Poster H-53

Using latent semantic indexing (LSI) to evaluate the similarity of sets of sequences without multiples alignments character-by-character

Authors:
Braulio RGM Couto (*UNI-BH*)
Ana Paula Ladeira (*UNI-BH*)

Short Abstract: We present an algorithm that estimates relatedness between large numbers of biomolecules without the requirement of multiples alignments. Proteins recoded as n-peptide frequency values using all possible overlapping n-peptides, which generates a matrix, reduced by SVD. Cosine and Euclidean distance between proteins vectors are used as similarity measure.

Long Abstract:
In this paper, we present an algorithm that estimates relatedness between large numbers of biomolecules without the requirement of multiples sequences alignment. Protein sequences are recoded as n-peptide frequency values using all possible overlapping n-peptides, which generates a frequency matrix that is reduced by SVD - singular value decomposition. Matrices built using one, two, three and four peptides were tested. These sparse matrices have dimensions of 20xm, 400xm, 8,000xm and 160,000xm, respectively, where m is the number of sequences studied. Each protein is analyzed as a vector in a high-dimensional space. The similarity among sequences can be done by using the cosine or the Euclidean distance between proteins vectors. The first database evaluated had 64 vertebrate mitochondrial genomes composed of 832 total proteins from 13 known gene families (ATP6, ATP8, COX1, COX2, COX3, CYTB, ND1, ND2, ND3, ND4, ND4L, ND5 and ND6). The second database was composed by a sample of sequences from nine types of proteins, randomly retrieved from GenBank: 100 sequences of each type of protein (globin, cytochrome, histone, cyclohydrolase, pyrophosphatase, ferredoxin, keratin and collagen) and 200 other proteins from lymphocytes and bacteriophages, totaling 1,000 sequences. Firstly, we assessed the correlation between the cosine, the Euclidean distance and a sequence alignment measure: 208 sequences from the first database and 200 from the second set were randomly selected and compared by using the global edit distance between each pair of sequence and respective cosines and Euclidean distances. This pairwise analysis generated 41,428 similarity measures. Despite the fact that we worked with quite different methods (LSI and global distance alignment), the correlation between the cosine and edit distance was -0.29 (p-value < 0.01) and between the Euclidean distance and edit distance was +0.70 (p-value < 0.01), which indicates that Euclidean distance is better than the cosine to evaluate the similarity of sequences without multiples alignments character-by-character. In order to validate the ability of LSI to classify the sequences according to their categories, we used a sample of 202 randomly sequences from the 13 genes families as queries and the rest of proteins (630) were used to generated the n-peptide frequency matrix. The best result was achieved with a 3-peptide frequency matrix (size of 8,000 rows and 630 columns) , reconstruct by SVD with 28 terms: all 202 queries were correctly classified in each 13 genes

families (accuracy = 100%). For the second database, 735 sequences were selected to build the n-peptide frequency matrices and 265 proteins were randomly selected as queries. By using a 3-peptide frequency matrix (size of 8,000 rows and 735 columns), reconstruct by SVD with 32 terms, we got a global accuracy of 71% in classifying the 265 queries in one of the nine proteins categories. We had 100% accuracy for cytochrome, 92% for histone, 85% for keratin, 80% for globin, 74% for collagen, 66% for cyclohydrolase, 55% for pyrophosphatase, 52% for ferredoxin, and 61% for other proteins. These results show that to apply latent semantic indexing (LSI) to evaluate the similarity of sets of sequences is a promising method and very attractive, because sequence alignments are neither generated nor required. Euclidean distance between SVD vectors has a higher correlation with edit distance between sequences than the cosine. In order to achieve similar results as that observed by using edit distance analysis, we recommend that Euclidean distance must be used as a similarity measure for proteins sequences in LSI methods.