The search of sequence variants using a constrained protein evolution simulation approach

Pierre Tuffery, Sjoerd de Vries

The search of sequence variants using a constrained protein evolution simulation approach

Pierre Tuffery[1] * and Sjoerd de Vries[1]

1: Université de Paris, BFA, UMR 8251, CNRS, ERL U1133, Inserm, RPBS, F-75013 Paris, France
*: to whom correspondance should be sent.

Abstract

Protein engineering or candidate therapeutic peptide optimization are processes in which the identification of relevant sequence variants is critical. Starting from one amino-acid sequence, the choice of the substitutions must meet the objective of not disrupting the structure of the protein, not impacting the main functional properties of the starting entity, while also meeting the condition to enhance some expected property such as thermal stability, resistance to degradation, ... Here, we introduce a new approach of sequence evolution that focuses on the objective of not disrupting the structure of the initial protein by embedding a point to point control on the preservation of the local structure at each position in the sequence. For 6 mini-proteins, we find that, starting from a single sequence, our simple approach intrinsically contains information about site-specific rate heterogeneity of substitution, and that it is able to reproduce sequence diversity as can be observed in the sequences available in the Uniref repository. We show that our approach is able to provide information about positions not to substitute and about substitutions not to perform at a given position to maintain structure integrity. Overall, our results demonstrate that point to point preservation of the local structure along a sequence is an important determinant of sequence evolution.

1. Introduction

Protein or peptide engineering or design, as well as therapeutic candidate peptide identification and optimization are processes in which the identification of relevant sequence variants is critical. Indeed, even not considering the insertion of un-natural amino-acids, the choice of substitutions must face the objective of not disrupting the structure of the protein, while meeting the condition to enhance some expected property such as stability, resistance to degradation, reduced immune response, among others. Ways to identify such substitutions include experimental techniques such as phage display (1), directed mutagenesis (2), and more recently deep mutation scanning (3). Due to progress in understanding the determinants of protein folding and increasing amount of data about related sequences, computational approaches to assist the design have also been developed. Among these, the simulation of sequence evolution along a phylogenetic tree has been the subject of intense efforts during the past decades (4-7). Such simulation techniques are now able to deal with different rates of substitution along sequences (8), consider co-evolution for some (9), and models to manage indels have emerged (4). Apart from the fundamental goal of deciphering the rules underlying protein family sequence evolution, these approaches also have implications for enhanced sequence alignment for instance (10-12).

These techniques based on the analysis of sequences alone can however reveal to be limited for protein engineering. In the context of sequence optimization, the search for sequence variants might escape the rules of natural evolution, and phylogenetic inference is not always possible in the context of *de novo* protein design, e.g. for the design of un-natural proteins. Finally, structural constraints, although implicitly contained along sequence evolution might require more direct consideration. Explicit account for structure has led to the field of computational protein design, in which the impact of candidate substitutions on the free energy of folding of the protein is usually questioned to drive the process (13). Indeed, various techniques to question the relationship between the sequence space and

the structure space, have been developed in the perspective of protein design or engineering. These include sophisticated protocols (see 14), such as Rosetta-design (15), protocols based on the dead end elimination theorem (e.g. 16), protocols focusing accurately on the relationship between evolution and side chain-packing (17), coarse grained models (e.g. 18), to cite some, or a combination of the simulation of protein sequence evolution under the constraint of explicit structural context (19). Of note, 3D based-techniques have led to successful applications in protein design (20-22).

In the context of sequence optimization of therapeutic peptide/protein candidates, a primary objective, beyond the goals of enhancing peptide stability and bioactivity, is the preservation of the 3D conformation among the sequence variants, not to disrupt the geometry of the functional interactions with a target. Thus means to identify which substitutions are likely to disrupt or to preserve those interactions are highly desirable. In particular, it is well-known that some sites of the protein sequence are much more sensitive for this than others. For example, in protein folding, a few key residues force the chain to adopt a rudimentary native-like architecture (23). In sequence evolution, the phylogenetic tree is characterized by a few tree-invariant positions - detectable by Evolutionary Tracing (24), and tree determinants (25) that correspond to functionally important residues in the structure. Therefore, it is essential to identify those key residues and delineate the sequence space that preserves the 3D conformation in their vicinity. Although protein design techniques coupled to molecular simulations techniques can fulfill such a goal, such protocols are hardly tractable to test the complete space of sequences that rapidly growths with peptide size.

Here, we introduce a mid term strategy which does not rely on sophisticated phylogenetic inference nor on explicit structure consideration. We use the concept of structural alphabet (26), a discrete model of local structure, to constrain a very simple protocol for sequence evolution. Our previous works have demonstrated the identification and use of the concept of structural alphabet, and the fact that divergent profiles tend to correspond to divergent structures, as illustrated by a fragment search strategy exploiting this trend (27). This is the first time that our structural alphabet is applied to the effect of mutations of the same structure. We show that it provides, at a limited computational cost, sequence sampling consistent to that observed in the subset of the natural sequences identified and that it provides information about critical positions, ie. conserved positions not to substitute or position specific disruptive amino-acids substitutions.

2. Theory

2.1 Structural alphabet (SA)

A structural alphabet (SA) can be considered as a generalized secondary structure in which each region of a protein can be associated to one specific conformation, or "SA-letter" of the alphabet. Thus, a structure can be transposed into an optimal string of SA-letters expected to describes accurately enough the series of the local conformations - two different strings should be able to distinguish between small conformational changes and hopefuly provide a basis to regenerate the 3D conformation. Here, we use a structural alphabet made of 27 SA-letters that have been identified using Hidden Markov Models (26). Each SA-letter corresponds to a fragment of four amino-acids of the protein, and is described by four distance descriptors and the SA-letters overlap by three amino-acids. Thus, the conformation of a protein of size L amino-acids can be described by a string a SA-letters of size $L-3$.

Given the amino-acid sequence of L amino-acids, it is possible to predict the probability that each series of 4 amino acid is associated with each of the 27 SA-letters, i.e. a SA-profile of size $27(L-3)$ values. In such profile, each of the 27 probabilities associated to position l correspond to estimates of

p(SA/aa), i.e. probabilities of the SA letters given information of the amino-acids. Briefly, for a given protein sequence, a sequence profile is first generated using PSI-BLAST. It is then used to predict an SA-profile. This prediction is performed by a Support Vector Machine (SVM) that was trained to reproduce the SA-profiles extracted from known structures. Here, we use a 2012' update of the SVM predictor that was described in (28), learnt over a collection of proteins of size more than 80 amino acids. Of note, those profiles have been demonstrated effective for the *de novo* modeling of peptides, using different strategies to exploit the information they contain (29).

Interestingly, the SA-profiles make a link between sequence and structure, and their variations are likely to reflect how the substitutions in the amino-acid sequence can impact the 3D structure of the protein. They contain information specific of the structure, the profiles of two proteins of dissimilar folds being dissimilar and the profiles of tow proteins of similar fold being similar. We have shown in a previous study (27) that using the Jensen Shannon divergence (JS) as a measure of profile dissimilarity, it is possible to identify, for low values of JS, protein fragments that have similar conformations. Our purpose here is to use SA-profile divergence as a control over the preservation of the structure of the target after an amino-acid substitution, to ensure a position by position preservation of the local conformation associated to the initial amino-acid sequence. For the purpose of illustration, Figure 1 shows the relationship between the JS and the RMSD for a fragment of 4 amino acids taken from the structure of the GCC-box binding domain (PDB(30): 3GCC), positions 151-155 aligned with random fragments of a collection of proteins. One clearly sees that small JS value (resp. large JS values) tend to be associated with small RMSD values (resp. large RMSD values). Note that there are a few cases for which the RMSD is large while the JS is low, which is in agreement with the well known observation that small fragment with identical sequences can adopt different conformations (31). Conversely some fragments come with a large JS value (different sequences) and small RMSD values, again consistent with previous observations. Importantly, in the present study, we do not consider unrelated fragments, but we make evolve one specific sequence, which is expected to prevent the occurrence of such singular behaviors. Figure 2 illustrates how a SA-profile can be affected by one single mutation. Here, we present the impact of the A53T mutation of the α-Synuclein, reported to impact the local conformation around position 53 (32). One clearly sees that just one single substitution results in large differences in the SA-profile around position 53, not affecting, as could be expected other parts of the profile (JS values of more than 0.13). Note that in our SA model, the probability distributions correspond in fact to overlapping fragments of 4 amino acids. Thus, one amino acid contributes to 4 consecutive probability distributions.

3. Methods

3.1 Sequence evolution

Figure 3 presents an overview of the procedure. The protocol to simulate sequence evolution is similar to a Monte Carlo procedure. Starting from the sequence of a target protein, it will generate substitution events that are accepted or not, simulating a random walk in the sequence space. For each substituted sequence, a SA-profile is predicted and compared to the SA-profile generated with the initial sequence. Acceptation of substitutions is performed according to the expected structural divergence as measured by JS.

The core simulation of sequence evolution is made according to the scheme proposed by PSeqGen (33) or CS-PseqGen (34). The JTT substitution matrix is used to allow for different rates of substitutions depending on the type of amino acids. Since we are interested in the substitution events disregarding any molecular clock or any phylogenetic model, evolutionary time is considered as infinite, meaning

each step results in selecting a substitution, similarly to a Monte Carlo inspired process. Site-specific rate heterogeneity is not considered either, since in the context of protein design, it is largely unknown. Substitution acceptance is controlled as detailed below. In total, simulations are defined by an initial sequence and a number of substitutions to generate.

Acceptance or reject of the substitutions is controlled by the estimation of the expected impact of a substitution on the preservation of the local structure of the target, i.e. on the dissimilarity of the SA-profiles before and after substitution, using JS as follows:

$$JS(SAP_{ref}, SAP_{cur}) < JS_{cut\text{-}off}: \text{ accept, reject otherwise}$$

where SAP_{ref} corresponds to the SA-profile of the initial sequence, SAP_{cur} to the SA-profile of the current sequence and the $JS_{cut\text{-}off}$ value corresponds to the value above which the divergence is considered as too large to guarantee structure preservation. Varying the value of this cut-off can be assimilated to varying the temperature of the Monte-carlo.

SA-profiles associated with an amino-acid sequence of size L consist in L-3 probability distributions, each of 27 values. For each pair of distributions belonging to SAP_{ref}^{l}, SAP_{cur}^{l}, , where l varies from 1 to $L\text{-}3$, we use $JS(l)$ as a mesure of the dissimilarity. As mentioned above, a particular position l in the sequence contributes to the SA-profiles of four positions (l .. $l+3$), which requires to check the JS at those four positions. For sake of simplicity, we measure the JS for all of the L-3 positions, and to combine the L-3 JS values, we consider:

$$JS(\mathrm{SAP_{ref}}, \mathrm{SAP_{cur}}) = max(JS(l),\ l = 1 \,..\, L\text{-}3)$$

where the maximum stands to ensure that, given a cutoff value, for no profile the deviation is more than this value.

Of note, the observed variability over independent simulations is, in our experience, rather weak and is not discussed any further here. Also note that for each substitution event, we perform a full SVM prediction which requires itself a psi-blast and thus simulations of 3000 substitution events take between one and two days each on a standard workstation.

3.2 Amino-acid profile divergence

Once simulations performed, one needs means to analyze the diversity of the sequence generated. To estimate the number of amino acids occurring at a given position l in the sequence, we use the number of equivalent amino-acids defined as:

$$\mathrm{Neq}(l) = 1/p_{max}(l)$$

where $p_{max}(l)$ corresponds to the maximum of the occurrence probabilities associated with the 20 amino-acids at position l in the sequence. These values are averaged over the L positions of the sequence to get an estimate of the diversity over all positions.

The comparison of the two profiles made of L distributions of 20 probabilities is a difficult issue. Here, we simply consider the correlation coefficient of the two vectors made of the $L\,.\,20$ values, which we found more intuitive than criteria such as the average JS or the average dot product over the L positions.

3.3 Test sets

To assess the performance of our procedure, we have chosen 6 peptides and mini-proteins of known structure, not intrinsically disordered, and of varied topologies. Table 1 provides details on their size, secondary structure, and sequence. Such small sized proteins have the advantage of not belonging to the range of size used to learn the SVM. We emphasize, however, that although we assess the protocol on such small proteins, the procedure is very general and applicable to larger proteins, as illustrated Figure 2 for the α-Synuclein, a protein of 140 amino acids.

As a reference for "naturally observed" sequence variation, we have considered the sequence variants made of the homologs of these targets found either in the Protein Data Bank, or in the Uniprot repository. For the later case, since our aim here is to estimate the acceptable substitutions in terms of amino-acids, we have build profiles considering entries of the Uniref 90 subset, to discard too identical sequences. The procedure was as follows:

1/ search uniref90 using blastp (35)
2/ retrieve the entries of the hits
3/ perform a multiple alignment of the hits using clustal omega (36)
4/ identify the region of the alignment matching the query and extract the corresponding profiles

Name	PDB	s2	L	Seq.
Transactivation domain of CRE-BP1/ATF-2	1bhi	α2β	38	MSDDKPFLCTAPGCGQRFTNEDHLAVHKHKHEMTLKFG
N-terminal leucine-repeat region of hepatitis delta antigen	1by0	α	27	RKKLEELERDLRKLKKKIKKLEEDNPW
C-terminal UBA domain of the human homologue of RAD23A	1dv0	α3	47	GSQEKEAIERLKALGFPESLVIQAYFACEKNENLAANFLLSQNFDDE
Bomain X of measle phosphoprotein	2k9d	α3	44	VIRSIIKSSRLEEDRKRYLMTLLDDIKGANDLAKFHQMLVKIIM
FAF-1 UBA Domain	3e21	α3	45	GSMDREMILADFQACTGIENIDEAITLLEQNNWDLVAAINGVIPQ
first WW domain of Nedd4-2	1wr3	β3	36	GSPPLPPGWEEKVDNLGRTYYVNHNNRSTQWHRPSL

Table 1: Name: the name of the protein. PDB: PDB identifier of the structure. s2: secondary structure topology. L: size (amino-acids). Seq.: amino acid sequence.

4. Results and discussion

4.1 Sequence divergence depends on the JS cut-off value

Figure 4 depicts, for the transactivation domain of CRE-BP1/ATF-2 (PDB: 1bhi), the evolution of the JS and of the sequence identity as a function of the number of substitution events. As could be expected, one observes a fluctuation of JS between low and large values (top inset, red) whereas JS of the accepted sequences remains under the limit imposed of 0.12 (top inset, black). Interestingly, one also observes that the sequence identity of the sequences generated, when compared to the initial sequence, rapidly decreases down to values close to only 10-20% (black), which corresponds to the twilight zone in terms of homology. Small discrepancies are observed between the sequences accepted and those generated, on average, as can be observed for instance around step 200. A control simulation using the same procedure, but not applying any JS constraint shows that rapidly, sequence diverge in a random manner (gray).

Varying the value of the JS cut-off impacts, as expected, the degree of divergence of the sequences. Indeed, lower value of JS cut-off results (JS cut-off 0.08, green) in exploring sequences with higher sequence identities, given that a more strict control on the impact of substitutions on the predicted local conformations results in rejecting more substitutions.

Similar behaviors are observed for the other targets of the test set, and are summarized in Table 1 that reports, for all targets, the lowest sequence identity reached during the simulations. Overall, the results show that the procedure is able to reach explore sequences diverging down to close to 10-20 % sequence identity.

4.2 Constrained simulated sequence evolution intrinsically embeds site-specific rate heterogeneity of substitution

Figure 5 presents, for 1bhi, a logo representation of the distribution of the 20 amino-acids per site. Figure 5B-D present logos corresponding to subsets of sequences obtained from a simulation of 3000 events driven by a JS cut-off of 0.12, using as second step filters JS cut-off values of 0.12, 0.10 and 0.08, respectively. The size of the subsets is of 892, 256 and 29 sequences, respectively. As a reference, Figure 5A presents the results of an unconstrained simulation of 892 events, i.e. a size identical to the subset depicted Figure 5B. Figures 5E and 5F correspond to the logos obtained from the PDB (13 sequences only) and uniref90 (323 sequences), using blastp (see methods).

A first observation is that compared to the unconstrained simulation , that leads to a rather flat profile, the logo profiles differ largely. In Figure 5B, positions 6, 14, 18 and 31, associated with amino acids P, C, F and H, respectively, appear more conserved than other positions ($Neq < 1.3$). Interestingly, considering lower JS cut-off values (Figure 5C-D), the number of positions with $Neq < 1.3$ increases to encompass progressively also positions 7, 19, 22. According to the design of our procedure, this suggests that these amino acids are probably critical to maintain the local conformation at these positions. Looking at the structure of 1bhi (Figure 6), one notes that these positions are located at the extremities of secondary structure elements. Positions 9, 14, 27 and 31 correspond to the cysteines and histidines involved in the coordination of the zinc ion (not present in the structure).

Looking at the profile generated using homologues of the PDB (Figure 5E), only positions 9, 14, 18, 20, 22, 27 and 31 are associated with *Neq* values less than 1.3. Despite the weak number of sequences, those positions match rather well the results of the simulation. Note that, since we could not identify homologues with known structure for all 6 cases, we do not discuss the PDB profiles for the other targets.

Looking at the distributions obtained for uniref90, one observes that not only positions 6-7, 14, 18, 22 and 31 are conserved, but in fact positions 6-7, 9, 11-14, 16-18, 21-24, 27, and 29-37. Apart from C-ter residues 32-37 that are involved in the functional interaction with kinases, as reported by ELM (37), those residues correspond to amino acids located at the interface between the secondary structure elements, most being involved in long range interactions stabilizing the overall conformations. Hence, the observed heterogeneity obtained from simulations constrained by JS seems to effectively contain information about key residues that are required for the preservation of the secondary structure elements.

4.3 Sequence divergence is target dependent.

We now consider all proteins of the test set. Table 2 reports for all targets the number of sequences accepted for different JS cut-off values, and the average *Neq* values of the profiles. Note that all simulations have been performed using JS of 0.12. For JS values below 0.12, the numbers correspond to the subsets of the simulated sequences matching the JS condition. It is striking that the numbers depend on each target, meaning target sequence/topology has an impact on the simulated evolution.

	1bhi		**1by0**		**1dv0**		**2k9d**		**3e21**		**1wr3**	
JS	*#*	*Neq*	*#*	*Neq*	*#*	*Neq*	*#*	*Neq*	*#*	*Neq*	*#*	*Neq*
0.12	892	3.12	836	1.98	1048	3.02	1307	3.02	677	2.30	1126	3.05
0.10	256	2.89	632	2.00	566	3.	356	2.89	186	2.23	535	3.04
0.08	58	2.20	405	1.98	178	2.87	80	2.33	27	1.75	79	2.45
0.06	6	1.13	172	1.85	29	2.08	11	1.22	8	1.06	11	1.13
0.04	0	-	46	1.75	4	1.03	9	1.07	2	1.02	3	1.02
uniref90	323	1.36	171	1.38	332	1.16	18	1.74	232	2.19	237	1.05
JS_u	0.075		0.032		0.047		0.075		0.12		0.06	
JS_2	0.08 (#:649)		0.04 (#: 608)		0.05 (#: 538)		0.08 (#: 1110))		0.12 {#: 677)		0.07 (#: 887)	
JS_{u2}	0.05 (#:28)		0.025 (#: 38)		0.025 (#: 18)		0.05 (#: 96)		0.12 (#: 677)		0.035 (#: 16)	
r	0.84		0.87		0.86		0.63		0.77		0.77	

Table2: #: Number of sequences accepted at the given JS value, and the associated Neq. Simulations of 3000 substitution events were performed using JS=0.12. Uniref90: corresponding values observed among homolog sequences of uniref90. JS_u: JS values for which simulated Neq is identical to that of uniref90. JS_2: JS cut-off values used for the second stage simulations (section 4.4). JS_{u2}: JS_u for the second stage simulations. *r*: Correlation coefficient between second stage and uniref90 profiles.

If, for a JS-cutoff value of 0.12, the observed variation in the number of sequences is rather limited (acceptance rate of substitution events between 22 and 43 %), larger variations are observed for lower JS cut-off values. For instance for 0.08, the number of sequence varies between 405 and 27 (acceptance rate between 13 and 0.9%), and the Neq values vary between 2.87 and 1.75, without any obvious relation to the number of accepted substitutions.
Finally, if we consider the values of Neq obtained for the uniprot90 homologs, they look rather small (between 1.05 and 2.19) compared to the values obtained from the simulations. For the simulations, values of Neq similar to those observed for uniprot90 are obtained for rather low values of JS - between

0.03 and 0.12 (JS_u), but again one observes a strong dependence on the specifics of each particular case.

4.4 JS controlled sequence evolution can mimmick sequence divergence observed in natural sequences.

In order to get a better agreement between the simulations and the observations of uniref90, we have performed second stage simulations using JS values for the Neq were close to that of uniref90 (Table 2, JS_{u2} values). Note that these new cut-off values guide the search into new regions of the sequence space, and thus, we do not expect the Neq values to be preserved, thus JS_2 values were chosen slightly larger than the JS_u values. A first observation about these simulations is that indeed, the acceptance rate is, as expected much larger for lower JS values (between 18 and 37% – see Table2) than that observed for the initial simulations, which confirms a better sampling of the sequence space for the targeted divergence, the minimum number of accepted substitution events being of 538, when it was below 100, for the first series of simulations. The minimal sequence identities for the accepted sequence were of 23, 37, 32, 20, 31 and 25% for 1bhi, 1by0, 1dv0 2k9d, 3e21 and 1wr3, respectively.

Figure 7 shows for each of the 6 targets, the logos obtained for the JS_{u2} values, i.e. for average *Neq* values similar to those obtained for uniref90. There is a good visual agreement between the logos of the simulations and uniref90, and the *r* values vary between 0.63 and 0.77 (see Table 2), which is highly significant ($p < 10^{-10}$). Note however that the number of sequences, even if much larger than that of the first series of simulations, remains low, except for 2k9d and 3e21 for which the r values, even if a bit lower remain very significant. Of note, this low number of accepted sequences is also conditioned by the weak Neq values observed for uniref90. We have further verified that for series of 1000 tests permuting the series of 20 probabilities for one of the profiles, the distribution of r values is close to 0-0.1 on average. Overall, this suggests that our simple procedure is able to mimmick accurately enough the sequence fluctuations as observed in uniref90.

Looking at the positions highly conserved (*Neq* < 1.1) it is striking that most of them are located at the termination of secondary structures, which suggest that the approach is probably sensitive enough to detect that these regions are critical for structure preservation. This is even more true when looking at residues conserved for slightly larger values of JS cut-offs (magenta in Figure 7). Overall, the positions highly conserved identified by our protocol correspond to 31 over 55 (56 %) of the positions highly conserved observed in the uniref90 distributions. As illustrated upper for the 1bhi case, constraints on sequence can occur due to several reasons not related to the preservation of the local structure.

Finally, another outcome of the approach is that it also provides for each position in the sequence, information about the substitutions that have been rejected. Figure 7 also presents, for all 6 cases the logos of the amino acids that were not accepted by the JS cut-off, and were excluded at least once with a JS value of more than 0.3 (chosen large enough compared to the JS cut-off values to get confident enough a local conformational change is expected). It is interesting to compare these amino acids to those present in the uniref sequences. For instance, at 1by0 conserved position 22, D was rejected several times by our protocol, whereas E was preserved. D is not present either in the uniref profile. At position 10 of 1dv0, the uniref profile has a conserved position with R, when our simulations report occurrences of R,K and M, and A, L, T and V where rejected. For 3e21 position 10, the uniref contains occurrences of D, N, S, E, accepted amino acids of the simulations contain D, N, E and G, A, H, R, S, T, Y were rejected. Another such similar example is observed at 1wr3 position 28. In few cases, however, some rejected amino acids are observed in the uniref profiles. Such a situation occurs for

2k9d position 10, where the simulations accepted R, H, K, uniref contains occurrences of N, H, K, R and rejected amino-acids include N. Several causes can explain such events, including epistatis effects (occurrence of N conditioned by the occurrence of another amino-acid close in the sequence to preserve local structure), or insufficient accuracy of the SVM prediction for some sequence motifs, to mention two of them. Such cases are however rare, and concern mostly amino acids occurring with weak frequencies in the uniref90 profiles. Over the 6 targets of our test set, the average cumulated probabilities of such amino acids if of only 6.7%.

5. Discussion and perspectives

In the present study, we have introduced a new paradigm to simulate protein sequence evolution. It builds on the concept of structural alphabet – a kind a generalized secondary structure – used as a means to constrain a very basic procedure to sequence evolution. The underlying hypothesis is that the limitation of the divergence between two structural alphabet profiles ensures the point to point preservation of the local conformation along the sequence. Surprizingly, we find that this very simple procedure is able, for the mini-proteins of our test set, to result in generating sequences which significantly fit sequence diversity as observed among uniref90. Indeed, we find that our procedure is able to grab the specifics of the heterogeneity of rate among sites, and of the nature of the amino-acids at those sites. This being only induced by the constraint on the divergence between the structural alphabet profiles implies that the point to point control of the preservation of the local structure of a target is an important determinant of sequence evolution. To the best of our knowledge, our procedure is the first procedure to directly address this question, and this clearly opens the door to further investigations on the amount by which sequence evolution is controlled by such local constraints versus longer range interactions. One has however to consider carefully the present results. For one part, we have tested so far a limited number of targets, and in addition, some optimizations can already be foreseen. A special point is about the heterogeneity in the JS cut-off values depending on the targets. Probably, it could be of interest to study the convergence of our simulations for a larger number of cycles of simulations – here, only two were performed. A comparison and/or combination with the results of simulation procedure along a phylogenetic tree could also reveal informative.

It is difficult to compare the results of our protocol to those of former approaches described to measure the impact of mutations, as their objectives differ largely. The aim of the approaches developed so far has usually been to predict the impact of substitutions on protein stability. This includes local conformation preservation considerations, but also longer range effects involved in the stabilization of the structure, such as the interactions between secondary structure elements, or more complex effects. Here our main focus is only about the preservation of the local structure, disregarding longer range effect. It is thus expected that our results are more specific than those of other approaches. In addition, most of former approaches, such as Backrub (38), Rosetta-Design (15) or POP-Music (39) require the 3D coordinates of the protein while we consider here only the sequence. We discuss briefly our results in the light of the results obtained by the INPS server (40) – based on sequence and the very recent SAAMBE-3D server (41) – based on structure. Considering for instance the 1by0 target, both INPS and SAAMBE-3D predict that the subsitution E22D is slightly destabilizing, while it is not accepted by our protocol, nor observed in unriref90, which seems consistent. At position 10 of 1dv0, SAAMBE-3D predicts all substitutions to be destabilizing - which is consistent the full preservation of K in uniref90, while INPS predicts that substitutions of R into A, T, K or M would decrease the stability of the protein and that substitutions into L or V would increase the stability. Our results suggest that substitutions of K into R or M are compatible with local structure preservation, occurrences of A, L, T or V being never observed. For 2K9D position 10, the same kind of observations can be done. R10H, R10N, R10K substitutions are all predicted to decrease the stability by both INPS and SAAMBE-3D, while our

procedure suggests that R10K and R10H preserve local structure, and all are observed in uniref90. In fact, it is fully possible that some substitutions that destabilize the global structure do not affect the local structure and are still compatible with the preservation of the local structure. It is also to be noted that servers such as INPS or SAAMBE-3D consider the impact of single mutations, all other amino acids being preserved, while our procedure makes the complete sequence evolve, and thus, some substitutions can be accepted conditional to previous events in the neighborhood of a site. We also recall that our procedure focusing on the preservation of 3D-structure comes with no quantification of the impact of the substitutions on stability or on function.

Indeed, a motivation for developing this procedure was the perspective to assist sequence optimization, as can be needed for the development of a candidate therapeutic peptide or mini-protein. Two questions can be posed in such context. The first is to identify which positions in the sequence should not be modified to ensure the preservation of the conformation of the candidate. To this respect, it is interesting that the observed heterogeneity of substitution rates over positions obtained from our simulations make it possible, from a single sequence, to identity positions at which substitutions seem risky. A second is to propose, for positions at which we find some diversity, which substitutions are likely. Numerous protocols and prediction approaches have been developed to this aim (see for instance 42-43). Here, a particularity of our approach, which is not quantitative, is probably to return information about which residue substitutions were not accepted at a given position. Although our results clearly show that rejected substitutions depend on the context of the neighboring residues since false rejections were observed at a frequency of close to 7%, we put emphasis on the fact that our procedure is based on single mutation events only, whereas more sophisticated ways to simulate evolution could be setup, for instance considering co-evolution. All together, we however hope that our procedure of sequence evolution under the constraint of the point to point preservation of local structure already meets the objective of assisting sequence optimization.

Funding: INSERM U1133 recurrent funding

Acknowledgments: The authors wish to thank J. Chomilier and Ph. Derreumaux for useful discussions.

References

1. Hoess, R. H. (2001). Protein design and phage display. *Chemical Reviews*, *101*(10), 3205-3218.
2. Bornscheuer, U. T., & Pohl, M. (2001). Improved biocatalysts by directed evolution and rational protein design. *Current opinion in chemical biology*, *5*(2), 137-143.
3. Fowler, D. M., & Fields, S. (2014). Deep mutational scanning: a new style of protein science. *Nature methods*, *11*(8), 801.
4. Fletcher, W., & Yang, Z. (2009). INDELible: a flexible simulator of biological sequence evolution. *Molecular biology and evolution*, *26*(8), 1879-1888.
5. Sipos, B., Massingham, T., Jordan, G. E., & Goldman, N. (2011). PhyloSim-Monte Carlo simulation of sequence evolution in the R statistical computing environment. *BMC bioinformatics*, *12*(1), 104.
6. Spielman, S. J., & Wilke, C. O. (2015). Pyvolve: a flexible Python module for simulating sequences along phylogenies. *PloS one*, *10*(9).
7. Low, A., Rodrigue, N., & Wong, A. (2017). COMPASS: the COMPletely Arbitrary Sequence Simulator. *Bioinformatics*, *33*(19), 3101-3103.
8. Shapiro, B., Rambaut, A., & Drummond, A. J. (2006). Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Molecular biology and evolution*, *23*(1), 7-9.
9. Camenares, D. (2019). Simulating protein and nucleic acid sequence co-evolution. *The FASEB Journal*, *33*(1_supplement), 642-2.
10. Anderson, C. L., Strope, C. L., & Moriyama, E. N. (2011). SuiteMSA: visual tools for multiple sequence alignment comparison and molecular sequence simulation. *BMC bioinformatics*, *12*(1), 184.
11. Löytynoja, A., Vilella, A. J., & Goldman, N. (2012). Accurate extension of multiple sequence alignments using a phylogeny-aware graph algorithm. *Bioinformatics*, *28*(13), 1684-1691.
12. Katoh, K., & Standley, D. M. (2016). A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics*, *32*(13), 1933-1942.
13. Xia, Y., & Levitt, M. (2004). Simulating protein evolution in sequence and structure space. *Current opinion in structural biology*, *14*(2), 202-207.
14. Gainza, P., Nisonoff, H. M., & Donald, B. R. (2016). Algorithms for protein design. *Current opinion in structural biology*, *39*, 16-26.
15. Liu, Y., & Kuhlman, B. (2006). RosettaDesign server for protein design. *Nucleic acids research*, *34*(suppl_2), W235-W238.
16. Allouche, D., André, I., Barbe, S., Davies, J., de Givry, S., Katsirelos, G., ... & Traoré, S. (2014). Computational protein design as an optimization problem. *Artificial Intelligence*, *212*, 59-79.
17. Perron, U., Kozlov, A. M., Stamatakis, A., Goldman, N., & Moal, I. H. (2019). Modeling Structural Constraints on Protein Evolution via Side-Chain Conformational States. *Molecular biology and evolution*, *36*(9), 2086-2103.
18. Grahnen, J. A., Nandakumar, P., Kubelka, J., & Liberles, D. A. (2011). Biophysical and structural considerations for protein sequence evolution. *BMC evolutionary biology*, *11*(1), 361.
19. Grahnen, J. A., & Liberles, D. A. (2012). CASS: Protein sequence simulation with explicit genotype-phenotype mapping. *Trends in Evolutionary Biology*, *4*(1), e9-e9.
20. Bhardwaj, G., Mulligan, V. K., Bahl, C. D., Gilmore, J. M., Harvey, P. J., Cheneval, O., ... & Huang, P. S. (2016). Accurate de novo design of hyperstable constrained peptides. *Nature*, *538*(7625), 329-335.
21. Huang, P. S., Boyken, S. E., & Baker, D. (2016). The coming of age of de novo protein design. *Nature*, *537*(7620), 320-327.
22. Dou, J., Vorobieva, A. A., Sheffler, W., Doyle, L. A., Park, H., Bick, M. J., ... & Carter, L. (2018). De novo design of a fluorescence-activating β-barrel. *Nature*, *561*(7724), 485-491.
23. Dobson, C. M. (2003). Protein folding and misfolding. *Nature*, *426*(6968), 884-890.

24. Lichtarge, O., Bourne, H. R., & Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *Journal of molecular biology*, *257*(2), 342-358.
25. del Sol Mesa, A., Pazos, F., & Valencia, A. (2003). Automatic methods for predicting functionally important residues. *Journal of molecular biology*, *326*(4), 1289-1302.
26. Camproux, A. C., Gautier, R., & Tuffery, P. (2004). A hidden markov model derived structural alphabet for proteins. *Journal of molecular biology*, *339*(3), 591-605.
27. Shen, Y., Picord, G., Guyon, F., & Tuffery, P. (2013). Detecting protein candidate fragments using a structural alphabet profile comparison approach. *PloS one*, *8*(11).
28. Maupetit, J., Derreumaux, P., & Tufféry, P. (2010). A fast method for large-scale De Novo peptide and miniprotein structure prediction. *Journal of computational chemistry*, *31*(4), 726-738.
29. Lamiable, A., Thévenet, P., Rey, J., Vavrusa, M., Derreumaux, P., & Tufféry, P. (2016). PEP-FOLD3: faster de novo structure prediction for linear peptides in solution and in complex. *Nucleic acids research*, *44*(W1), W449-W454.
30. Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., ... & Fagan, P. (2002). The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*, *58*(6), 899-907.
31. Sternberg, M. J., & Islam, S. A. (1990). Local protein sequence similarity does not imply a structural relationship. *Protein Engineering, Design and Selection*, *4*(2), 125-131.
32. Kang, L., Wu, K. P., Vendruscolo, M., & Baum, J. (2011). The A53T mutation is key in defining the differences in the aggregation kinetics of human and mouse α-synuclein. *Journal of the American Chemical Society*, *133*(34), 13465-13470.
33. Grassly, N. C., Adachj, J., & Rambaut, A. (1997). PSeq-Gen: an application for the Monte Carlo simulation of protein sequence evolution along phylogenetic trees. *Bioinformatics*, *13*(5), 559-560.
34. Tufféry, P. (2002). CS-PSeq-Gen: simulating the evolution of protein sequence under constraints. *Bioinformatics*, *18*(7), 1015-1016.
35. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., & Wu, C. H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, *23*(10), 1282-1288.
36. Sievers, F., Wilm, A., Dineen D., Gibson, T.J., Karplus, K., Li, W., ... & Higgins, D.G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol., 7*, 539.
37. Dinkel, H., Michael, S., Weatheritt, R. J., Davey, N. E., Van Roey, K., Altenberg, B., ... & Jödicke, L. (2012). ELM—the database of eukaryotic linear motifs. *Nucleic acids research*, *40*(D1), D242-D251.
38. Lauck, F., Smith, C. A., Friedland, G. F., Humphris, E. L., & Kortemme, T. (2010). RosettaBackrub—a web server for flexible backbone protein structure modeling and design. *Nucleic acids research*, *38*(suppl_2), W569-W575.
39. Dehouck, Y., Kwasigroch, J. M., Gilis, D., & Rooman, M. (2011). PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC bioinformatics*, *12*(1), 151.
40. Fariselli, P., Martelli, P. L., Savojardo, C., & Casadio, R. (2015). INPS: predicting the impact of non-synonymous variations on protein stability from sequence. *Bioinformatics*, *31*(17), 2816-2821.
41. Pahari, S., Li, G., Murthy, A. K., Liang, S., Fragoza, R., Yu, H., & Alexov, E. (2020). SAAMBE-3D: Predicting Effect of Mutations on Protein–Protein Interactions. *International Journal of Molecular Sciences*, *21*(7), 2563.
42. Steinbrecher, T., Zhu, C., Wang, L., Abel, R., Negron, C., Pearlman, D., ... & Sherman, W. (2017). Predicting the effect of amino acid single-point mutations on protein stability—large-scale validation of MD-based relative free energy calculations. *Journal of molecular biology*, *429*(7), 948-963.
43. Quan, L., Lv, Q., & Zhang, Y. (2016). STRUM: structure-based prediction of protein stability changes upon single-point mutation. *Bioinformatics*, *32*(19), 2936-2946.

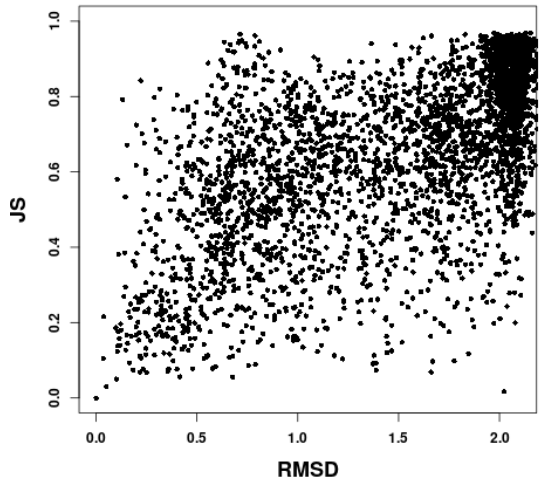


Figure 1: JS as a function of RMSD, using as seed the fragment 151-154 - 4 amino acids - of the GCC-box binding domain (PDB: 3GCC) with all fragments of 77 small proteins.

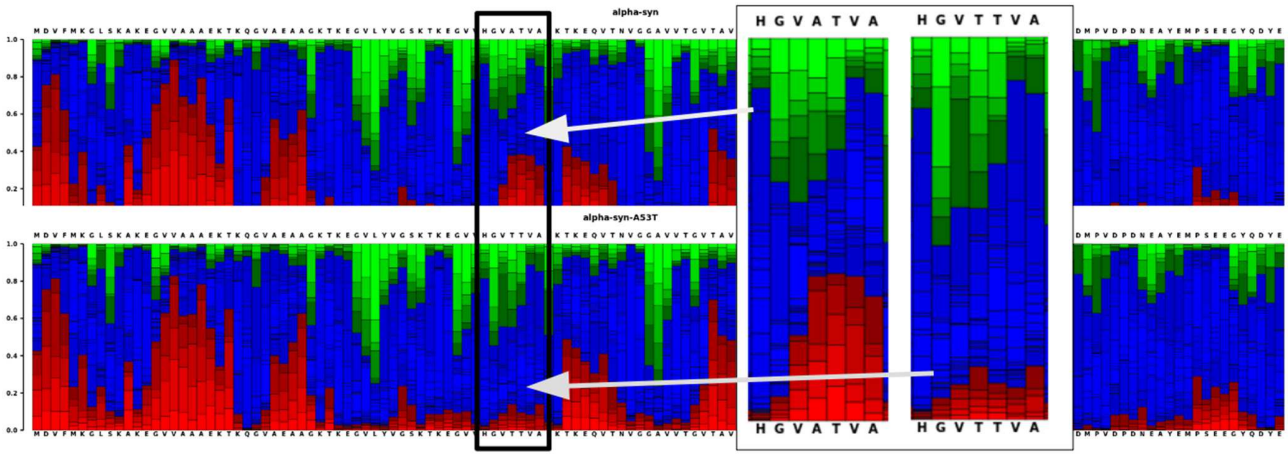


Figure 2: SA-profiles of the α-Synuclein, wild type (top) and A53T mutation (bottom). Each column corresponds to the probability distribution of the 27 letters. SA-letters are sorted from most helical (red) to most extended conformations (green). Details of the probability distributions around position 53 are provided as insets. Of note, SA letters correspond to fragments of 4 amino-acids. Thus the A53 position is associated with HGVA, GVAT VATV and ATVA fragments, facing the H, G, V and A columns, respectively.

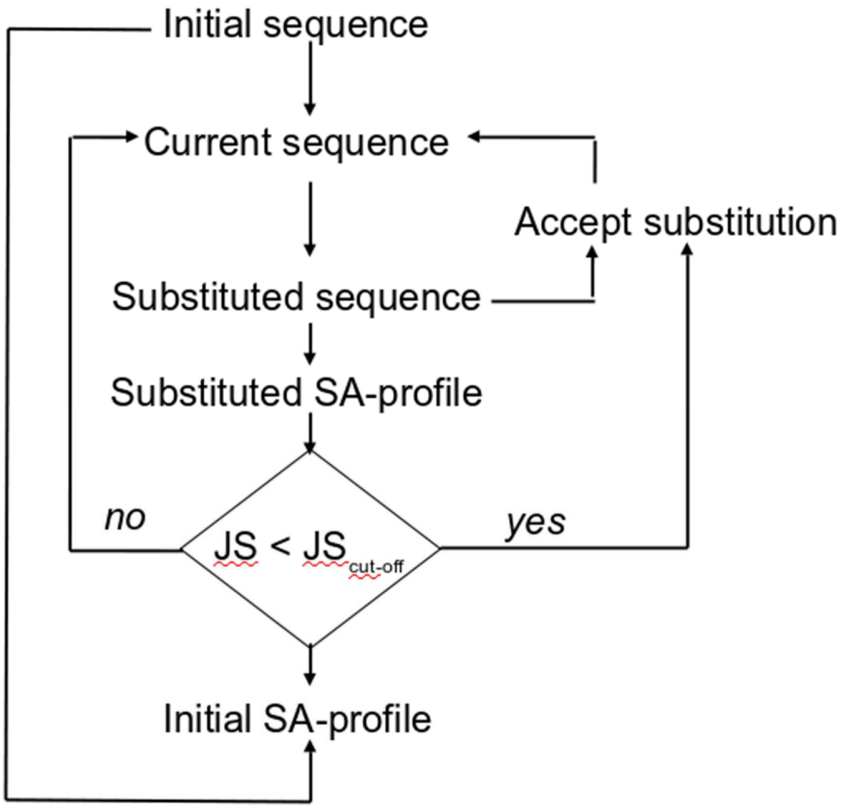


Figure 3: Flowchart of a simulation

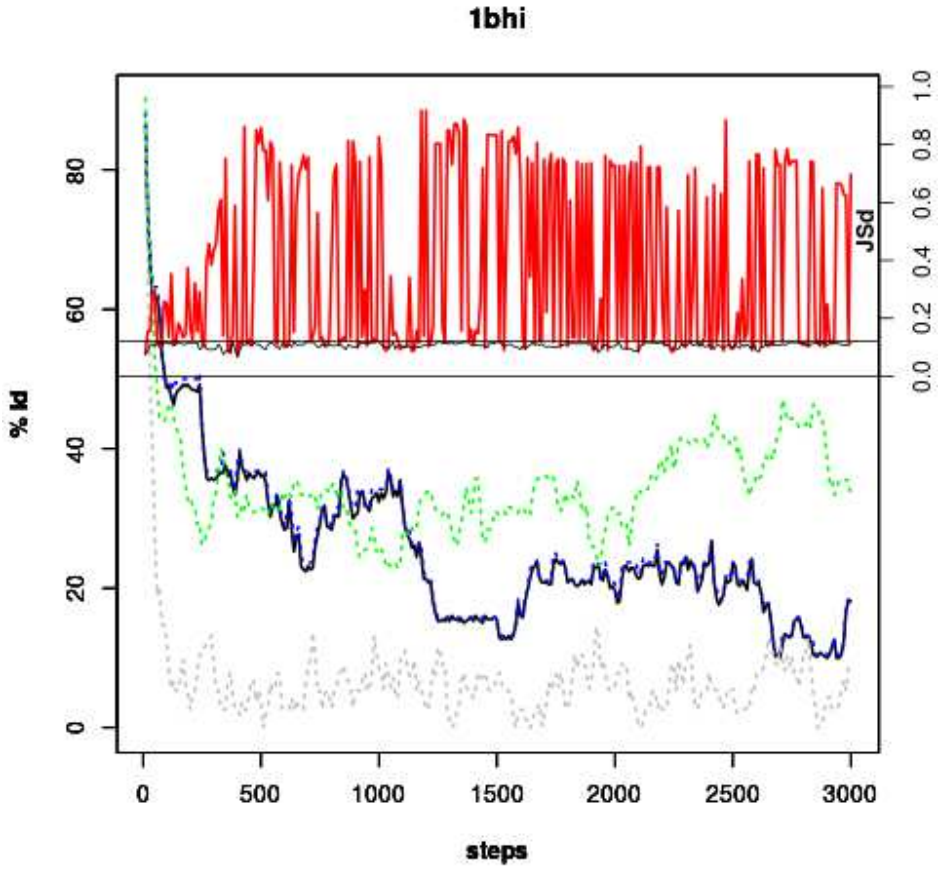


Figure 4: Evolution of the sequence identity to the initial sequence as a function of the number of substitution events. Gray: control without the use of the JS constraint (JS = 1.) Black: JS = 012, accepted sequences. Blue: JS = 12, all sequences. Green: JS = 0.08, accepted sequences.

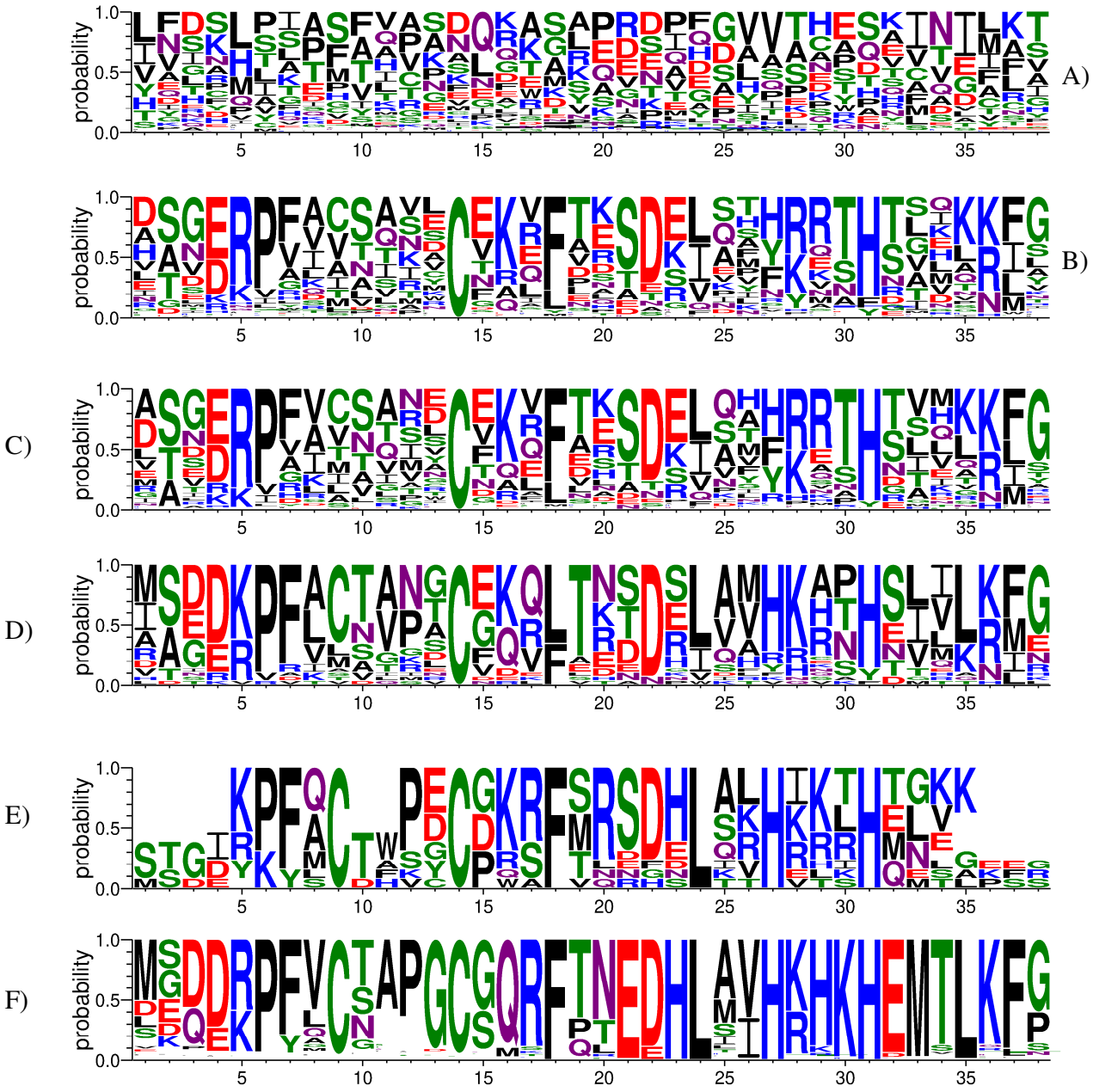


Figure 5: Logo representations of the accepted sequence variants of 1bhi. A: control simulation, JS = 1. B: JS = 0.12. C, D, subsets of sequence of B for JS=0.10, 0.08, respectively. E: homologs of known structure deposited in the PDB. F: homologs identified in uniref90

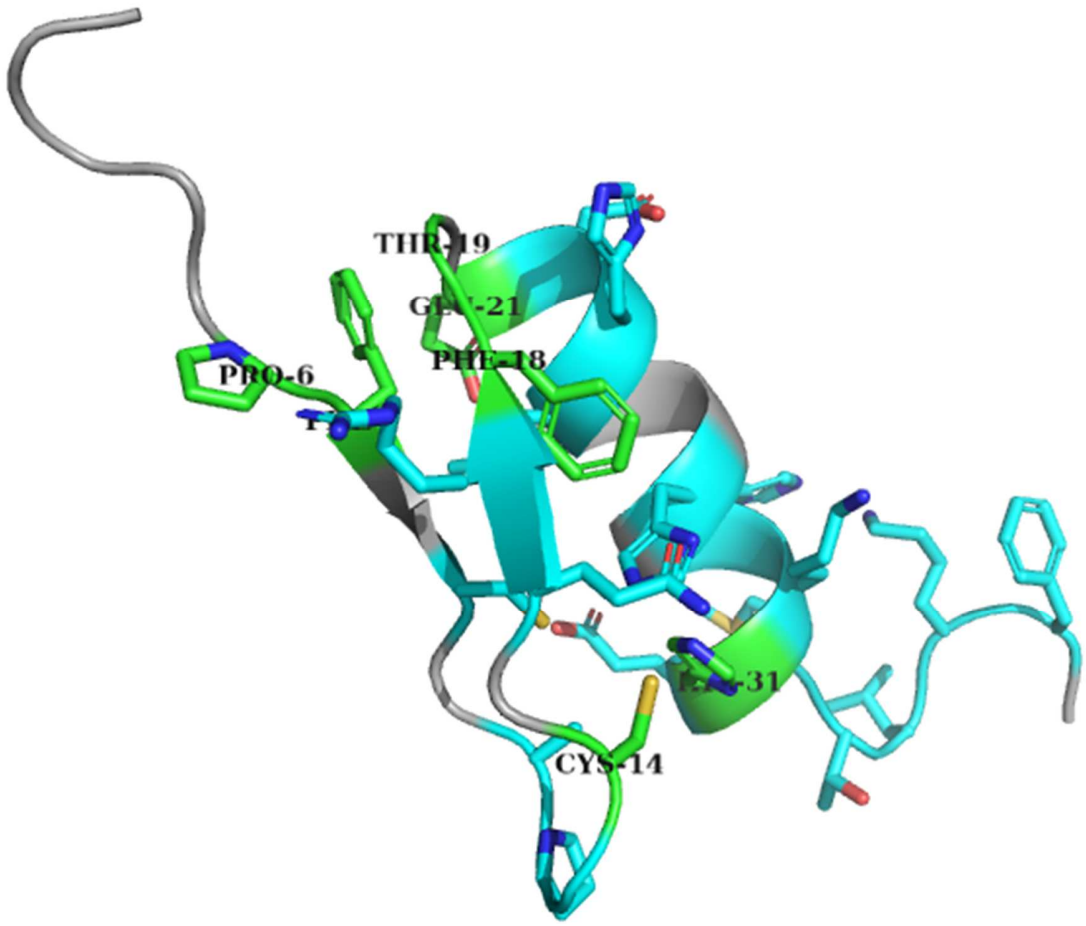
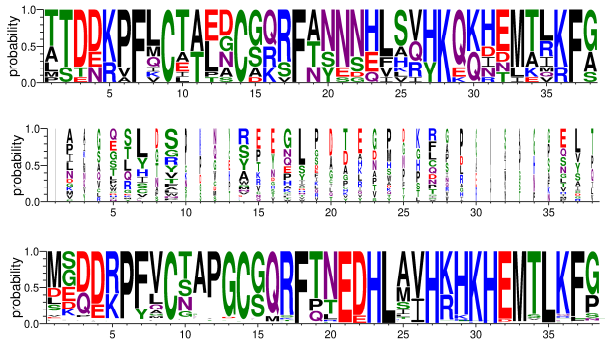
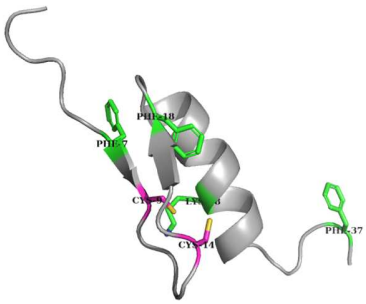


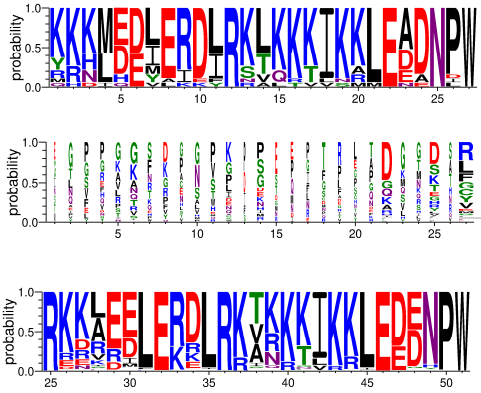
Figure 6: Example of 1bhi. Simulations using JS=0.12. The amino acids at the conserved positions (Neq < 1.3) are depicted in green. Amino acids conserved in the uniref90 profiles are depicted in blue.

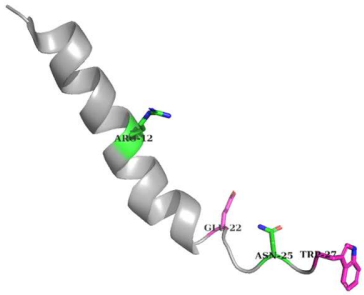
Figure 7



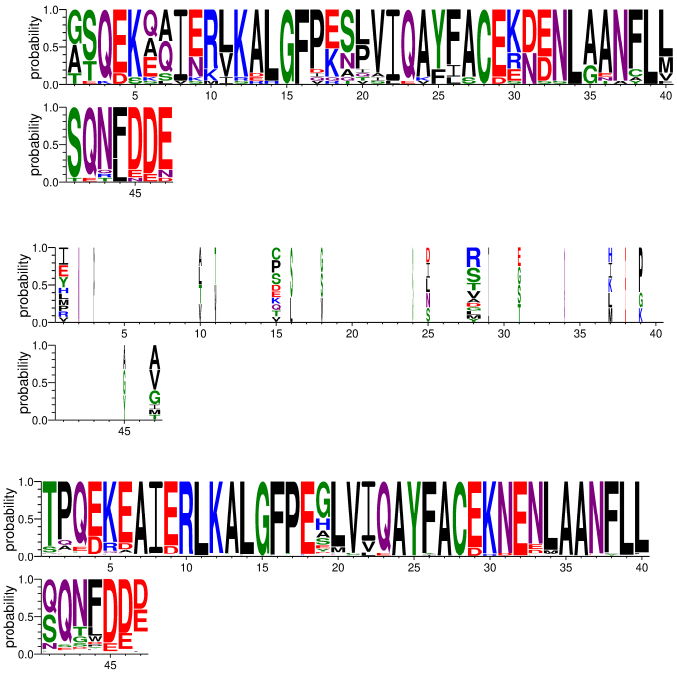


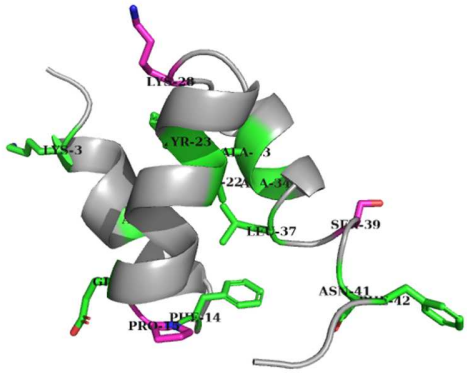
1bhi



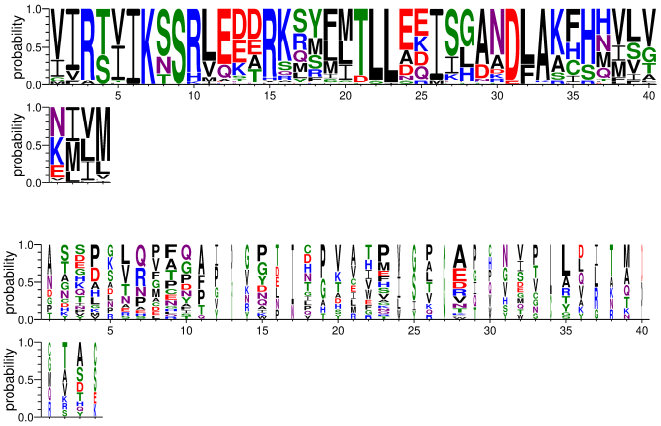


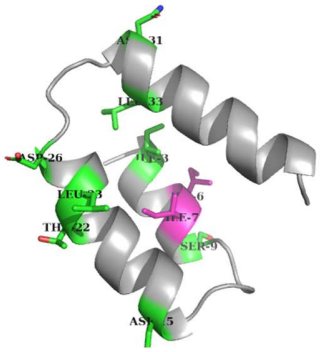
1by0



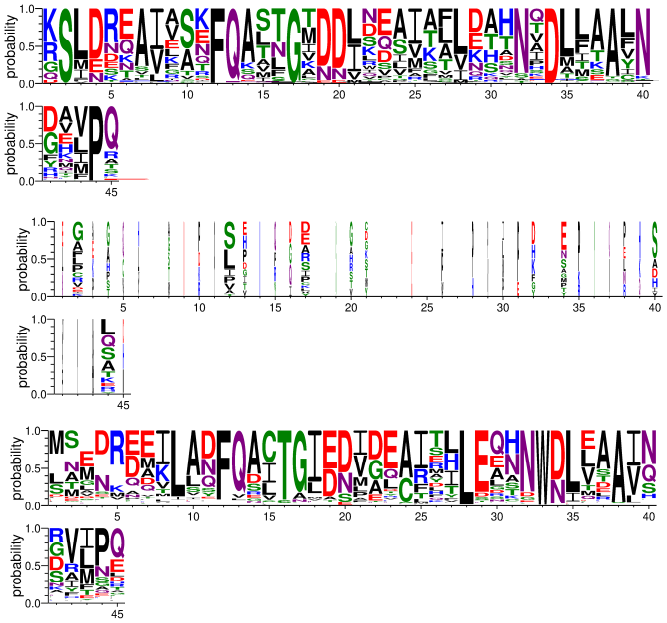


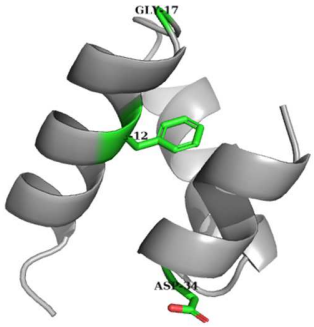
1dv0


probability
1.0
0.5
0.0
5
10
15
20
25
30
35
40

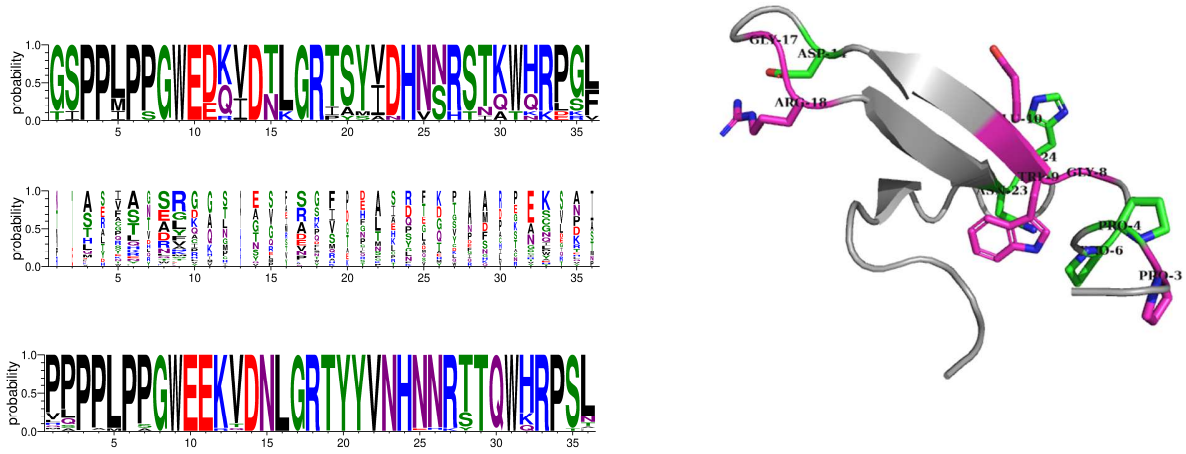


2k9d


probability
1.0
0.5
0.0
5
10
15
20
25
30
35
40
45


GLY-17

3e21



1wr3

Figure 7: Results obtained for simulations calibrated according to the uniref90 Neq.
Left: For each target, we present le logo of the accepted sequences of the simulation (top), the uniref90 logo (bottom). The middle logo reports the distribution of the amino acids that have been systematically rejected during the simulations. For sake of clarity, the first residues of the targets are numbered as 1.
Right: Conserved (Neq < 1.1) residues as obtained from the simulations. Cyan residues depict residues that are still conserved in the subset of sequences using JS + 0.02, when possible (i.e. except for 3e21)

RKKLEELERDLRKLKKKIKKLEEDNPW