



HAL
open science

Investigating the Formation of Structural Elements in Proteins Using Local Sequence-Dependent Information and a Heuristic Search Algorithm

Alejandro N Estaña, Malik Ghallab, Pau Bernadó, Juan Cortés

► **To cite this version:**

Alejandro N Estaña, Malik Ghallab, Pau Bernadó, Juan Cortés. Investigating the Formation of Structural Elements in Proteins Using Local Sequence-Dependent Information and a Heuristic Search Algorithm. *Molecules*, 2019, 24 (6), pp.1150. 10.3390/molecules24061150 . hal-02080026

HAL Id: hal-02080026

<https://laas.hal.science/hal-02080026v1>


Submitted on 26 Mar 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Article

Investigating the formation of structural elements in proteins using local sequence-dependent information and a heuristic search algorithm

Alejandro Estaña ^{1,2}, Malik Ghallab ¹, Pau Bernadó ² and Juan Cortés ¹ *

¹ LAAS-CNRS, Université de Toulouse, CNRS, Toulouse, France

² Centre de Biochimie Structurale. INSERM, CNRS, Université de Montpellier, France

* Correspondence: juan.cortes@laas.fr; Tel.: +33-561336345

Version March 26, 2019 submitted to *Molecules*

Abstract: Structural elements inserted in proteins are essential to define folding/unfolding mechanisms and partner recognition events governing signaling processes in living organisms. Here, we present an original approach to model the folding mechanism of these structural elements. Our approach is based on the exploitation of local, sequence-dependent structural information encoded in a database of three-residue fragments extracted from a large set of high-resolution experimentally determined protein structures. The computation of conformational transitions leading to the formation of the structural elements is formulated as a discrete path search problem using this database. To solve this problem, we propose a heuristically-guided depth-first search algorithm. The domain-dependent heuristic function aims at minimizing the length of the path in terms of angular distances, while maximizing the local density of the intermediate states, which is related to their probability of existence. We have applied the strategy to two small synthetic polypeptides mimicking two common structural motifs in proteins. The folding mechanisms extracted are very similar to those obtained when using traditional, computationally expensive approaches. These results show that the proposed approach, thanks to its simplicity and computational efficiency, is a promising research direction.

Keywords: proteins; structural elements; conformational transitions; structural database; heuristic search algorithms

1. Introduction and related work

Proteins are biomacromolecules that perform essential functions in living organisms. They are composed of chains of amino acid residues¹ (also called polypeptide chains) that, in most of the cases, fold into functional three-dimensional structures. The amino acid sequence determines the three-dimensional structure and its stability. The sequence also determines the frequency and the transition rate between unfolded and folded states. Understanding the mechanisms of protein folding and unfolding as a function of the amino acid sequence is of paramount importance, giving their relevance in biological processes [1]. Furthermore, numerous diseases are related to the inability of proteins to fold correctly or to form insoluble amyloidogenic aggregates due to mutations or metabolic deregulation [2,3].

Intensive research efforts over several decades, using both experimental and computational approaches, have yielded important bricks of knowledge on the underlying mechanisms of protein

¹ In the following, we will use the word *residue* to refer to an *amino acid residue*.

30 folding, unfolding and other conformational transitions [4–9]. Nevertheless, we still lack of a complete
31 understanding of these mechanisms. Some theories about protein folding give more importance to
32 interactions between the protein side-chains, whereas others consider that the propensity of protein
33 backbone fragments to form secondary structural elements, such as α -helices, β -sheets and turns, is the
34 most important mechanism for protein folding. Note that, in addition to their importance in the overall
35 protein folding process, small structural elements may play key roles in molecular recognition in
36 intrinsically disordered proteins (IDPs). These elements, the so called Molecular Recognition Elements
37 (MOREs), are partially folded fragments inserted into otherwise disordered chains [10,11]. MOREs
38 recognize with high specificity their globular partners while displaying a moderate affinity, explaining
39 their fundamental role in signalling, metabolic regulation and homeostasis [12].

40 We believe that local, sequence-dependent structural preferences are essential to drive the
41 formation of structural elements, while other phenomena such as hydrophobic effects or electrostatic
42 forces help stabilizing the overall structure. Following this hypothesis, we propose a theoretical
43 approach to compute conformational transitions using local structural information extracted from
44 experimental data. Interactions between distant residues are (explicitly) neglected for the exploration of
45 transition paths, with the exception of collisions that would lead to unrealistic conformations. However,
46 as further explained below, non-bonded interactions associated with local structural preferences are
47 implicitly considered, and can be propagated along the sequence thanks to the application of constraints
48 within the path search algorithm.

49 Information extracted from experimentally determined protein structures is frequently used in
50 computational biology. The usual usage is the prediction of the conformation of the protein side-chains,
51 using the so-called *rotamer* libraries [13], which encode the most frequent values of the side-chain
52 dihedral angles for each amino acid type. The construction of protein backbone structural databases is
53 less straightforward than for the side-chains as it requires to subdivide proteins into fragments. The
54 length of the fragments and considerations regarding the amino acid sequence may depend on the
55 specific application. Statistics about the most frequent values of the backbone dihedral angles of amino
56 acid types have been frequently used to explore the conformational sampling of highly-flexible proteins
57 or regions [14–16]. However, such minimalistic single-residue fragments neglect the effects exerted by
58 neighboring residues. Structural libraries involving larger fragments (usually, from 3 to 14 residues)
59 have been shown to be powerful tools for the prediction of probable (stable) conformations of globular
60 proteins and peptides [17–20]. Fragment libraries can also be used to investigate conformational
61 transitions in proteins. In a recent work, local moves using a fragment library were combined with
62 other types of structural perturbations to compute transitions between several folded states of a
63 protein [21]. Since the aforementioned fragment libraries were mainly conceived for protein structure
64 prediction, they are focused on the most probable conformations of small and medium-sized fragments.
65 As a consequence, they are not exhaustive enough for the study of conformational transitions. This
66 limitation is more evident when the length of the fragments increases. Fragments involving three
67 consecutive amino acid residues (called *tripeptides* from now on) represent a good trade-off between
68 sequence-dependent structural preferences and exhaustiveness. Indeed, tripeptides contain relevant
69 structural information [22] and are sufficiently small to capture the conformational variability of the 20
70 proteinogenic amino acids in their sequence context. Recently, we showed that an extensive database
71 of tripeptides allows to accurately sample the conformational variability of IDPs [23]. Here, we exploit
72 the combination of this type of local structural information with a path search algorithm to compute
73 conformational transitions in small proteins and protein fragments corresponding to relevant structural
74 elements.

75 A protein cannot exhaustively explore its huge conformational space to seek transition pathways.
76 This idea, referred to as the Levinthal's paradox [24,25], is widely accepted. Indeed, a protein performs
77 some search process to find the most efficient folding and transition pathways. We can say that the
78 protein follows a powerful *heuristic* to avoid exploring an astronomically large number of possible
79 pathways. This heuristic is not well understood yet, but, as mentioned above, we believe that local

80 sequence-dependent structural preferences play an important role in it. Our contribution investigates
81 this open question, and proposes a simple, heuristically-guided search algorithm, inspired from
82 Artificial Intelligence (AI) and Robotics, to compute conformational transitions. AI and Robotics
83 planning representations and techniques have been found valuable for solving several computational
84 biology problems [26–28]. This paper illustrates through an original approach their effectiveness in
85 modeling folding mechanisms of structural elements in proteins.

86 The approach presented herein is very different from the ones in related work. First, the structural
87 information is collected and used in a different way, and secondly, the algorithmic approach is
88 totally different. Concretely, we use a heuristically guided depth-first algorithm, adapted from
89 search techniques in constraint satisfaction problems over finite sets (CSP) and in automated task
90 planning [29]. In our case, the state variables are the protein tripeptides, which range over finite
91 sets of conformations extracted from a global database. The equivalent of an *action* is a constrained
92 local change in a state variable. The algorithm relies on *adjacency graphs* of the state variables [30],
93 which are computed at preprocessing time and are essential for efficiently testing the feasibility of
94 transitions and for calculating the heuristic, which is based on statistical physics considerations. Our
95 approach tends to favor paths going through high-density states, which are the most probable ones
96 according to experimental observations recorded in the structural database. In other words, if we
97 assume that the probability of the observed states for each tripeptide follows a Boltzmann distribution,
98 we can say that the path search tends to follow the valleys of the free-energy landscape [31]. The
99 search process also gives priority to short paths, which should correspond to faster transitions. The
100 structural preferences for a tripeptide (*i.e.* at the state variable level) tend to be propagated along
101 the sequence due to constraints imposed on the bond angles in the state transition validation, which
102 reinforces neighbor-dependent structural preferences encoded in the database (see Section S2 in
103 supplementary material for details). Thus, the path search process incorporates in an implicit way
104 non-local interactions along the sequence such as backbone hydrogen bonds in α -helices.

105 We applied our approach to two synthetic mini-proteins, Chignolin [32] and DS119 [33], which
106 were particularly designed to fold into well-defined structural motifs present in natural proteins.
107 These two molecules have been investigated in recent years using different methods [34,35]. The
108 results reported in this paper are consistent with respect to those described in related literature, and
109 already show the interest of the proposed approach, which is extremely fast when compared with
110 currently-used computational methods based on molecular dynamics (MD) simulations [36]. Indeed,
111 MD simulations of large-amplitude protein motions require *ad-hoc* computer architectures [8] or
112 massively-distributed computing [37]. The efficiency of our approach allows to widely investigate,
113 with modest computational resources, the effect of mutations on protein folding and unfolding, or on
114 other functionally-important conformational transitions.

115 2. Results and Discussion

116 This section presents results obtained with the proposed approach for the analysis of the folding
117 process of two synthetic mini-proteins, Chignolin and DS119, which were designed to fold into
118 structural motifs present in natural proteins. First, we present a deeper analysis for Chignolin and
119 two point mutants. Then, results presented for DS119 show that the approach is general and can be
120 applied to the investigation of different structural elements.

121 2.1. Chignolin

122 Chignolin is a synthetic polypeptide consisting of 10 residues [32]. Despite its small size, Chignolin
123 behaves as a macromolecular protein from structural and thermodynamic points of view: it folds
124 into a well-defined structure in water, and shows a cooperative thermal transition between unfolded
125 and folded states [39]. The folded conformation of Chignolin corresponds to a β -hairpin motif, which
126 can be found in many natural proteins (Figure 1.d). Therefore, elucidating the folding mechanism of
127 Chignolin helps to understand the folding patterns of more complex proteins. This has motivated

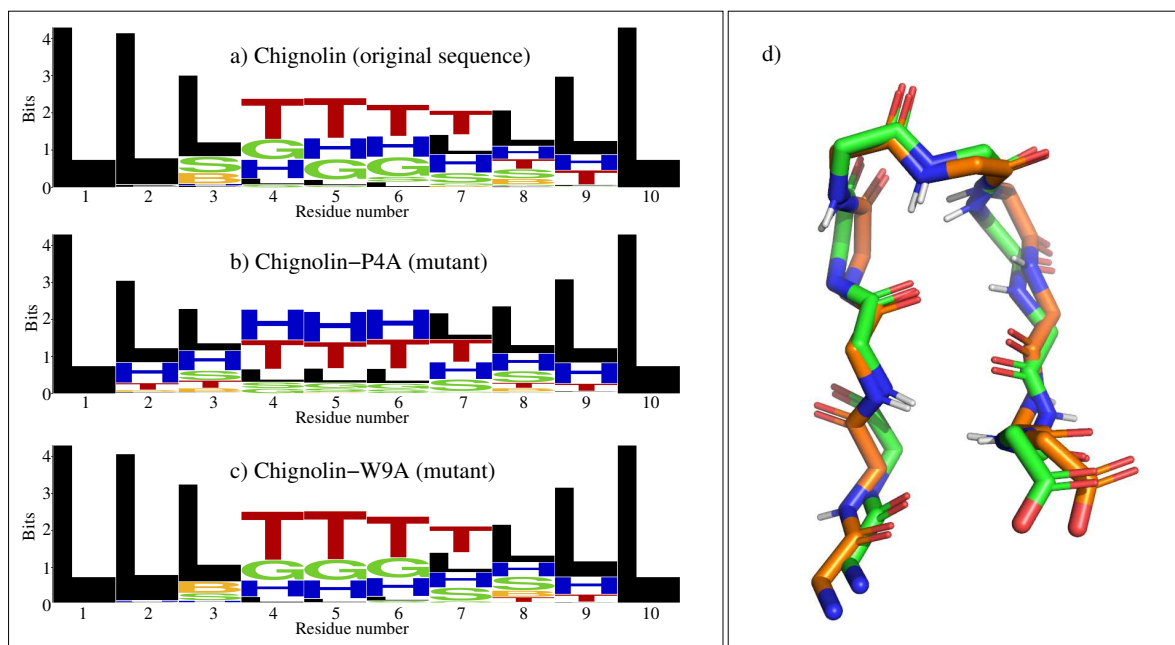


Figure 1. The left side panel represents the structural propensities at the residue level observed from a set of 1,000 conformations randomly generated from the structural database. Each plot displays the DSSP structural classes using the WebLogo format for (a) Chignolin, and two mutants: (b) Chignolin-P4A, and (c) Chignolin-W9A. (d) Structural representation of Chignolin: superposition of an experimentally determined structure (with carbon atoms in green) and the closest one in the set of 1,000 sampled conformations (with carbon atoms in orange). For clarity, only the protein backbone is represented, using PyMOL [38].

128 several experimental and computational studies on Chignolin in recent years. Here, we compare our
 129 results with those of Enemark et al. [34], which are based on extensive molecular dynamics simulations,
 130 and provide detailed information at the single-residue level.

131 Table 1 provides the number of conformations (*i.e.*, number of values of state variables) contained
 132 in our database for the eight overlapping tripeptides composing Chignolin. The search space size
 133 is upper-bounded by $\prod_i |D_i| \approx 4 \times 10^{23}$, which is huge when compared to the extremely focused
 134 explorations performed by our algorithm. Thanks to the search guidance of its heuristics, we observed
 135 a manageable complexity growth, as explained in Section 3.3 and in the supplementary material.

136 In a first experiment, we assessed the ability to obtain realistic conformations of Chignolin using
 137 the structural information encoded in our tripeptide database. We generated an ensemble of 1,000
 138 Chignolin states by randomly sampling values of the state variables one by one, in an incremental
 139 manner, enforcing the consistency with neighbor state variables, and rejecting those leading to collisions
 140 between atoms. Interestingly, several states in this relatively small ensemble are close to the folded

Tripeptide sequence	Nb conformations
Gly-Tyr-Asp	994
Tyr-Asp-Pro	710
Asp-Pro-Glu	1541
Pro-Glu-Thr	1030
Glu-Thr-Gly	1446
Thr-Gly-Thr	1779
Gly-Thr-Trp	545
Thr-Trp-Gly	240

Table 1. Number of conformations (*i.e.*, number of values of state variables) for the eight overlapping tripeptides composing Chignolin.

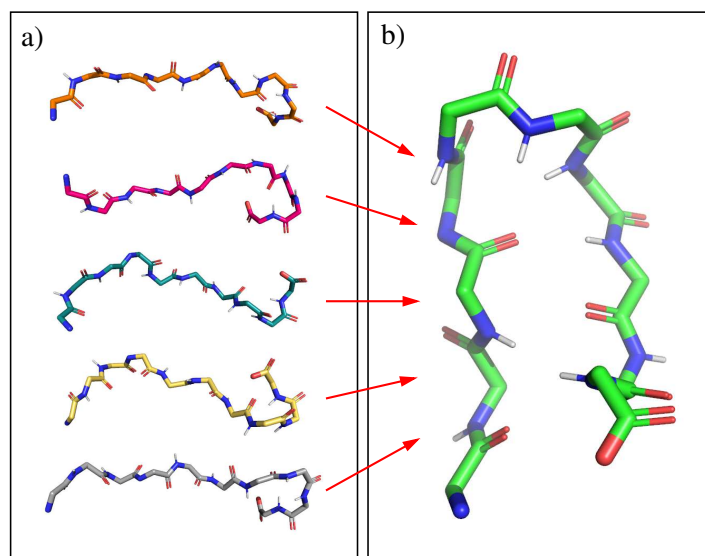


Figure 2. Structural representation of Chignolin. (a) A set of extended conformations involving the initial turn at the C-terminal side. (b) Folded conformation. Only the protein backbone is represented, using PyMOL [38].

141 conformation of Chignolin [32]. Indeed, 240 over the 1,000 sampled states have an angular RMSD
142 distance to the folded conformation below 0.5 radian, the closest one being around 0.2 radians (see
143 Figure 1.d). This confirms that the most important regions of the conformational space can be sampled
144 by building states from the tripeptide database.

145 In order to better characterize the conformational ensemble, secondary structure types for each
146 state were identified at the single residue level using DSSP [40]. DSSP distinguishes eight types of
147 structural classes, labeled with a letter: H for α -helix, B for β -bridge, E for strand, G for helix-3, I for
148 helix-5, T for turn, S for bend, and "blank" (here labeled as L) for coil/loop. We used the WebLogo
149 tool [41] to display the structural propensities in the ensemble. WebLogo is usually applied to analyze
150 results of multiple sequence alignment, but it can be used in a different context, as we did. Each
151 logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of
152 the stack indicates the conservation of the DSSP structural class at that position, while the height of
153 symbols within the stack indicates the relative frequency of each class at that position. The results in
154 Figure 1.a clearly show the propensity of the central residues to adopt a turn conformation. The rest of
155 the molecule tends to be more extended, although turns are also formed in the C-terminal region. As
156 discussed in detail below, these turns in residues 8 and 9 play a key role in the folding mechanism of
157 Chignolin. Conversely, turns are not observed in the N-terminal side. These observations are fully
158 consistent with the original study [34], and show that the states sampled using the tripeptide database
159 are structurally relevant.

160 We repeated the experiment for two mutants of Chignolin: Chignolin-P4A (Pro4 replaced by
161 Ala) and Chignolin-W9A (Trp9 replaced by Ala). Figure 1.b shows that, for Chignolin-P4A, the turn
162 propensity slightly decreases in the central region, and that it increases in the N-terminal side. For
163 Chignolin-W9A, Figure 1.c shows that the propensity to form turns in the central region is similar to
164 that of the native Chignolin molecule. However, it decreases in the C-terminal region, which may have
165 consequences for the efficiency of the folding process. Overall, these observations are very similar to the
166 results reported in [34], which use computationally expensive molecular dynamics simulations; they
167 show the strong influence of single modifications in the sequence on the conformational preferences of
168 the molecule, and that our approach captures these perturbations.

169 It has been suggested that the turn in Chignolin originates in the C-terminal region, and then
170 propagates along the chain until reaching the middle residues [34]. This has been called the "roll-up"
171 mechanism. To investigate this mechanism, we selected (among the set of 1,000 conformations) 15

172 conformations of Chignolin presenting turns in residues 8 and 9, and with a relatively extended
173 conformation for the rest of the chain. These conformations were used as initial states to compute
174 folding paths, as illustrated in Figure 2. The goal state was defined as the closest conformation to the
175 experimental structure of Chignolin built from values contained in the tripeptide database. These
176 two conformations are very similar, with an angular RMSD of 0.1 radians. The HDFS algorithm was
177 applied 20 times to solve each of these 15 problems (i.e. 300 runs in total). On average, the algorithm
178 required around 10 seconds to find folding pathways (1st column in Table 2), which is extremely fast.²
179 Intermediate states along each path were selected with a step-size corresponding to 1/10th of its total
180 length. The left side panel in Figure 3 shows the structural propensities at the residue level for these
181 intermediate states. It can be observed that the turns in the C-terminal residues tend to disappear,
182 while these structural elements appear in the middle residues. This "roll-up" mechanism can also be
183 observed in the right side panel in Figure 3, which represents several intermediate states along one
184 of the folding paths. The first frames (starting from the top) show that the curvature of the molecule,
185 initially involving residues 8 and 9, rapidly propagates to residues 6 and 7. Then, residues 5 and 4 also
186 bend successively, and the molecule tends to form a hairpin-like structure. Finally, the two terminal
187 parts adopt a relatively extended conformation.

188 As explained in related work [39], the folding process of Chignolin may lead to misfolded states,
189 which are characterized by interactions between residue pairs Tyr2-Thr8 and Asp3-Gly7, rather than
190 Tyr2-Trp9 and Asp3-Thr8, as in the correctly folded structure. We generated a representative model of
191 a misfolded state, and we computed conformational transitions from initial conformations with the
192 C-terminal turn (C-ter T) to this state. We also computed transitions from fully-extended conformations
193 to folded and misfolded states. The results are summarized in the top part of Table 2. This table
194 provides average values (over 300 runs) for: the computing time required by the HDFS algorithm to find
195 a path; the number of recursions and backtracks; the number of steps in the solution path; the length
196 of the solution path, computed as the sum of the lengths associated to edges in the adjacency graphs;
197 the density of the solution paths, computed as the average of the density of all the state variables
198 along the path. The most meaningful numbers in this table are those associated with the density, since
199 they reflect the probability of existence of each pathway. Compared to the extended→folded pathway,
200 the C-ter T→folded pathway goes across more dense and probable regions. This may explain why
201 Chignolin efficiently folds from unfolded states involving this structural feature. In both cases, starting
202 from C-ter T or fully-extended states, the transitions to misfolded states seem to be much less probable.
203 This may explain why the misfolded state of Chignolin is much less frequently observed than the
204 correctly folded state [42].

205 We repeated the experiments for the mutant Chignolin-W9A. The results are summarized in
206 the bottom part of Table 2. As mentioned above, the set of conformations generated for these two
207 molecules look structurally similar (see Figure 1 and the associated comments). The figures in Table 2
208 also show a very similar behavior of the HDFS algorithm when computing transition paths for this
209 mutant compared to the original Chignolin. Interestingly, the main difference is observed for the
210 density of the path extended→misfolded. This path is significantly more favorable in the case of
211 the mutant. Our results complement the study of Enemark et al. [34], which suggested that the
212 replacement of Trp9 by Ala facilitates a "roll-back" mechanism, acting against the "roll-up" mechanism,
213 hindering the formation of the native turn in the middle residues. We show another possible effect of
214 this mutation, favoring the formation of misfolded states in competition with the native structure.

215 2.2. DS119

216 DS119 is another synthetic polypeptide, consisting of 36 amino acid residues, which was designed
217 to fold into a $\beta\alpha\beta$ motif [33] (see last frame in Figure 4). The folding process of DS119 has been studied

² CPU time was measured with an Intel® Core™ i7 processor at 2.8 GHz, using a single core.

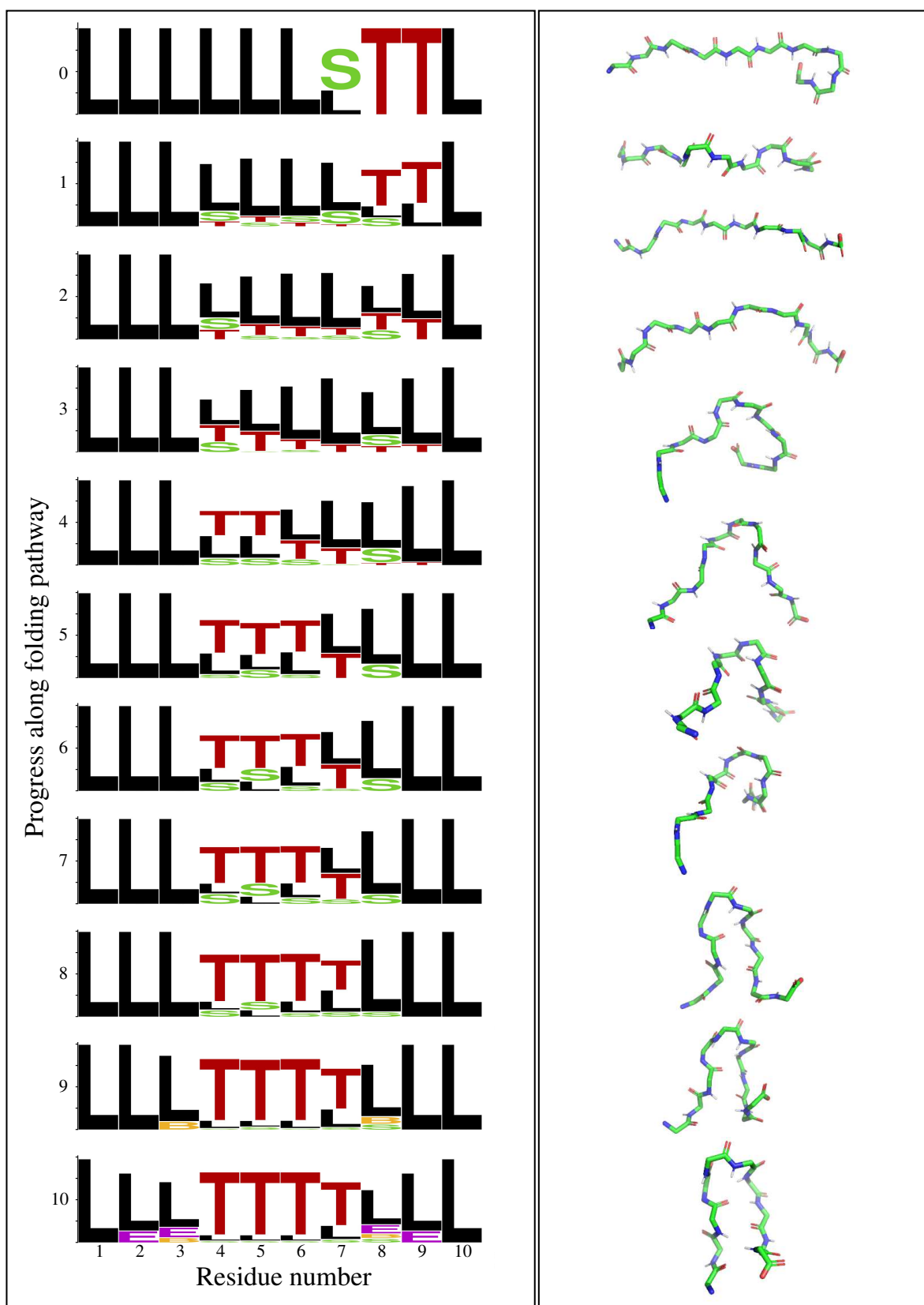


Figure 3. The left side panel represents the evolution of the structural propensities at the residue level along Chignolin folding pathway (see Figure 1 and the associated comments for additional explanations about this representation). The right side panel shows some intermediate states along one of the computed folding paths. Only the protein backbone is represented, using PyMOL [38].

	chignolin (original sequence)			
	C-ter T→folded	C-ter T→misfolded	extended→folded	extended→misfolded
CPU time (s)	11.1	8.7	5.2	3.5
# states	5416.4	2587.7	2800.1	849.5
# backtracks	234.6	136.6	124.6	39.2
Path length (# steps)	133.8	54.5	106.3	48.7
Path distance (rad)	8.8	5.1	6.0	7.0
Path density	31.9	5.5	23.3	4.5

	chignolin-W9A (mutant)			
	C-ter T→folded	C-ter T→misfolded	extended→folded	extended→misfolded
CPU time (s)	12.2	8.8	5.6	5.1
# states	4943.6	2567.8	2317.0	2946.0
# backtracks	219.6	139.0	101.3	126.3
Path length (# steps)	140.3	51.3	103.0	125.7
Path distance (rad)	8.2	9.0	5.8	8.2
Path density	31.2	4.6	23.4	23.8

Table 2. Performance indicators of the HDFS algorithm to compute different conformational transitions of Chignolin (top) and the mutant Chignolin-W9A (bottom). CPU time was measured with an Intel® Core™ i7 processor at 2.8 GHz, using a single core.

218 using molecular dynamics simulations [35]. This previous work showed that the N-terminal side of
 219 the central helix tends to form very quickly. Then, the C-terminal side of the helix starts to form, and
 220 the full helix is finally stabilized. The relatively extended fragments at the two ends of the molecule
 221 tend to come together at the end of the folding process.

222 To investigate the folding mechanism of DS119, we applied a similar procedure as for Chignolin.
 223 In this case, we selected 15 relatively extended conformations, involving only the L DSSP structural
 224 class for all the residues, from a set of 1,000 randomly generated conformations using the tripeptide
 225 database. These conformations were used as initial states for the HDFS algorithm. As final state, we
 226 used the closest conformation to the experimentally solved structure of DS119 (PDB ID: 2KI0) built
 227 from values contained in the tripeptide database. These two conformations are very similar, with an
 228 angular RMSD of 0.06 rad. The algorithm was applied 20 times to solve each of these 15 problems (i.e.
 229 300 runs in total).

230 Figure 4 illustrates the results obtained by the HDFS algorithm. The left side panel shows the
 231 evolution of the structural propensities along the folding path, using logos based on DSSP classes.
 232 The right side panel represents several intermediate states along one of the solution paths. For clarity
 233 purposes, only a few intermediate states are shown using a "cartoon" representation of the backbone,
 234 where the helical fragments can be easily identified. It can be observed that, starting from an extended
 235 conformation, the protein backbone rapidly starts to bend around residues 12-13. Recall that the S letter,
 236 for "bend", corresponds to a highly curved protein backbone. Hydrogen bonds required to stabilize the
 237 helical conformation are not yet identified by DSSP at this early stage. Next, curved/helical fragments
 238 start to appear in all central residues (from residue 14 until residue 27), as well as in three residues
 239 in the N-terminal side (residues 3-5). The central helix continues to fold, and it is almost completely
 240 formed at the 7th intermediate frame. In the final part of the path, the extended fragments at the
 241 two ends get close to each other, nearly forming a parallel β -sheet. This description of the folding
 242 process strongly resembles the one reported in the literature, based on computationally-expensive
 243 simulations [35].

244 Table 3 presents numbers (averaged over the 300 runs) concerning the performance of the HDFS
 245 algorithm to compute folding paths of DS119. The required CPU time (and the number of recursions)
 246 is only about three times the one required to compute folding paths for Chignolin. This shows that,
 247 despite the theoretical (worst-case) exponential complexity, in practice, the computing time scales
 248 approximately linearly with the number of variables. This tendency has been confirmed by preliminary

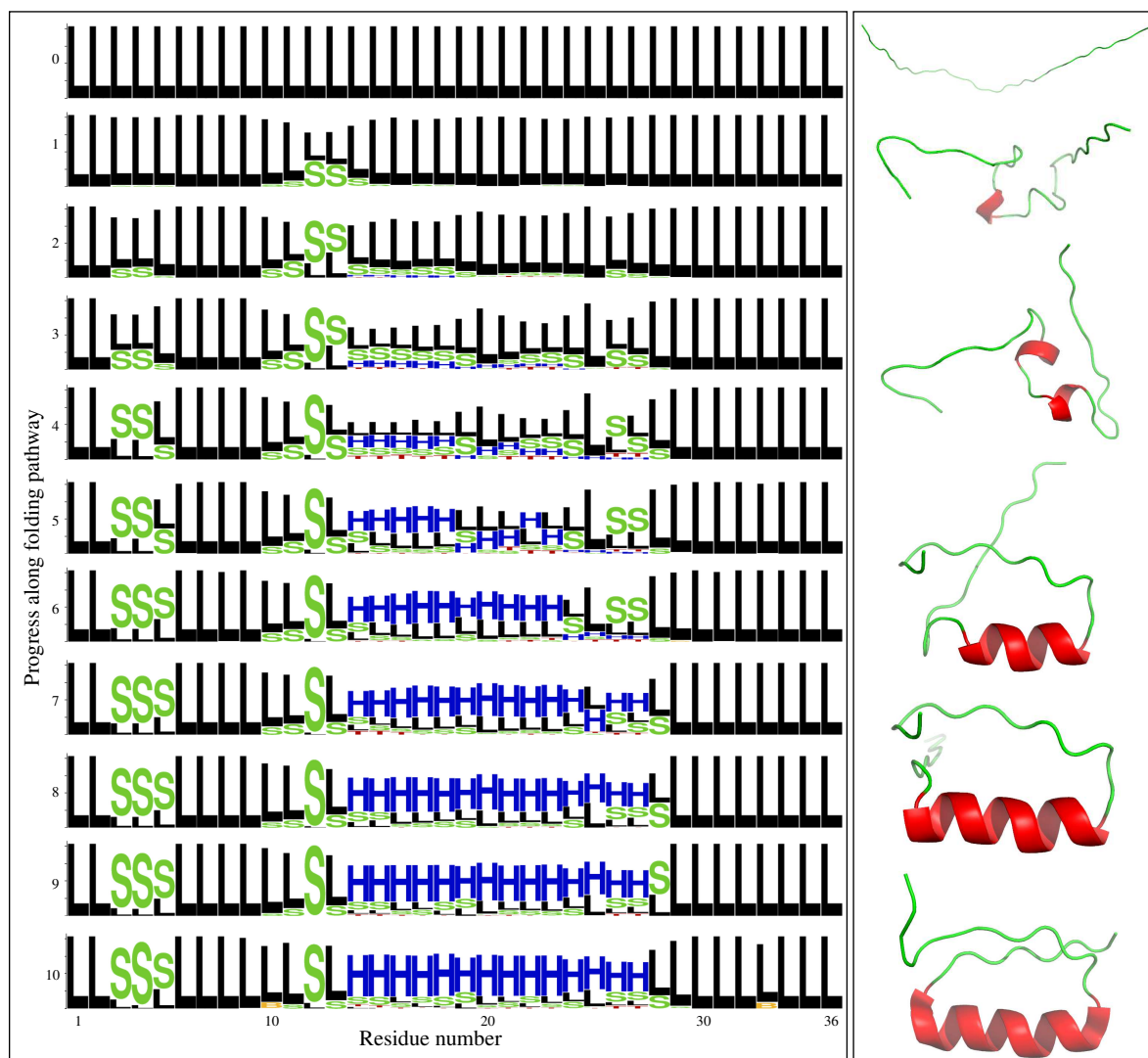


Figure 4. The left side panel represents the evolution of the structural propensities at the residue level along DS119 folding pathway. The right side panel shows some intermediate states along one of the computed folding paths. The "cartoon" representation clearly shows the formation of the helix. PyMOL [38] was used for the structural visualization.

249 tests for larger molecules (not presented in this paper). Once again, we insist that computing time
 250 is orders of magnitude faster than traditional molecular dynamics simulation methods. The higher
 251 density of the path compared to Chignolin can be explained by the larger number of conformations
 252 for some of the tripeptides, particularly for those composing the middle helix. Table 4 provides the
 253 numbers of conformations (*i.e.*, number of values of state variables) contained in our database for the
 254 34 overlapping tripeptides composing DS119.

255 3. Materials and Methods

256 The proposed approach relies on a large database of protein structures, represented as sequences of
 257 partially overlapping tripeptides. As stressed above, tripeptides are the minimal structurally-relevant
 258 units in proteins. The problem is formalized as a search in a space of tripeptide conformations for
 259 a feasible path from an initial state to a target state of a protein. The state variables correspond to
 260 tripeptides; their values are the conformations of tripeptides actually observed and recorded in the
 261 database. A state variable in the sequence describing a protein shares its first two residues with its
 262 predecessor and its last two with its successor state variables in the sequence (see Figure 6). A transition

	DS119 : extended → folded
CPU time (s)	25.2
# states	70558.2
# backtracks	8210.4
Path length (# steps)	158.2
Path distance (rad)	11.3
Path density	124.4

Table 3. Performance indicators of the HDFS algorithm on DS119.

Tripeptide sequence	Nb conformations	Tripeptide sequence	Nb conformations
Gly-Ser-Gly	3727	Lys-Lys-Leu	2286
Ser-Gly-Gln	1118	Lys-Leu-Lys	1996
Gly-Gln-Val	1294	Leu-Lys-Glu	3100
Gln-Val-Arg	607	Leu-Glu-Glu	1631
Val-Arg-Thr	970	Glu-Glu-Ala	2591
Arg-Thr-Ile	757	Glu-Ala-Lys	1514
Thr-Ile-Trp	181	Ala-Lys-Lys	1714
Ile-Trp-Val	180	Lys-Lys-Ala	1629
Trp-Val-Gly	279	Lys-Ala-Asn	1009
Val-Gly-Gly	2443	Ala-Asn-Ile	1010
Gly-Gly-Thr	2510	Asn-Ile-Arg	647
Gly-Thr-Pro	1428	Ile-Arg-Val	998
Thr-Pro-Glu	1738	Arg-Val-Thr	1351
Pro-Glu-Glu	1752	Val-Thr-Phe	888
Glu-Glu-Leu	3433	Thr-Phe-Trp	151
Glu-Leu-Lys	2378	Phe-Trp-Gly	192
Leu-Lys-Lys	2528	Trp-Gly-Asp	257

Table 4. Number of conformations (i.e. number of values of state variables) for the eight overlapping tripeptides composing DS119.

263 between two values of a state variable is feasible if it meets a consistency constraint with respect to
 264 the predecessor and successor state variables, and if the corresponding conformation of the protein
 265 is collision free. The search algorithm seeks a feasible path using a heuristically-guided depth-first
 266 search schema. The heuristic function is a weighted sum of the distance between two conformations,
 267 an estimate of the distance to the target and a density term to advantage energetically favorable states.

268 We present next the construction of the structural database, then the statement of the
 269 conformational transition problem as a discrete path search problem; we detail the proposed algorithm
 270 and the heuristics used to solve this problem.

271 3.1. Structural database

272 A tripeptide database was built from a large set of high-resolution experimentally-determined
 273 protein structures. We generated this set from SCOPe (release 2.06) [43], avoiding redundancies
 274 in protein sequence and structure. The total number of tripeptides extracted from these protein
 275 structures is 5,630,271. The tripeptides are characterized by their amino acid sequence. Since natural
 276 proteins involve 20 types of amino acids, the total number of tripeptides is $20^3 = 8,000$. The database
 277 construction process is illustrated in Figure 5.a-c. All the 8,000 tripeptides appear in our database. The
 278 number of their instances ranges between 9 for the less frequent tripeptide (Cys-Cys-Trp) to 4,512 for
 279 the most frequent one (Ala-Ala-Ala).³ The average number of instances is about 688.

280 It is important to highlight that the database includes fragments extracted from coil regions, which
 281 have been shown to be useful elements to model unfolded or disordered proteins [23,44]. Therefore,

³ These standard three-letter abbreviations stand respectively for Cysteine, Tryptophan and Alanine.

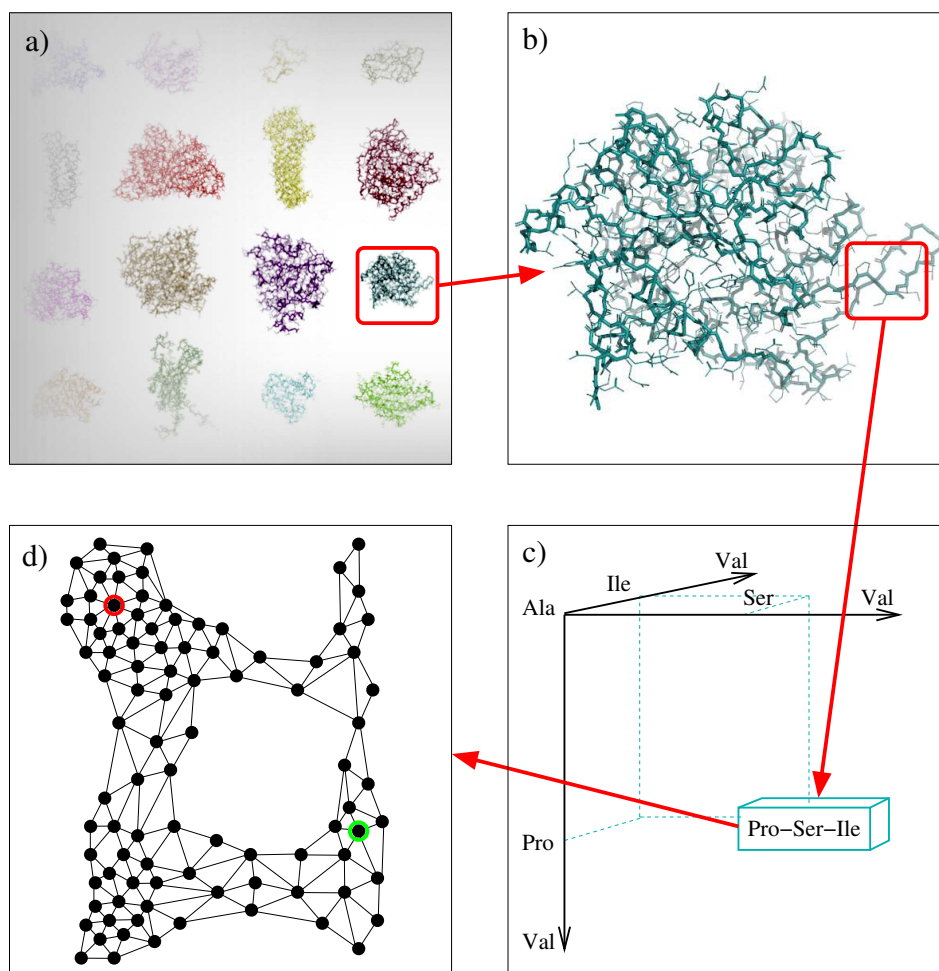


Figure 5. Construction of the tripeptide database: (a) A non-redundant set of experimentally-determined protein structures is used as input. (b) For each protein, fragments of three consecutive residues (called tripeptides) are analyzed. (c) The structural information is stored in a database containing one record for each tripeptide (8,000 in total). (d) For each tripeptide, the conformations recorded in the database are related with a proximity criterion and structured into an adjacency graph (the figure shows a simplified representation of this graph for tripeptide Pro-Ser-Ile).

282 we assume that the structural information encoded in the database is not limited to folded states, and
 283 that it can be useful to investigate folding processes.

284 We adopt a rigid geometry simplification [45], which assumes constant bond lengths and bond
 285 angles. Indeed, the standard deviation for the bond lengths and the bond angles in our database
 286 is two orders of magnitude smaller than their average value, and therefore, we can neglect their
 287 variation. In addition, as usually done to simplify protein modeling, we assume that the torsion angles
 288 corresponding to peptide bonds (*i.e.*, the bonds connecting consecutive residues) are constant. This is
 289 also a reasonable assumption given that this angle slightly fluctuates around a value of 0 or π radians
 290 (that is, the *cis* and *trans* conformations), with a standard deviation of around 0.1 radians. Therefore
 291 the only variables required to determine the conformation of a protein backbone correspond to the ϕ
 292 and ψ dihedral angles of each amino acid residue. The database stores these angular values for each
 293 tripeptide extracted from the ensemble of protein structures (*i.e.*, 6 angles for each tripeptide). Figure 6
 294 represents a protein fragment involving 5 residues, from which 3 tripeptides are extracted. The angles
 295 defining the conformation of each residue are represented on the corresponding bonds.

296 In this work, we do not consider an all-atom model of the protein side-chains, but a simplified
 297 model involving a pseudo-atom for each side-chain. The pseudo-atom is centered at the position of

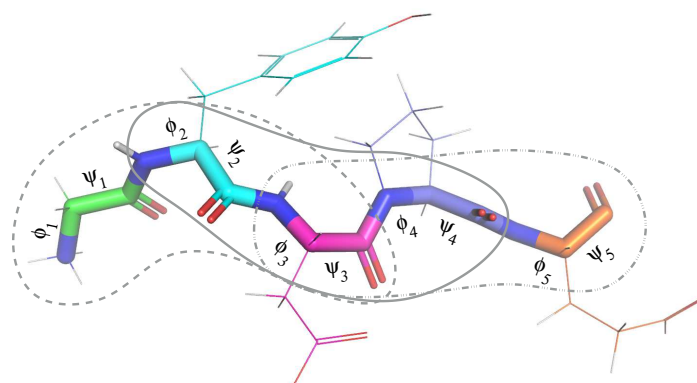


Figure 6. Illustration of a protein fragment involving 5 residues. Each residue is represented using a different colors for the carbon atoms. The backbone is represented using thicker lines. Considering constant bond lengths, bond angles and peptide bond torsions, the protein backbone conformation can be defined from a pair of angles (ϕ and ψ) for each residue. The gray lines indicate the 3 overlapping tripeptides composing this 5-residue fragment.

298 the β -carbon atom, and the size depends on the amino acid type, as originally proposed by Levitt [46].
 299 Therefore, no additional variables are required to represent the side-chains.

300 Let \mathcal{X} be the set of all 8,000 tripeptides. An element $x_i \in \mathcal{X}$ is a state variable in our representation.
 301 Let D_i be the set of all the conformations of x_i recorded in our database. The conformation of x_i is
 302 characterized by the six backbone dihedral angles of the three residues in the tripeptide, denoted
 303 $\phi_{i,j}$ and $\psi_{i,j}$, for $1 \leq j \leq 3$. Although a conformation is characterized by an angular vector of 6
 304 real numbers, for the purpose of our search algorithm over biologically observed conformations, we
 305 consider that the range of each state variable x_i is the finite set D_i of the recorded conformations in the
 306 database. We write $x_i = v_i$ for some $v_i \in D_i$.

The distance $d(v_i, v'_i)$ between two values v_i and v'_i is defined as the angular root-mean-square deviation (RMSD) between the two corresponding angular vectors. More precisely:

$$d(v_i, v'_i) = \sqrt{1/6 \sum_{j=1}^3 ((\phi_{i,j} - \phi'_{i,j})^2 + (\psi_{i,j} - \psi'_{i,j})^2)}$$

307 We also define the central distance $d_c(v_i, v'_i)$ with an identical formula for $j = 2$ solely, *i.e.*, restricted to
 308 the central amino acid residue of x_i . The idea is to compute a feasible path in the conformations of a
 309 protein as a sequence of elementary transitions focused on the central residue of each tripeptide.

310 These distances d and d_c allows us to structure the finite range D_i of each state variable as an
 311 *adjacency graph*, as illustrated in Figure 5.d. Its vertices are the elements in D_i . There is an edge
 312 (v_i, v'_i) when $d_c(v_i, v'_i) < \theta$ and $d(v_i, v'_i) < \theta + \zeta$, where θ is a variable adjacency threshold and ζ
 313 is a small constant tolerance margin. The adjacency threshold θ represents a tradeoff between a
 314 fully connected graph (no transition constraints between conformations) and an unconnected one
 315 (unreachable conformations), both cases being unrealistic. We set the threshold such that the adjacency
 316 graph of each tripeptide has a single connected component with moderate edge connectivity. This
 317 threshold θ is slightly different for different tripeptides, with an average value around 1.0 radian. The
 318 value of ζ was set to 0.35 radians in all the cases.

The vertices are also characterized by a density function defined as follows:

$$\rho(v_i) = 1 + |\{v'_i \mid v'_i \text{ connected to } v_i \text{ and } d(v_i, v'_i) < \zeta\}|.$$

319 The threshold ζ has to be smaller than the adjacency threshold θ . Here, we set $\zeta = 0.2$ radians for all the
 320 tripeptides. The density ρ is related to the probability of existence of the corresponding conformation

of the tripeptide. Considering basic principles in statistical physics (*i.e.*, the Boltzmann distribution), this probability depends on the energy of the state of the molecule. Thus, the most dense regions in the adjacency graph are also the most energetically-favorable ones.

3.2. Formal statement of the conformation path finding problem

A protein (or protein region) of interest is defined by a sequence of state variables $\langle x_1, \dots, x_i, \dots, x_n \rangle$, with overlaps. For example, the mini-protein Chignolin is a sequence of 10 amino acid residues: $\langle \text{Gly-Tyr-Asp-Pro-Glu-Thr-Gly-Thr-Trp-Gly} \rangle$; it is defined with 8 state variables $x_1 = \text{Gly-Tyr-Asp}$, $x_2 = \text{Tyr-Asp-Pro}$, ... $x_8 = \text{Thr-Trp-Gly}$. Hence, the state variables are not independent: a transition in a state variable may or may not be consistent with another transition in the previous or following state variables in the sequence.

For a given conformational state of the protein $s = \langle (x_1 = v_1), \dots, (x_i = v_i), \dots, (x_n = v_n) \rangle$, the overlap between consecutive state variables means that a tripeptide x_i shares its first two residues with its predecessors in the sequence and its last two with its successors; that is:

$$\phi_{i,1} = \phi_{i-1,2} = \phi_{i-2,3}, \quad \phi_{i,2} = \phi_{i-1,3} = \phi_{i+1,1}, \quad \text{and} \quad \phi_{i,3} = \phi_{i+1,2} = \phi_{i+2,1}, \quad (1)$$

and similarly for the ψ angles.

An elementary state transition with respect to x_i , from the value v_i to an adjacent value v'_i , involves a conformational change mainly in the central residue of x_i (by construction of the adjacency graph). This entails constraints on x_{i-1} and x_{i+1} with respect to their current values in state s . We express these constraints as inequalities with a tolerance margin as follows:

$$\begin{aligned} |\phi'_{i,2} - \phi_{i-1,3}| < \epsilon, \quad |\phi'_{i,2} - \phi_{i+1,1}| < \epsilon, \\ |\psi'_{i,2} - \psi_{i-1,3}| < \epsilon, \quad |\psi'_{i,2} - \psi_{i+1,1}| < \epsilon. \end{aligned} \quad (2)$$

where the angles for the last and first residues of x_{i-1} and x_{i+1} correspond to their current values v_{i-1} and v_{i+1} . These constraints can be relaxed during the search by dynamically adjusting the value of ϵ , as explained below. Here, we set initially $\epsilon = 0.35$ radians.

Definition 1 (Feasible transition). A transition in the conformation of a protein from a state s where $x_i = v_i$ to a state s' where $x_i = v'_i$ is said to be a *feasible transition* if and only if:

- (i) the values v_{i-1} and v_{i+1} meet the constraints of [Equation 2](#), and
- (ii) there are no collisions between the atoms of the protein in the state s' .

A *feasible path* is a sequence of feasible transitions.

Let $\gamma(s, (v_i \rightarrow v'_i))$ denotes the state s' corresponding to this transition when it is feasible, otherwise γ is undefined.

The conformation path finding problem can be formally stated as follows: given \mathcal{X} and the adjacency graphs of all the state variables in a protein, and given an initial state s_0 and a goal state s_g , the problem is to find a feasible path that transforms the protein conformation from s_0 into s_g , if there exists such a path.

3.3. Search algorithm

To generate a feasible path from s_0 to s_g , we rely on a heuristically-guided depth-first search in the space $\prod_i D_i$, over all state variables x_i in the protein. To ease the presentation, the algorithm is stated in the pseudo-code of [Figure 7](#) as a simple recursive nondeterministic search procedure called HDFS. The initial call is $\text{HDFS}(s_0, \langle s_0 \rangle)$. The *nondeterministic choice* (step labelled \triangleleft) is a convenient notation meaning that the algorithm makes at this point a branching decision; it explores potentially all possible options, expressed here as the set \mathcal{E} ; it stops on the first path which succeeds or it returns


```

HDFS(s, Path)
  if s = sg then return(Path · s)
   $\mathcal{E} \leftarrow \emptyset$ 
  for each state variable  $x_i$  in s do
     $\mathcal{E} \leftarrow \mathcal{E} \cup \text{Transition-Filter}(s, x_i, \textit{Path})$ 
  if  $\mathcal{E} = \emptyset$  then return(failure)
  else do
    Nondeterministically choose in  $\mathcal{E}$  a transition ( $v_i \rightarrow v'_i$ ) ◁
     $s' \leftarrow \gamma(s, (v_i \rightarrow v'_i))$ 
    HDFS(s', Path · s)

Transition-Filter(s,  $x_i$ , Path)
   $v_i \leftarrow$  value of  $x_i$  in s
   $\mathcal{A} \leftarrow$  set of values adjacent to  $v_i$  in adjacency graph  $D_i$ 
  for each  $v'_i \in \mathcal{A}$  do
    if  $\gamma(s, (v_i \rightarrow v'_i))$  is undefined or
    if it is a state already in Path
    then remove  $v'_i$  from  $\mathcal{A}$ 
  return( $\mathcal{A}$ )

```

Figure 7. Main procedure as a recursive nondeterministic best-first search. The choice (in step ◁) is guided with the heuristic *cost* function used to order the set \mathcal{A} . In the case of failure, backtracking is performed at this step to other remaining options in the set \mathcal{E} , which is computed incrementally.

353 failure if all paths fail.⁴ The deterministic implementation of HDFS makes at this step a heuristic choice
 354 over which it backtracks in case of failure; if needed, this is repeated as long as an option in \mathcal{E} remains
 355 unexplored. The heuristic driving this choice is detailed below.

356 The algorithm iterates over all tripeptides in the protein to find their feasible transitions. For a
 357 given state variable $x_i = v_i$ in *s*, procedure Transition-Filter checks the values adjacent to v_i in graph D_i .
 358 Unfeasible transitions are disregarded, as well as transitions that loop back into a circuit of the search
 359 space. The set \mathcal{E} is the union of all retained transitions ($v_i \rightarrow v'_i$) over all state variables. When \mathcal{E} is
 360 empty, then *s* is a dead end; a backtracking is performed.

361 In our more efficient and deterministic implementation of the algorithm, \mathcal{E} is computed
 362 incrementally. \mathcal{E} starts with the transitions of a single state variable, which has feasible transitions. \mathcal{E}
 363 is augmented with respect to new state variables when backtracking requires alternative options. In
 364 our current code, the ordering of the state variables in the HDFS loop is not heuristically guided. The
 365 effects of state variable ordering heuristics, such as the proximity to the goal or the average density in
 366 the adjacency graph, remain to be investigated.

367 Heuristic guidance function

For the results presented in this paper, the search is guided through the ordering in procedure Transition-Filter of the set \mathcal{A} of feasible values. \mathcal{A} is ordered with the following cost function:

$$\textit{cost}(v_i, v'_i) = d(v_i, v'_i) + w_1 \times h(v'_i, v_i^g) + w_2 / \rho(v'_i),$$

368 where *d* and ρ are the distance and density functions defined earlier, v_i^g is the value of x_i in the goal state
 369 s_g , *h* is the shortest path in the transition graph to the goal, and w_1 and w_2 are weight parameters. The
 370 first term seeks to minimize the distance between consecutive states along the path (*i.e.*, to maximize
 371 the continuity of the path). The second term is the sum of the distances of a minimal path from v'_i to the
 372 goal. The third term intends to maximize the density of the states along the path, which, as explained

⁴ The metaphor to help explain a nondeterministic specification of an algorithm is that of a machine able to multiply itself at each branching point into identical copies, each copy pursuing the search in parallel until one finds a solution or all fail.

373 earlier, are the most energetically favorable ones. The weights w_1 and w_2 permit a tuning of the three
374 components; their proper setting remains to be investigated. Here, we simply set $w_1 = w_2 = 1$. Note
375 that h is a lower bound for the remaining *cost* from v' to v^s , since a path in the transition graph, minimal
376 with respect to the distance d , relaxes the feasibility constraints of Definition 1 and cannot be longer
377 than a feasible path.

378 In order to speedup the search, a preprocessing of the adjacency graphs labels edges with their
379 distance d and computes for every vertex the shortest path to the goal as well as the density of every
380 node in each graph. This is done with a standard graph search algorithm.

381 The test of collision-free states is computed using a variant of the classical Cell Linked-List (CLL)
382 algorithm [47]. A pair of non-bonded (pseudo-)atoms is considered to be in collision if their distance
383 is less than 65% of the sum of their radii. In this work, we considered the radii values proposed by
384 Bondi [48] for the backbone atoms, and those proposed by Levitt [46] for the side-chains pseudo-atoms.

385 Note that the feasibility constraints in Equation 2 are too conservative. A more flexible definition
386 would also accept as feasible the transitions for which either the current values of x_{i-1} and x_{i+1} , or
387 some of their respectively adjacent values v'_{i-1} and v'_{i+1} , meet these constraints. In that case, the state
388 $s' = \gamma(s, (v_i \rightarrow v'_i))$ involves changes in x_i but also in its predecessor and successor state variables. The
389 cost function driving the search would naturally be extended to $cost(s, s')$ over entire states. Instead,
390 we have implemented a simpler mechanism to locally relax this constraint if the search process gets
391 blocked : if state transitions fail f consecutive times ($f = 5$ in our implementation), the tolerance
392 value ϵ is increased to 0.7 radians. ϵ is reset to 0.35 radians after a successful transition. The next
393 section shows that, even with such a simplified implementation, the proposed approach already gives
394 meaningful results.

395 Properties of HDFS

396 The algorithm is *sound*; that is, it returns a path which is feasible, in case of success. This is because
397 each transition meets Definition 1. HDFS is also *complete*; that is, it finds a feasible path if one exists
398 with respect to the transitions in the adjacency graphs of the state variables. This is the case since in
399 each search state s , \mathcal{E} is the entire set of feasible transitions over all state variables, loops are avoided,
400 and backtracking is systematic.

401 As for any backtrack search algorithm, the worst case complexity is exponential, in $O(\prod_i |D_i|)$.⁵
402 A more useful complexity model is in $O(d^b)$, where d is the depth of the search (i.e., the length of the
403 found path), and b is the branching factor. An upper bound on the branching factor is $n \times p$, where n is
404 the length of the protein and p is the maximum degree of vertices over all adjacency graphs. However,
405 thanks to the search guidance of its heuristics, we observed a manageable complexity growth. Our
406 experiments with seven proteins, ranging in length $10 \leq n \leq 67$ residues, show that b does not
407 grow with n ; it is constant and very small, about $b \simeq 1.04$. The overall search complexity has a
408 low polynomial growth in n . Furthermore, we confirmed that, as expected for a local propagation
409 mechanism, the computation time required for each search state is not a function of n , but a quite small
410 constant, of about 0.9 ms per state on a standard CPU. The Section S1 in the supplementary material
411 details this analysis as well as a discussion contrasting the scalability of our approach with that of MD
412 methods.

⁵ It is possible to compute the total size of the search space for each given problem (using Dynamic Programming and taking into account state variable dependencies); but this information is not very useful since in practice the algorithm explores a very small fraction of the search space.

4. Conclusion

Despite the simplicity of both the algorithm and the heuristic, the results presented in this paper show that the proposed approach constitutes a promising new research direction towards the identification of relevant protein folding pathways. The structural analysis of the folding mechanisms of Chignolin and DS119 are consistent with respect to descriptions provided in the literature. Note however that a more detailed and quantitative comparison between the paths obtained with other methods and trajectories obtained from MD simulations would not be very meaningful, since the aims of both methods are different: The paths provided by our algorithm are an approximation, from which interesting information about folding mechanisms can already be obtained, but that should be refined (using other methods and models) to get access to accurate information at the atomic level (as provided by MD simulations). On the other hand, our algorithm is orders of magnitude faster than atomistic MD simulations.

Overall, the results highlight the importance of local structural preferences, which are encoded in our tripeptide database. They also suggest that interactions between distant residues in the sequence, even though they can be essential for stabilization of the final fold, are less important at an earlier stage to drive the formation of structural elements.

The good results obtained with the implementation presented in this paper motivate us to continue in this research direction. Several points remain to be further investigated. One important question is about the possibility to include non-local interactions in the heuristic cost function. Although this does not seem to be necessary for structural elements or small proteins, interactions between distant residues in the sequence can be essential to study folding processes of larger molecules, or aspects related to stability. We also plan to implement and evaluate transitions over several state variables, as well as different heuristics for variable ordering. More sophisticated, tree-based search algorithms [29] can improve the quality and the diversity of the solutions, particularly for large proteins. Finally, let us mention the limitations imposed by the information contained in the structural database. Structural information is very limited in some regions of the conformational space corresponding to states of low probability, but which may be relevant for an accurate modeling of conformational transitions. With the increasing number of experimentally-determined high-resolution protein structures, we expect that more extensive and higher-quality tripeptide databases will be constructed in the future. Alternatively, these sparsely populated transition regions can be identified using our approach and subsequently explored using physics-based molecular models and (continuous) motion planning algorithms [49].

Supplementary Materials: The following are available online at <http://www.mdpi.com/1420-3049/xx/1/5/s1>: Section S1 - Scalability analysis; Section S2 - Neighbor-dependent structural preferences.

Author Contributions: A.E., M.G. and J.C. conceived the methods; A.E. implemented the methods; P.B. and J.C. conceived and designed the experiments; A.E. performed the experiments and analyzed the data; all the authors wrote the paper.

Funding: This work was supported by the European Research Council under the H2020 Programme (2014-2020) *chemREPEAT* [648030], and Labex EpiGenMed (ANR-10-LABX-12-01) awarded to P.B. The CBS is a member of the French Infrastructure for Integrated Structural Biology-FRISBI (ANR-10-INSB-05).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Vendruscolo, M.; Zurdo, J.; Macphree, C.; M Dobson, C. Protein folding and misfolding: A paradigm of self-assembly and regulation in complex biological systems. *Philosophical Transactions. Series A, Mathematical, Physical, and Engineering Sciences* **2003**, *361*, 1205–22.
2. Valastyan, J.S.; Lindquist, S. Mechanisms of protein-folding diseases at a glance. *Disease Models & Mechanisms* **2014**, *7*, 9–14.
3. Knowles, T.; Vendruscolo, M.; Dobson, C. The amyloid state and its association with protein misfolding diseases. *Structure* **2014**, *15*, 384–396.
4. Baldwin, R.L. Protein folding: Matching speed and stability. *Nature* **1994**, *369*, 183–184.

- 462 5. Wolynes, P.; Onuchic, J.; Thirumalai, D. Navigating the folding routes. *Science* **1995**, *267*, 1619–1620.
- 463 6. Dobson, C.M. Protein folding and misfolding. *Nature* **2003**, *426*, 884–890.
- 464 7. Rose, G.D.; Fleming, P.J.; Banavar, J.R.; Maritan, A. A backbone-based theory of protein folding. *Proceedings*
465 *of the National Academy of Sciences of the United States of America* **2006**, *103*, 16623–16633.
- 466 8. Lindorff-Larsen, K.; Piana, S.; Dror, R.O.; Shaw, D.E. How fast-folding proteins fold. *Science* **2011**,
467 *334*, 517–520.
- 468 9. Best, R.B. Atomistic molecular simulations of protein folding. *Current Opinion in Structural Biology* **2012**,
469 *22*, 52–61.
- 470 10. Pancsa, R.; Fuxreiter, M. Interactions via intrinsically disordered regions: What kind of motifs? *IUBMB*
471 *Life* **2012**, *64*, 513–520.
- 472 11. Tompa, P.; Davey, N.E.; Gibson, T.J.; Babu, M.M. A Million Peptide Motifs for the Molecular Biologist.
473 *Molecular Cell* **2014**, *55*, 161–169.
- 474 12. Tompa, P.; Schad, E.; Tantos, A.; Kalmar, L. Intrinsically disordered proteins: emerging interaction
475 specialists. *Current Opinion in Structural Biology* **2015**, *35*, 49 – 59.
- 476 13. Dunbrack, R. Rotamer libraries in the 21st century. *Current Opinion in Structural Biology* **2002**, *12*, 431–440.
- 477 14. Smith, L.J.; Bolin, K.A.; Schwalbe, H.; MacArthur, M.W.; Thornton, J.M.; Dobson, C.M. Analysis of
478 main chain torsion angles in proteins: Prediction of NMR coupling constants for native and random coil
479 conformations. *Journal of Molecular Biology* **1996**, *255*, 494 – 506.
- 480 15. Jha, A.K.; Colubri, A.; Freed, K.F.; Sosnick, T.R. Statistical coil model of the unfolded state: Resolving the
481 reconciliation problem. *Proceedings of the National Academy of Sciences* **2005**, *102*, 13099–13104.
- 482 16. Bernadó, P.; Blanchard, L.; Timmins, P.; Marion, D.; Ruigrok, R.W.H.; Blackledge, M. A structural model
483 for unfolded proteins from residual dipolar couplings and small-angle X-ray scattering. *Proceedings of the*
484 *National Academy of Sciences of the United States of America* **2005**, *102*, 17002–17007.
- 485 17. Kolodny, R.; Koehl, P.; Guibas, L.; Levitt, M. Small libraries of protein fragments model native protein
486 structures accurately. *Journal of Molecular Biology* **2002**, *323*, 297 – 307.
- 487 18. Rohl, C.A.; Strauss, C.E.; Misura, K.M.; Baker, D. Protein structure prediction using Rosetta. In *Numerical*
488 *Computer Methods, Part D*; Academic Press, 2004; Vol. 383, *Methods in Enzymology*, pp. 66 – 93.
- 489 19. Baeten, L.; Reumers, J.; Tur, V.; Stricher, F.; Lenaerts, T.; Serrano, L.; Rousseau, F.; Schymkowitz, J.
490 Reconstruction of protein backbones from the BriX collection of canonical protein fragments. *PLOS*
491 *Computational Biology* **2008**, *4*, 1–11.
- 492 20. Maupetit, J.; Derreumaux, P.; Tufféry, P. A fast method for large-scale De Novo peptide and miniprotein
493 structure prediction. *Journal of Computational Chemistry* **2010**, *31*, 726–738.
- 494 21. Molloy, K.; Shehu, A. A general, adaptive, roadmap-based algorithm for protein motion computation.
495 *IEEE Transactions on NanoBioscience* **2016**, *15*, 158–165.
- 496 22. Huang, J.R.; Ozenne, V.; Jensen, M.R.; Blackledge, M. Direct prediction of NMR residual dipolar couplings
497 from the primary sequence of unfolded proteins. *Angewandte Chemie-International Edition* **2013**, *52*, 687–690.
- 498 23. Estaña, A.; Sibille, N.; Delaforge, E.; Vaisset, M.; Cortés, J.; Bernadó, P. Realistic ensemble models of
499 intrinsically disordered proteins using a structure-encoding coil database. *Structure* **2019**, *27*, 381–391.e2.
- 500 24. Levinthal, C. How to fold graciously. *Mössbauer Spectroscopy in Biological Systems Proceedings* **1969**, pp.
501 22–24.
- 502 25. Rooman, M.; Dehouck, Y.; Kwasigroch, J.; Biot, C.; Gilis, D. What is paradoxical about Levinthal paradox?
503 *Journal of Biomolecular Structure & Dynamics* **2003**, *20*, 327–9.
- 504 26. Al-Bluwi, I.; Siméon, T.; Cortés, J. Motion planning algorithms for molecular simulations: A survey.
505 *Computer Science Review* **2012**, *6*, 125–143.
- 506 27. Gipson, B.; Hsu, D.; Kavraki, L.; Latombe, J.C. Computational models of protein kinematics and dynamics:
507 Beyond simulation. *Annual Review of Analytical Chemistry* **2012**, *5*, 273–91.
- 508 28. Shehu, A.; Plaku, E. A survey of computational treatments of biomolecules by robotics-inspired methods
509 modeling equilibrium structure and dynamic. *Journal of Artificial Intelligence Research* **2016**, *57*, 509–572.
- 510 29. Ghallab, M.; Nau, D.; Traverso, P. *Automated Planning: Theory and Practice*; Morgan Kaufmann Publishers,
511 Elsevier, 2004.
- 512 30. Richter, S.; Westphal, M. The LAMA planner: Guiding cost-based anytime planning with landmarks.
513 *Journal of Artificial Intelligence Research* **2010**, *39*, 127–177.

- 514 31. Wales, D. *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*; Cambridge University Press, 2003.
515
- 516 32. Honda, S.; Yamasaki, K.; Sawada, Y.; Morii, H. 10 residue folded peptide designed by segment statistics.
517 *Structure* **2004**, *12*, 1507–1518.
- 518 33. Liang, H.; Chen, H.; Fan, K.; Wei, P.; Guo, X.; Jin, C.; Zeng, C.; Tang, C.; Lai, L. De novo design of a $\beta\alpha\beta$
519 Motif. *Angewandte Chemie International Edition* **2009**, *48*, 3301–3303.
- 520 34. Enemark, S.; Kurniawan, N.A.; Rajagopalan, R. β -hairpin forms by rolling up from C-terminal: Topological
521 guidance of early folding dynamics. *Scientific Reports* **2012**, *2*.
- 522 35. Qi, Y.; Huang, Y.; Liang, H.; Liu, Z.; Lai, L. Folding simulations of a de novo designed protein with a $\beta\alpha\beta$
523 fold. *Biophysical Journal* **2010**, *98*, 321 – 329.
- 524 36. Rapaport, D.C. *The art of molecular dynamics simulation*; Academic Press, 2007.
- 525 37. Snow, C.; Zagrovic, B.; Pande, V. The Trp-cage: Folding kinetics and unfolded state topology via molecular
526 dynamics simulations. *Journal of the American Chemical Society* **2003**, *124*, 14548–9.
- 527 38. The PyMOL Molecular Graphics System, Version 2.0, Schrödinger, LLC.
- 528 39. Satoh, D.; Shimizu, K.; Nakamura, S.; Terada, T. Folding free-energy landscape of a 10-residue mini-protein,
529 chignolin. *FEBS Letters* **2006**, *580*, 3422–3426.
- 530 40. Kabsch, W.; Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded
531 and geometrical features. *Biopolymers* **1983**, *22*, 2577–2637.
- 532 41. Crooks, G.E.; Hon, G.; Chandonia, J.M.; Brenner, S.E. WebLogo: a sequence logo generator. *Genome*
533 *Research* **2004**, *14*, 1188–1190.
- 534 42. Kührová, P.; De Simone, A.; Otyepka, M.; Best, R.B. Force-field dependence of Chignolin folding and
535 misfolding: Comparison with experiment and redesign. *Biophysical Journal* **2012**, *102*, 1897–1906.
- 536 43. Fox, N.K.; Brenner, S.E.; Chandonia, J.M. SCOPe: Structural classification of proteins—extended, integrating
537 SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research* **2014**, *42*, D304–D309.
- 538 44. Jha, A.K.; Colubri, A.; Freed, K.F.; Sosnick, T.R. Statistical coil model of the unfolded state: Resolving the
539 reconciliation problem. *Proceedings of the National Academy of Sciences of the United States of America* **2005**,
540 *102*, 13099–13104.
- 541 45. Scott, R.A.; Scheraga, H.A. Conformational analysis of macromolecules. III. Helical structures of
542 polyglycine and poly-L-alanine. *The Journal of Chemical Physics* **1966**, *45*, 2091–2101.
- 543 46. Levitt, M. A simplified representation of protein conformations for rapid simulation of protein folding.
544 *Journal of Molecular Biology* **1976**, *104*, 59–107.
- 545 47. Ruiz de Angulo, V.; Cortés, J.; Porta, J.M. Rigid-CLL: Avoiding constant-distance computations in cell
546 linked-lists algorithms. *Journal of Computational Chemistry*, *33*, 294–300.
- 547 48. Bondi, A. Van der Waals Volumes and Radii. *Journal of Physical Chemistry* **1964**, *68*, 441–451.
- 548 49. Devaurs, D.; Molloy, K.; Vaisset, M.; Shehu, A.; Siméon, T.; Cortés, J. Characterizing energy landscapes
549 of peptides using a combination of stochastic algorithms. *IEEE Transactions on NanoBioscience* **2015**,
550 *14*, 545–552.