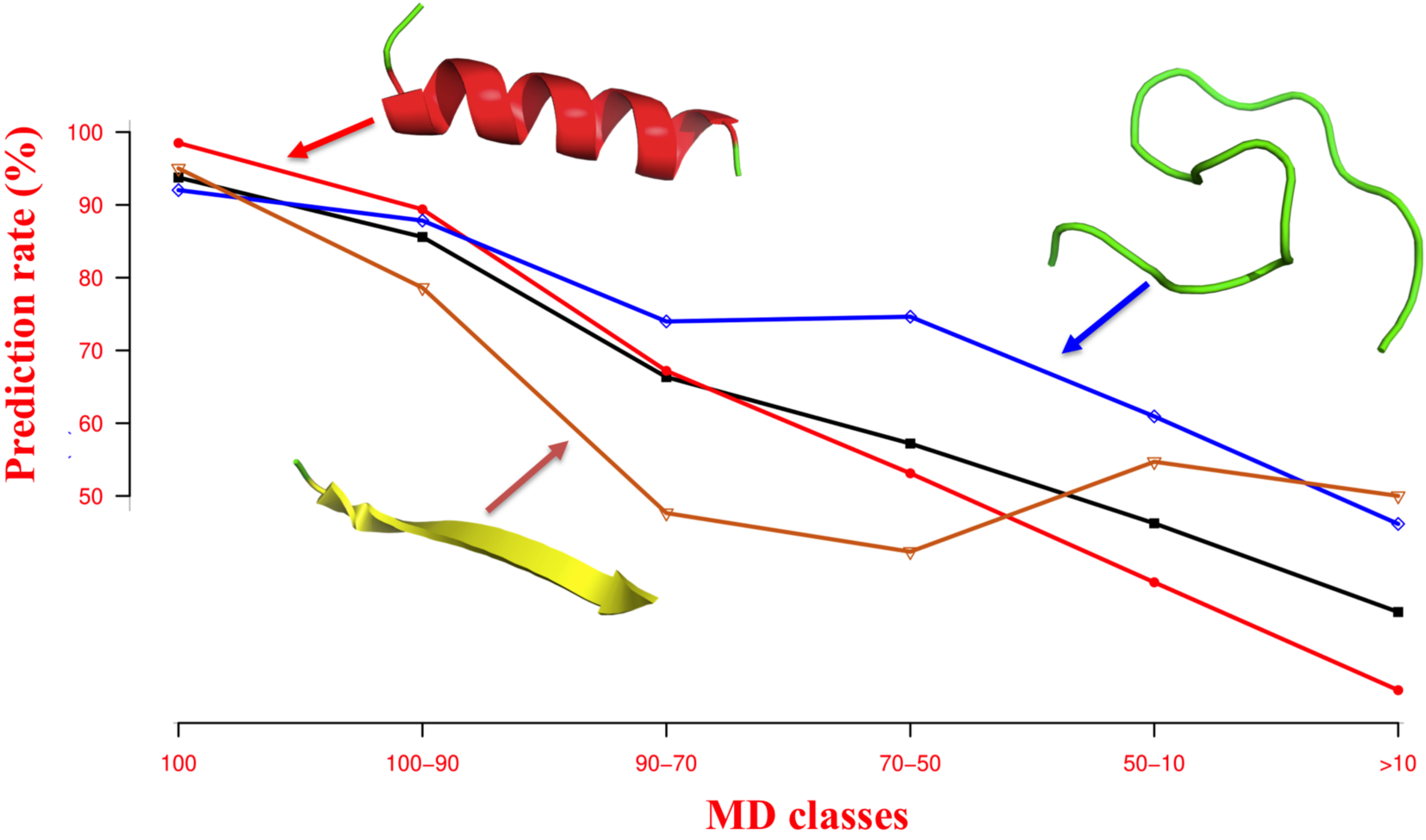
(c)

SSpro8to3	DSSP	H	C	E	
H	35.5	**86.2**	10.9	2.8	
C	39.0	10.5	**75.8**	13.7	
E	25.5	2.9	15.9	**81.2**	
Q_3					**80.8**
inad.					1.7

Table 1. *Secondary structure prediction assessment*. (a) PSIPRED, (b) SSpro3 and (c) SSpro8to3. SSpro3 and SSpro8 are used without the help of homologous structures. Is provided the corresponding frequency as assigned by DSSP and the confusion matrix between observed and predicted secondary structures. Finally is given the global prediction rate Q_3 and percentage of highly inadequate prediction (named *inad.* for helix residue predicted as sheet and inversely)

	DSSP MD	H	C	E	
H	33.1	**84.0**	14.9	1.1	
C	42.7	6.3	**86.1**	7.6	
E	24.2	0.7	18.4	**80.8**	
Q_3					**84.1**
inad.					0.54

Table 2. *Secondary structure prediction when the assignment is taken as the most frequent ones observed during MD simulations*. See Table 1 for legend.


Prediction rate (%)
100
90
80
70
60
50
100
100–90
90–70
70–50
50–10
>10
MD classes