Why is There a Limited Number of Protein Folds in Nature?

Abram Magner, Yifeng Yang**, Daisuke Kihara****
*Mathematics/Computer science, **Biological sciences,
***Biological sciences/Computer science, Purdue University

Abstract

It has long been known that only a limited number of amino acid sequences and tertiary protein structures are realized, and their frequency distributions are highly skewed. We investigated the relationship between sequence and structure through the lens of information theory to determine a possible explanation of this phenomenon. We hypothesized that the dense mapping from sequence to structure is a result of an efficient channel coding process. By exhaustively enumerating all 16-node compact structures/sequences, we constructed a table of structure probabilities conditioned on sequences, which was viewed as a noisy channel. The capacity of this channel was computed, and the optimizing sequence distribution was compared with the observed distribution of sequences and structures in nature. These distributions were found to be highly skewed and similar in shape to one another. These results lend credibility to the conclusion that the dense sequence-structure mapping arises as an efficient encoding of proteins necessary to sustain life.

Methods

All possible compact, 16-node structures were enumerated on a 4x4 lattice (69 of these).

- All possible {H, P} strings of length 16 (65536 of these) were enumerated and used to generate a table of conditional probabilities P(S|X), where S is a given structure and X a given sequence. P(S|X) was viewed as a noisy channel mapping sequences to structures.
- The Arimoto-Blahut algorithm was used to compute an optimizing capacity C, sequence distribution X*, and structure distribution S^*_{exp}.

H-P Lattice Model

- Frequently used in biophysics.
- Energy for a sequence-structure combination:

$$E(seq, struct) = \sum_{x \in seqnodes} e(x)$$

  - e(x) is computed as a sum of scores for all contacts involving node x.
  - Scoring table:

	H	P
H	-2.3	-1
P	-1	0

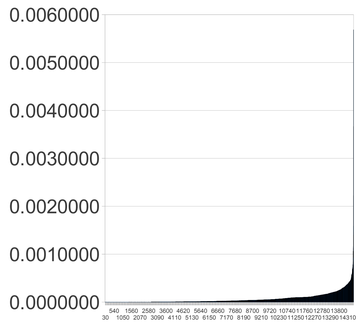
- Probability of structure S conditioned on sequence X:

$$Pr(S|X) = \frac{e^{-E(S,X)}}{\sum_{s \in structs} e^{-E(s,X)}}$$

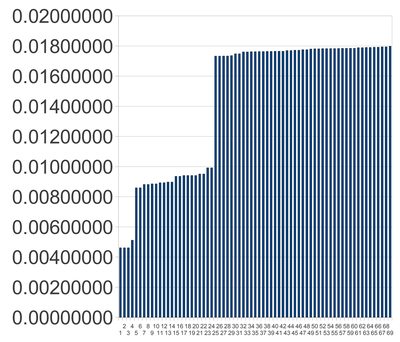
- Example sequence (H is black, P is white):

Graphs of X^* and S^*_{exp}

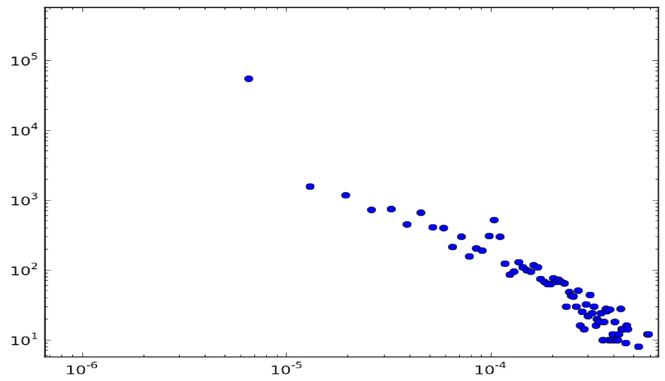
Optimal {H, P} Sequence Probabilities



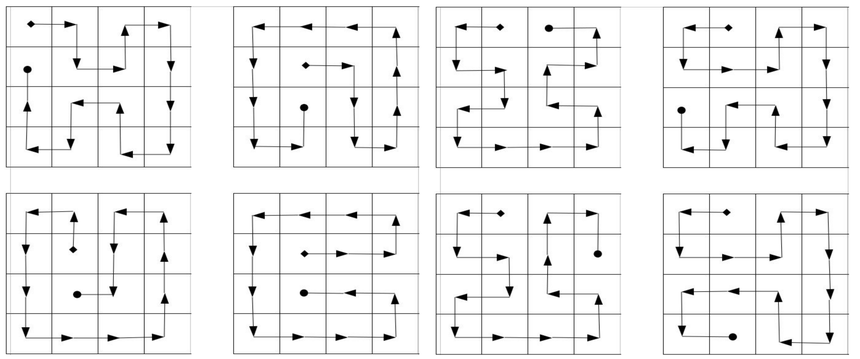
Optimal Structure Distribution



X^* Obeys a Power Law



Top/Bottom 4 Lattice Walks



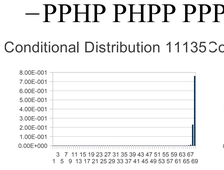
Information Theoretic Quantities

$$H(p) = -\sum_{i=1}^{n} p(x_i) log\, p(x_i)$$
$$I(X;Y) = H(X) - H(X||Y)$$
$$C = \max_{p(x)} I(X;Y)$$

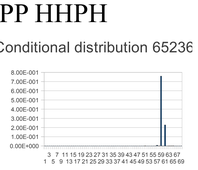
Results

- X^* obeyed a power law.
- The skewness of structure distributions conditioned on sequence decreased as the ranks of the sequences approached 1.
- Same results when the experiment is repeated with semi-compact structures.
- Most probable sequences:
  - HHHH HHHP HHPH HPHP
  - PHHH PHHH HHHH PHPP
  - PHPP HPPP PPPP HPPH
  - PPHP PHPP PPPP HHPH

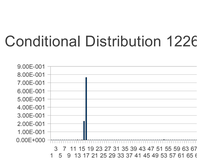
Conditional Distribution 11135



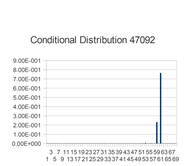
Conditional distribution 65236



Conditional Distribution 1226

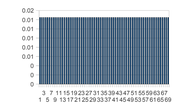


Conditional Distribution 47092

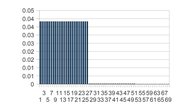


- Least probable sequences:
  - HHHH HHHH HHHH HHHH
  - PPPP PPPP PPPP PPPP
  - HHHH HHHH HHHH HHHP
  - HHHH HHHH HHHH HHPH

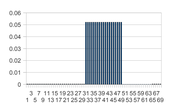
Conditional distribution 0



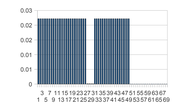
Conditional Distribution 1



Conditional Distribution 2



Conditional Distribution 3



Conclusions

The relationship between the rank of a sequence and the skewness of its conditional distribution lends credibility to the conclusion that the dense mapping from amino acid sequence to protein structure arises as an efficient encoding of proteins necessary to sustain life.

References

Cover, T., and Thomas, J. (2006), *Elements of Information Theory*, Second Edition, John Wiley & Sons, Inc.

Nakamura, H.K., & Sasai, M. (2001). Population analyses of kinetic partitioning in protein folding. *Proteins: Structure, Function, and Genetics.* 43, 280-291.