Does Shannon Entropy Convergence Spell Doom for All Living Things?

Ilse Bettina Schmitz Haoran Xiang Oliver Brown Vlad Radostev
Samuel Alampay

June 2024

Abstract

A controversial hypothesis suggests that genomes are slowly losing information due to the accumulation of mutations over long periods of evolutionary time. We observe how the Shannon entropy of Amino Acid Distributions change as we simulate the accrument of Mutations over time. We simulate mutations by iteratively running Amino Acid Distributions through a Markov chain. We plot the entropy of these distributions as they accrue mutations, and observe whether patterns are reflected by the controversial hypothesis. We observe that tested initial distributions converge to a common final distribution, with entropy neither approaching zero nor the maximum for a twenty possible event distribution. Thus we conclude this data does not indicate doom for the informational content of our DNA strings.

1 Introduction

A controversial hypothesis put forth by Sanford and Baumgardner suggests that the genomes of all life forms are slowly degenerating due to the inexorable accumulation of these slightly harmful mutations over long periods of evolutionary time. This "genetic entropy" theory posits that in the absence of perfect repair mechanisms to prevent all mutations, the overall fitness and complexity of organisms will trend toward decline and eventual extinction.

At the heart of this hypothesis lies the fundamental process of genetic mutation. Mutations are changes to the nucleotide sequence of an organism's DNA that can occur due to errors during DNA replication, exposure to mutagens, or simply random chance events. While many mutations are neutral or even beneficial under certain environmental contexts, providing the raw material for adaptive evolution, others can be deleterious by disrupting the encoded information required to produce viable proteins and sustain essential biological functions.

1.1 Introduction to Mutation

Proteins are large biomolecules consisting of long chains of amino acids linked by peptide bonds. The specific sequence of these 20 standard amino acids encodes the information required for each protein to fold into a precise three-dimensional structure capable of performing its designated role, be it structural, enzymatic, regulatory, or otherwise. Mutations in the DNA sequence of a protein-coding gene can lead to alterations in the amino acid sequence of the encoded protein. Depending on the location and severity of the mutation, this can range from having no effect on protein structure and function to partially or completely inactivating the protein.

It is this potential for mutations to degrade the information content encoded in genomic sequences that underlie the genetic entropy concept. From this view, harmful mutations represent losses of functional information over generations by corrupting the carefully optimized amino acid sequences of constitutive proteins required for organismal survival and fitness.

1.2 Introduction to the Mutation Model

To quantify this process, principles of information theory can be applied to biological sequences like proteins. The information entropy of a sequence provides a measure of its informational content or complexity. Higher entropy values correspond to more random, disordered arrangements while lower values indicate greater organization and structure to the patterning of elements.

By modeling the mutations accumulating in a protein-coding gene as a Markov chain process over successive generations, inspired by Thorvaldsen's mutation model derived from first principles of the Genetic Code [6], one can analyze how the information entropy of the encoded amino acid sequence changes over time due to the gradual fixation of mutations. A Markov chain provides a stochastic framework for modeling how a system transitions between states based on a set of probabilities dependent only on the current state. In this case, the "states" are the possible amino acids at each position in the sequence, and the transition probabilities govern the likelihood of a specific mutation occurring per generation.

Applying Markov chain analysis allows one to derive the steady-state distribution of amino acids at each position, from which the overall sequence entropy can be calculated. This provides a way to test the genetic entropy hypothesis by seeing if the information content of the sequence exhibits a general, irreversible decline towards higher entropy values over multitudinous generations as posited.

1.3 Simplifying Assumptions

Of course, the genetic entropy concept and its implications remain hotly debated within the scientific community. Critics argue that the hypothesis oversimplifies the complex dynamics of mutation, natural selection, and genome evolution. They contend that while slightly deleterious mutations may accumulate, other mechanisms like purifying selection against severely harmful mutations, genetic recombination generating new variation, and horizontal gene transfer can introduce new information to counteract information losses.

Additionally, some question whether information/entropy measures derived from information theory are truly applicable to the biological complexity encoded in genomes and proteins. The mapping between nucleotide and amino acid sequences to their expressed phenotypes is profoundly degenerate, contextual and non-linear. Simple metrics may fail to fully capture the functional information embedded in these systems.

Nonetheless, the genetic entropy hypothesis has spurred fascinating research into quantifying the information dynamics of evolving genetic systems. Markov models provide a versatile mathematical framework to investigate these processes, even if their underlying assumptions of site independence and time homogeneity may not always hold.

In this study, we assumed several simplifying assumptions so that the system could be implemented: 1) Only one substitution mutation per codon (triplet of DNA coding for a single amino acid) in order for the Markov Chain to hold as a workable model. 2) All frameshift mutation is not sustainable in the long term and is considered not viable. 3) All the viable mutants have the same reproductive efficiency, that is the survival fitness and reproductive rate are not affected.

1.4 Prospective of the Project

By applying Markov chain analysis to simulate the mutational degradation of a protein's amino acid sequence over deep evolutionary timescales, this study aims to shed light on the trajectory of informational change in a fundamental biological polymer. Tracking how sequence entropy values change as mutations steadily accrue can grant insights into whether the genetic entropy view aligns with the long-term fate of this essential class of biomolecules.

The first biomolecule we chose to simulate the long-term behavior of was human Insulin. It was ideal for a variety of reasons, the two most important being that Insulin is made up of a relatively small chain of amino acids, and is a particularly relevant protein to current day medicine. Insulin is made up of 51 amino acids, and while our various programs of choice, were very powerful, when recreating and testing a model, having a smaller protein to practice with means the program has a potentially shorter runtime, as well as potentially running into less bugs. Insulin is also a key player in diseases like diabetes, which has been an epidemic since 1994, a notable hot topic of American health and becoming more relevant worldwide.

2 Thesis Statement

Of particular interest is whether the amino acid sequence exhibits an inevitable, unidirectional increase in entropy as postulated by the genetic entropy theory. Or whether the patterns reveal more complex informational dynamics, with entropy increases and decreases reflecting the perpetual information flows and buffering mechanisms proposed by critics of the hypothesis.

3 Understanding Shannon Entropy

3.1 Sample Spaces and Distributions

A basic introduction to statistics for students in AMATH 383 who may have not taken a statistics course:

A sample space is a set of possible random events, each of which with an assigned chance of happening, and the probability distribution is the function that maps events to their respective probabilities. The sum of the probabilities of all events in a sample space must be 1 for the space to be valid. Discrete sample spaces have a countable number of possible events.

For example, the sample space for a single coin flip contains two events, heads and tails. Their respective probabilities are 0.5. This is intuitive because it would not make sense for the same weighted coin to flip heads 90 percent of the time, and also flip tails 90 percent of the time.

3.2 Profitable Coin Flipping

Consider your friend offers you the opportunity to win some money in a game of change. You flip a fair coin, if the coin lands heads you get paid 2 dollars, and if it lands tails you make nothing. How much money should you pay to play the game? Since half the time you make 2 dollars and the other half you make nothing, you can expect to make an average of a dollar per coin flip. You cannot get exactly a dollar back for a single coin flip, but if you played the game 20 times, you should expect to see about 20 dollars made.

Thus, you should definitely play your friend's game if they are charging less than a dollar to play.

3.3 Expectation

The expectation function finds the average return of a probability distribution by considering the total sum of any value's return multiplied by its probability of occurring. Mathematically speaking the equation looks as follows: $E(x) = (x \in X) \sum (x * P(x))$

3.4 Introducing Shannon Entropy

The Shannon Entropy of a given probability distribution is as follows $SE(X) = (x \in X) \sum (\log(\frac{1}{P(x)}) * P(x))$ Notice how the function looks just like the expectation equation, however, now x has been replaced by the value $\log(\frac{1}{P(x)})$.

$\log(\frac{1}{P(x)})$ is called the information content of an outcome in our probability space, or the event's surprise.

3.5 Interpreting the Shannon Entropy of an Amino Acid Distribution

DNA strings are extremely long, and it can be difficult to analyze and compare entire strings of nucleotides. Thus instead we compare the distribution of Amino Acids contained in the DNA string. This simplifies the process of modeling future proportions of Amino Acids following mutation rates.

Let's say we knew that every Methionine, given it mutates, converts to Isoleucine a third of the time. We cannot say after mutating Methionine becomes one-third part of Isoleucine, and two-thirds parts other amino acids. Thus the simplification of distributions is preferred when using Markov, state-changing, simulations.

Consider if we were to randomly choose an element in our amino acid distribution, if one-twentieth of those amino acids are Methionine, then the chance we choose a Methionine is also 1/20th. This makes every distribution a valid probability distribution for a random choice of amino acid, meaning it also has an associated Shannon Entropy. So what does the Shannon Entropy of this distribution represent, and why is it a valuable measure?

3.6 Encoding DNA sequences

Shannon Entropy being the minimum expected amount of questions needed to deduce an outcome in a probability distribution translates directly into optimizing space efficiency of encoding DNA sequences.

Consider that for our given DNA sequence we know that somehow 90 percent of the Amino Acids in our sequence are Methionine. Then instead of maybe having 5-bit strings, representing numbers between 0-31, and assigning Amino Acids to the first 20 possible numbers. We could potentially do 6 bit strings for every Amino Acid except Methionine, and have Methionine simply be the bit 1. This reduces the average bit length of our Amino Acids from 5 to 1.5, saving a lot of space in memory for our DNA sequences.

This is the intuitive strategy behind Huffman Encoding, and low Shannon Entropy for an Amino Acid distribution indicates a high level of compressibility.

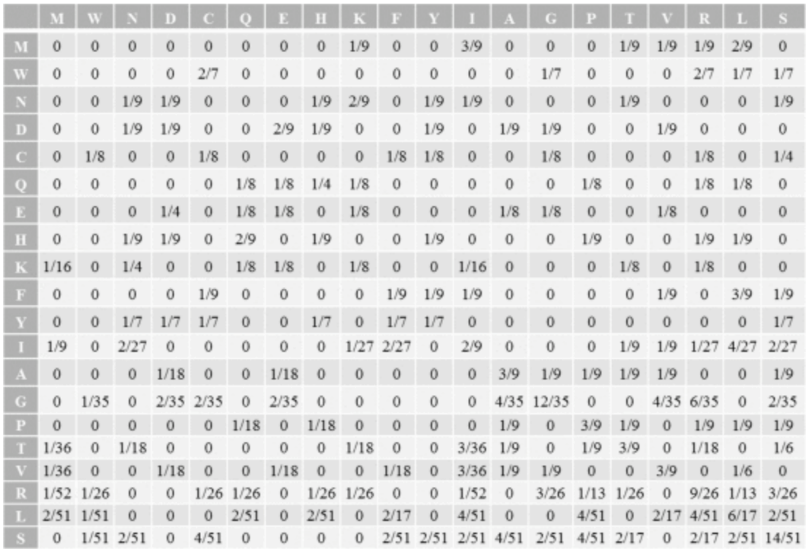
4 Method

4.1 Introduction to Markov Chain

Markov Chains model state changes of elements in a system, and are used for modeling the long-term expected states of elements in this system. The reasoning for the basis of our model comes from a paper[6] covering the long-term behavior of mutations of coding genes at an amino acid level. This paper utilizes Markov Matrices to project the long-term behavior of proteins. Markov chains have a plethora of names, including Markov matrices, transition Matrices, Markov processes, and stochastic processes. For this paper, they will be referred to as Markov Chains. They are a common statistical tool. In particular, they relate to the probability of future events depending only on present circumstances. To iterate in a statistics sense, Markov chains are built off the probability $P(A|B)$, given A and B are separate events.

Thorvaldsen's model works off the same principle, that the probability any one amino acid will mutate in the next cycle of replication is based only on the amino acid's current state. While protein replication is dependent on a massive number of factors, all of varying complexity, Thorvaldsen expresses there is precedence for simplifying the quality of mitosis to depend only on the current state of protein DNA.

When taking a larger view of the human body, the fact that proteins can be built from chains of hundreds of amino acids, and a significant portion of the human body is made of proteins, including bones, muscles, and organs, it becomes incredibly important in the context of mutations and the frequency with which mutations occur. Despite being a small number, it's a small number multiplied across millions of cells. The basis of the theory of evolution is that there is a chance of DNA mutating and thus having noticeable effects on daily life or human reproduction. This makes such research valuable to understanding.



	M	W	N	D	C	Q	E	H	K	F	Y	I	A	G	P	T	V	R	L	S
M	0	0	0	0	0	0	0	0	1/9	0	0	3/9	0	0	0	1/9	1/9	1/9	2/9	0
W	0	0	0	0	2/7	0	0	0	0	0	0	0	0	1/7	0	0	0	2/7	1/7	1/7
N	0	0	1/9	1/9	0	0	0	1/9	2/9	0	1/9	1/9	0	0	0	1/9	0	0	0	1/9
D	0	0	1/9	1/9	0	0	2/9	1/9	0	0	1/9	0	1/9	1/9	0	0	1/9	0	0	0
C	0	1/8	0	0	1/8	0	0	0	0	1/8	1/8	0	0	1/8	0	0	0	1/8	0	1/4
Q	0	0	0	0	0	1/8	1/8	1/4	1/8	0	0	0	0	0	1/8	0	0	1/8	1/8	0
E	0	0	0	1/4	0	1/8	1/8	0	1/8	0	0	0	1/8	1/8	0	0	1/8	0	0	0
H	0	0	1/9	1/9	0	2/9	0	1/9	0	0	1/9	0	0	0	1/9	0	0	1/9	1/9	0
K	1/16	0	1/4	0	0	1/8	1/8	0	1/8	0	0	1/16	0	0	0	1/8	0	1/8	0	0
F	0	0	0	0	1/9	0	0	0	0	1/9	1/9	1/9	0	0	0	0	1/9	0	3/9	1/9
Y	0	0	1/7	1/7	1/7	0	0	1/7	0	1/7	1/7	0	0	0	0	0	0	0	0	1/7
I	1/9	0	2/27	0	0	0	0	0	1/27	2/27	0	2/9	0	0	0	1/9	1/9	1/27	4/27	2/27
A	0	0	0	1/18	0	0	1/18	0	0	0	0	0	3/9	1/9	1/9	1/9	1/9	0	0	1/9
G	0	1/35	0	2/35	2/35	0	2/35	0	0	0	0	0	4/35	12/35	0	0	4/35	6/35	0	2/35
P	0	0	0	0	0	1/18	0	1/18	0	0	0	0	1/9	0	3/9	1/9	0	1/9	1/9	1/9
T	1/36	0	1/18	0	0	0	0	0	1/18	0	0	3/36	1/9	0	1/9	3/9	0	1/18	0	1/6
V	1/36	0	0	1/18	0	0	1/18	0	0	1/18	0	3/36	1/9	1/9	0	0	3/9	0	1/6	0
R	1/52	1/26	0	0	1/26	1/26	0	1/26	1/26	0	0	1/52	0	3/26	1/13	1/26	0	9/26	1/13	3/26
L	2/51	1/51	0	0	0	2/51	0	2/51	0	2/17	0	4/51	0	0	4/51	0	2/17	4/51	6/17	2/51
S	0	1/51	2/51	0	4/51	0	0	0	0	2/51	2/51	2/51	4/51	2/51	4/51	2/17	0	2/17	2/51	14/51

Information Entropy of Insulin Over Generations

4.2 Introduction to Information Entropy Analysis

Of particular interest is whether the amino acid sequence exhibits an inevitable, unidirectional increase in entropy as postulated by the genetic entropy theory. Or whether the patterns reveal more complex informational dynamics, with entropy increases and decreases reflecting the perpetual information flows and buffering mechanisms proposed by critics of the hypothesis.

The entropy is a measure of the degree of evenness of amino acid distribution; i.e., the larger this value, the more evenly amino acids are distributed. [3]. Observing changes in entropy should indicate changes in the informational complexity of resulting amino acid distributions. We are interested in finding which distributions converge in entropy as mutations occur, whether they share points of convergence, what Shannon entropies distributions converge to, and if generally entropy grows or diminishes as we simulate time passing and more mutations occur. Entropy growth indicates a more even distribution, while entropy diminishment indicates a less even distribution, implying more predictability, implying more pattern-following content in amino acid sequences.

5 Data & Results

5.1 Recreating Experiments

Model experiments were conducted in Matlab and R, and were conducted using the following mathematical models. The central 3 components for experimentation were:

- 20 dimensional vectors representing Amino Acid Distributions. For experimenting with random distributions, we generated random 20-dimensional vectors with all positive components and the sum of its components was 1.

$$\{\overrightarrow{X} \in R^{20} | \ \forall i \ x_i > 0 \ , \sum_{i=1}^{20} x_i = 1\}$$

- 20 x 20 Transition Matrix for Amino Acid Distributions. Codons rarely mutate, however, so we multiply the guaranteed mutation matrix, M, by the probability of mutation, μ. And then add the resulting matrix to the Identity matrix multiplied by $(1 - \mu)$. We experiment with the size of μ due to rounding concerns considering research suggested the paper's μ being 10^{-10}

$$A_{20\times 20} = M_{20\times 20} * \mu + I_{20\times 20} * (1 - \mu)$$

- Function that computes Shannon Entropy using probability distribution vectors

$$SE(\overrightarrow{X}) = \sum_{x_i \in \overrightarrow{X}} x_i \log_2(x_i)$$

5.2 What is the Distribution of Amino Acids of Insulin?

”M = 0.021708” ”W = 0.014472” ”N = 0.052098” ”D = 0.050651” ”C = 0.034009” ”Q = 0.027496” ”E = 0.0767” ”H = 0.026773” ”K = 0.04848” ”F = 0.042692” ”Y = 0.036903” ”I = 0.047757” ”A = 0.047757” ”G = 0.06657” ”P = 0.059334” ”T = 0.049928” ”V = 0.062952” ”R = 0.059334” ”L = 0.098408” ”S = 0.075977”

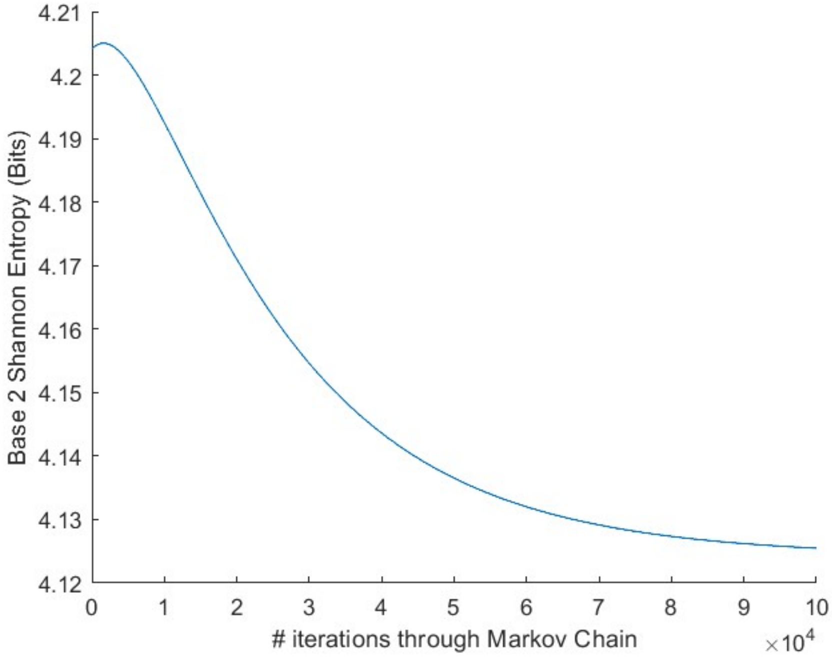
5.3 What is the Shannon Entropy of the Insulin Amino Acid Distribution?

4.2041 bits

On its own, this value does not share a lot of information, we are more curious about how the entropy changes with mutations. Then we can make conclusions about whether we are losing average information content, or gaining average information content.

5.4 How does the Shannon Entropy of Insulin change as it mutates?

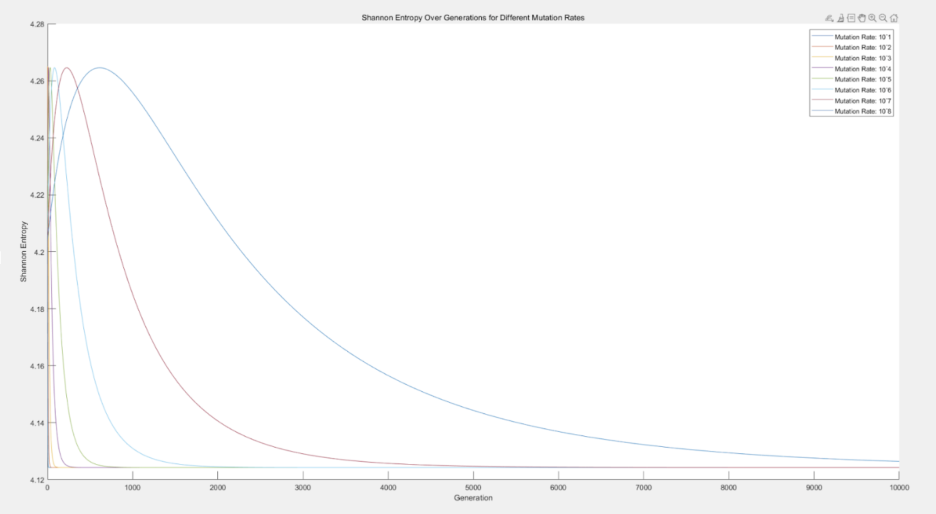
Using a mutation rate of 10^{-4}, as opposed to the suggested 10^{-10}, we get the following convergence over 10^5 iterations.



To address the far larger probability of mutation, we must compare how convergence changes for different orders of magnitude of mutation chance.

5.5 Do different mutation rates affect the point of convergence? Do different mutation rates affect the rate of convergence of Shannon Entropy?

We notice that higher mutation rates do not change the point of convergence, and simply require more iterations before the entropy fully converges. Additionally, because we have finite levels of accuracy for double-type numbers in the computer's memory, thus we can approximate the long-term behavior of Amino Acid Distributions by increasing the probability of mutation.

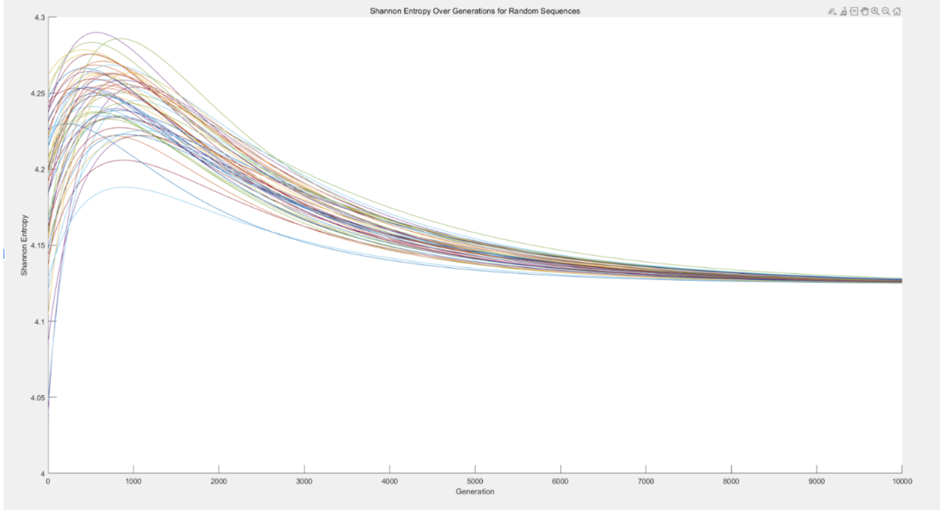


Shannon Entropy of Different Mutation Rates over Time

While it did not provide any insight into the behavior of insulin, this graph was important to provide for a number of reasons. Notably, most of the mutation rates above 10^4 converge to the same value, just above 4.12, and reach the same maximum, just above 4.26. This indicates that the Shannon Entropy of these mutation rates is either sped up or slowed down without affecting the minimum and maximum entropy values, just the generation at which they reach it. For ease of readability, we decided to use 10^7 for most of our mutation rates, as it gave a relatively fast convergence over enough generations to get clean, readable data. It was chosen semi-arbitrarily but was relatively close to what was used in Thorvaldsen's paper, $5 \cdot 10^{-5}$.

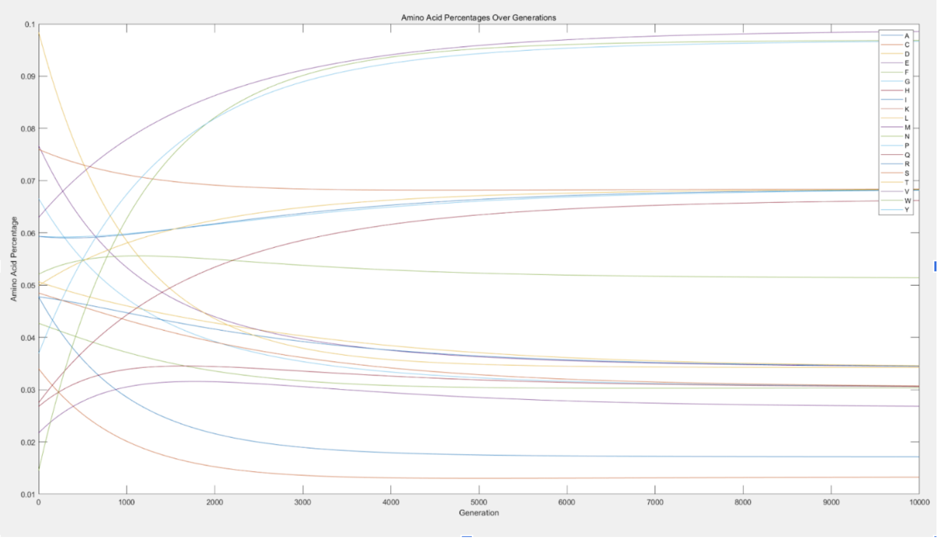
5.6 How do Shannon Entropy convergence rates change for randomly generated distributions of Amino Acids?

For a mutation rate of $= 10^{-3}$, we can observe no commonly shared point of convergent Shannon Entropy

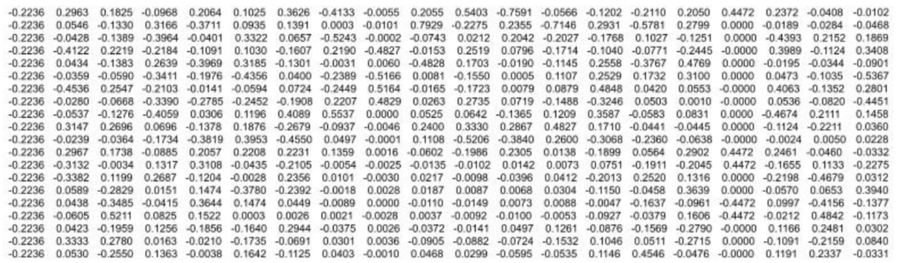


Information entropy of different randomly generated distributions

5.7 Markov-Chain-Specific Analysis



Analysis of Percentage of Amino Acids of a Single Protein (Insulin) Over Time



-0.2236	0.2963	0.1825	-0.0968	0.2064	0.1025	0.3626	-0.4133	-0.0055	0.2055	0.5403	-0.7591	-0.0566	-0.1202	-0.2110	0.2050	0.4472	0.2372	-0.0408	-0.0102
-0.2236	0.0546	-0.1330	0.3166	-0.3711	0.0935	0.1391	0.0003	-0.0101	0.7929	-0.2275	0.2355	-0.7146	0.2931	-0.5781	0.2799	0.0000	-0.0189	-0.0284	-0.0468
-0.2236	-0.0428	-0.1389	-0.3964	-0.0401	0.3322	0.0657	-0.5243	-0.0002	-0.0743	0.0212	0.2042	-0.2027	-0.1768	0.1027	-0.1251	0.0000	-0.4393	0.2152	0.1869
-0.2236	-0.4122	0.2219	-0.2184	-0.1091	0.1030	-0.1607	0.2190	-0.4827	-0.0153	0.2519	0.0796	-0.1714	-0.1040	-0.0771	-0.2445	-0.0000	0.3989	-0.1124	0.3408
-0.2236	0.0434	-0.1383	0.2639	-0.3969	0.3185	-0.1301	-0.0031	0.0060	-0.4828	0.1703	-0.0190	-0.1145	0.2558	-0.3767	0.4769	0.0000	-0.0195	-0.0344	-0.0901
-0.2236	-0.0359	-0.0590	-0.3411	-0.1976	-0.4356	0.0400	-0.2389	-0.5166	0.0081	-0.1550	0.0005	0.1107	0.2529	0.1732	0.3100	0.0000	0.0473	-0.1035	-0.5367
-0.2236	-0.4536	0.2547	-0.2103	-0.0141	-0.0594	0.0724	-0.2449	0.5164	-0.0165	-0.1723	0.0079	0.0879	0.4848	0.0420	0.0553	-0.0000	0.4063	-0.1352	0.2801
-0.2236	-0.0280	-0.0668	-0.3390	-0.2785	-0.2452	-0.1908	0.2207	0.4829	0.0263	0.2735	0.0719	-0.1488	-0.3246	0.0503	0.0010	-0.0000	0.0536	-0.0820	-0.4451
-0.2236	-0.0537	-0.1276	-0.4059	0.0306	0.1196	0.4089	0.5537	0.0000	0.0525	0.0642	-0.1365	0.1209	0.3587	-0.0583	0.0831	0.0000	-0.4674	0.2111	0.1458
-0.2236	0.3147	0.2696	0.0696	-0.1378	0.1876	-0.2679	-0.0937	-0.0046	0.2400	0.3330	0.2867	0.4827	0.1710	-0.0441	-0.0445	0.0000	-0.1124	-0.2211	0.0360
-0.2236	-0.0239	-0.0364	-0.1734	-0.3819	0.3953	-0.4550	0.0497	-0.0001	0.1108	-0.5206	-0.3840	0.2600	-0.3068	-0.2360	-0.0638	-0.0000	-0.0024	0.0050	0.0228
-0.2236	0.2967	0.1738	-0.0885	0.2057	0.2208	0.2231	0.1359	0.0016	-0.0602	-0.1986	0.2305	0.0138	-0.1899	0.0564	0.2902	0.4472	0.2461	-0.0460	-0.0332
-0.2236	-0.3132	-0.0034	0.1317	0.3108	-0.0435	-0.2105	-0.0054	-0.0025	-0.0135	-0.0102	0.0142	0.0073	0.0751	-0.1911	-0.2045	0.4472	-0.1655	0.1133	-0.2275
-0.2236	-0.3382	0.1199	0.2687	-0.1204	-0.0028	0.2356	0.0101	-0.0030	0.0217	-0.0098	-0.0396	0.0412	-0.2013	0.2520	0.1316	0.0000	-0.2198	-0.4679	0.0312
-0.2236	0.0589	-0.2829	0.0151	0.1474	-0.3780	-0.2392	-0.0018	0.0028	0.0187	0.0087	0.0068	0.0304	-0.1150	-0.0458	0.3639	0.0000	-0.0570	0.0653	0.3940
-0.2236	0.0438	-0.3485	-0.0415	0.3644	0.1474	0.0449	-0.0089	0.0000	-0.0110	-0.0149	0.0073	0.0088	-0.0047	-0.1637	-0.0961	-0.4472	0.0997	-0.4156	-0.1377
-0.2236	-0.0605	0.5211	0.0825	0.1522	0.0003	0.0026	0.0021	-0.0028	0.0037	-0.0092	-0.0100	-0.0053	-0.0927	-0.0379	0.1606	-0.4472	-0.0212	0.4842	-0.1173
-0.2236	0.0423	-0.1959	0.1256	-0.1856	-0.1640	0.2944	-0.0375	0.0026	-0.0372	-0.0141	0.0497	0.1261	-0.0876	-0.1569	-0.2790	-0.0000	0.1166	0.2481	0.0302
-0.2236	0.3333	0.2780	0.0163	-0.0210	-0.1735	-0.0691	0.0301	0.0036	-0.0905	-0.0882	-0.0724	-0.1532	0.1046	0.0511	-0.2715	0.0000	-0.1091	-0.2159	0.0840
-0.2236	0.0530	-0.2550	0.1363	-0.0038	0.1642	-0.1125	0.0403	-0.0010	0.0468	0.0299	-0.0595	-0.0535	0.1146	0.4546	-0.0476	-0.0000	0.1191	0.2337	-0.0331

Transition Matrix Eigenvectors

6 Conclusion

6.1 What distributions of Amino Acids converge the quickest?

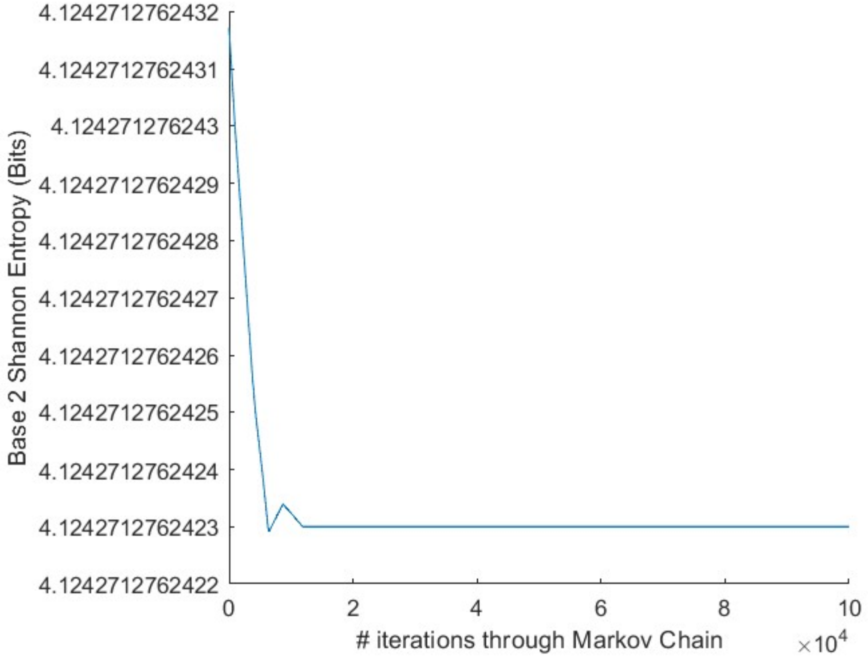
Since all Amino Acid Distributions seemed to converge to a specific entropy, this implies there is an associated distribution that distributions converge to.

To approximate this distribution, we multiplied the Transition Matrix by itself many times, and multiplied the insulin vector by this longer term mutation matrix.

$$\lim_{n \to \infty} vA^n = v_\infty$$

We are curious if any real DNA sequences have this distribution of Amino Acids. The ordering is the same as the previously described Insulin Vector

0.0171, 0.0133, 0.0342, 0.0342, 0.0304, 0.0304, 0.0304, 0.0342, 0.0304, 0.0342, 0.0266, 0.0513, 0.0684, 0.0665, 0.0684, 0.0684, 0.0684, 0.0989, 0.0970, 0.0970



Noticing the scale on the left, we can say that this distribution nearly converges quickest.

As a retrospective, this model has the capacity to be a versatile tool, given enough time and focus. It could certainly be refined beyond the limitations of this paper, and be focused to better suit the specific circumstances that come with various different proteins.

We began this project by utilizing the model from a crucial research paper and utilizing Matlab, R, and Python, and while programming the basics went smoothly, refining the program was anything but. Multiple iterations across various platforms meant trial and error was abundant, but we ended up with a working model in the end.

6.2 Does the data fit the hypothesis?

For the first set of generations, the beginning portion of the data generally fits the hypothesis. Our study found that the Shannon entropy of the insulin amino acid sequence initially increased over time, indicating a rise in randomness and a decrease in informational content due to mutations. This aligns with the genetic entropy hypothesis proposed by Sanford and Baumgardner.

However, it is important to note that the entropy values eventually flattened out, reaching a point of convergence above the initial entropy level. This suggests that while the informational content decreases as mutations accumulate, the process eventually stabilizes rather than continuing indefinitely toward higher entropy.

This differs from our hypothesis, as we expected entropy to increase in a unidirectional manner, towards some positive asymptote with no additional deviation. The fluctuations of the level of entropy support the idea that genomes experience a decline in informational content due to the accumulation of slightly harmful mutations, but they also indicate the presence of mechanisms that prevent entropy from increasing indefinitely. This stabilization could be due to factors such as natural selection or other evolutionary dynamics not fully captured in the model.

7 Discussion

7.1 Limitation of the model

Although our model provides valuable insights into the process of genetic entropy, it is important to acknowledge several limitations:

Simplifying Assumptions: Our model assumes that each codon can only undergo one substitution mutation and excludes frameshift mutations. Additionally, it assumes all viable mutants have the same reproductive efficiency. These simplifications cannot fully capture the complexity of actual biological systems where multiple types of mutations and varying fitness effects occur, both for various humans and animals and for individual protein groups of those people and animals.

Mutation Rates: As an extension of the previous part, we used a fixed mutation rate for our simulations, but in reality, mutation rates can vary depending on environmental factors and different genomic regions. Despite our model's evidence showing otherwise, there is a real possibility the actual mutation rate could affect the rate and extent of entropy changes observed. This is doubly true when considering reproduction, as a fast mutation rate in an organism could reduce the chance of reaching the age of reproduction.

Ignoring Natural Selection: Our model primarily focuses on the accumulation of mutations and their impact on entropy. It does not account for the effects of natural selection, which can play a significant role in removing deleterious mutations and promoting beneficial ones, thereby affecting the overall entropy of the sequence.

Computational Limits: Due to the finite precision of floating-point arithmetic in computers, especially when dealing with very small mutation rates, rounding errors could influence the results. This limitation could affect the accuracy of the long-term behavior and convergence points of the entropy calculations.

This model is useful for the long-term behavior of isolated protein replication, making it sub-optimal for analysis for general information on human health. It assumes mutation rates of Amino acids are completely independent from factors beyond the individual amino acid's initial state as a specific amino acid. Given the relatively recent development of being able to analyze individual genetic sequences, this method would be much more suited to highly specific tailoring, to better match the most likely circumstances and context that occur within and around a specific protein.

8 Bibliography

References

[1] A. Chao, et al., *"Expected Shannon Entropy and Shannon Differentiation between Subpopulations for Neutral Genes under the Finite Island Model,"* PLoS ONE (2015)

[2] S. Copling, et al., *"Understanding Model Independent Genetic Mutations through Trends in Increase in Entropy,"* Scientific Research Publishing (2022)

[3] M. Hasegawa, and TA. Yano, *"Entropy of the Genetic Information and Evolution,"* Origins Life Evol Biosphere (1975)

[4] J. Sanford, *"Genetic Entropy,"* FMS Publications (2014)

[5] W. Sherwin, *"Entropy and Information Approaches to Genetic Diversity and its Expression: Genomic Geography,"* Entropy (2010)

[6] S. Thorvaldsen, *"A Mutation Model from First Principles of the Genetic Code,"* IEEE/ACM Trans Comput Biol Bioinform (2016)

[7] Shannon, C.E. *"A Mathematical Theory of Communication."* Bell System Technical Journal, 27, 379-423.(1948)

[8] Harchol-Balter, M. *"Introduction to Probability for Computing."* Cambridge: Cambridge University Press. (2023)

[9] *"P06213 · INSR_HUMAN"* UniProt