Mapping Protein Sequences to Feature Spaces for Efficient Indexing

Ahmet Sacan
ahmet@ceng.metu.edu.tr
Dept. of Computer Eng.
Middle East Technical University

Searching for homologous sequences in genomic and proteomic databases has become a frequent task of the molecular biologist. As the size of these databases increase rapidly, rendering the current methods costly in terms of time and memory requirements, it is desirable to build efficient index structures allowing fast similarity searches. In previous studies we've developed efficient index structures and sequence transformations to achieve fast querying of the genome data sets. In this work, we propose methods that can be used to map the protein sequences to feature vector spaces in order to facilitate the building of index structures that have successful filtering capabilities. The proposed methods for sequence indexing exploit the similarities between amino acid residues and the biochemical properties of the amino acids, which we expect would give better results in homology searching.

Motivation

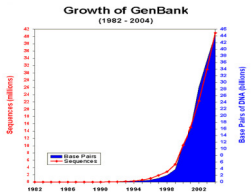
Tasks in Bioinformatics:

- 35.2%: Sequence Similarity
  - 40.5%: Protein sequence search
  - 33.4%: DNA sequence search
  - 26.1%: unspecified sequence type

. S. C. Goble, P. Baker, and A. Brass. A classification of tasks in bioinformatics. Bioinformatics, 17:180–188, 2001.

Challenges (esp. for Proteins)

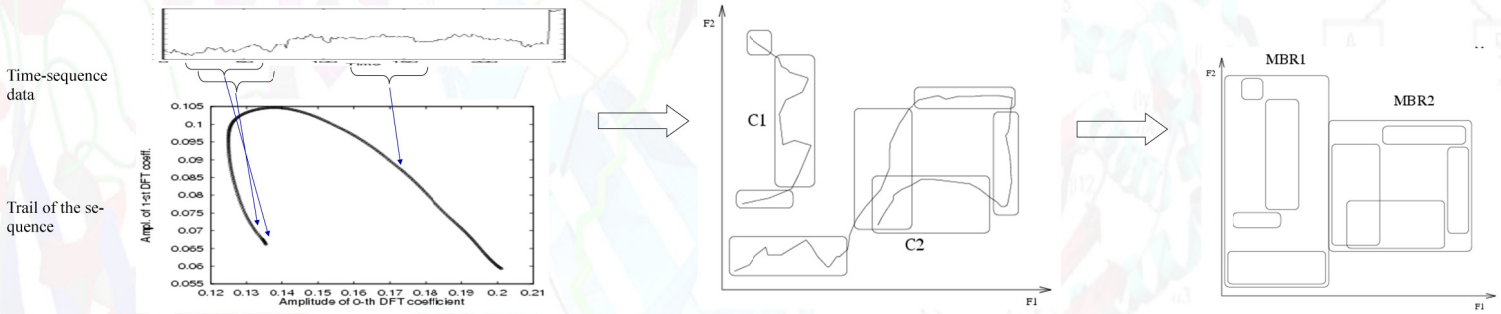
- Handle growing size of the sequence databases
- Accommodate bigger alphabet size of the proteins
- Model inter-residue similarity
- Find local similarity (not whole-query matching)
- Find distant homologs without incurring high costs
- Perform all of the above efficiently and accurately



NCBI Genbank Statistics, Feb 2004, http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

Inspiration: Subsequence Matching in Time-Series Databases

- Obtain frequency-domain trail of the sequence using sliding window over the time series data
- Divide trails into subtrails
- Construct MBR's covering the subtrails
- Store MBR's using Spatial Access Methods (e.g., R*-tree)



. Faloutsos, M. Ranganathan, and Y. Manolopoulos. Fast Subsequence Matching in Time- Series Databases. Proceedings of ACM SIGMOD, Minneapolis, MN, pages 419-429, May, 1994.

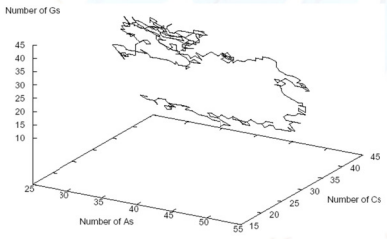
Previous Work on DNA Sequences

Frequency Features

- Use frequency vectors (count of each nucleotide within the window)
- Define a Frequency Distance (FD) from frequency vectors, which always underestimates Edit Distance (ED).

Tamer Kahveci and Ambuj K. Singh, An Efficient Index Structure for String Databases, VLDB, 2001, September, Roma, Italy.

Sequence Trail in Feature Space



Refining Features & Distances

a. Wavelet Transformation (Kahveci and Singh, 2001)
   - Defined similar to Haar wavelet
   - A distance function (WD) defined from 1st and 2nd wavelet coefficients.
   - WD also underestimates ED.

b. Can use n-tuple alphabet to increase sensitivity. (O. Ozturk and H. Ferhatosmanoglu. Effective Indexing and Filtering for Similarity Search in Large Biosequence Databases. O. IEEE Int. Symp. BIBE '03, pp. 359-366. Washington, DC. March 2003.)

Problems with Previous Work

a. Inter-residue similarity not modeled
b. Whole-query matching is used
   - Most biological questions seek significant local alignments

c. Edit Distance approximation is not sufficiently tight
   - Impractical for detecting distant homologs

d. Have only been applied to DNA data

Proposed Methods for Mapping Proteins

A. Compressed Amino Acid Alphabet

- e.g., e.g., 6-letter alphabet reduced from BLOSUM62 (Li,T., Fan,K., Wang,J. and Wang,W. (2003) Reduction of protein sequence complexity by residue grouping. Protein Eng., 16, 323-30)
- Reduces complexity
- Incorporates the amino-acid similarities into search
- Better chances to detect longer matching regions

B. Weighted Frequency Counting

- Shift/scale each row of the similarity matrix to obtain a sum of 1.
- Update the frequency vector of the window based on the new matrix

	A	B	C
A	3	1	0
B	1	2	1
C	0	1	4

⇩

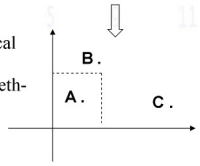
	A	B	C
A	0.75	0.25	0
B	0.25	0.50	0.25
C	0	0.20	0.80

C,D. Converting Proteins into Numerical Strings

C. Convert similarity matrix to low dimensional space using distance-preserving mapping
D. Represent amino-acids by a vector of chosen physico-chemical properties

	A	B	C
A	0	1	3
B		0	2
C			0

⇩



- When we have numerical strings:
  - Can apply other methods in time series datasets

Contributions

- First feature-mapping based indexing for proteins
- Optimal encoding of proteins for similarity searches
- High-throughput feature extraction from proteins
  - Can be used for motif-extraction
  - Protein structure/function classification