integration strategies as well as the ability to create specialized data warehouses initialized from community databases. Our integration strategy is based upon the concept of "mediators", which serve a group of related applications by providing a uniform structural interface to the relevant data sources. This approach is cost-effective in terms of query development time and maintenance. We have examined in detail methods for optimizing queries such as "retrieve all known human sequence containing an Alu repeat in an intragenic region" where the data sources are heterogeneous and distributed across the Internet.

Transformation of data resources, that is the structural reorganization of a data resource from one form to another, arises frequently in genome informatics. Examples include the creation of data warehouses and database evolution. Implementing such transformations by hand on a case by case basis is time consuming and error prone. Consequently there is a need for a method of specifying, implementing and formally verifying transformations in a uniform way across a wide variety of different data models. Morphase is a prototype system for specifying transformations between data sources and targets in an intuitively appealing, declarative language based on Horn clause logic. Transformations specification in Morphase are translated into CPL and executed in the Kleisli system. The data-types underlying Morphase include arbitrarily nested records, sets, variants, lists and object identity, thus capturing the types common to most data formats relevant to genome informatics, including ASN.1 and ACE. Morphase can be connected to a wide variety of data sources simultaneously through Kleisli. In this way, data can be read from multiple heterogeneous data sources, transformed using Morphase according to the desired output format, and inserted into the target data source.

We have tested Morphase by applying it to a variety of different transformation problems involving Sybase, ACE and ASN.1. For example, we used it to specify a transformation between the Sanger Center's Chromosome 22 ACE database (ACE22DB) and a Chromosome 22 Sybase database (Chr22DB), as well as between a portion of GDB and Chr22DB. Some of these transformations had already been hand-coded without our tools, forming a basis for comparison.

Once the semantic correspondences between objects in the various databases were understood, writing the transformation program in Morphase was easy, even by a non-expert of the system. Furthermore, it was easy to find conceptual errors in the transformation specification. In contrast, the hand-coded programs were obtuse, difficult to understand, and even more difficult to debug.

DOE Grant No. DE-FG02-94ER61923.

Relevant Publications

P. Buneman, S.B. Davidson, K. Hart, C. Overton and L. Wong,"A Data Transformation System for Biological Data Sources," in Proceedings of VLDB, Sept. 1995 (Zurich, Switzerland). Also available as Technical Report MS-CIS-95-10, University of Pennsylvania, March 1995.

S.B. Davidson, C. Overton and P. Buneman, "Challenges in Integrating Biological Data Sources," J. Computational Biology 2 (1995), pp 557-572.

A. Kosky, "Transforming Databases with Recursive Data Structures," PhD Thesis, December 1995.

S.B. Davidson and A. Kosky, "Effecting Database Transformations Using Morphase," Technical Report MS-CIS-96-05, University of Pennsylvania.

A. Kosky, S.B. Davidson and P. Buneman, "Semantics of Database Transformations," Technical Report MS-CIS-95-25, University of Pennsylvania, 1995.

K. Hart and L. Wong, "Pruning Nested Data Values Using Branch Expressions With Wildcards," In Abstracts of MIMBD, Cambridge, England, July 1995.

Las Vegas Algorithm for Gene Recognition: Suboptimal and Error-Tolerant Spliced Alignment

Sing Hoi Sze and **Pavel A. Pevzner**[1]
Departments of Computer Science and [1]Mathematics;
University of Southern California; Los Angeles, CA 90089
Pevzner: 213/740-2407, Fax: -2424
ppevzner@hto.usc.edu
http://www-hto.usc.edu/software/procrustes

Recently, Gelfand, Mironov, and Pevzner (Proc. Natl. Acad. Sci. USA, 1996, 9061-9066) proposed a spliced alignment approach to gene recognition that provides 99% accurate recognition of human gene if a related mammalian protein is available. However, even 99% accurate gene predictions are insufficient for automated sequence annotation in large-scale sequencing projects and therefore have to be complemented by experimental gene verification. 100% accurate gene predictions would lead to a substantial reduction of experimental work on gene identification. Our goal is to develop an algorithm that either predicts an exon assembly with accuracy sufficient for sequence annotation or warns a biologist that the accuracy of a prediction is insufficient and further experimental work is required. We study suboptimal and error-tolerant spliced alignment problems as the first steps towards such an algorithm, and report an algorithm which provides 100% accurate recognition of human genes in 37% of cases (if a related mammalian protein is available). For 52% of genes, the algorithm predicts at least one exon with 100% accuracy.

DOE Grant No. DE-FG03-97ER62383.