PLOS ONE

RESEARCH ARTICLE

SAFlex: A structural alphabet extension to integrate protein structural flexibility and missing data information

Ikram Allam[1,2,3,4], Delphine Flatters[1,4], Géraldine Caumes[1,4], Leslie Regad[1,4], Vincent Delos[2,3,4], Gregory Nuel[2,3,4☯]*, Anne-Claude Camproux[1,4☯]*

1 Molécules thérapeutiques in silico (MTi), INSERM UMR-S973, University Paris Diderot, Paris 7, France, **2** Probability Statistique and Biology (PSB), LPMA laboratory, CNRS INSMI UMR 7599, University Pierre et Marie Curie, Paris 6, France, **3** Mathématiques Appliquées, MAP5 laboratory, CNRS UMR 8145, University Paris Descartes, Paris 5, France, **4** Sorbonne Paris Cité, Paris, France

☯ These authors contributed equally to this work.
* Gregory.Nuel@math.cnrs.fr (GN); Anne-Claude.Camproux@univ-paris-diderot.fr (AC)

Citation: Allam I, Flatters D, Caumes G, Regad L, Delos V, Nuel G, et al. (2018) SAFlex: A structural alphabet extension to integrate protein structural flexibility and missing data information. PLoS ONE 13(7): e0198854. https://doi.org/10.1371/journal.pone.0198854

Editor: Manuela Helmer-Citterich, Universita degli Studi di Roma Tor Vergata, ITALY

Received: October 18, 2017

Accepted: May 25, 2018

Published: July 5, 2018

Data Availability Statement: All relevant data are in the Protein Data Bank (2hba.pdb, 3h8z.pdb, 1gme.pdb) or within the paper and its Supporting Information files.

Funding: This work was supported by a USPC grant (SA-Flex) and an ANR grant (ANR-10-BINF-0003, BIP-BIP).

Competing interests: The authors have declared that no competing interests exist.

Abstract

In this paper, we describe SAFlex (Structural Alphabet Flexibility), an extension of an existing structural alphabet (HMM-SA), to better explore increasing protein three dimensional structure information by encoding conformations of proteins in case of missing residues or uncertainties. An SA aims to reduce three dimensional conformations of proteins as well as their analysis and comparison complexity by simplifying any conformation in a series of structural letters. Our methodology presents several novelties. Firstly, it can account for the encoding uncertainty by providing a wide range of encoding options: the maximum a posteriori, the marginal posterior distribution, and the effective number of letters at each given position. Secondly, our new algorithm deals with the missing data in the protein structure files (concerning more than 75% of the proteins from the Protein Data Bank) in a rigorous probabilistic framework. Thirdly, SAFlex is able to encode and to build a consensus encoding from different replicates of a single protein such as several homomer chains. This allows localizing structural differences between different chains and detecting structural variability, which is essential for protein flexibility identification. These improvements are illustrated on different proteins, such as the crystal structure of an eukaryotic small heat shock protein. They are promising to explore increasing protein redundancy data and obtain useful quantification of their flexibility.

Introduction

Over the past two decades, the notion of a structural alphabet (SA) has attracted much attention. SA encodes protein fragments into structural letters (SL). SA encoding plays a key role in compressing the three-dimensional (3D) protein conformations into a one-dimensional (1D) SL representation, thereby allowing for a simplified protein structure analysis [1–5]. This approach also dramatically simplifies the comparison of 3D conformations by using well-

Abbreviations: ENT, marginal posterior entropy; HMM, hidden Markov model; MAP, Maximum a Posteriori; NEFF, effective number of structural letters; PDB, Protein Data Bank; POST, marginal posterior distribution; SA, structural alphabet; SL, structural letters; 1D, one-dimensional; 3D, three-dimensional.

known sequence comparison algorithms (ex: local score from Smith and Waterman, [6]) on SL sequences.

Many studies have developed SAs, based on mixture models [7], classification methods such as AutoANN [8], SOM [9] and K-Nearest Neighbor [10]: Structural Building Blocks [11], Protein Blocks [4], SABD [12] and USA [13], M32K25 [14]; and hidden Markov model (HMM): HMM-SA [2, 3, 15]. The choice between these methods and models plays a major part in the construction of an accurate SA. They have been applied in the past to protein structure analysis, including multiple structure alignment, structure mining [16, 17], protein fold classification [18], dynamic molecular analysis [19–21], structure fast comparison [17] and generation of 3D peptide conformations [22, 23]. The SA approach also appears to be promising to characterize structural variability [24, 25], to explore the local backbone deformation involved in protein-protein interactions [26, 27] and to predict local protein flexibility [28, 29].

However, current SAs have not been trained to take into account the wealth of available protein structure data: their uncertainty (ex: missing data) and redundancy (ex: multiple homomers chains). The growth and speed of macromolecular structure determination techniques (protein crystallography and NMR spectroscopy) results in a considerable increase to 3D structures in the Protein Data Bank (PDB, [30, 31]), which currently has more than 130,000 3D protein structures. More than half of PDB structures share at least 95% sequence identity. Even if this redundancy is considered valuable in investigating families of homologous sequences [32, 33], the dominant approach for data mining the PDB considers redundancy as non-informative [34], resulting in an artificial reduction in the variability of the structural space. Yet, protein redundancy analysis is crucial for protein flexibility insight. Proteins are highly flexible macromolecules and their 3D folding and dynamic properties are essential in many biological processes [32, 34]. Integrating PDB redundancy has potential to improve understanding of protein intrinsic flexibility [35]. PDB files include monomers corresponding to individual protein chain but also homomeric complexes formed by the assembly of multiple copies of a single type of polypeptide chain, and heteromic complexes formed from multiple distinct polypeptide chains [36]. Different PDB files can also correspond to a same protein in different conditions, called multi-conformations. For instance, in 2015, the non-redundant snapshot of protein crystal structures contained 7,972 monomers and 9,206 homomers and 2,677 heteromers [37]. The authors concluded, 87% of crystal structures involve only a single type of polypeptide chain, and a slight majority (54%) of these self-assemble into homomers. Thus, it is a very important point to be able to model this multiple chain data.

Another source of uncertainty in PDB data analysis is that most PDB structures have some missing parts which strongly impact the determination of their accurate protein folding and lead to important difficulty for protein structure and function interpretation [38]. This kind of issue is very serious, from missing side chains, entire loop regions, to whole domain. For instance, it was demonstrated in 2007 by [39] that $\sim$10% of 16,370 PDB X-ray structure files contain regions of more than 30 missing or ambiguous amino acids and $\sim$40% have missing or ambiguous regions between 10 and 30 amino acids. These missing parts can result from some resolution difficulty or from intrinsic flexibility of proteins [38, 39]. The absence of the coordinates of the alpha-carbon, which makes it possible to connect to the peptide skeleton, poses a serious problem because it effectively prevents knowledge of the secondary structure and of the 3D folding. These missing parts often relate to the ends of the protein or loops that are the most flexible regions of proteins and involved in protein interactions and function. Thus the detection and modeling of these missing data could have a very appealing impact for protein structure analysis. However, they have not been explicitly modeled by different SA approaches.

In this paper, we extend a previously published SA, HMM-SA [3, 15, 40], in SAFlex (Structural Alphabet Flexibility) to encode 3D conformations of proteins in case of missing residues or uncertainties with the aim to better explore increasing protein 3D structure information. HMM-SA was modeled using HMM, which provides a very precise description of protein structures, particularly loop regions [41] known to play important roles in protein function. One major contribution of HMM is that this model implicitly takes the SL sequential connections into account. For example, this markovian modeling allows for efficient extraction of functional motifs [5, 41]. The paper is organized as follows. The "Materials and Methods" section provides descriptions of the improvements based on our HMM modeling. It presents a collection of technical advances including a wide range of encoding options (the maximum a posteriori, the marginal posterior distribution, and the effective number of letters at each given position), the robustness of missing data in the PDB files, and the ability to encode a monomer or an heteromer as well as different replicates of a single protein with multiple chains (homomers) leading to a consensus encoding. This approach is however not yet able to directly take into account protein with multi-conformations (*i.e.* several PDB files). The "Results" section illustrates the application of our new approach on different PDB structures of interest. Finally, the "Conclusion" section summarizes the manuscript and discusses potential development and application of SAFlex for addressing structural biology challenges.

Materials and methods

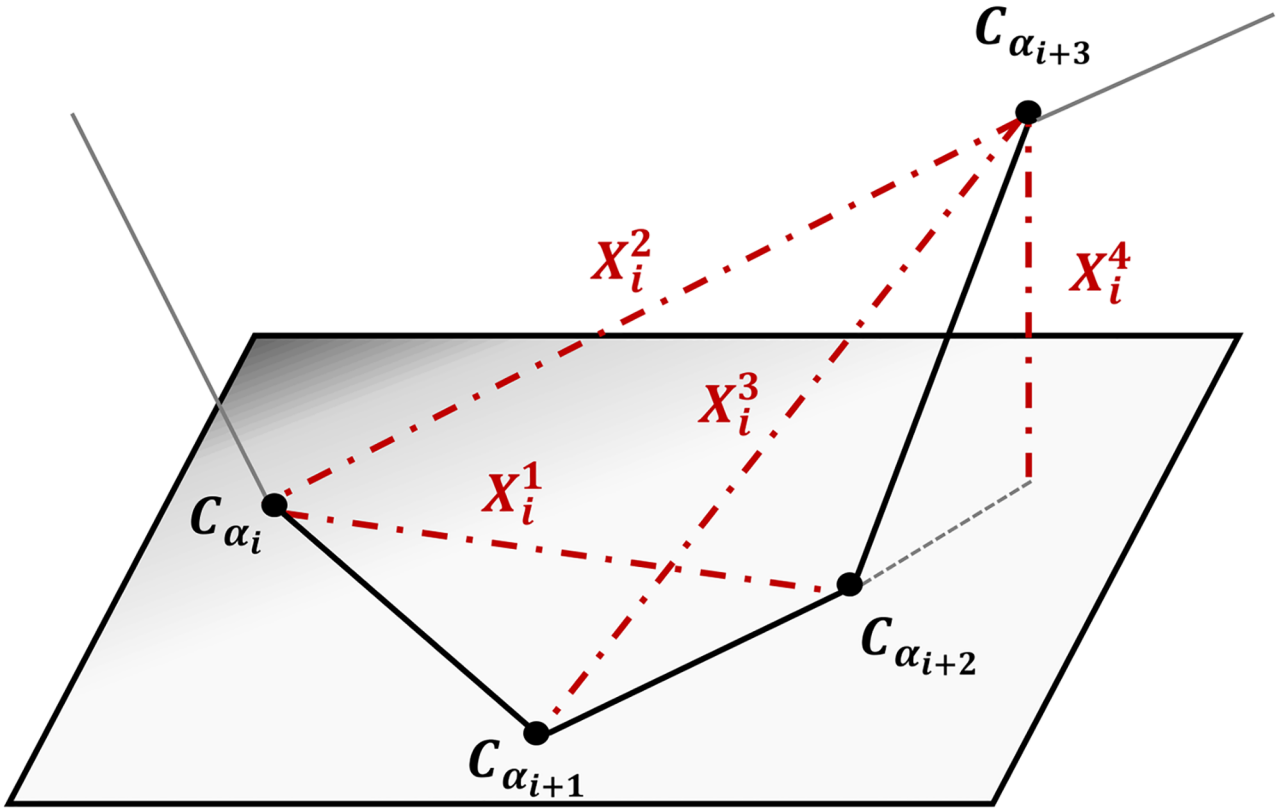
In this section, we describe the model used to encode protein structures (PDB files) into SL sequences. To that aim, we introduce an HMM in which the spatial conformation of the protein is the observation and the underlying structural sequence is the hidden part.

Structural fragments

The PDB file input, obtained from the worldwide Protein Data Bank (wwPDB) (http://www.wwpdb.org/) [42, 43], contains the atomic coordinates describing the 3D structure of the protein. Since the original observation (3D structures from PDB files) is highly complex, we start by reducing this complexity to a sequence of numeric descriptors as in the original HMM-SA publications [2, 3]. Starting from the 3D positions of the alpha carbons (denoted a C_α), "*fragment*" is a succession of four consecutive C_α and we use the four descriptors in Fig 1. Formally, for the i^{th} fragment we have: $X_i^1 = D(C_{\alpha_i}, C_{\alpha_{i+2}})$, $X_i^2 = D(C_{\alpha_i}, C_{\alpha_{i+3}})$, $X_i^3 = D(C_{\alpha_{i+1}}, C_{\alpha_{i+3}})$, and $X_i^4 = \eta D(C_{\alpha_{i+3}}, H)$ where D is the Euclidian distance, H is the orthogonal projection of $C_{\alpha_{i+3}}$ on the plane $(C_{\alpha_i}, C_{\alpha_{i+1}}, C_{\alpha_{i+2}})$, and where $\eta = +1$ (resp. -1) if the cross-product of vector $C_{\alpha_i} \rightarrow C_{\alpha_{i+2}}$ and vector $C_{\alpha_i} \rightarrow C_{\alpha_{i+1}}$ has the same (resp. opposite) direction than vector $H \rightarrow C_{\alpha_{i+3}}$. Since a structural fragment is formed by four consecutive alpha carbons, a sequence of n+ 3 alpha carbons will have only a total of n fragments with successive fragments overlapping on three alpha carbons. Hence the Fragment i corresponds to the four alpha carbons: $C_{\alpha_i}, C_{\alpha_{i+1}}, C_{\alpha_{i+2}}, C_{\alpha_{i+3}}$. Of course, numerous other structural alphabets are based on different geometrical descriptors such as angles and torsions descriptors but we can note several studies such as [1, 44–51] focus on geometrical descriptors based on RMSD, on cRMD between alpha carbons, or, like our own descriptors, using Euclidean distances between alpha carbons.

A hidden markov model

Our idea is to consider the sequence $X_{1:n} \in \mathbb{R}^{n\times 4}$ of n structural fragments as the observed states of an HMM where the hidden states are the SL $S_{1:n} \in \{1, \ldots, m\}^n$. A Markov dependency



$$X_i = (X_i^1, X_i^2, X_i^3, X_i^4) \in \mathbb{R}^4$$

Fig 1. The four descriptors $X_i = (X_i^1, X_i^2, X_i^3, X_i^4) \in \mathbb{R}^4$ for the i^{th} fragment (alpha carbons C_{α_i} to $C_{\alpha_{i+3}}$).

https://doi.org/10.1371/journal.pone.0198854.g001

is assumed among the m SL. A (conditional) Gaussian model is used for the fragment descriptor distribution. The resulting model, represented in Fig 2, has the following probability distribution:

$$\mathbb{P}(X_{1:n}, S_{1:n}) = \underbrace{\mathbb{P}(S_1)\prod_{i=2}^{n}\mathbb{P}(S_i|S_{i-1})}_{\text{Markov part}} \times \underbrace{\prod_{i=1}^{n}\mathbb{P}(X_i|S_i)}_{\text{Emission part}} \quad (1)$$

Assuming that the Markov chain has a uniform starting distribution, a homogeneous transition matrix $\pi \in \mathbb{R}^{m\times m}$, and that we denote $e_i(S_i) = \mathbb{P}(X_i|S_i)$ we get the following simplified equation:

$$\mathbb{P}(X_{1:n}, S_{1:n}) = \frac{1}{m}\prod_{i=2}^{n}\pi(S_{i-1}, S_i)\prod_{i=1}^{n}e_i(S_i) \quad (2)$$

For the emission distribution, we simply assume that fragment descriptors are Gaussian distributed, (with a specific mean vector $\mu_s \in \mathbb{R}^4$ and covariance matrix $\Sigma_s \in \mathbb{R}^{4\times4}$), for each

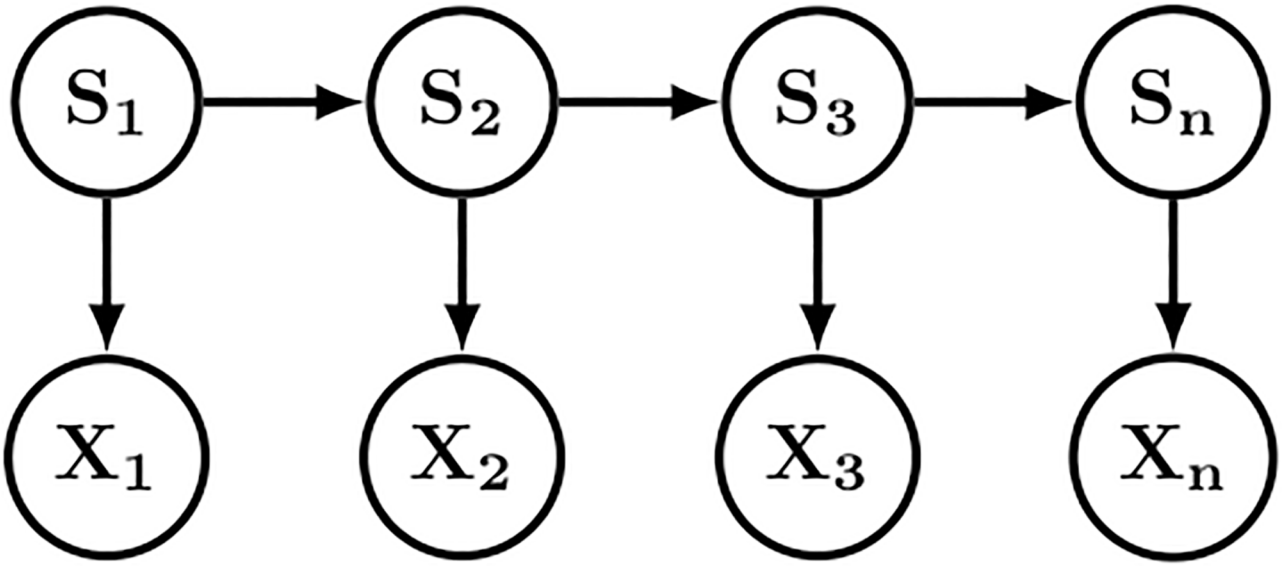


Fig 2. The HMM-SA model. $X_i \in \mathbb{R}^4$ are the fragment descriptors, and $S_i \in \{1, 2, \ldots, m\}$ are the structural letters.

https://doi.org/10.1371/journal.pone.0198854.g002

structural letter $s \in \{1, \ldots, m\}$ which hence gives:

$$\log e_i(s) = \text{cst.} - \frac{1}{2}\log\det\Sigma_s - \frac{1}{2}(X_i - \mu_s)^T\Sigma_s^{-1}(X_i - \mu_s). \tag{3}$$

One should note that the choice of a uniform starting distribution (rather than an estimation as in the case of original HMM-SA publications [2, 4]) has little impact on the encoding (Markov has a short range dependency), but is quite useful in terms of parsimony, since the resulting model has less parameters to estimate.

Protein encoding

Forward and backward. In order to perform exact inference in hidden Markov models, it is a common practice to introduce the so-called forward and backward quantities. The expression of these quantities along with the mathematical results allows for computing and deriving probabilistic quantities of interest. This can be found in classical reference textbooks (see [52] for example). For completeness, we present a brief form of the forward/backward quantity definition and the key mathematical results.

The Forward and Backward quantities are defined for all $i = 2 \ldots n$ and for all SL s by:

$$\begin{cases} F_i(s) = \sum_{S_{1:i-1}} \mathbb{P}(X_{1:i}, S_{1:i-1}, S_i = s) = \mathbb{P}(X_{1:i}, S_i = s) \\ B_{i-1}(s) = \sum_{S_{i:n}} \mathbb{P}(X_{1:i}, S_{i:n} | S_i = s) = \mathbb{P}(X_{1:i} | S_i = s) \end{cases} \tag{4}$$

with $F_1(s) = e_1(s)/m$ and $B_n(s) = 1$.

These quantities can be computed recursively through a recursion on $i = 2 \dots n$ (resp. $i = n \dots 2$) for forward (resp. backward) given for all SL r and s by:

$$\begin{cases} F_i(s) = \sum_r F_{i-1}(r)\pi(r,s)e_i(s) \\ B_{i-1}(r) = \sum_s \pi(r,s)e_i(s)B_i(s) \end{cases} \tag{5}$$

Marginal posterior distribution. First, we focus on the marginal posterior distribution (POST) which can be immediately derived from the forward/backward quantities:

$$\mathrm{POST}_i(s) = \mathbb{P}(S_i = s|X_{1:n}) \propto F_i(s)B_i(s) = \frac{F_i(s)B_i(s)}{\sum_r F_i(r)B_i(r)}. \tag{6}$$

This POST also can be used to quantify the level of uncertainty of the SL encoding by computing the marginal posterior entropy (ENT) and effective number of SL (NEFF).

$$\mathrm{ENT}_i = -\sum_s \mathrm{POST}_i(s) \log \mathrm{POST}_i(s) \quad \text{and} \quad \mathrm{NEFF}_i = \exp(\mathrm{ENT}_i) \tag{7}$$

ENT_i is a measure of disorder for Fragment i. If the posterior distribution is a Dirac, $\mathrm{ENT}_i = 0$, the minimal entropy, if the encoding uncertainty is maximum with $\mathrm{POST}_i(s) = 1/m$, $\mathrm{ENT}_i = \log m$ is maximal. The effective number of SL $\mathrm{NEFF}_i \in [1, m]$ provides a simpler interpretation of the entropy as the effective number of SL acceptable for Fragment i.

Maximum a posteriori. The task of encoding a 3D structure (sequence of n fragments) into a structural sequence can be achieved by computing the Maximum a Posteriori (MAP) defined by:

$$\mathrm{MAP}_{1:n} = \arg\max_{s_{1:n}} \mathbb{P}(S_{1:n}|X_{1:n}) = \arg\max_{s_{1:n}} \mathbb{P}(X_{1:n}, S_{1:n}). \tag{8}$$

For computing the MAP, we need to introduce the max-forward and max-backward quantities which are defined and recursively computed by simply replacing all 'Σ' occurrences by 'max' in Eqs (4) and (5). Once the max-forward and max-backward quantities are computed, it is possible to obtain the MAP immediately with:

$$\mathrm{MAP}_i = \arg\max_s F_i^{\max}(s)B_i^{\max}(s) \quad \text{for all } i = 1 \dots n. \tag{9}$$

Note that MAP_i is not necessarily equal to $\arg\max_s \mathrm{POST}_i(s)$ since POST is a *marginal* posterior distribution. However, the two quantities are often the same when the posterior distribution is 'sharp' enough.

Missing data

When dealing with 3D proteins structures, it is quite common that some alpha carbon positions are missing, which results in several missing data in fragment descriptors. Therefore it is necessary to deal with these missing data efficiently, which is quite straightforward under the Gaussian assumption. Let the subset of non-missing descriptors for fragment i be denoted by $J \subset \{1, 2, 3, 4\}$. $J = \emptyset$ if all descriptors are missing, and $J = \{1, 2, 3, 4\}$ if none are missing. Then the emission probability of fragment j for the structural letter s is obtained by considering the restriction $X_i[J] \sim \mathcal{N}(\mu_s[J], \Sigma_s[J, J])$. By convention, $e_i(s) = 1$ for all s if $J = \emptyset$ meaning that the totally missing fragment i is totally uninformative.

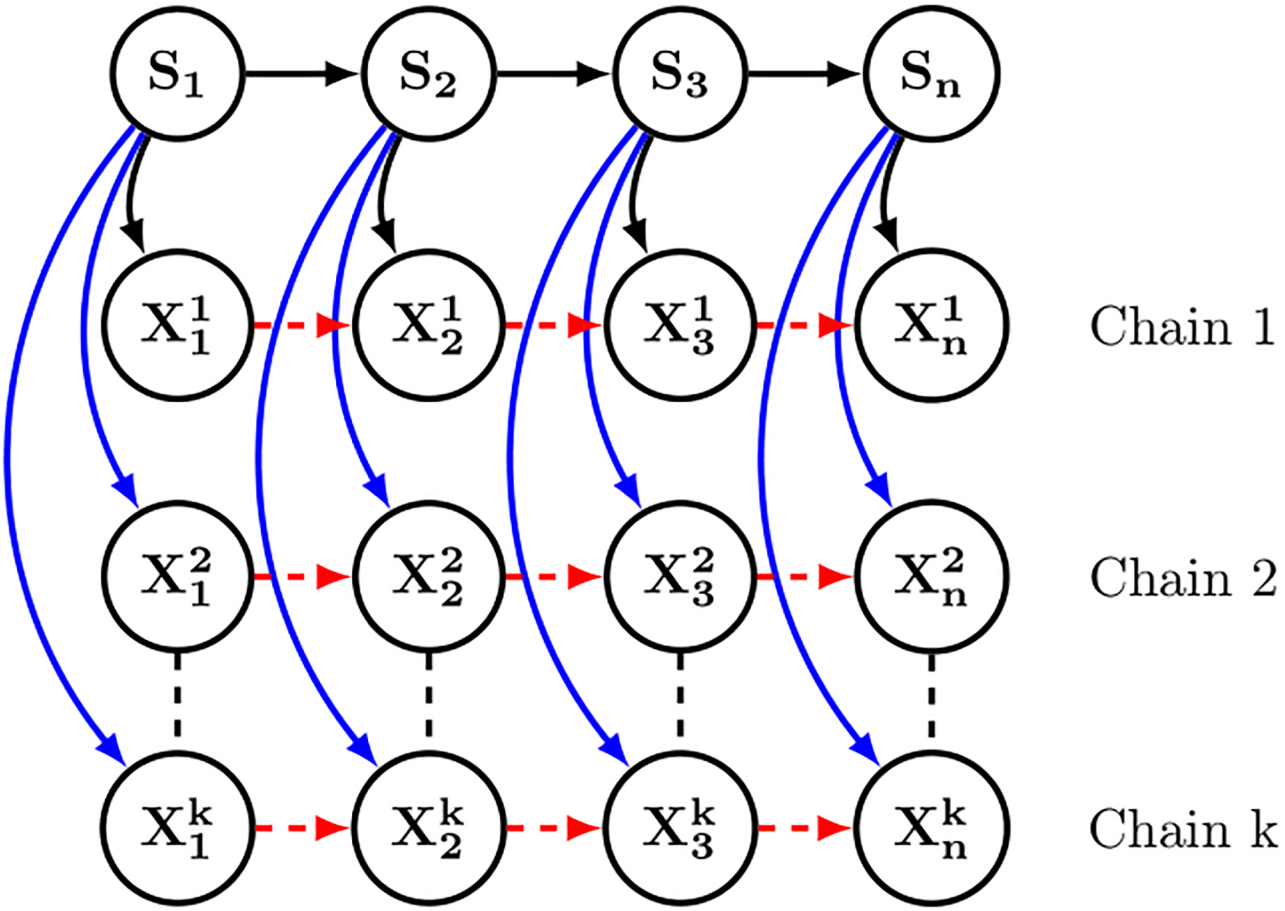


Fig 3. The SAFlex model with k chains.

https://doi.org/10.1371/journal.pone.0198854.g003

Multiple chains

In the PDB, some proteins are represented several times. They could be simple replicates, or 3D structures in different conditions (e.g. associated or not with various partners, with or without mutated positions). It is obviously possible to perform one encoding for each of these replicates, but there is also another possibility: build a consensus encoding from the whole replicate set. This can be done easily by considering the model in Fig 3 which results in the following likelihood:

$$\mathbb{P}(X_{1:n}^{1:k}, S_{1:n}) = \underbrace{\mathbb{P}(S_1)\prod_{i=2}^{n}\mathbb{P}(S_i|S_{i-1})}_{\text{consensus encoding}} \times \prod_{j=1}^{k}\underbrace{\prod_{i=1}^{n}\mathbb{P}(X_i^j|S_i)}_{\text{emission for Chain } j} \tag{10}$$

The previous encoding algorithm can be adapted easily to this new context by slightly changing the expression of $e_i(s)$:

$$\log e_i(s) = \text{cst.} - \sum_{j=1}^{k}\left(\frac{1}{2}\log\det\Sigma_s + (X_i^j - \mu_s)^T\Sigma_s^{-1}(X_i^j - \mu_s)\right). \tag{11}$$

Note that missing data can also be easily taken into account in this context. In the particular case when a large portion of data are missing in multiple chains, the model can easily cope with this situation as long as all the chains are properly aligned using a common reference index. In such a situation, a particular position would typically be informative only for a small portion of the chains, which is not a problem for the model.

SAFlex structural alphabet implementation

An SAFlex preliminary web server including dynamic pages (php/javascript/html/css) is made freely available to the scientific community. The backend was developed in C++ language as a high performance encoding program. All computations (evidence, forward, backward and posterior marginal) are performed in logarithmic scale and thus allow for low probabilities. The SAFlex server is able to complete encoding and indication of data uncertainties in less than one second (in most cases). The SAFlex web server has been successfully tested in the latest version of Chrome, Firefox and Safari. SAFlex is available at the following URL http://saflex.rpbs.univ-paris-diderot.fr/SA-Encoder.php.

Results

A new structural alphabet encoding: SAFlex

Here we propose to extend the SA-based approach, HMM-SA ([3, 15]) in SAFlex, to take into account 3D protein flexibility and redundancy as well as missing data. For completness, we briefly recall the HMM-SA construction methodology. The backbone of protein structures was split in overlapping fragments of four residues with each one described by the four descriptors illustrated in Fig 1. HMM-SA was estimated using a collection of non-redundant globular proteins, presenting less than 30% of sequence identity. Only proteins of at least 30 amino acids long, no chain breaks, and obtained by X-ray diffraction with a resolution greater than 2.5 Å were retained. This resulted in a collection of 1,429 protein chains, a total of 336,780 amino acids and 332,493 four-residue fragments. The optimal structural alphabet model was selected by comparing structural alphabets of different number of SL using the Bayesian Information Criterion, which balances the log-likelihood of the model and a penalty term related to the number of parameters of the model and the sample size. HMM-SA results in $m = 27$ SL: four SL specific to helices, five SL specific to strands and the remaining 18 SL that describe loops [15].

SAFlex corresponds to 27 HMM-SA SL but in order to improve the interpretability of the 27 SL, a novel nomenclature is applied in SAFlex. This structural letter assignment now depends on the secondary structure type, as identified using the STRIDE software [53] in [3]. Hence, we use only three letters ('A','B','C') which refer to the class of SL (*Helices*, *Strands* or *Coils* respectively) as indicated in Table 1. Each letter assignment is followed by a number indicating its frequency in the data set after unsupervised learning. Thus, SL are ranked in descending order according to their frequencies. For example: the letter 'A1' is more frequent than the letter 'A2'. The correspondence with the previous HMM-SA nomenclature is provided in Table 1. SA nomenclature update and corresponding frequency, effective number of acceptable SL in input or output are also indicated. The covariance matrix $\mathbf{\Sigma}$ of each SL is provided together with the four descriptors $X = (X^1, X^2, X^3, X^4) \in \mathbb{R}^4$ in supplementary Data, (S1). For completeness, the SAFlex 27 representative fragments are illustrated in Fig 4.

To illustrate the interrelationship of the 27 SL, Fig 5 provides a graphical representation of the main Markovian transitions. Unsurprisingly, the transition structure is highly asymetrical, and most transitions occur within SL from belonging to the same secondary structure class.

Table 1. SAFlex structural description. Nomenclature correspondence between HMM-SA and SAFlex SL. Frequency corresponds to associated frequency observed in the HMM-SA training data set, NEFF_output (resp. NEFF_input) corresponds to the effective number of acceptable SL in output (resp. input).

	27 structural letters																									
	helices				coils																		strands			
SAFlex	**A1**	**A2**	**A3**	**A4**	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**C8**	**C9**	**C10**	**C11**	**C12**	**C13**	**C14**	**C15**	**C16**	**C17**	**C18**	**B1**	**B2**	**B3**	**B4**	**B5**
HMM-SA	**A**	**V**	**W**	**a**	**B**	**Z**	**P**	**K**	**Q**	**G**	**S**	**I**	**H**	**D**	**E**	**J**	**U**	**Y**	**F**	**C**	**R**	**O**	**M**	**L**	**N**	**X**	**T**
Frequency %	12.6	5.6	5.3	2.6	4.7	4.5	4.4	4.1	4.1	3.4	3.2	2.9	2.7	2.0	2.0	2.0	2.0	2.0	1.9	1.8	1.7	1.5	5.3	5.1	4.9	4.7	3.0
NEFF_output	2.4	3.6	3.7	2.8	9.6	11.7	11.8	10.4	10.7	12.6	10.4	9.3	9.6	4.3	12.4	11.7	8.5	15.5	8.9	11.6	11.8	13.0	7.9	10.8	9.6	9.2	7.6
NEFF_input	2.3	4.5	3.4	3.0	9.0	12.4	13.3	10.5	12.2	17.5	8.3	7.1	9.7	9.6	11.2	8.8	12.5	10.8	13.5	6.2	12.4	9.8	8.0	10.8	7.0	10.4	7.7

https://doi.org/10.1371/journal.pone.0198854.t001

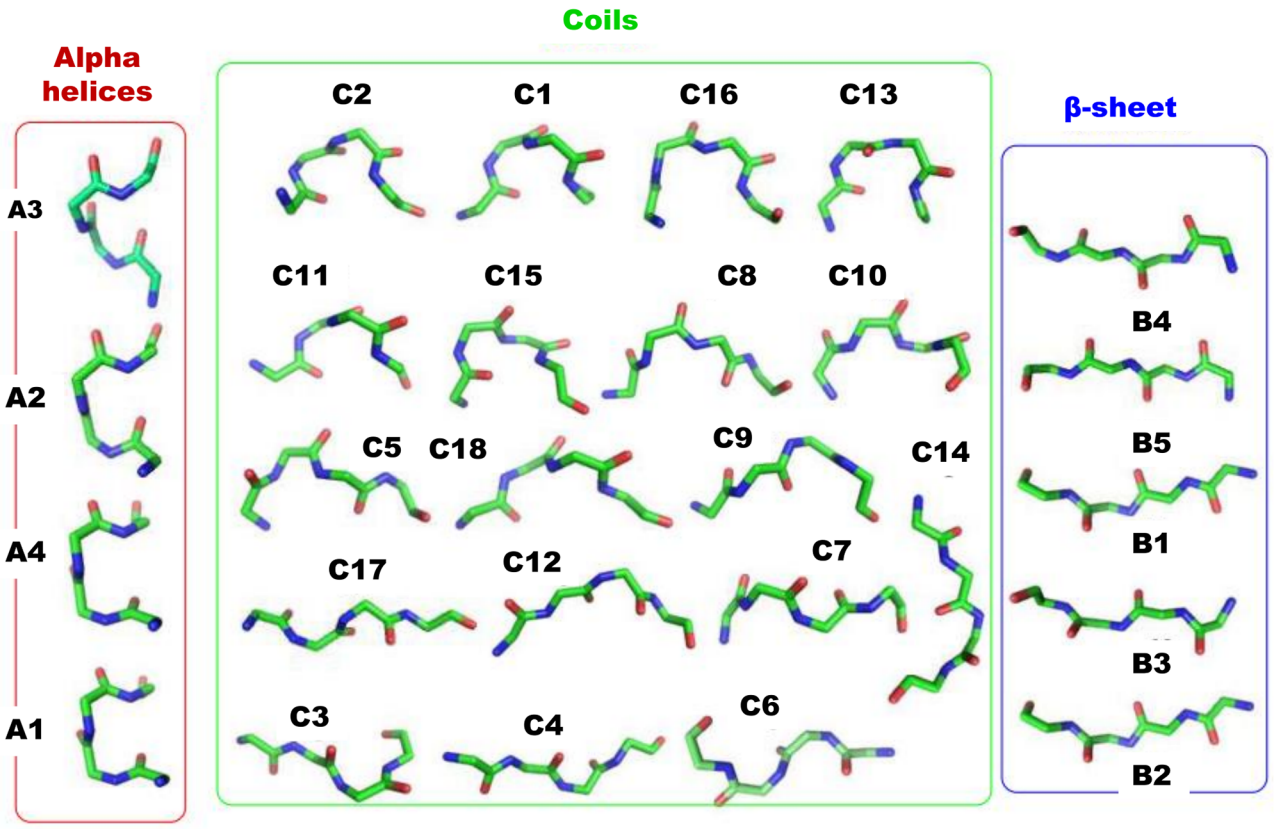


Fig 4. Structural letter prototypes of SAFlex. The 27 representative 3D structural fragments associated with 27 SL of SAFlex. The SL are classified into three groups relative to their secondary structure correspondence: alpha-helices, coils and beta-strands, as described in [3]. The alpha-helices correspond to 4 SL, the coils to 16 SL and beta-strands to 5 SL. Main trajectories between 27 SL of SAFlex.

https://doi.org/10.1371/journal.pone.0198854.g004

Nevertheless, as previously pointed out [15], there are recurrent transitions from secondary structure classes through specific SL but indirect transitions between Helices and Strands. The complete transition matrix of the SA is described in [3].

As explained in the "Materials and Methods" section, our model provides three different encoding information for each chain: 1) the MAP offers the most probable SL sequence fully taking account the complex dependence structure between the SL (including transition probabilities). This is typically the primary encoding used by experimentalists for structure analysis and comparison; 2) in order to account for encoding uncertainty, the POST provides the weighted distribution of the possible SL at each given position. This information is particularly useful for the structural regions where the encoding is difficult or variable. It could be more representative of the 3D structure than the MAP; 3) for better interpretation, the NEFF of SL at each given position is also derived from the entropy of POST. This NEFF is typically close to 1 when the encoding is highly certain and can reach 27 for totally uncertain positions. Therefore this NEFF provides a convenient measurement of the encoding certainty.

We can see the 3D chain A of the 2hba PDB and the three corresponding outputs of SAFlex in Fig 6. This structure corresponds to the N-terminal domain of the ribosomal protein L9, (NTL9), a small alpha-beta protein of 52-residue mixed protein. NTL9 has been widely used as a model system for experimental and computational studies of protein folding and for investigations of the unfolded state [54]. In the left panel of Fig 6, the 3D structure of 52 residues is

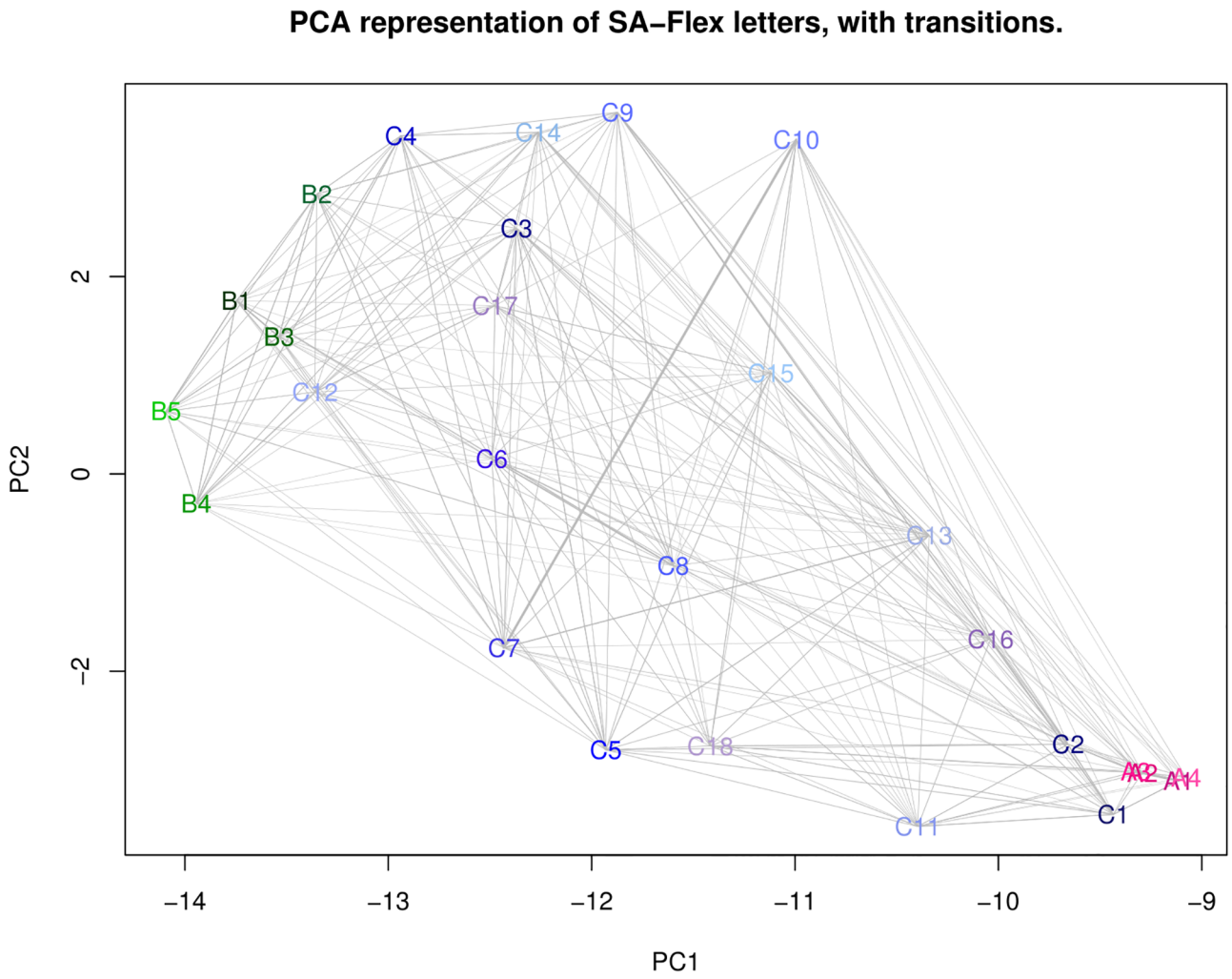


Fig 5. Main Markovian transitions between 27 SL of SAFlex. The SL are projected in the PCA (97.45% of explained variance with the two first axis) space formed by the expected value of the geometrical descriptors of each SL. Orientated Markovian (non reflexive) transitions above 20% probability (resp. between 10% and 20% probability) are represented with a solid (resp. dashed) arrow. There is no direct transition between SL associated with Helices and with Strands.

https://doi.org/10.1371/journal.pone.0198854.g005

represented, which gives 49 overlapping structural fragments colored according to the 27 SL encoding. In the right panel, MAP, POST and NEFFare represented from top to bottom. The MAP clearly corresponds to a alpha-beta mixed protein with few coil links. In the POST, we can see that most encoding positions are highly certain, even if a few positions display some uncertainty. We observe on the protein three regions of relative uncertainty in fragment positions [7-11], [21-24], and the end region [44-49]. For example, Fragment 10 posterior distribution highlights three possible coil letters: C2 (prob = 0.726), C15 (prob = 0.167), C8 (prob = 0.093), associated with a NEFF = 2.244. Finally, the NEFF graph provides a representation of the encoding uncertainty at each position.

One of the interesting features of structural alphabet encoding is that, by design, it accounts only for the local 3D structure. Similarities between structural sequences (ex: using local or global pairwise alignment with suitable parameters) might then highlight local 3D similarities that global RMSD comparisons might totally miss (ex: two proteins with two very similar

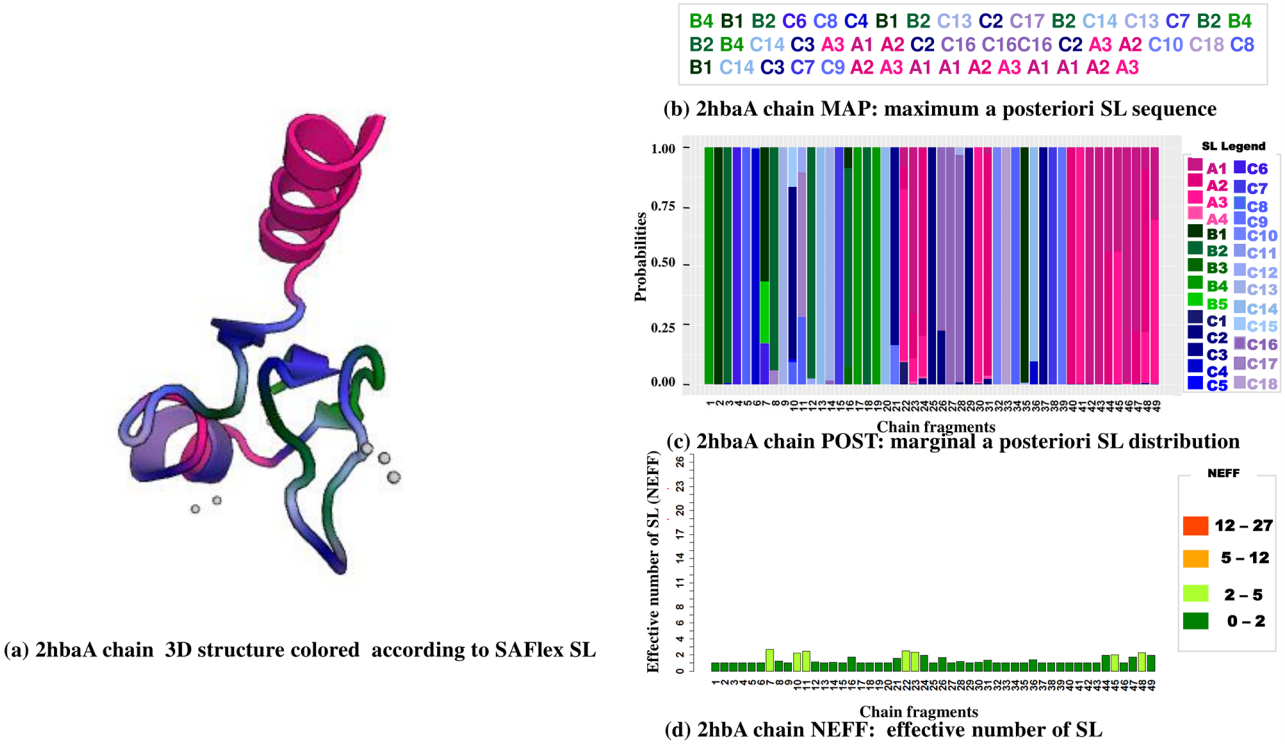


Fig 6. SAFlex encoding of the 2hba pdb structure, corresponding to the N-terminal domain of the ribosomal protein L9 (NTL9): (a) the 2hba 3D structure itself is represented, colored according to the 27 SAFlex SL, (b) the 2hba corresponding SAFlex MAP, (c) the 2hba POST encoding colored according to the 27 SAFlex SL and (d) the 2hba NEFF values.

https://doi.org/10.1371/journal.pone.0198854.g006

domains but different torsion between them). Nevertheless, the calibration of SL similarity score and gap cost for 3D structure comparison is its own research subject for a forthcoming publication.

This uncertainty could typically come from the presence of missing data (see Section "Missing data") and/or of multiple chains (see Section "Multi-chains"). However, we might observe some encoding uncertainty even in the situation of a single chain for a given protein and no missing data (as in Fig 6). This could be due to several reasons: 1) poor crystallographic quality; 2) properties of the protein such as disordered [55] or flexibility; 3) the fragment might be compatible with several SL due to the training limitations of the SA. However, since our SA is only an approximation of the structure composition of proteins, it is therefore not surprising that some ambiguity remains regardless of the accuracy of the SA. Therefore, this intrinsically leads to a posterior probabilistic distribution of encoding trajectories which is the precise purpose of the "Protein encoding" section.

Missing data

Many structures in the PDB have missing parts. For example, a representative data set, PDBselect [56], includes 8,565 PDB files with 75% of chains faced with a problem of missing data: either missing coordinates of the complete residues or alpha-carbon atoms.

As explained in the "Materials and Methods" section, our new model rigorously takes into account missing information through probabilistic computations, which depend on the missing pattern of descriptors at the corresponding positions. When descriptors are missing, the

local likelihood can still be computed using the marginal Gaussian emission probabilities and this partial information can be combined with the local context to provide some estimations. When only few descriptors are missing, the local likelihood can still be computed using the marginal Gaussian emission probabilities and this partial information can be combined with the local context to provide more reliable estimations in terms of exact SL. When all four descriptors are unavailable, the position is totally uninformative (local likelihood of 1.0 for all SL) but the context of the position is still accounted thanks to the Markov dependence of the model through the transition matrix.

In order to quantify the number of SL mismatches induced by the presence of missing residues in the chain in the case of MAP encoding, we designed the following numerical experiment. First, we selected a total of 39 PDB chains (The 39 PDB ids: 1XMK 1MZ9 1QSA 3IIS 1GXM 4K12 4HI8 4DEQ 1UUN 4GV5 2AYD 2O9S 2EVB 3WJT 2E5Y 4A02 3NBC 2DPF 1VMO 3DCL 3C7X 1TL2 1PJX 2XDW 1W6S 1M8N 4E2V 1IGD 1EWF 4BEU 1VBM 2HBA 1A9X 1DL5 1HQ0 1UD9 2CI1 1J0P 1V54) with no missing residue and representative of the variability of available structures in terms of secondary structure types. At most, we used one chain of mainly alpha, mainly beta, alpha-beta or with few regular secondary structure proteins. Then we randomly selected (uniformly) one of these chains, removed the residue information for k consecutive residues (random uniform position) with $k \in \{1, 2, \ldots, 10\}$, and then compared the original SL encoding (using the MAP) to the encoding obtained with the chain with missing residues. The experiment was repeated 10, 000 times. The number of SL mismatches and class mismatches (helices, beta-strands, coils) are reported in Fig 7.

Not suprisinsgly, we see that both the mismatches (left) and class-mismatches (right) increase with the number of (consecutive) missing residues in a roughly linear tendency and that we obtain fewer class-mismatches than SL mismatches. We can note one unique missing residue impacts four consecutive fragments, due to their overlapping on three alpha carbons (corresponding to one to three missing descriptors). Quantitatively, for $k = 1$ (resp. $k = 2$) missing residues, we obtain an average of 1.872 SL mismatches, 1st quartile 1, median 2, 3rd quartile 3 (resp. 3.308 with 1st quartile 2, median 4, 3rd quartile 5). In terms of secondary structure information, for $k = 1$ (resp. $k = 2$) missing residues, we obtain a weak average of 0.479 secondary structure SL mismatches, 1st quartile 0, median 0, 3rd quartile 1 (resp. and 1.127, 1st quartile 0, median 1, 3rd quartile 2. These figures are small compared to the average length of 199.3 SL, (1st quartile 67.0, median 151.0, 3rd quartile 317.0) of the 10, 000

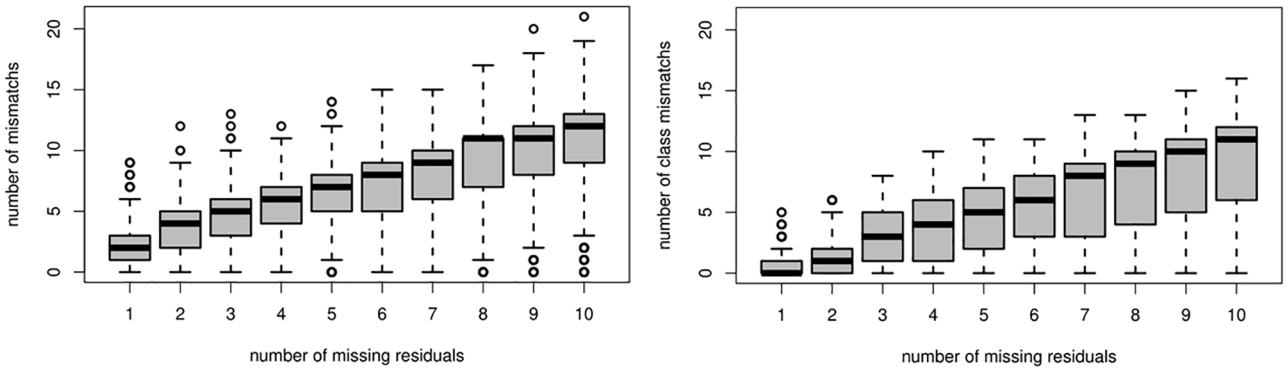


Fig 7. Average mismatches induced by missing residuals on 10,000 simulations: average number of SL mismatches (left) and secondary class (helices, beta-strands, coils) mismatches (right) as a function of the number of consecutive missing residues.

https://doi.org/10.1371/journal.pone.0198854.g007

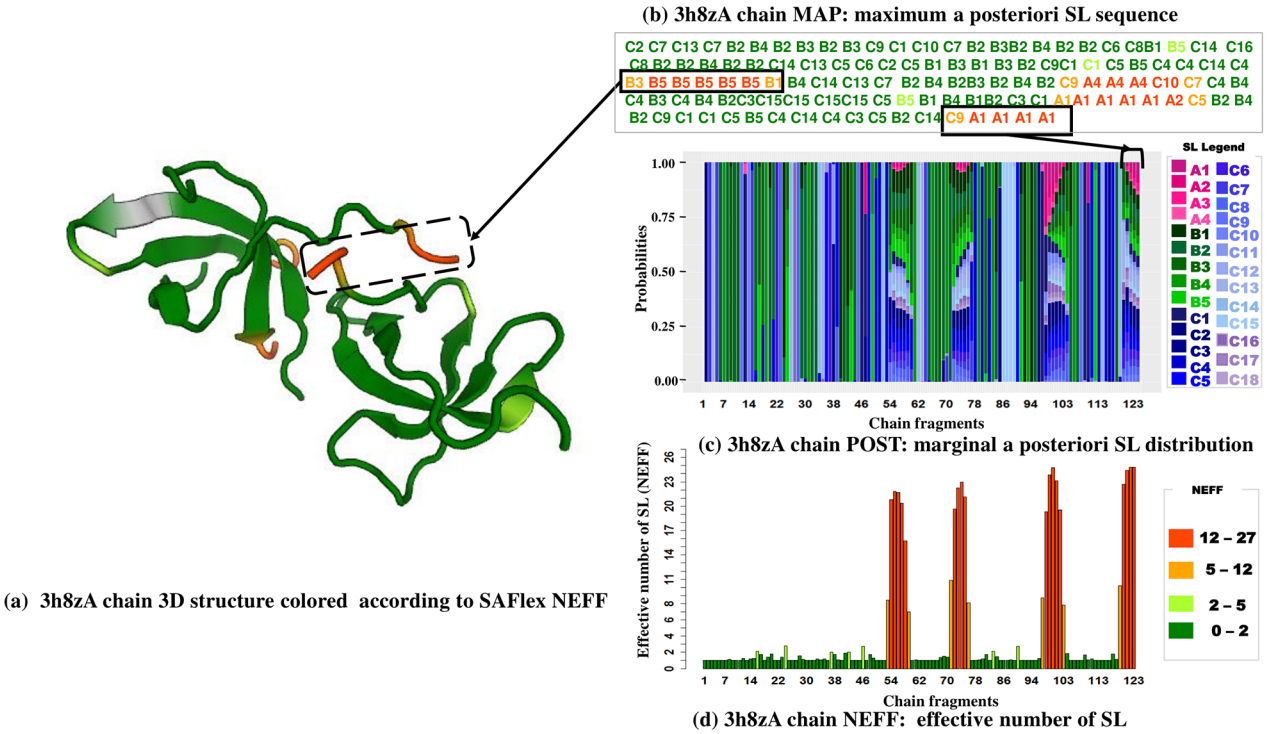


Fig 8. Missing conformation detection of the 3h8z pdb entry, corresponding to the homo sapiens fragile X mental retardation syndrome-related protein 2 associated with FXR2 gene: (a) the 3h8z 3D structure is presented colored according to the NEFF value legend, (b) the 3h8z corresponding SAFlex MAP colored according to the NEFF value legend, (c) the 3h8Z POST encoding and (d) the 3h8z NEFF values.

https://doi.org/10.1371/journal.pone.0198854.g008

considered chains. Unsurprisingly, the prediction of the secondary structure is more robust to missing data than the SL prediction itself.

Finally, this confirms the resulting encoding is highly uncertain when repeated missing residues appear in a given region of the 3D structure, but interestingly this information is provided by SA-Flex by high NEFF values. To illustrate the interest of different encoding proposed by SAFlex, in case of missing data, we consider the protein 3h8z corresponding to the homo sapiens fragile X mental retardation syndrome-related protein 2 associated with FXR2 gene. The FMRP, FXR1 and FXR2 proteins comprise a small family of highly conserved proteins that appear to be important in translational regulation, particularly in neuronal cells [57]. In Fig 8, we can see the crystal structure of the tudor domains from FXR2 (left panel) from the PDB 3h8z and the SAFlex outputs from top to bottom: MAP, POST, and NEFF (right panel). This 3D structure corresponds to 123 overlapping structural fragments colored according to their NEFF values. In Fig 8(a) and 8(b), we observe that a big part of the protein correspond to beta-strand SL linked by short coil SL regions, associated with weak uncertainties, as illustrated in Fig 8(c) and 8(d). This is coherent with the fact Tud1 domain forms a canonical tudor barre comprising five highly twisted antiparallel beta-strands. However, we clearly observe the presence of four regions of high uncertainty (NEFF close to 27). These four regions correspond to the 16 residues with missing 3D coordinates information in the PDB file with the first region [53-59] predicted as disorder using psipred website [58, 59]. Despite this high uncertainty, these positions being associated to a NEFF close to 27, one should note that the MAP suggests encodings for these four regions. This further illustrates the interest of the multiple outputs of SAFlex.

Multi-chains

In recent decades, many protein complex structures containing multiple chains have been determined: homomers, heteromers or different PDB files corresponding to the same protein in different conditions. If heteromers are naturally encoded as different structural sequences, homomers can be considered as replicates of the same underlying structure. SAFlex proposes to encode homomers either as independent structures (one structural sequence per chain) or as a single consensus one, where a single hidden structural sequence is shared by all homomeric chains. The resulting consensus encoding hence represents the variability of the homomer across the chains. This variability is either due to measurement uncertainty or to intrinsic flexibility. One illustration is presented on the small Heat Shock Proteins (sHSPs), which are important in stress tolerance and play an essential role in preventing aggregation of target proteins [60]. They participate in protecting, maintaining and regulating specific protein functions. The PDB entry 1gme corresponds to HSP16.9, a member of the sHSPs, that assembles into a dodecameric double disk. In the PDB file, an available tetramer can be used to reconstruct the dodecamer by symmetry operations [61]. The monomer's residue length is 151 which gives 148 overlapping structural fragments. The four monomers have a global common structure, called the alpha crystallin domain signature [60] but have differences in some regions: the 42 N-terminal residues are missing in the two monomers B and D, whereas the N-terminal arm in A and C monomers is fully resolved and composed of helices connected by random coils. In Fig 9, we can see the values of NEFF for the four chains (upper panel) and for the consensus encoding (lower panel). The two missing regions of chains B and D are clearly highlighted (NEFF close to 27 on positions [1, 42]). The overall uncertainty of encoding along the four chains is quite large with an average value of NEFF $\simeq$ 4.4 on all the regions and of NEFF $\simeq$ 1.3 when excluding missing regions. In the lower panel, we can see that the consensus encoding on all regions has a much lower uncertainty of NEFF $\simeq$ 1.1. This illustrates the interest of the consensus approach which not only provides an encoding for the complete protein despite the missing patterns of chains B and D, while taking advantage of the replicates to refine the structural encoding.

In Fig 10, we see (a) the 3D structure of the four chains of 1gme (left panel), (b) the multiple alignment of the MAP for the four chains and the consensus encoding as well as (c) the POST encoding for the consensus encoding. The missing regions in chains B and D appear clearly in the MAP with long runs of the SL-A4; this is coherent in the absence of any additional information since this structural letter has the highest probability of self-transition in the SA (see transition matrix in the supplementary data). However, chains A and C provide informative MAPs for the corresponding positions and the result is clearly consistent with the consensus' MAP. For most of the remaining positions, we observe a strong concordence among all encodings reflecting low structural variability. However, for some regions in fragment positions [29-32], [85-92], [110-116] and [136-142], there is higher variability across the MAP of the fours chains, indicated by different SL for the four chains. This suggests that some of these positions could correspond to intrinsic flexible chain positions or to resolution uncertainties. In this context, the consensus encoding tries to find the most adequate common structural letter to reflect this variability and selects the SL-C15 (former F in HMM-SA) which is known to correspond to the fuzzy coil state of the alphabet [40] and is associated with very high posterior probabilities (close to 1), Fig 10(c). This result is clearly consistent with the assumption of a common underlying structure among the different chains. However, it also shows its limits in case of conformation plasticity. In this case, the independent chain encodings can be carrefully explored to detect variable positions and SL changes potentially due to intrinsic flexilibity, to partner binding or to sequence mutation effects [25].

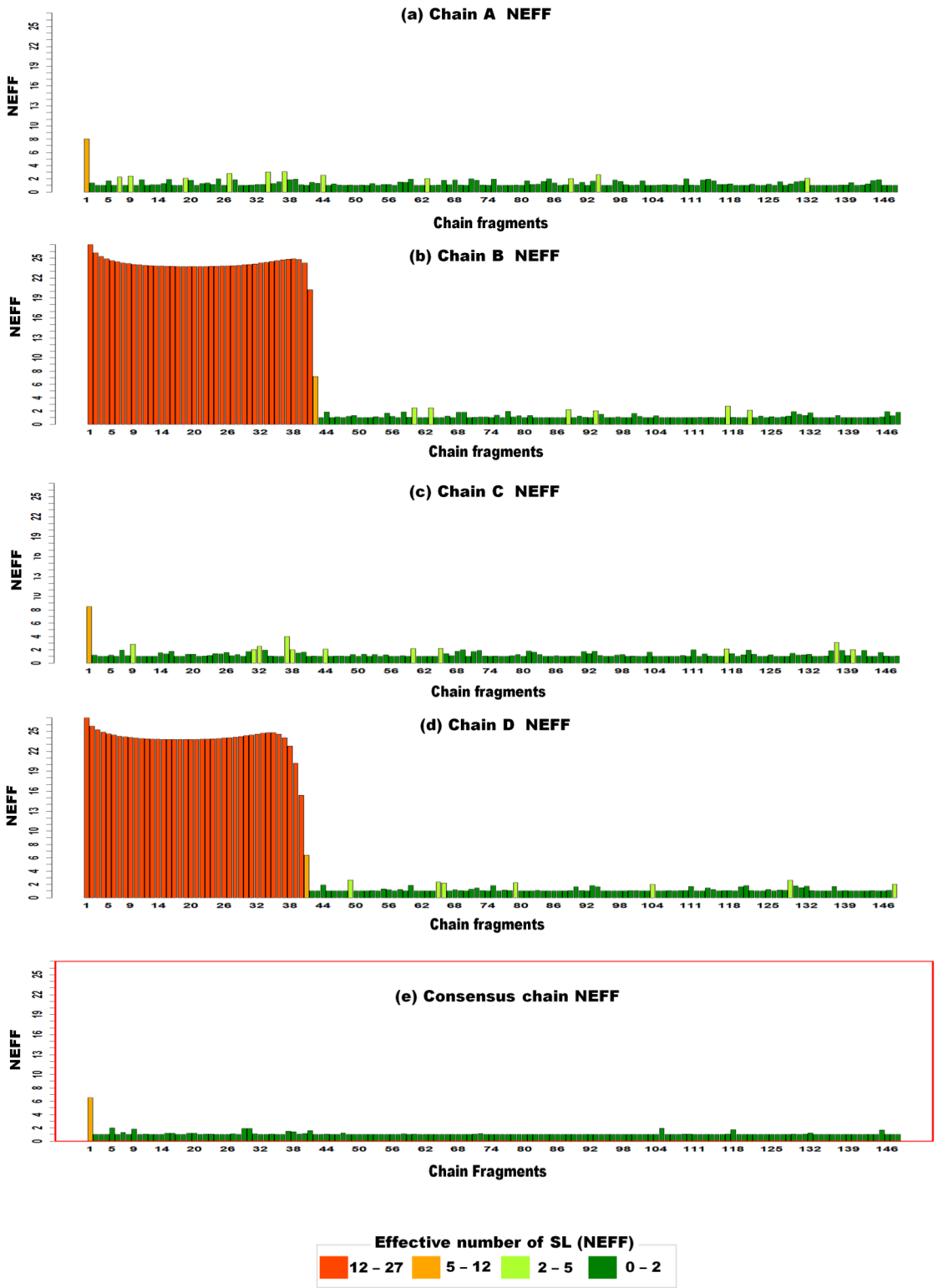


Fig 9. NEFF of the four monomers and consensus chain from the 1gme pdb, corresponding to the small Heat Shock Proteins (sHSPs). The x-axis of the charts correspond to the 148 fragment numbers and the y-axis to the NEFF values, from 1 to 27. Each bar is colored according to the NEFF value legend. The charts **(a), (b), (c), (d)** correspond to the NEFF values for the four monomer chains A, B, C and D and the chart **(e)**, in a red box to the consensus encoding of the four monomers.

https://doi.org/10.1371/journal.pone.0198854.g009

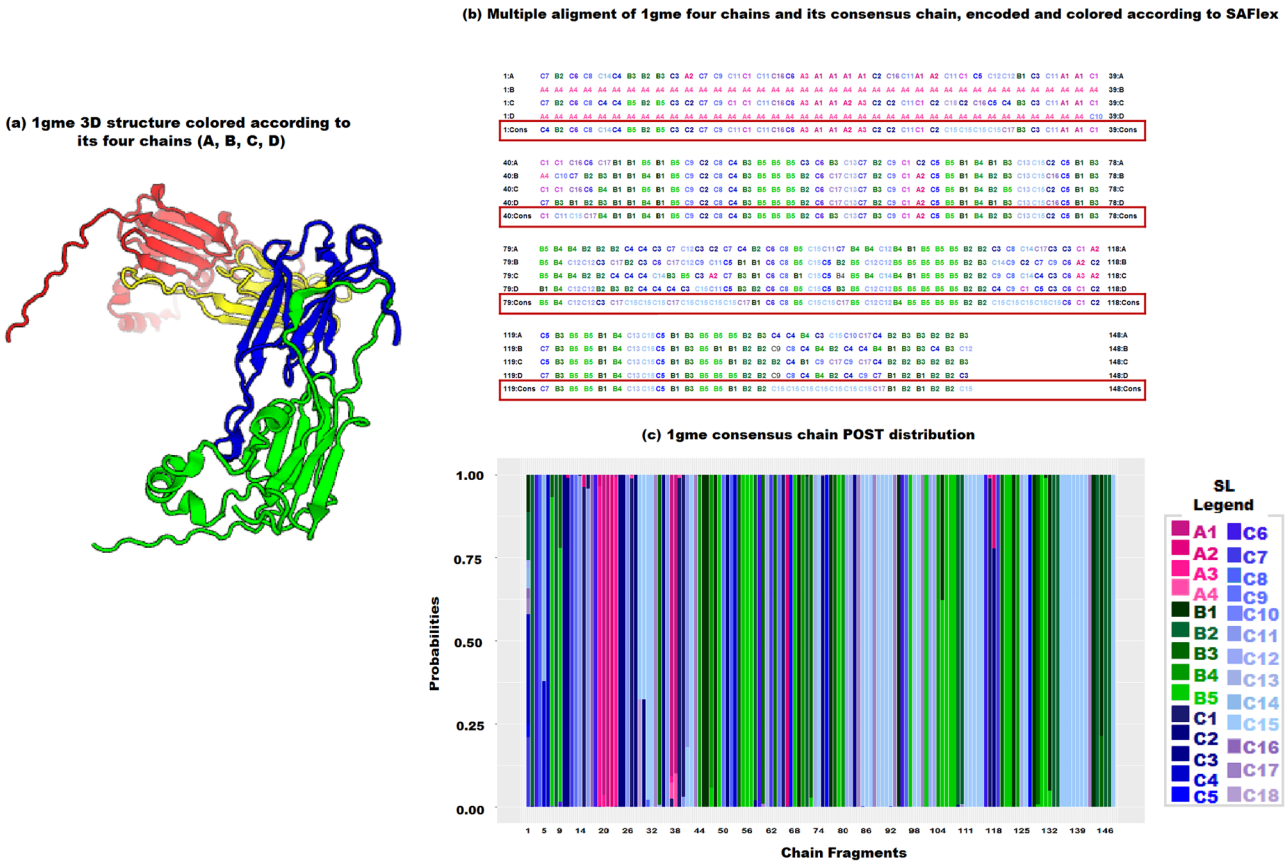


Fig 10. SAFlex independent and consensus encoding of the 1gme four chains: (a) the 3D structure of 1gme is displayed by textcolorredPV, TO DEFINE, colored according to its four chains: A in red, B in yellow, C in green and D in blue. Each monomer corresponds to 148 overlapping structural fragments; (b) the multiple alignment of the MAP of the four encoded chains and the consensus encoding in a red box. Each letter is colored according to SAFlex SL; (c) the POST encoding for the consensus chain. The x-axis corresponds to the fragment numbers and the y-axis represents the posterior probabilities.

https://doi.org/10.1371/journal.pone.0198854.g010

Conclusion

Conceiving protein structures has moved beyond static representations to include dynamic aspects of quaternary structures, like conformational changes upon binding and structural fluctuations occurring within fully assembled complexes [38].

SA encodings are known to perform well for fine study of their structural properties. However, SA have to be improved to better understand uncertainties such as missing data and intrinsic flexibility observed between different available replicates in the PDB. This has an impact on our knowledge of protein functions and their disordered regions, which contribute to the protein capacity to establish interactions with different partners.

In this paper, we presented SAFlex, extended from HMM-SA (with similar fragment descriptors, number of letters (27), Gaussian distribution per letter, and transition matrix) to provide structure encoding in terms of missing residues as well as uncertainties. SAFlex has the following three main novelties: 1) allows for three different encoding outputs (MAP, marginal posterior distribution, and entropy-related statistics); 2) new implementation is robust for any missing data pattern in the PDB file; and 3) new model can take into account replicates and include a new consensus encoding for homomers. These correspond to important improvements as there are many chains with missing data (e.g. 75% of PDBselect chains

faced with a problem of missing data [56]) and most of available structural information on protein complexes concerns homomers [36]. All of these improvements are freely available to the public through a web server application.

Concerning missing data, as pointed out by our experiments, when there are too many consecutive missing residues, the loss of information eventually leads to many encoding errors. It could hence be interesting to include primary or secondary structures into the alphabet in order to exploit this additional information in the context of large portion of missing residues in the PDB chains.

Another interesting features of our structural alphabet encoding is that, by design, it accounts only for the local 3D structure. Similarities between structural sequences (ex: using local or global pairwise alignment with suitable parameters) might then highlight local 3D similarities that global RMSD comparisons might totally miss (ex: two proteins with two very similar domains but different torsion between them). Nevertheless, the calibration of SL similarity score and gap cost for 3D structure comparison is its own research subject for a forthcoming publication.

Having implemented a basic version of SAFlex, our next step would be to use up-to-date large scale data to train a new SA with the following characteristics: more parsimony with uniform starting distribution, model selection using penalized approaches (e.g. adaptive ridge [62]) for the (very sparse) transition matrix, more and/or different descriptors to ensure a bijection between the 3D fragment conformation and the descriptor space, and model extensions to allow for multi-conformations in addition to the multi-chains feature.

Supporting information

S1 Table. The table contains the Gaussian emission parameter of SAFlex 27. Rows correspond to the SAFlex SL. The SL are classified by group: alpha-helices, coils and beta-strands. For each structural letter is given the four descriptors $X = (X^1, X^2, X^3, X^4) \in \mathbb{R}^4$ and the covariance matrix Σ. Rows and colons of the matrix corresponds to descriptors.
(TEX)

Acknowledgments

This work was supported by an USPC grant (SA-Flex) and an ANR grant (ANR-10-BINF-0003, BIP-BIP). We are grateful to M. Petitjean for helpful discussions.

Author Contributions

Conceptualization: Gregory Nuel, Anne-Claude Camproux.

Data curation: Ikram Allam, Delphine Flatters, Géraldine Caumes, Leslie Regad.

Formal analysis: Ikram Allam, Delphine Flatters, Géraldine Caumes, Leslie Regad.

Funding acquisition: Gregory Nuel, Anne-Claude Camproux.

Investigation: Ikram Allam, Delphine Flatters, Géraldine Caumes, Leslie Regad, Gregory Nuel, Anne-Claude Camproux.

Methodology: Géraldine Caumes, Vincent Delos, Gregory Nuel, Anne-Claude Camproux.

Resources: Delphine Flatters, Géraldine Caumes, Vincent Delos, Gregory Nuel, Anne-Claude Camproux.

Software: Ikram Allam, Delphine Flatters, Gregory Nuel, Anne-Claude Camproux.

Supervision: Gregory Nuel, Anne-Claude Camproux.

Validation: Ikram Allam, Delphine Flatters, Gregory Nuel, Anne-Claude Camproux.

Visualization: Ikram Allam, Gregory Nuel.

Writing – original draft: Ikram Allam, Gregory Nuel, Anne-Claude Camproux.

References

1. Unger R, Harel D, Wherland S, Sussman JL. A 3D building blocks approach to analyzing and predicting structure of proteins. Proteins: Structure, Function, and Bioinformatics. 1989; 5(4):355–373. https://doi.org/10.1002/prot.340050410
2. Camproux AC, Tufféry P, Chevrolat JP, Boisvieux JF, Hazout S. Hidden Markov model approach for identifying the modular framework of the protein backbone. Protein Engineering. 1999; 12(12):1063–1073. https://doi.org/10.1093/protein/12.12.1063 PMID: 10611400
3. Camproux AC, Gautier R, Tufféry P. A hidden markov model derived structural alphabet for proteins. Journal of Molecular Biology. 2004; 339(3):591–605. https://doi.org/10.1016/j.jmb.2004.04.005 PMID: 15147844
4. De Brevern AG, Etchebest C, Hazout S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. Proteins: Structure, Function, and Bioinformatics. 2000; 41(3):271–287. https://doi.org/10.1002/1097-0134(20001115)41:3%3C271::AID-PROT10%3E3.0.CO;2-Z
5. Nuel G, Regad L, Martin J, Camproux AC. Exact distribution of a pattern in a set of random sequences generated by a Markov source: applications to biological data. Algorithms for Molecular Biology. 2010; 5(1):15. https://doi.org/10.1186/1748-7188-5-15 PMID: 20205909
6. Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of Molecular Biology. 1981; 147(1):195–197. PMID: 7265238
7. Wang S, Zheng WM. CLePAPS: fast pair alignment of protein structures based on conformational letters. Journal of Bioinformatics and Computational Biology. 2008; 6(02):347–366. https://doi.org/10.1142/S0219720008003461 PMID: 18464327
8. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. The Bulletin Of Mathematical Biophysics. 1943; 5(4):115–133. https://doi.org/10.1007/BF02478259
9. Kohonen T. Self-organized formation of topologically correct feature maps. Biological Cybernetics. 1982; 43(1):59–69. https://doi.org/10.1007/BF00337288
10. Fix E, Hodges JL Jr. Discriminatory analysis-nonparametric discrimination: consistency properties. California Univ Berkeley; 1951.
11. Zhang X, Fetrow JS, Rennie WA, Waltz DL, Berg G. Automatic derivation of substructures yields novel structural building blocks in globular proteins. In: ISMB. vol. 1; 1993. p. 438–446.
12. Tung CH, Huang JW, Yang JM. Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for rapid search of protein structure database. Genome Biology. 2007; 8(3):R31. https://doi.org/10.1186/gb-2007-8-3-r31 PMID: 17335583
13. Tung CH, Nacher JC. A Complex Network Approach for the Analysis of Protein Units Similarity Using Structural Alphabet. International Journal of Bioscience, Biochemistry and Bioinformatics. 2013; 3(5):433–437. https://doi.org/10.7763/IJBBB.2013.V3.250
14. Pandini A, Fornili A, Kleinjung J. Structural alphabets derived from attractors in conformational space. BMC Bioinformatics. 2010; 11(1):97. https://doi.org/10.1186/1471-2105-11-97 PMID: 20170534
15. Camproux AC, Tufféry P. Hidden Markov model-derived structural alphabet for proteins: the learning of protein local shapes captures sequence specificity. Biochimica et Biophysica Acta (BBA)-General Subjects. 2005; 1724(3):394–403. https://doi.org/10.1016/j.bbagen.2005.05.019
16. Gautier R, Camproux AC, Tufféry P. SCit: web tools for protein side chain conformation analysis. Nucleic Acids Research. 2004; 32(suppl 2):W508–W511. https://doi.org/10.1093/nar/gkh388 PMID: 15215438
17. Guyon F, Camproux AC, Hochez J, Tufféry P. SA-Search: a web tool for protein structure mining based on a Structural Alphabet. Nucleic Acids Research. 2004; 32(suppl 2):W545–W548. https://doi.org/10.1093/nar/gkh467 PMID: 15215446
18. Deschavanne P, Tufféry P. Enhanced protein fold recognition using a structural alphabet. Proteins: Structure, Function, and Bioinformatics. 2009; 76(1):129–137. https://doi.org/10.1002/prot.22324

19. Pandini A, Fornili A. Using local states to drive the sampling of global conformations in proteins. Journal of Chemical Theory and Computation. 2016; 12(3):1368–1379. https://doi.org/10.1021/acs.jctc.5b00992 PMID: 26808351

20. Pandini A, Fornili A, Fraternali F, Kleinjung J. GSATools: analysis of allosteric communication and functional local motions using a structural alphabet. Bioinformatics. 2013; 29(16):2053–2055. https://doi.org/10.1093/bioinformatics/btt326 PMID: 23740748

21. Mahajan S, de Brevern AG, Offmann B, Srinivasan N. Correlation between local structural dynamics of proteins inferred from NMR ensembles and evolutionary dynamics of homologues of known structure. Journal of Biomolecular Structure and Dynamics. 2014; 32(5):751–758. https://doi.org/10.1080/07391102.2013.789989 PMID: 23730714

22. Lamiable A, Thevenet P, Tufféry P. A critical assessment of hidden markov model sub-optimal sampling strategies applied to the generation of peptide 3D models. Journal of Computational Chemistry. 2016; 37(21):2006–2016. https://doi.org/10.1002/jcc.24422 PMID: 27317417

23. Lamiable A, Thévenet P, Rey J, Vavrusa M, Derreumaux P, Tufféry P. PEP-FOLD3: faster de novo structure prediction for linear peptides in solution and in complex. Nucleic Acids Research. 2016; 44(W1):W449–W454. https://doi.org/10.1093/nar/gkw329 PMID: 27131374

24. Craveur P, Joseph AP, Esque J, Narwani TJ, Noël F, Shinada N, et al. Protein flexibility in the light of structural alphabets. Frontiers in Molecular Biosciences. 2015; 2:20. https://doi.org/10.3389/fmolb.2015.00020 PMID: 26075209

25. Regad L, Chéron JB, Triki D, Senac C, Flatters D, Camproux AC. Exploring the potential of a structural alphabet-based tool for mining multiple target conformations and target flexibility insight. PLOS ONE. 2017; 12(8):1–29. https://doi.org/10.1371/journal.pone.0182972

26. Martin J, Regad L, Lecornet H, Camproux AC. Structural deformation upon protein-protein interaction: a structural alphabet approach. BMC Structural Biology. 2008; 8(1):12. https://doi.org/10.1186/1472-6807-8-12 PMID: 18307769

27. Baussand J, Camproux AC. Deciphering the shape and deformation of secondary structures through local conformation analysis. BMC Structural Biology. 2011; 11(1):9. https://doi.org/10.1186/1472-6807-11-9 PMID: 21284872

28. de Brevern AG, Bornot A, Craveur P, Etchebest C, Gelly JC. PredyFlexy: flexibility and local structure prediction from sequence. Nucleic Acids Research. 2012; 40(W1):W317–W322. https://doi.org/10.1093/nar/gks482 PMID: 22689641

29. Dong Q, Wang K, Liu B, Liu X. Characterization and prediction of protein flexibility based on structural alphabets. BioMed Research International. 2016; 2016(ID 4628025):7.

30. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, et al. The protein data bank: A computer-based archival file for macromolecular structures. Archives of Biochemistry and Biophysics. 1978; 185(2):584–591. https://doi.org/10.1016/0003-9861(78)90204-7 PMID: 626512

31. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Research. 2000; 28(1):235–242. https://doi.org/10.1093/nar/28.1.235 PMID: 10592235

32. Goh CS, Milburn D, Gerstein M. Conformational changes associated with protein–protein interactions. Current Opinion in Structural Biology. 2004; 14(1):104–109. https://doi.org/10.1016/j.sbi.2004.01.005 PMID: 15102456

33. Grünberg R, Leckner J, Nilges M. Complementarity of structure ensembles in protein-protein binding. Structure. 2004; 12(12):2125–2136. https://doi.org/10.1016/j.str.2004.09.014 PMID: 15576027

34. Lensink MF, Mendez R. Recognition-induced conformational changes in protein-protein docking. Current Pharmaceutical Biotechnology. 2008; 9(2):77–86. https://doi.org/10.2174/138920108783955173 PMID: 18393864

35. Marsh JA, Teichmann SA. Protein flexibility facilitates quaternary structure assembly and evolution. PLOS Biology. 2014; 12(5):1–11. https://doi.org/10.1371/journal.pbio.1001870

36. Bergendahl LT, Marsh JA. Functional determinants of protein assembly into homomeric complexes. Scientific reports. 2017; 7(1):4932. https://doi.org/10.1038/s41598-017-05084-8 PMID: 28694495

37. Marsh JA, Teichmann SA. Structure, dynamics, assembly, and evolution of protein complexes. Annual Review of Biochemistry. 2015; 84:551–575. https://doi.org/10.1146/annurev-biochem-060614-034142 PMID: 25494300

38. Sormanni P, Piovesan D, Heller GT, Bonomi M, Kukic P, Camilloni C, et al. Simultaneous quantification of protein order and disorder. Nature Chemical Biology. 2017; 13(4):339–342. https://doi.org/10.1038/nchembio.2331 PMID: 28328918

39. Gall TL, Romero PR, Cortese MS, Uversky VN, Dunker AK. Intrinsic Disorder in the Protein Data Bank. Journal of Biomolecular Structure and Dynamics. 2007; 24(4):325–341. https://doi.org/10.1080/07391102.2007.10507123 PMID: 17206849

40. Regad L, Guyon F, Maupetit J, Tufféry P, Camproux AC. A Hidden Markov Model applied to the protein 3D structure analysis. Computational Statistics & Data Analysis. 2008; 52(6):3198–3207. https://doi.org/10.1016/j.csda.2007.09.010

41. Regad L, Martin J, Nuel G, Camproux AC. Mining protein loops using a structural alphabet and statistical exceptionality. BMC Bioinformatics. 2010; 11(1):75. https://doi.org/10.1186/1471-2105-11-75 PMID: 20132552

42. Berman H, Henrick K, Nakamura H. Announcing the worldwide protein data bank. Nature Structural & Molecular Biology. 2003; 10(12):980–980. https://doi.org/10.1038/nsb1203-980

43. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Research. 2007; 35(suppl_1):D301–D303. https://doi.org/10.1093/nar/gkl971 PMID: 17142228

44. Rooman MJ, Rodriguez J, Wodak SJ. Automatic definition of recurrent local structure motifs in proteins. Journal of Molecular Biology. 1990; 213(2):327–336. https://doi.org/10.1016/S0022-2836(05)80194-9 PMID: 2342110

45. Micheletti C, Seno F, Maritan A. Recurrent oligomers in proteins: An optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. Proteins: Structure, Function and Genetics. 2000; 40(4):662–674. https://doi.org/10.1002/1097-0134(20000901)40:4%3C662::AID-PROT90%3E3.0.CO;2-F

46. Kolodny R, Koehl P, Guibas L, Levitt M. Small libraries of protein fragments model native protein structures accurately. Journal of Molecular Biology. 2002; 323(2):297–307. https://doi.org/10.1016/S0022-2836(02)00942-7 PMID: 12381322

47. Sander O, Sommer I, Lengauer T. Local protein structure prediction using discriminative models. BMC Bioinformatics. 2006; 7(1):14. https://doi.org/10.1186/1471-2105-7-14 PMID: 16405736

48. Dong QW, Wang XL, Lin L. Methods for optimizing the structure alphabet sequences of proteins. Computers in Biology and Medicine. 2007; 37(11):1610–1616. https://doi.org/10.1016/j.compbiomed.2007.03.002 PMID: 17493604

49. Baeten L, Reumers J, Tur V, Stricher F, Lenaerts T, Serrano L, et al. Reconstruction of protein backbones from the BriX collection of canonical protein fragments. PLOS Computational Biology. 2008; 4(5):1–11. https://doi.org/10.1371/journal.pcbi.1000083

50. Budowski-Tal I, Nov Y, Kolodny R. FragBag, an accurate representation of protein structure, retrieves structural neighbors from the entire PDB quickly and accurately. Proceedings of the National Academy of Sciences. 2010; 107(8):3481–3486. https://doi.org/10.1073/pnas.0914097107

51. Kalev I, Habeck M. HHfrag: HMM-based fragment detection using HHpred. Bioinformatics. 2011; 27(22):3110–3116. https://doi.org/10.1093/bioinformatics/btr541 PMID: 21965821

52. Rabiner LR. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE. 1989; 77(2):257–286. https://doi.org/10.1109/5.18626

53. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins: Structure, Function, and Bioinformatics. 1995; 23(4):566–579. https://doi.org/10.1002/prot.340230412

54. Cho JH, Meng W, Sato S, Kim EY, Schindelin H, Raleigh DP. Energetically significant networks of coupled interactions within an unfolded protein. Proceedings of the National Academy of Sciences. 2014; 111(33):12079–12084. https://doi.org/10.1073/pnas.1402054111

55. Dunker AK, Silman I, Uversky VN, Sussman JL. Function and structure of inherently disordered proteins. Current opinion in structural biology. 2008; 18(6):756–764. https://doi.org/10.1016/j.sbi.2008.10.002 PMID: 18952168

56. Griep S, Hobohm U. PDBselect 1992–2009 and PDBfilter-select. Nucleic Acids Research. 2009; 38(suppl_1):D318–D319. https://doi.org/10.1093/nar/gkp786 PMID: 19783827

57. Adams-Cioaba MA, Guo Y, Bian C, Amaya MF, Lam R, Wasney GA, et al. Structural studies of the tandem Tudor domains of fragile X mental retardation related proteins FXR1 and FXR2. PLOS ONE. 2010; 5(11):1–9. https://doi.org/10.1371/journal.pone.0013559

58. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology. 1999; 292(2):195–202. https://doi.org/10.1006/jmbi.1999.3091 PMID: 10493868

59. Buchan DW, Minneci F, Nugent TC, Bryson K, Jones DT. Scalable web services for the PSIPRED Protein Analysis Workbench. Nucleic Acids Research. 2013; 41(W1):W349–W357. https://doi.org/10.1093/nar/gkt381 PMID: 23748958

60. Poulain P, Gelly JC, Flatters D. Detection and architecture of small heat shock protein monomers. PLOS ONE. 2010; 5(4):1–10. https://doi.org/10.1371/journal.pone.0009990

61. van Montfort RL, Basha E, Friedrich KL, Slingsby C, Vierling E. Crystal structure and assembly of a eukaryotic small heat shock protein. Nature Structural & Molecular Biology. 2001; 8(12):1025–1030. https://doi.org/10.1038/nsb722

62. Frommlet F, Nuel G. An Adaptive Ridge Procedure for L0 Regularization. PLOS ONE. 2016; 11(2):1–23. https://doi.org/10.1371/journal.pone.0148620