Clustering Sequences in a Metric Space

The MoBIoS Project

Rui Mao, Daniel P. Miranker,
Jacob Neal Sarvela and Weijia Xu

{rmao,miranker,sarvela,xwj}@cs.utexas.edu

Department of Computer Sciences
University of Texas
Austin, TX 78712

The exponential growth of sequence data requires the development of scalable database index structures. It is insufficient to merely store sequence data in a database and use utilities for sequence alignment (e.g., BLAST). The sequences must serve as index keys.

By analogy, the B-tree indexing mechanisms for linearly ordered business data were crucial to the adoption of relational databases. Similarly, variations of R-trees index 2- and 3-dimensional data in geographic information systems [1]. It is anticipated that metric-space index structures will be the foundation for multimedia databases [4].

Definition: A *metric space* is a set with a binary distance function, d, satisfying the following for every three objects x, y & z:

1) (*Positivity*) d(x,y) >= 0 and d(x,y) = 0 iff x = y;

2) *(Symmetry*) d(x,y) = d(y,x);

3) (*Triangle Inequality*) d(x,y) + d(y,z) >= d(x,z).

We are developing a [Molecular] *Biological Information System* (***MoBIoS***) based on metric space indices. Unfortunately, common similarity measures for sequence alignment do not form a metric-distance function. This is particularly vexing since the usual definition of edit distance does form a metric. Most clearly, the use of PAM log-odds matrices [2] yields higher similarity scores for more closely related sequences, an intuitively appealing result that reverses metric order. Further, log-odds scoring matrices contain negative values that can yield negative global alignment scores. This violates positivity. Use of PAM matrices also can violate symmetry and the triangle inequality.

Since biological databases must embody an evolutionary model, we revisited the original definition of the accepted point mutation model. We reworked the mathematics while maintaining metric properties. The results are *metric* PAM matrices (mPAM, see Figure 1). We then took the contents of a protein sequence database and divided it into database records, each containing a small fixed-length segment of the original database. We loaded the records into a metric-space indexing package [3]. Our results using mPAM-250 suggest significant clustering for word sizes of ~10 amino acids.

A metric-space index look-up may replace BLAST's directly-addressed hotspot arrays while permitting longer word sizes and gaps. We are exploring a number of other applications of MoBIoS including hierarchically clustering sequences to mine promoter regions with respect to regulatory pathways [4] and protein identification. Commercial database indices are now extensible, so our results may be the basis for scalable indexing of biological data types on commercial database platforms [5].

[1] V. Gaede and O. Gunther. "Multidimensional access methods," ACM Computing Surveys, Volume 30, Number 2, June 1998.

[2] M.O. Dayhoff, R. Schwartz and B.C. Orcutt. “Atlas of protein sequence and structure,” Vol. 5, Suppl. 3, Ed. M. O. Dayhoff, 1978.

[3] P. Ciaccia, M. Patella and P. Zezula. "M-tree: an efficient access method for similarity search in metric spaces," Proc. Conf. On Very Large Databases (VLDB), 1997.

[4] J. L. DeRisi, V. R. Iyer and P. O. Brown. “Exploring the metabolic and genetic control of gene expression on a genomic scale,” Science 278:680-686, 1997.

[5] P. M. Aoki. “Generalizing ``search" in generalized search trees,” Proc. 14th Int'l Conf. on Data Eng., Orlando, FL, 380-389, Feb. 1998.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	*	_
A	0	3	3	3	4	3	3	2	3	3	3	3	3	4	2	3	3	5	4	3	6	7
R	3	0	3	3	5	3	3	3	3	3	4	2	3	5	3	3	3	4	4	3	6	7
N	3	3	0	3	4	3	3	2	3	3	3	3	3	4	3	3	3	5	4	3	6	7
D	3	3	3	0	4	3	2	2	3	3	4	3	3	5	3	3	3	5	4	3	6	7
C	4	5	4	4	0	5	4	4	4	4	5	4	5	5	4	4	4	6	4	4	6	7
Q	3	3	3	3	5	0	3	3	3	3	3	3	3	4	3	3	3	5	4	3	6	7
E	3	3	3	2	4	3	0	2	3	3	4	3	3	5	3	3	3	5	4	3	6	7
G	2	3	2	2	4	3	2	0	3	3	3	3	3	4	2	2	2	5	4	3	6	7
H	3	3	3	3	4	3	3	3	0	4	3	3	3	4	3	3	3	5	4	3	6	7
I	3	3	3	3	4	3	3	3	4	0	2	3	2	3	3	3	3	5	3	2	6	7
L	3	4	3	4	5	3	4	3	3	2	0	3	2	2	3	3	3	4	3	2	6	7
K	3	2	3	3	4	3	3	3	3	3	3	0	3	5	3	3	3	5	4	3	6	7
M	3	3	3	3	5	3	3	3	3	2	2	3	0	3	3	3	3	5	4	2	6	7
F	4	5	4	5	5	4	5	4	4	3	2	5	3	0	4	4	4	4	1	3	6	7
P	2	3	3	3	4	3	3	2	3	3	3	3	3	4	0	3	3	5	4	3	6	7
S	3	3	3	3	4	3	3	2	3	3	3	3	3	4	3	0	3	5	4	3	6	7
T	3	3	3	3	4	3	3	2	3	3	3	3	3	4	3	3	0	5	4	3	6	7
W	5	4	5	5	6	5	5	5	5	4	5	5	5	4	5	5	5	0	4	5	6	7
Y	4	4	4	4	4	4	4	4	4	3	3	4	4	1	4	4	4	4	0	4	6	7
V	3	3	3	3	4	3	3	3	3	2	2	3	2	3	3	3	3	5	4	0	6	7
*	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	6	0	7
_	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	7	0

Figure 1. mPAM-250 matrix

Research supported in part by the Texas Higher Education Coordinating Board, Texas Advanced Research Program