Perplexity And Statistical Energy: Bridging the Gap Between Language and Proteins

Authors: Francisco McGee, Marija Stanojevic, Avik Biswas

Language-based machine learning approaches are redefining the state of the art in generative protein sequence modeling. While seemingly quite effective, there are many unanswered questions regarding the shared features of datasets composed of sequences of words, or alternatively sequences of amino acids, allowing them to be learned by the same generative models. Here, by correlating perplexity measurements from natural language processing (NLP) with statistical energy *E(S)* measurements from biophysics, we explore the notion that it is information entropy that links language to proteins, providing novel insights into the grammar of life.

Perplexity measurements for sentences are well-established in NLP, whereas *E(S)* for proteins in structural bioinformatics is a more recent development. While both measurements are entropy-based, their possible equivalency has not yet been demonstrated. Here, on two synthetic datasets, HIV proteases and the kinase superfamily of proteins, each generated from pairwise Potts Hamiltonian models, we demonstrate the correlation between these two entropy-based measurements in a comparative analysis of generative protein sequence models (GPSMs), at the fundamental level of first- and second-order terms. Each GPSM was trained on the same synthetic multiple sequence alignment (MSA), and made to generate an evaluation MSA. Perplexity measurements were computed between target and evaluation MSAs. *E(S)* represents the predicted probability (“prevalence”) of individual sequences in the dataset. We show that unigram perplexity (*PP-U)* correlates most strongly with *E(S)* of a single-site protein model (Indep), whereas the combined unigram+bigram perplexity (*PP-UB)* correlates strongly with that of the pairwise Potts Hamiltonian protein model (Mi3). For Indep and Mi3, these findings are mathematically explained by the functional forms of their Hamiltonians. In the pairwise instance of Mi3, it is the sum of the sitewise "fields" and pairwise "coupling" terms, reflecting the functional form of *PP-UB*, while for Indep, the Hamiltonia includes only field terms, reflecting *PP-U*.

Correspondence between perplexity and *E(S)* provides much needed explanatory power for the aforementioned trend in GPSMs. Having established that proteins can be measured as sentences at the most fundamental level, it is implied that more sophisticated perplexity measurements might be applicable to the optimization of GPSMs. Conversely, metrics and methods specific to GPSMs might be applicable to NLP. This convergence of language and proteins, demonstrated by our simple and intuitive comparative analysis, suggests some amount of fungibility between the two types of data, and provides essential theoretical foundation for future developments into the grammar of life.