.EvoSim Comparison: Benchmarking INDELible with Novel Evolutionary Tools Analysis

Taqwa Azba

Supervisors: Naiel Jabareen and Prof. Tal Pupko

Date: June 2025

Introduction

Zipf. Imagine you're at a massive music festival, and you're curious about which songs people are vibing to the most. The headlining artist's hit track is playing on repeat everywhere you go—it's the *"most frequent"* sound in the air. The second-most popular song? Played a lot, but not nearly as much. And that obscure indie tunes your friend loves? You might hear it once, if at all. This imbalance in popularity isn't just a festival quirk—it's a phenomenon beautifully captured by the Zipf distribution, named so after linguist George Zipf (fig. 1), who first noticed this pattern studying word frequencies. The Zipf distribution is all about how a few things dominate, while the rest barely make a dent. George Zipf discovered that the most common word in English ("the") is used substantially more than the second-most common ("of"), and so on.

Figure 1. George Zipf (1902-1950). Zipf's discovered the Zipf's law in 1935.

Mathematical model for the Zipf's distribution. Let k denote the rank of a word. Assume that the probability to find this word in a text is proportional to its rank, raised to a power s (the exponent parameter). Mathematically: $P(k) \propto \frac{1}{k^s}$. In order for this to be a distribution, the sum over all ranks should equal 1, and thus, we normalize over all the words that are found in the text. Assume that the total number of words is N. We obtain:

$$P(k) = \frac{\frac{1}{k^s}}{\sum_{i=1}^{N} \frac{1}{i^s}}$$

where s>0 controls the skewness of the distribution. The observed frequencies may slightly differ from the theoretical probabilities, but if there is a good fit between the frequencies derived from the data and the computed probabilities, we infer that the data follow a power-law distribution, dictated by the above equation.

Scale-Invariance. For many types of data, the fit to the power-law distribution remains unchanged regardless of scale—whether examining an entire language or a single text. The ranking proportions remain consistent, a phenomenon called scale-invariance.

Why Does This Happen? The Zipf distribution is nature's way of saying, "Winner's win." Once something becomes popular, it keeps getting more attention—a sort of a snowball effect. This happens because:

- People follow trends: "Everyone's listening to this song, so I will too!"
- Resources are limited: You can't give every book or song equal time.
- Pareto Principle (80/20 Rule): A small fraction of things always accounts for most of the action.

An Example: Distribution of Words in English Text. The following sentence represents the top 20 most common English words by their order:

"THE OF AND TO A IN IS I THAT IT FOR YOU WAS WITH ON AS HAVE BUT BE THEY"

Whether the words are ranked across an entire language, or just one book or an article, almost every time the following pattern emerges. The second most used word will appear around half as often as the most used, the third one third as often. The fourth, $\frac{1}{4}$ as often, The fifth $\frac{1}{5}$ as often and so on...all the way down (fig. 2)

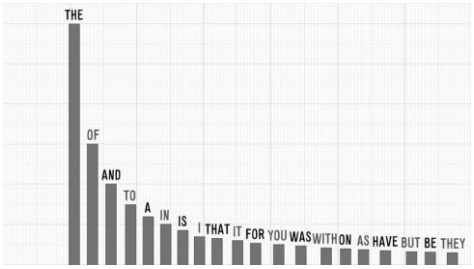


Figure 2. The distribution of English words in a large number of English tests.

More generally, the number of times a word is used is proportional to $\frac{1}{rank}$. Thus, word frequency and ranking fit "a power-law distribution with $s = 1$. In a log-log graph, one obtains a straight line (fig 3). It is this relationship that is called the Zipf's law. Of note, this phenomenon does not only apply to English. Rather, it is valid for many other languages, even ancient ones that were not yet translated.

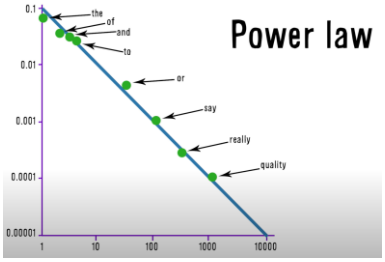
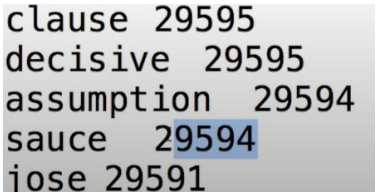


Figure 3. The X axis represents the rank of items, while the Y axis represents the frequency or probability.

According to wordcount.org which ranks words as found in the British national corpus, "sauce" is the 5555th most common English word (fig 4a). Consider the frequency of every word in Wikipedia and in the entire Gutenberg corpse of thousands of public domain books. The most frequent word," the", shows up about 181 million times (fig 4b). Knowing this, we can estimate that the word sauce should appear ($f(r)=f(1)x\frac{1}{r}$= 181 million x $\frac{1}{5,555}$) = 30,000 times on Wikipedia and Gutenberg combined and it pretty much does it appears 29,254 times.

Zipfs law does not only describe language behavior but also found in city population, solar flare intensities, protein sequences and a lot more.

(a)



(b)

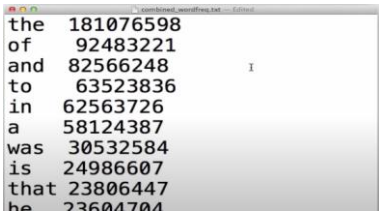


Figure 4. (a) The frequency of the word "sauce"; (b) The most common words.

Zipf Distribution: a Biological Perspective. The Zipf distribution was shown to fit various distributions in biology.

1. Genomics and DNA Sequences. In the study of DNA and genomes, the Zipf distribution emerges when looking at the frequency of specific nucleotide sequences. For example, some codons are used substantially more frequently than others to encode amino acids in proteins, a phenomenon called codon usage, e.g., the codons for common amino acids like leucine or serine are overrepresented compared to others. In addition, in non-coding regions of DNA, certain repeated sequences, e.g., ATATAT or GCGCGC occur more frequently, while many others are rare. The understanding that these patterns follow the Zipf's law may help elucidate the uneven distribution of sequences and may help biologists identify regions of functional importance, such as highly conserved genes or regulatory elements.

2. Protein Frequencies. Proteins follow a Zipf-like distribution in terms of their abundance in cells. A few proteins, such as enzymes critical for metabolism, are produced in large quantities. Most proteins, including some signaling molecules or regulatory proteins, are present in much smaller amounts. This helps cells allocate resources efficiently, focusing on what's most needed for survival.

3. Word Frequency in Genetics. Gene expression levels behave like "words" in a text. Highly expressed genes (the "the" or "and" of the genome) dominate transcription. Genes are expressed at much lower levels, like rare words in a novel. This idea is particularly useful in transcriptomics, where researchers study which genes are turned on or off in different conditions.

4. Microbiome. The gut microbiome is a classic example. A few bacterial species, like *Bacteroides* and *Firmicutes*, are incredibly abundant. Most other species exist in tiny amounts, forming a "long tail" of rare microbes. Understanding this distribution helps researchers identify dominant species responsible for health or disease.

5. Neuroscience. In the brain, the Zipf distribution appears in neuron firing rates. A few neurons fire frequently to handle critical tasks, while most remain relatively quiet. In addition, the Zipf's law is found in cognitive processes. Like language (where Zipf was first discovered), neural activity in tasks like memory retrieval or problem-solving exhibits power-law behavior.

6. Population Genetics. In populations of organisms, Zipf shows up when studying gene variants: a few alleles (versions of a gene) dominate within a population, while most are rare (Rosenberg 2003, Reed and Hughes 2002). In addition, the Zip's law also appears when studying adaptive traits: Certain traits are highly selected and widespread, while others are more niche or less advantageous.

Phylogeny. A phylogenetic tree is a diagram that represents the evolutionary relationships among various species or other entities based on their shared ancestry. It illustrates how species or genes have evolved from a common ancestor over time. The Root Represents the common ancestor of all species (or sequences) in the tree. The Branches are the lines connecting nodes in the tree. Branches indicate evolutionary pathways and can have lengths proportional to the amount of evolutionary change (e.g., genetic differences or time). Nodes can be divided in two types. Internal nodes represent common ancestors that gave rise to descendant species. External nodes (leaves or tips) represent the species, individuals, or genes being studied. In some trees, branch lengths correspond to evolutionary distances or time since divergence. An Evolutionary distance measures how different two species (or genes) are in terms of their evolutionary history. It tells us how long ago two organisms shared a common ancestor or how many genetic changes (mutations) occurred between them (fig 5).

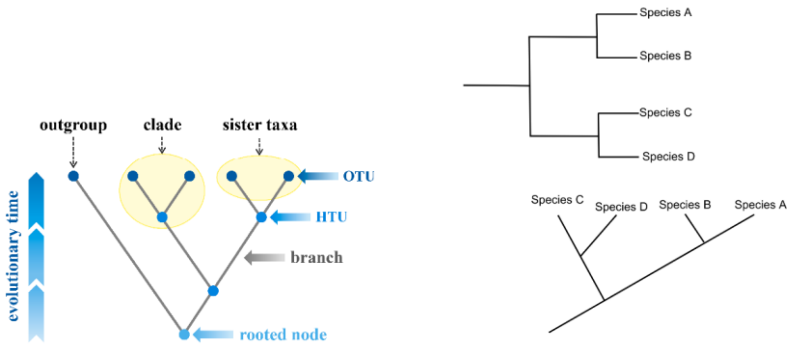


Fig 5. <write a short description pls>

Commented [TP1]: missing

Evolutionary Models. Substitution models incorporate exchanges among nucleotides or amino acids. Several models exist, e.g., the Jukes-Cantor (JC69), Kimura 2-parameter, and others. They are used to simulate changes in the sequence over time, without changing its length. In contrast, indel models are used to introduce indels using customizable rate parameters and length distributions. e.g., Zipfian. Sequences evolve along the branches of a user-defined phylogenetic tree.

INDELible. INDELible (Fletcher and Yang 2009) is a computer program used to simulate the evolution of biological sequences. It is specifically designed for generating realistic DNA, RNA, or protein sequences by modeling evolutionary processes, including substitutions, insertions, and deletions (collectively known as indels). INDELible

generates sequences at the leaves of a phylogenetic tree, by simulating how they evolved along the tree. The process starts when the program generates a root sequence. This sequence then evolves along the branches of the tree based on specified evolutionary models. The program produces a Multiple Sequence Alignment (MSA) file in FASTA format that contains the sequences of all species (or taxa) at the tree leaves. Simulated MSA can be used for downstream analyses, such as testing phylogenetic inference methods or evaluating alignment accuracy.

References:

Rosenberg, N. A. (2003) — *The shapes of neutral gene genealogies in two species: probabilities of monophyly, paraphyly, and polyphyly in a coalescent model* (Evolution, 57(7), pp. 1465–1477):

Reed, W. J. & Hughes, B. D. (2002) — *From gene families and genera to incomes and internet file sizes: Why power laws are so common in nature* (Physical Review E 66, 067103): pubmed.ncbi.nlm.nih.gov+15link.aps.org+15scholar.google.com.au+15

Fletcher, W., & Yang, Z. (2009). INDELible: A flexible simulator of biological sequence evolution. *Molecular Biology and Evolution, 26*(8), 1879–1888. https://doi.org/10.1093/molbev/msp098