A Mathematical Model for Secondary Structure in Proteins

Alexey Nikolaev [1] Saad Mneimneh [2]

***Abstract*— We propose a new mathematical model for secondary structure in proteins. Our model is inspired by percolation theory on binary strings. What sets us apart from similar work on the subject is our attempt to deviate from a data mining approach (which is mostly the trend is science these days). Therefore, in predicting secondary structures, we make it our challenge to adhere to sequence information alone, in a non ad-hoc way, with only minimal information extracted from databases of known structures. Initial results show that our model captures some essential aspects of structure formation, notably a de novo discovery of hydrophobicity from an optimization perspective. A comparison of our prediction algorithm to similar methods shows improved performance. In addition, some evolutionary algorithms using our model exhibit convergences that are consistent with information obtained from structural biology.**

I. Introduction

The prediction of protein secondary structure is an important and well developed area in Computational Biology. State of the art methods in this area, e.g. [10] and [13], rely heavily on large databases of experimentally observed secondary structures and known proteins' homologs; and they use this data with remarkable success! Nevertheless, the problem of secondary structure prediction is still far from being solved, and there is a room for new ideas and approaches. In this work, we make a new attempt to move away from data-driven methods, and revisit the problem from a solely mathematical perspective. In doing so, we rely on two premises.

First, sequences of biological origin, e.g. peptide chains (proteins), are the result of a long evolutionary process. When treated as strings, these sequences will reveal patterns (substrings) that are not likely to be seen in randomly generated strings. Such unusual patterns may just be the right candidates for significant structures. The claim that all structures conform to this paradigm is of course an overstatement: some may still look like a simple random string. But without an extensive database, they are not identifiable, and one can only hope to turn to such probabilistic arguments.

Second, interactions among amino acids in a protein exhibit a certain locality. This locality is especially apparent in secondary structures where "neighboring" amino acids form helices and strands. It is this particular feature that stands behind the 1969 marked speech of Levinthal: "protein folding is speeded and guided by the rapid formation of local interactions which then determine the further folding of the peptide" [1]. Perhaps the periodicity in a secondary structure will best exemplify this notion of neighborhood. For instance, α-helices, with a period of 3.6 residues (amino acids) per turn, form by making hydrogen bonds to every 4^{th} amino acid. This loosely determines a neighborhood of 4. Other types of helices such as 3_{10}-helices, π-helices, and Polyproline II helices have different periods. Strands have a period of 2.

TABLE I

SOME PROTEIN SECONDARY STRUCTURES [15], [20].

	residues per turn	hyd. bond C=O$\cdots$HN	neighborhood
α-helix	3.6	$i \rightarrow i+4$	4
3_{10}-helix	3	$i \rightarrow i+3$	3
π-helix	4.4	$i \rightarrow i+5$	5
Polyproline II helix	3		3
β-strand	2		2

We will explore structures by identifying the unusual patterns that emerge in binary strings (of '1's and '0's), through a process of cluster formation similar to the one found in percolation theory [7]. In our model, '1's that are separated by at most $k-1$ consecutive '0's, for some integer k (the neighborhood), are considered to belong to the same cluster. We will show that our model, though restricted to binary strings, captures some relevant properties of biological structures when amino acids are binarized. Moreover, it lends itself to a simple algorithm for predicting secondary structure. While our algorithm requires almost no information extracted from structural databases, comparison to similar works that do, such as the Chou Fassman method [2] and a recent improvement of it [6], reveals similar or better performance.

II. What sets us apart from similar work

Both binary strings and percolation theory have been previously considered as themes in connection to protein structure, though not together. Some early work considers protein folding as a combinatorial problem on binary strings, by mapping hydrophobic amino acids to '1' and the remaining amino acids to '0', e.g. [16], [17], and [18]. Hydrophobic Cluster Analysis (HCA), e.g. in [9], [12], and [14], provides a primitive construct for cluster formation in strings using a similar mapping, but lacks a rigorous probabilistic framework for distinguishing clusters. A similar attempt can be found in [11]. The work in [8] explores percolation theory, but only in the context of extracting a resemblance of existing protein structures to random graphs and, therefore, relies heavily on an advance knowledge of the structure. Our attempt to avoid as much as possible the use of structural information means that we should adhere to sequence information alone. To that end, some ad-hoc

Supported by NSF Award CCF-AF 1049902 and a CUNY GC Science Fellowship.

Supported by NSF Award CCF-AF 1049902.

The authors contributed equally to this work.

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

attempts like [3], [4], [5], and [6] have been characterized as methods that rely on sequence information alone; however, they all make use of information extracted from patterns of known structures, e.g. frequency of amino acids in a given secondary structure. We will intentionally refrain from such ways, and to be specific, when a database is used, we only extract a mapping of the amino acids to binary.

III. Model

A. Proteins as binary strings

We think of an amino acid sequence as a binary string consisting of two kinds of symbols: Structure forming ('1's) and structure indifferent ('0's) (this by no means is a claim that such amino acids are biologically indifferent to structure). We will motivate the need for a third kind, structure breakers, which will be later added to the model. Probabilities of '1's and '0's are assumed to be known. We model structures as clusters that form by local interactions of '1's. Informally, '1's that are "neighbors" belong to the same cluster, and we use k, an integer, to quantify this closeness. Formally, for a given binary string, consider a hierarchy of clusters formed by the following merging process:

Definition 1: Every '1' is called a cluster at level $k = 0$. Two or more clusters merge and form a bigger cluster at level $k > 0$, if they are separated by no more that $k - 1$ consecutive '0's (See Figure 1).

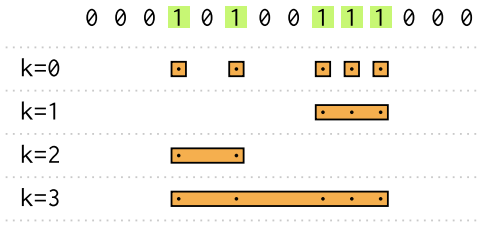

Fig. 1. Clusters in a binary string for different values of k.

Definition 2: The size of a cluster is the number of '1's in it.

If the probability of '1's is $0 < p < 1$, and the probability of '0's is $q = 1 - p$, we can quantify how unusual each cluster is. We start by defining the following probability.

Definition 3: Let $w_{k,s}$ be the probability that a given '1' falls in a cluster of size $s \geq 1$ at level $k \geq 0$.
It can be shown that:

$$w_{k,s} = \begin{cases} 0^{s-1} & k = 0 \\ (\beta_{k,s} - \beta_{k-1,s}) \cdot q^{2k} & k > 0 \end{cases}$$

where $\beta_{k,s} = s(p\alpha_k)^{s-1}$ and $\alpha_k = \sum_{i=0}^{k-1} q^i$. Observe that $\sum_{s=1}^{\infty} w_{k,s} = 1 - q^2$ for $k > 0$; there is a probability of q^2 that for a given '1' no cluster exists at level $k > 0$.

The reader may think of this model is an extension of the one dimensional percolation theory (percolation on a line when $k = 1$) [7]. However, while in percolation theory one is mostly concerned with critical properties of the system, here we look at the clusters themselves, and their likelihood.

To capture the unusual-ness of a cluster, we define its weight as its tail probability. Thus, unusual cluster have smaller weights.

Definition 4: The weight of a cluster of size s at level k is

$$W(k,s) = \frac{1}{\zeta_k} \min\Big(\sum_{t=1}^{s} w_{k,t}, \sum_{t=s}^{\infty} w_{k,t}\Big)$$

where $\zeta_k = \sum_{t=1}^{\infty} w_{k,t}$ which is 1 when $k = 0$ and $1 - q^2$ when $k > 0$.

What is the best cluster? For a given '1', we say that the best cluster is the one with the smallest weight (most unusual). Consider an algorithm that, for a given '1', sequentially examines levels $k = 0, 1, 2, \ldots$ in an attempt to find the best cluster. Every time a better cluster is found, the algorithm outputs the corresponding value of k.

Algorithm 1
$W \leftarrow \infty$
for $k \in \{0, 1, 2 \ldots\}$
  if $\exists$ a cluster of size s at level k s.t. $W(k,s) < W$
    then $W \leftarrow W(k,s)$
      output k

Definition 5: Given a '1' in a random binary string of infinite length, let $P(k)$ be the probability that k is the last output performed by the above algorithm.

Why are we interested in $P(k)$? When the best cluster is at level k, k may be used to distinguish the type of structure. We do not make a claim that there is a one-to-one correspondence between k and the neighborhood as listed in Table I, but a probabilistic argument is in accordance with our line of thought. Therefore, we hope that $P(k)$ will reflect a distribution that is reasonable when k is interpreted as the value in Table I. The following result shatters the hope but motivates an interesting approach described in the following section.

Theorem 1: $P(k) = 0$ for all k.

Proof: First observe that when $p > 0$, the probability that the algorithm will stop finding clusters is 0. Assume that the cluster at level k has size s and that the best cluster up to level k has weight W^*.

Now we show that there is a probability greater than a positive constant to find a better cluster at level $k+1$. Either $W(k+1, s+\Delta s) < W^*$ for all $\Delta s > 0$, or there exists an $s_0 > s$ such that $W(k+1, s_0) \geq W^*$ and $W(k+1, s_0+t) < W^*$ for all $t > 0$. In the former case, the best cluster will change with probability $1 - q^2$ (the probability that there is a cluster at level $k + 1$). In the latter case, consider the probability $t_{s,k}^{\Delta s}$ that the size of the cluster at level $k + 1$ is $s + \Delta s$. We can show (omitted here) that this probability is bounded as follows:

$$t_{k,s}^{\Delta s} \geq p^2 w_{k+1,\Delta s-1}$$

For $s + \Delta s > s_0$ i.e. $\Delta s > s_0 - s$, we have

$$\sum_{\Delta s > s_0 - s} t_{k,s}^{\Delta s} \geq p^2 \sum_{\Delta s > s_0 - s} w_{k+1,\Delta s-1} = p^2 \sum_{\Delta s \geq s_0 - s} w_{k+1,\Delta s}$$

$$> p^2 \sum_{\Delta s \geq s_0} w_{k+1,\Delta s} \geq p^2(1-q^2) W(k+1, s_0) \geq p^2(1-q^2) W^*$$

which is a constant given W^*.

Therefore, for every cluster weight W^*, there is a constant non-zero probability that the best cluster will change. This proves the theorem. ■

Since the distribution $P(k)$ does not exist, it imposes several difficulties when we look for the best clusters in finite strings. Specifically, the best clusters will tend to be larger in long strings, and smaller in short ones. The problem can be solved by introducing a third symbol that "breaks" clusters, that is, prevents them from merging at higher levels. Indeed, such breakers exist in biological structures; for instance, Proline and Glycine are typically considered to be structure breakers in helices: with a rigid loop side chain, Proline lacks the flexibility to conform its ϕ and ψ angles to the typical structures (but not the structures that require this unique rigidity of Proline, e.g. polyproline helices). Glycine, on the other hand, has the most flexible side chain and, hence, is more likely than others to mediate a change in structure.

B. The introduction of breakers

We now assume that an amino acid sequence is a string over the alphabet {'1','0','π'}, where 'π' stands for structure breaker. Let $\pi > 0$ also denote the probability of breakers.

$$p + q + \pi = 1$$

With the introduction of breakers, the merging of clusters may now stop for different reasons. The probabilities and weights are updated accordingly. A cluster at level k may end with:

$$\text{no breakers: } w_{k,s}^{(0\pi)} = \begin{cases} (\beta_{k,s} - \beta_{k-1,s}) \cdot q^{2k} & (k > 0) \\ 0^{s-1} & (k = 0) \end{cases}$$

$$\text{breaker on one side: } w_{k,s}^{(1\pi)} = (\beta_{k,s} - \beta_{k-1,s}) \cdot 2q^k \alpha_k \pi$$

$$\text{on both sides: } w_{k,s}^{(2\pi)} = (\beta_{k,s} - \beta_{k-1,s}) \cdot (\alpha_k \pi)^2$$

$$W^{(x\pi)}(k,s) = \frac{1}{\zeta_k} \min\Big(\sum_{t=1}^{s} w_{k,t}^{(x\pi)}, \sum_{t=s}^{\infty} w_{k,t}^{(x\pi)}\Big) \text{ for } x \in \{0,1,2\}$$

where $\zeta_k = \sum_{t=1}^{\infty}(w_{k,t}^{(0\pi)} + w_{k,t}^{(1\pi)} + w_{k,t}^{(2\pi)})$ which is 1 when $k = 0$ and $[(1-p\alpha_k)^{-2} - (1-p\alpha_{k-1})^{-2}](q^k + \alpha_k\pi)^2$ when $k > 0$.

The probability $P(k)$ as defined in the previous section is now greater than 0 for every k because Algorithm 1 will stop finding clusters with probability 1 (when breakers are encountered on both sides of the given '1').

We can obtain $P(k)$ by running Algorithm 1 on random sequences. In Figure 2, we show $P(k)$ when $p = 0.33$, $q = 0.62$, and $\pi = 0.05$. These are the approximate probabilities when hydrophobic amino acids are mapped to '1' and Proline is mapped to 'π' (the convention used in HCA).

IV. DSSP DATA AND SECONDARY STRUCTURE PREDICTION

The (non-redundant) DSSP database [21] is a repository of annotated protein secondary structures: α-helices, 3_{10}-helices, and π-helices are marked by "H", "G", and "I" respectively. Extended β-strands are marked by "E". For each amino acid in the sequence, we would like to predict whether it belongs to a helix, a strand, or neither (a coil). When predicting three states, a standard measure of accuracy is the Q_3 score, defined the number of correctly predicted amino acids divided by the number of amino acids in the sequence (the sequence length). Therefore, to compute our Q_3 score, we consider amino acids marked by "H", "G", and "I" to belong to a helix, those marked by "E" to belong to a strand, and the remaining amino acids (not marked) to belong to a coil.

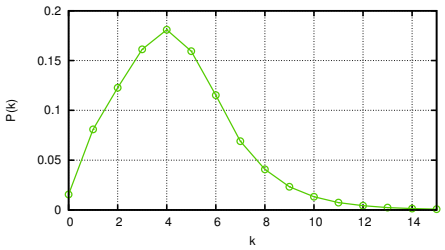

Fig. 2. Result of simulation on random sequences for $P(k)$ with $p = 0.33$, $q = 0.62$ and $\pi = 0.05$.

A. Random search for the best mapping

To successfully apply our model to real protein data, we need to know the best way to map 20 amino acids to {'1', '0', 'π'}. Then with the knowledge of the amino acids' frequencies and this mapping alone, we can obtain the probabilities p, q, and π.

We use a standard genetic algorithm to find the best mapping. A mapping is represented by a string of 20 symbols over the alphabet {'1', '0', 'π'}. First, we generate a random initial population of mappings. Then, we select random sequences from the DSSP database and run a prediction algorithm (described below) on these sequences for each mapping in the population. The fitness of a mapping is the Q_3 score it achieves. The best 50% of the population are saved to the next generation, and the remaining 50% are produced by mutations and recombination of the saved ones.

B. Prediction algorithm

Given a mapping of the 20 amino acids to {'1', '0', 'π'}, we compute the probabilities p, q, and π. For every '1' in the string, we find the best cluster using Algorithm 1. When finding those clusters in finite sequences, we assume imaginary "breakers" on both ends of the sequence. This choice is not critical in terms of performance, but a way to maintain consistency of the method. Figure 3 shows an example cluster coverage for the HCA mapping of amino acids {V, I, L, F, Y, M, W}→'1' and {P}→'π'.

Once all the best clusters have been found, we make a prediction guided by Table I in the following way (there should be better ways to make predictions by also considering the size of a cluster, but we stick to this one for now):

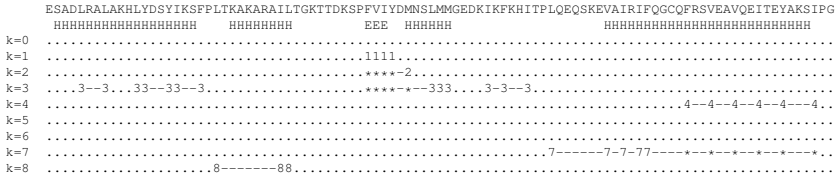

Fig. 3. Sequence is shown with the DSSP annotation to illustrate the cluster coverage of structures. A '-' denotes a *hydrophilic* amino acid in the cluster. A number in the cluster (always equal to its level) indicates that this cluster is the best for the corresponding *hydrophobic* amino acid, and a '*' means that it is not.

Algorithm 2
for every amino acid
  if covered by a cluster at level $1 \leq k \leq 2$
    then mark it as Strand
    else if covered by a cluster at level $k \geq 3$
      then mark it as Helix
      else mark it as Coil

C. Prediction heatmaps

The prediction algorithm described above was initially conceived from information in Table I, and our premise that structures with a neighborhood of k should probabilistically reveal the best clusters at level k. We back up this intuition with actual data from DSSP. Figure 4 shows heatmaps for how many times the best cluster is at level k with size s, for '1's in strands, helices, and coils. We used the best mapping obtained by the genetic algorithm described earlier.

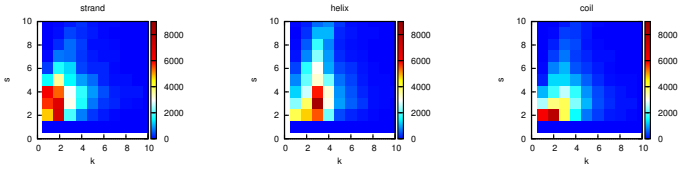

Fig. 4. Heatmaps of best clusters at level k with size s for strands, helices, and coils.

In addition, Figure 5 shows heatmaps that justify the distinction of strands, helices, and coils.

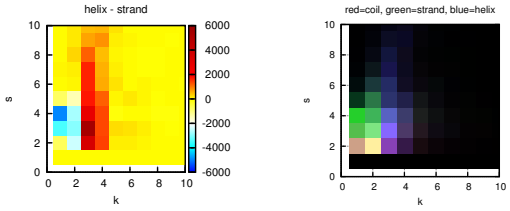

Fig. 5. Left: the difference between the first two heatmaps in Figure 4 (helix minus strand). Right: the superposition of all three heatmaps in Figure 4 with an RGB scheme.

V. De novo discovery of hydrophobicity and some results

We obtain {V, I, L, F, Y, M}→'1' and {P, G}→'π' as the best mapping (given the Q_3 fitness). As argued previously, Proline (P) and Glycine (G) are the most biologically reasonable breakers. The other six amino acids are known to be highly hydrophobic.

If we don't care to distinguish between strands and helices (thus reverting to a similarly defined Q_2 fitness), our genetic algorithm gives {V, I, L, F, Y, M, W, A}→'1' or {V, I, L, F, Y, M, W, A, Q}→'1', and {P, G}→'π' as the best mappings. The inclusion of Trypyophane (W), Alanine (A), and Glutamine (Q) as the next hydrophobic amino acids is known to be a biologically reasonable choice.

We emphasize here that the algorithm itself is not explicitly exploring hydrophobicity. This confirms that a binary system (with a breaker) is adequate to model the evolution of structures, and it happens that hydrophobicity plays that role very well. Clearly, if our model were just a computational artifact, one would expect to obtain arbitrary mappings into {'1','0','π'}. As it turns out, however, our model infers hydrophobicity as a key property for guiding the formation of secondary structure. Some results using the best mappings mentioned above are shown in Table II.

VI. Evolutionary algorithms and some insights

We explore the possibility of designing evolutionary mechanism based on our model that lead to the convergence of p and π (and thus q as well) to their typical values. For this, we consider two evolutionary algorithms in which we envision the biological system as an entity that learns. Given initial values for p and π, p^0 and π^0, it recomputes those probabilities, p^i and π^i for every iteration i, conditioned on the best clusters. This is repeated until the process converges.

A. Micro evolution from the view of a '1'

The main object of this evolutionary algorithm is the best cluster for a given '1'. However, in order to recompute both p and π, we extend a cluster at level k with a trailing region on both sides consisting of either k '0's or until we encounter a 'π' (this is naturally dictated by the system).

Define $E[p] = \sum_{k=0}^{\infty} p_k P(k|p^{i-1}, \pi^{i-1})$ ($E[\pi] = \sum_{k=0}^{\infty} \pi_k P(k|p^{i-1}, \pi^{i-1})$), where p_k (π_k) is the probability of '1' ('π') in an infinite string consisting of only random extended clusters at level k (given p^{i-1}, q^{i-1}, and π^{i-1}), and the notation $P(k|p, \pi)$ is simply $P(k)$ when p and π (and q) are given. This evolves as described in Algorithm 3 by computing new values for p and π at iteration i as $E[p]$ and $E[\pi]$, respectively. The new values are then weighted by probability p^{i-1} (the current probability of a '1') and the current values by $1 - p^{i-1}$. The weights produce smoother curves, but simply making $p^i \leftarrow E[p]$ and $\pi^i \leftarrow E[\pi]$ results in the exact same behavior.

TABLE II

FITNESS COMPARISON FOR OUR APPROACH (IMPROVED FROM 0.55 TO 0.57 WITH SOME HEURISTICS) AGAINST OTHERS ON SIMILAR DATA SETS, E.G. CB513 AND CB396 [6]. THE BLIND STRATEGY OF CALLING EVERY AMINO ACID BY THE MOST FREQUENT CATEGORY (COIL IN THE FIRST ROW AND NOT COIL IN THE SECOND) IS ALSO INCLUDED FOR COMPLETENESS.

	Our approach	Chou-Fasman	Improved C-F*	GOR†	State of the art‡	Blind
strand, helix, coil	0.57	0.46	0.56	0.62	0.8	0.44
coil, not coil	0.67	0.61	-	-	-	0.56

* includes some data analysis based on wavelets.
† GOR III uses known (non independent) probabilities of amino acids in α-helices and β-strands.
‡ State of the art uses alignment of homologs and neural networks.

Algorithm 3
$[p^0, \pi^0] \leftarrow$ any such that $p^0 + \pi^0 < 1$
for $i \in \{1, 2, 3, \ldots\}$
$p^i \leftarrow p^{i-1}E[p] + (1 - p^{i-1})p^{i-1}$
$\pi^i \leftarrow p^{i-1}E[\pi] + (1 - p^{i-1})\pi^{i-1}$

The simulation of Algorithm 3 is depicted in Figure 6. Regardless of (p^0, π^0), (p^∞, π^∞) drifts to $(1, 0)$ (the lower right corner of Figure 6). If we envision a control mechanism against unlimited growth in p, then π would stabilize at the point where the curve becomes horizontal. For instance, a variation of Algorithm 3 that makes $p^i \leftarrow \min(p^i, 1-q-\pi^i)$ after every update, confirms a convergence of π around 0.047 for every fixed $q \in [0.4, 0.65]$. This shows the resilience of π in the face of small perturbations on p, and provides a compelling argument for the existence of typical probabilities when Proline is considered as the only breaker ($\pi = 0.047$ is the probability of Proline).

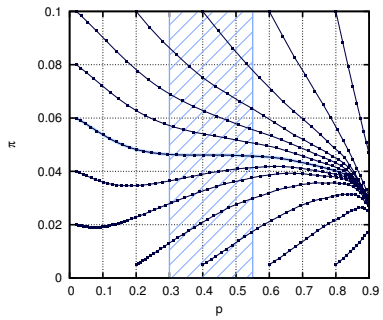

Fig. 6. The stability of π around 0.047, if there is a control mechanism against unlimited growth in p that contains it in the range $0.3 < p < 0.55$.

B. Macro evolution from a global view

Starting with 'π', consider a string obtained by generating symbols in {'1','0','π'} with the probabilities p^{i-1}, q^{i-1}, and π^{i-1}, respectively, until the occurrence of another 'π'. Call such a string a *chunk*. Define the coverage of a chunk as the string consisting of all the bits in the chunk that belong to a best extended cluster (as defined in the previous section) at **any** level k. We compute p^i and π^i as the probabilities of '1's and 'π's, respectively, in an infinite string made of random coverages. We then control these computed values by a damping factor $0 < \alpha \leq 1$. This evolves as described by Algorithm 4.

Algorithm 4
$[p^0, \pi^0] \leftarrow$ any such that $p^0 + \pi^0 < 1$
for $i \in \{1, 2, 3, \ldots\}$
$[p^i, \pi^i] \leftarrow$ probabilities of '1' and 'π', respectively, in an infinite coverage
$[p^i, \pi^i] \leftarrow \alpha[p^i, \pi^i]$

The simulation of Algorithm 4 is depicted in Figure 7. For every fixed value of α, $[p^\infty, \pi^\infty]$ converge to a point on the curve. It is worth to note here that replacing $[p^i, \pi^i] \leftarrow \alpha[p^i, \pi^i]$ by $[p^i, \pi^i] \leftarrow \frac{1-q}{p^i+\pi^i}[p^i, \pi^i]$ for a fixed q results in the same convergence. Surprisingly, the curve passes through the point $[p, \pi] = [0.45, 0.12]$ corresponding to $q = 0.43$, which roughly represents the probabilities given by the mapping {V, I, L, F, Y, M, W, A, Q}→'1', and {P, G}→'π'. This is one of the best mappings obtained when prediction had to distinguish between regular structure (helices and strands) and non-regular structure (coils).

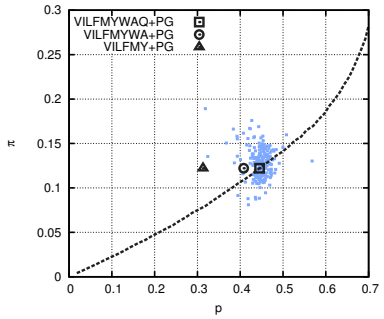

Fig. 7. Convergence of p and π for every fixed α. The curve passes through $[p, \pi] = [0.45, 0.12]$ corresponding to $q = 0.43$. The filled squares are the average $[p, \pi]$ for the 200 largest species in DSSP and the mapping {V, I, L, F, Y, M, W, A, Q}→'1', and {P, G}→'π'. The large open square, circle, and triangle are the average $[p, \pi]$ obtained from the whole DSSP database computed for three corresponding mappings.

In obtaining the above results, we make no explicit reference to structural biology or specific evolutionary mechanism. Yet, the two algorithms, Algorithm 3 and Algorithm 4, suggest that biological structures may have been the subject of similar evolution, and provide some evidence for the existence of typical probabilities from an evolutionary perspective. In additions, they show the distinctive roles of Proline and Glycine in structure formation: while Proline is a structure breaker at the micro level, Glycine acts, in addition, as a breaker between different types of structures at the macro level. This view is consistent with Ramachandran's classification of proteins in which Proline is the most rigid

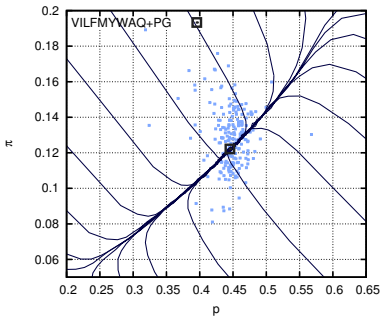

Fig. 8. Convergence in action to $[p, \pi] = [0.45, 0.12]$ corresponding to $q = 0.43$. The filled squares are the average $[p, \pi]$ for the 200 largest species in DSSP and the mapping {V, I, L, F, Y, M, W, A, Q}→'1', and {P, G}→'π'.

(hence a hard breaker of a structure) and Glycine is the most flexible (hence a soft breaker of a structure that possibly mediates a change in structure).

VII. Conclusion

Our model/algorithm succeeds in making initial non-trivial predictions with only minimal information obtained from structural biology. Compared to methods in the same class (those that do not require information of known structures or homologs), we observe improved (or similar) performance. It can be interesting to apply our model to more advanced prediction techniques and explore new ways of mapping sequences to binary, e.g. one may account for some dependence among the bits. We hope that our model or an extension of it can lead to a better understanding of folding mechanisms and structure formation from a puristic perspective, as our evolutionary algorithms might suggest.

References

[1] Levinthal, C.: How to Fold Graciously. Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois (1969)

[2] Chou, P. Y., Fasman, G. D. (1878) Prediction of the secondary structure of proteins from their amino acid sequence, *Adv Enzymol Relat Areas Mol Biol* **47**, 45-148.

[3] Garnier, J., Osguthorpe, D. J., Robson, B. (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins, *Journal of Molecular Biology*, **120**, 97-120.

[4] Garnier, J., Gibrat, J. F., Robson, B. (1996) GOR method for predicting protein secondary structure from amino acid sequence, *Methods Enzymol*, **266**, 540-53.

[5] Kabsch, W. and Sander,C. (1983) How good are predictions of protein secondary structure?, *FEBS Letters*.

[6] Chen, H., Gu, F., and Huang, Z (2006) Improved Chou-Fasman method for protein secondary structure prediction, *BMC Bioinformatics* **7**(Suppl 4).

[7] Stauffer, D. and Aharony, A. (1985) Introduction to Percolation Theorey, Second Edition, *CRC Press*.

[8] Deb, D., Vishveshwara, S., and Vishveshwara, S. (2009) Understanding protein structure from a percolation perspective, *Biophysical Journal* **97** 1787-1794.

[9] Gaboriaud, C., Bissery, V., Benchetrit T., and Mornon, J. P. (1987) Hydrophobic Cluster Analysis: An Efficient New Way to Compare and Analyse Amino Acid Sequences, FEBS Letters, **224**(1).

[10] Rost, B., Sander, C., Schneider, R. (1994) PHD - An Automatic Mail Server for Protein Secondary Structure Predication, *Computational Applied Bioscience*, **10**(1).

[11] West, M. W. and Hecht, M. H. (1995) Binary Patterning of Polar and Nonpolar Amino Acids in the Sequences and Structures of Native Proteins, *Protein Science*, **4**.

[12] Callebaut, I., Labesse, G., Durand, P., Poupon, A., Canard, L., Chomilier, J., Henrissat, B., and Mornon, J. P. (1997) Deciphering Protein Sequence Information Through Hydrophobic Cluster Analysis (HCA): Current Status and Perspectives, *Cellular and Molecular Life Sciences*, **53**.

[13] McGuffin, L. J., Bryson, K., Jones, D. T. (2000) The PSIPRED protein structure prediction server, *Bioinformatics*, **16**(4).

[14] Hennetin, J., Le, T. K., Canard, L., Colloc'h, N., Mornon, J. P., Callebaut, I. (2003) Non-intertwined binary patterns of hydrophobic/nonhydrophobic amino acids are considerably better markers of regular secondary structures than nonconstrained patterns, *Proteins*, **51**(2), 236-44.

[15] Lesk, A. M. (1999) Introduction to Protein Architecture, the structural biology of protein, *Oxford University Press*.

[16] Hart, W. E., Istrail, S. (1995) Fast Protein Folding in the Hydrophobic-hydrophilic Model Within Three-eights of Optimal, *STOC*.

[17] Agarwala, R., Batzoglou, S., Dancik, V., Decatur, S. E., Farach, M., Hannenhalli, S., Skiena, S. (1997) Local Rules for Protein Folding on a Triangular Lattice and Generalized Hydrophobicity in the HP Model, *SODA*.

[18] Berger,B. and Leighton, T. (1998) Protein Folding in the Hydrophobic-Hydrophilc (HP) Model is NP-Complete, *Journal of Computational Biology*, **5**(1).

[19] The UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt), *Nucleic Acid Research*, **40**.

[20] Cooley, R. B., Arp, D. J., Karplus, P. A. (2010) Evolutionary origin of a secondary structure: π-helices as cryptic but widespread insertional variations of α-helices enhancing protein functionality, *Journal of Molecular Biology*, **404**(2).

[21] Kabsch, W., Sander, C. (1983) Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-bonded and Geometrical Features, *Biopolymers*, **22**(12).