Differences in usage of local combinations of amino acids in various genomes. J. Klein-Seetharaman, M. Ganapathraju, A. Patel, R. Rosenfeld, J. Carbonell and R. Reddy. School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15217, USA.

Detection of linguistic features in biological sequence data based on the distribution of n-grams, i.e. sequences of amino acids of specified length, is controversial and the extent to which amino acid sequences can be modeled stochastically is not clear. Thus, the previously applied Zipf law and Shannon entropy calculations cannot determine alone the rules that distinguish protein sequences capable of folding to three-dimensional structures from those that cannot. Here we present a detailed analysis of the distribution of amino acid n-grams in 43 bacterial and archaeal genomes and the human draft sequence. For each genome, the top most frequently occurring n-grams were determined for n varying between 1 and 4, and compared with the corresponding frequency in the other 43 genomes. Striking differences were observed. For example, in Neisseria meningitidis the 4-grams MPSE, SDGI and GRLK occur amongst the top 20 most frequently used 4-grams, while their frequency is dramatically lower in all other organisms. This observation suggests that there are distinct features in the “languages” used by the different organisms, which may lead to developing novel antimicrobial agents based on specifically “talking” to a pathogen in the presence of its host.

Name of the presenting author	:	J. Klein-Seetharaman
Address	:	School of Computer Science Carnegie Mellon University Pittsburgh, PA 15217, USA
Email	:	judithks@cs.cmu.edu
Fax	:	412 683 5348