

**Comparative n-gram analysis of genome sequences. M. Ganapathraju, J. Klein-Seetharaman, J. Carbonell, R. Rosenfeld and R. Reddy. School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15217, USA.**

Understanding the relationship between primary sequence and tertiary structure of proteins is one of the major unsolved questions in bioinformatics research. Formulating rules that relate sequence and structure is conceptionally similar to text and meaning that characterize human languages. Critical to language analysis is n-gram word determination, which is used to generate statistical language models. Here we describe the development of a tool-kit that applies these methods to the study of protein sequences of the genomes of different organisms. The amino acids are treated as words, since each amino acid carries a chemical “meaning”. The difference in string length and vocabulary size between biological and human languages requires adaptation of the computational approaches to n-gram analysis of protein sequences. Optimum performance was achieved when the protein sequences of the different genomes were presorted and stored in suffix arrays. This allows rapid extraction of statistical information by using binary search. Thus, the distribution, length and position of n-gram exact and approximate matches can be determined, within individual organisms and in comparison between organisms. Analysis of the statistical data to extract information for future language models of genome sequences is underway (see accompanying abstract).

Name of the presenting author	:	M. Ganapathiraju
Address	:	School of Computer Science Carnegie Mellon University Pittsburgh, PA 15217, USA
Email	:	madhavi@cs.cmu.edu
Fax	:	412 683 5348