OMICS A Journal of Integrative Biology
Volume 7, Number 1, 2003

Metric-Space Indexes as a Basis for Scalable Biological Databases

DANIEL P. MIRANKER

ABSTRACT

Biochemical databases will be best served by the development of new specialized database management systems whose storage managers are based on metric-space indexing techniques and the development a database query languages that embody semantics derived from biochemical models of similarity and evolution. Important biochemical data types cannot be effectively mapped to low dimensional coordinate systems on which O(log n) indexing methods rely. It is clear from an abundance of bioinformatic discoveries that biochemical data is not random and exhibits interesting structure with respect to clustering. Metric-space indexing exploits a data set's intrinsic clustering to speed the execution of similarity queries, even when the data cannot be mapped to a coordinate system. Database management systems that seamlessly integrate semantically rich query languages with a metric-storage and retrieval mechanism will allow biologists to simply and concisely develop informatic studies that have traditionally been large and labor intensive.

INTRODUCTION

IT IS EASY TO MISTAKE a large biological-data web site as a database management system (DBMS) or to anticipate that the software underlying a biological web site is a DBMS. A DBMS is a body of software that provides an integrated set of services on a collection of data, a database. Many popular biology web sites provide a number of services and their web-masters are usually careful to provide a common look-and-feel to each web page. Thus, an end user witnesses an integrated system. The reality is that each service is provided by an ad-hoc collection of scripted utilities and the underlying administration of the system rarely embodies generalized data management facilities.

These systems cannot be expected to scale, neither in performance as the volume of data increase, nor in ongoing software development costs as functionality is increased. Fast access to data in a relational database management system (RDBMS) requires ordinal data (e.g. numbers, dates, names). Biochemical data of primary importance does not fit the relational paradigm. This includes biological sequences, mass spectra, protein and ligand structures, phylogenetic and pathway graphs, to name a few.

When RDBMSs are used to manage biochemical data, first class data is relegated to blob and unstructured text fields. Annotations of the data, such as organism name and protein family name, may exist as part of the record. Relational tests against the annotations may be used to filter the database. Ultimately, mining of the biochemical data types is accomplished by utilities outside the database. Even when object-

Department of Computer Science, University of Texas, Austin, Texas.

relational or semi-structured representations are used (e.g., ASN.1 and XML) there are few results in building persistent access paths capable of supporting sophisticated retrieval methods.

Consider two high profile examples, GenBank and TreeBASE (Genbank and Peil et al., 2000). Every web page in Genbank shares a common masthead. Argument structure and terminology is held consistent throughout the site. Thus, users witness an integrated set of services. Nevertheless sequence retrieval is accomplished using a utility, BLAST, which sequentially scans the entire sequence content of Genbank. Although Genbank is extensible through the use of ASN.1 descriptors, these descriptors face a manual curation process. When new tags are accepted by the curators, additional functionality does not automatically appear on the Genbank website.

TreeBASE serves as an archival repository of phylogenies. As in Genbank, a primary service of TreeBASE is to connect esoteric data to individual scientific articles that detail the investigation that derived the information. TreeBASE is built using an underlying relational database platform and embodies flexible and scalable query facilities for many needs. However, the phylogenies themselves are stored as a structured text strings known as Newick format (Felsenstein). One drawback of using the Newick format is that the database cannot directly support queries concerning the relationships between the taxa and the structure of the phylogeny. In TreeBASE, queries specific to the structure of the individual phylogenies are done outside the database (Shan et al., 2002). An initial query may derive a subset of the phylogenies, but then each of those phylogenies is exported and handled outside the database. A utility program that does just that has been integrated into the web site. The uninformed are unaware that the utility was not authored by the TreeBASE group or that it is running on a different server.

STORAGE AND SIMILARITY SEARCH IN METRIC-SPACES

The concept of a metric-space moves the attention away from the data to a distance function. Data structures that exploit metric space properties rely only on the relative distance of objects to each other.

Definition: A *metric space* is a set of objects, S, a *[metric] distance function*, d, such that, given any three objects, x, y, z in S, [Cha01]

(i) $d(x, y) >= 0$ and $d(x, y) = 0$ iff $x = y$. (*Positivity*)
(ii) $d(x, y) = d(y, x)$. (*Symmetry*)
(iii) $d(x, y) + d(y, z) >= d(x, z)$. (*Triangle Inequality*)

For metric-space indexes, there are two primary methods used to partition the data into disjoint subsets and thus mimic the behavior of binary tree search. The methods are called vantage point and generalized hyperplane. In vantage point methods, a bounding sphere is formed by identifying a center of the cluster (the vantage point) and a radius such that half the data is inside the sphere and half the data is out. In generalized hyperplanes, the equivalent of two center points are chosen. Membership of the remaining points in the partitions is determined by computing the distance of each point to the centers, and assigning the point to the closest center. The fan out may be increased by choosing many center points, in which case, in Euclidean space the data structure resembles a Voronoi partitioning (Brin, 1995; Chavez et al., 2001).

The value of a metric-space approach is that it is unnecessary to find a meaning for the data with respect to the axis of a coordinate system. The arguments to the distance function can be anything. Reusable database index code can be applied to any data provided there is a metric.

There is a clear connection between algorithms for hierarchical clustering and the construction of a tree-based index structures. One can say that a tree-based index structure of a metric space materializes a persistent representation of a hierarchical clustering of the data (Brin, 1995; Mao et al., 2003).

Partitioning methods to form metric index trees exploit the triangle inequality to quickly identify naturally occurring data clusters and allocate them to separate sub-trees. A data point and a radius specify a similarity query. At each interior vertex of the index tree, the query predicate is compared to the bounding predicate covering the data present in each subtree. The triangle-inequality is used to determine both if the query predicate overlaps the data in a subtree and to prune the number of distance calculations when visiting an interior node.

SIMILARITY VERSUS DISTANCE

A primary challenge in this thesis is the endemic use of similarity functions in biology rather than distances. By this we mean, in most biochemical models of similarity, objects that are more similar are rewarded higher scores, an intuitively appealing result that reverses metric order. In the case of sequence alignment, PAM and BLOSUM log-odds matrices contain negative values that can yield negative alignment scores. This violates positivity.

By using a main-memory metric index, Giladi et al. (2002) have already reported two and three order of magnitude improvements in execution speed of sequence look-up compared to BLAST. However, their metric is simple Hamming distance that is devoid of any model of evolution. By their own accounting, although fast, their result is limited in applicability to sequence assembly and studies between evolutionarily very close organisms.

"[N]othing in biology makes sense except in the light of evolution" (Dobzhansky, 1963). Thus, to exploit metric-space search biochemical models of similarity must be revisited and new models that are metrics invented. In some cases, such as sequence alignment this is challenging. There is no simple algebraic normalization that will convert PAM matrices to a metric. However, the same assumptions and raw data as originally used to develop the PAM models can be revisited and a metric evolutionary model of evolution defined (Mao et al., 2002). In other cases simple algebraic normalizations may be sufficient.

In proteomics, a common similarity measure used for database look-up of mass-spectra is shared peaks count. If we map mass-spectra to a vector-space model, one dimension for each resolvable peak, then the number of peaks shared by two spectra forms a similarity function computed by taking the inner product of two spectra. In text retrieval systems it is understood that the corresponding metric is cosine distance. To test if spectra would cluster in a manner that would benefit from metric space indexing we assembled a database of computed mass-spectrometer spectra for the yeast proteome following trypsin digestion (Zhang and Chait, 2000). We then computed the cosine distance between all pairs of spectra. The results, plotted in Figure 1, indicate that the spectra are clearly not uniformly distributed in the space of spectra and show several local maxima much closer to zero degrees than expected by chance, which is evidence of clustering (Brin, 1995).

DISCUSSION

In an obvious ploy to draw analogy to the semiconductor industry, it is often incorrectly stated that the growth of sequence data is exponential. The growth of biochemical data is hyper-exponential. The exponential improvements in semiconductors are a by-product of an established, regular cycle of process improvements. The genomic revolution is still in its infancy and the introduction of process improvements to automated laboratory apparatus is accelerating.

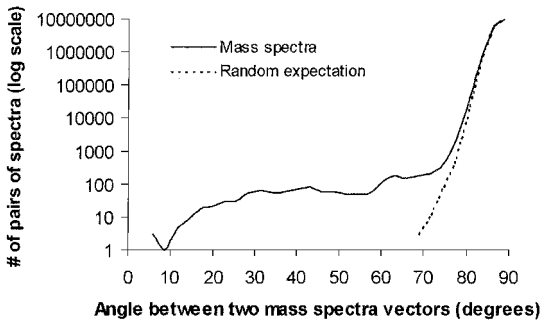

FIG. 1. Mass spectra of peptides from the known yeast proteins show considerably more clustering than expected by chance, indicating that spectra can be effectively organized in a metric-space model.

We recently passed a critical milestone. Where the doubling of the contents of Genbank has been reported as having a Moore's constant of 18 months, equal to the doubling time of processor speeds, the doubling time for Genbank has shrunk to 15 months. Where computer-processing capacity was increasing faster than the computing demands of Bioinformatics, then equal to them, we can now anticipate a widening gulf between computational capacity and the demand of Bioinformatics.

Storage and retrieval are the primary services required of a database management system. It is arguable that the advent of the B+-tree and it effectiveness for ordinal (one-dimensional) business data was critical to the initial success of relational database management systems (RDBMS). Subsequently, variations of R-trees and k-d trees (Gaede and Gunther, 1998) were developed to store and retrieve two- and three-dimensional data types and were instrumental in the creation of Geographic Information Systems (GIS). Other fields have also found critical support by virtue of specialized database management systems founded on different data models and their supporting data structures. Examples include Computer Aided Design (CAD), using object-oriented databases, and network management, using the hierarchical database component of LDAP.

These precedents suggest to cope with the avalanche of unconventional biochemical data specialized database management systems must be invented. Metric-space indexing methods with concomitant development of metric models of biochemical similarity offer a promising approach.

ACKNOWLEDGMENT

Research was supported in part by the Texas Higher Education Coordinating Board.

REFERENCES

BRIN, S. (1995). Near neighbor search in large metric spaces. Presented at the Very Large Database Conference.

FELSENSTEIN, J. (2003). The newick tree format. Available: http://evolution.genetics.washington.edu/phylip/newicktree.html.

CHAVEZ, E., NAVARRO, G., BAEZA-YATES, R., et al. (2001). Searching in metric spaces (Survey). ACM Computing Surveys **33(3),** 273–321.

GAEDE, V., and GUNTHER, O. (1998). Multidimensional access methods. ACM Computing Surveys **30(2),** 170–231.

GENBANK NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION. (2003). Available: www.ncbi.nlm.nih.gov/.

GILADI, E., WALKER, M.J., WANG, J.Z., et al. (2002). SST: an algorithm for finding near-exact sequence matches in time proportional to the logarithm of the database size. Bioinformatics **18,** 873–877.

MAO, R., MIRANKER, D.P., SARVELA, J.N., et al. (2002). Clustering sequences in a metric space. Presented at the ISMB 02, Edmonton, Canada.

MAO, R., XU, W., SINGH, W., et al. (2003). An assessment of a metric space database index to support sequence homology. Presented at the Third IEEE Conference on Bioinformatics and Bioengineering.

PIEL, W.H., DONOGHUE, M.J., and SANDERSON M.J. (2002) TreeBASE: a database of phylogenetic knowledge. In: Shimura, J., Wilson, K.L., and Gordon, D. (eds). *To the Interoperable "Catalog of Life" with Partners, Species 2000 Asia Oceania. Research Report from the National Institute for Environmental Studies No. 171.* (Tsukuba, Japan), 41–47.

SHAN, H., HERBERT, H.G., PIEL, W.J., et al. (2002). A structure-based search engine for phylogenetic databases. Presented at the 14th International Conference on Scientific and Statistical Database Management.

ZHANG, W., and CHAIT, B.T. (2000). ProFound—an expert system for protein identification using mass spectrometric peptide mapping information. Analytical Chemistry **72,** 2482–2489.

Address reprint requests to:
Dr. Daniel P. Miranker
Department of Computer Science
University of Texas
Room TAY 3.140B
Mail Code C0500
Austin, TX 78712

E-mail: miranker@cs.utexas.edu