Meeting report

Integromics: challenges in data integration

TV Venkatesh and Harry B Harlow

Address: Monsanto Co, Genomics, 800 North Lindbergh Boulevard, St. Louis, MO 63167, USA.

Correspondence: TV Venkatash. E-mail: t.v.venkatesh@monsanto.com

Published: 17 July 2002

Genome ***Biology*** 2002, **3(8)**:reports4027.1–4027.3

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/8/reports/4027

A report on Barnett International's 4th annual Bioinformatics and Data Integration conference, Philadelphia, USA, 7-8 March 2002.

Over the past two decades, advances in genomic technology have allowed laboratories to generate vast amounts of biological data. These data include, but are not limited to, gene sequences, protein structures, information on gene expression (transcripts, metabolites and proteins) and metabolic pathways. Automated instrumentation has enabled large volumes of data to be generated and automatically stored in computer databases, and this data has as many different formats as there are instruments. In addition to the new information gathered from genomic technologies, pharmaceutical and biotech companies have large amounts of 'legacy data' - data inherited from their own and other sources on chemical structures and properties of compounds, and clinical, phenotypic and toxicological information. Most of this is stored in older types of databases designed for the particular type of data, and a major computational challenge is to integrate the new genomic information with current database systems in order to facilitate decision-making.

Statistical approaches to data analysis and experimental design

John Weinstein (National Cancer Institute (NCI), Bethesda, USA) gave an overview of genomic and other 'omic' technologies and appropriately coined the new name 'integromics' for the data-integration issues associated with genomic research. He described a pharmaco-genomics database (Leadminer) developed at NCI to support high-throughput chemical screening using 60 cell lines and 70,000 chemical compounds. He presented a variety of data-analysis techniques, including the cluster-correlation image mapping (CIM) technique pioneered by his group in collaboration with that of Mike Eisen (University of California at Berkeley, USA). Cluster correlation provides a way of investigating gene-expression patterns associated with drug activity, characterizing both the cell lines and the putative modes of action of the chemical compounds using hierarchical clustering methods. He also presented results from the Medminer tool [http://discover.nci.nih.gov], which streamlines searching of the biomedical literature for annotation of genes used in microarray experiments. Weinstein demonstrated the success of his group at NCI in integrating the massive dataset comprising results from microarrays, proteomics, 70,000 chemical compounds and 60 cell lines.

Sherri Matis (Astra Zeneca Pharmaceuticals, Wilmington, USA) described the data-integration challenges associated with the analysis of expression-profiling data in molecular toxicology. She described the application of principal component analysis (PCA) for quality control, and the application of naive Bayesian analysis for clustering, as a prelude to identifying common promoter or enhancer elements. PCA, a statistical algorithm, was used to identify characteristic patterns of gene expression in the data, and to determine whether these patterns were informative and could be used to classify the samples into biologically appropriate groups. Once it was determined that the expression data was of acceptable quality, a Bayesian method of clustering was used to identify groups of genes sharing similar expression patterns. Once such a group was identified, the upstream regulatory regions were analyzed, using bioinformatics methods to identify over-represented elements. She discussed how clinical data, genomic and genetic information, motifs, and pathways can be used along with expression profiling to add additional annotation and prioritize genes in order of their biological importance. She showed how 'profiles' of promoters can be used in the analysis of tissue-specific, disease-specific or treatment-specific induction of gene expression.

We (T.V.V. and H.B.H.) discussed the expression-profiling process as it has been developed at Monsanto, as well as giving a brief description of genomics computer applications and an object-oriented bioinformatics framework (see below). We find that classical experimental design strategies can be used to integrate phenotypic, developmental and time-course expression data in genomics-based experiments with good results.

Mike Liebman (Abramson Family Research Center, Philadelphia, USA) described a systems-engineering approach to computational biology, focusing on the use of a combination of mechanistic and statistical models to analyze clinical data in oncology. He examined a variety of mechanistic models associated with breast cancer, which take account of its developmental progression and the existence of different clinical types. If the *a priori* information on disease progression and classification can be accounted for using such models, a significant amount of the clinical variation can be explained. He presented a variant of the standard pedigree model that included genealogy and an object-oriented data model for medical history, which enables the integration and analysis of a multidimensional set of clinical data, with the aim of improving treatment strategy.

Software tools for gene-expression analysis

Jeffery Schaffer (Omniviz, Maynard, USA) presented tools for analysis of free text, numerical data and genomic data. Omniviz functional genomics tools include a relaxation clustering method (Galaxy) for visualizing microarray data and integrates text and genomic annotation data with the expression data. The suite of software presented by C. Brett Jesse (Anvill Bioinformatics, Burlington, USA) for analyzing gene-expression data uses traditional statistical tools for assessing the quality of microarray data. Their proprietary visualization tool 'Radviz' can display hundreds of data points, each having thousands of attributes or descriptors. The software uses several clustering methods to classify the data.

Database integration

Ramesh Durvasul (Tripos Inc, St. Louis, USA) described four of the data-storage options currently available and their advantages and disadvantages in regard to integration of different types of data. The first, and least favored for integration, is a single database, typically managed under a relational database management system (where data collected are stored and presented as a series of relations and each relation is depicted in a table where columns are attributes and the rows represent entries). Although a relational database enables efficient access to data by means of a structured query language (SQL), which can be used to retrieve groups of data efficiently, it is cumbersome for data selection and manipulation in a semantically rich domain like molecular biology. His second example was the 'data warehouse', in which the core enterprise data (a collection of relational databases representing all data) are stored in a central store and which connects to other relational databases from which a subset of the data can be selectively extracted and loaded for analysis. The data warehouse places an emphasis on the ability to capture and copy data from a wide variety of diverse sources. The third example, the 'data mart', stores specialized data derived from a data warehouse for use in a particular analysis. The emphasis here is on content, presentation and ease of use in formats familiar to the specialized users. In the fourth example, the 'federated database', many databases are connected through a specialized network service shared by applications (databases and tools) and users, to create a virtual data warehouse. With good design, data warehouses and data marts can perform well in storing and retrieving biological data. But they have limited flexibility to accommodate to changing requirements and are expensive to implement. Federated databases are more flexible and less expensive, but it is often difficult to optimize their performance to deal with different types of data.

Richard Scott (DeNovo Pharmaceuticals, Cambridge, UK) discussed the advantages of a federated data warehouse strategy to support older systems from different domains and vendors. This approach allows new types of data to be plugged into the system and data from third-party systems to remain in their own application environments. A federated data warehouse allows users to use their own client applications and to share each other's data without learning each other's applications (for example, a biologist uses biological systems and a chemist uses chemical systems). The virtual data warehouse was described as needing an industrial-strength application server, a robust compute farm (a large group of interconnected computers each performing different parts of the same task), ample storage space, and a sound relational database management system. Scott also pointed out the advantages of employing open-source tools and utilities and building a strong search engine. He described the warehouse they have implemented to store data using technology developed in-house called SKELEGEN. This system deals with gigabytes of chemoinformatics screening data generated by their research program. Currently it contains more than 19,000 chemical structures and 60 projects with nearly 100,000 data files each.

Dave Parrish (Management Science Associates, Pittsburgh, USA) discussed the issues involved in building a centralized system for capturing laboratory information and presented a well thought-out data-integration process. He described a prototype system for flow cytometry data. In this example, the data system begins with the development of the protocol and establishing common data structures and dictionaries. The model is ontology based and integration is accomplished using an object-oriented (OO) approach, and collected data are subjected to normalization and transformation before they are put in the database. An object-oriented database is a generalization of a relational database that allows entries to be abstract objects or collections of objects and successive specialization of these objects. For example, an 'object' might be an abstract gene, and a specialization of a gene might be the gene encoding a specific enzyme such as a protein kinase. Objects can be thought of as collections of information, where the information may include data itself and/or scripts to display the data (protein structures, images, network maps, or even other objects). Using the gene example, a gene object might contain information on sequence, alleles, protein structure, and source organism along with scripts to display these data. Object-oriented databases give the scientist substantial latitude to store the diverse types of data that are encountered in the real world.

Parrish also described Protégé-2000, an OO frame-based system in which he and his colleagues have modeled domain ontology (concepts and relationships) and method ontology (for example, clinical guidelines and protocols) and created a domain knowledge base that contains known facts. The ontology they have developed extends beyond syntactic properties (such as would be found in a data dictionary) to include semantic characteristics (the hierarchical and process relationships between objects and the meaning of each entity). Data and procedures were packed into a common structure, and building blocks (classes, slots and facets) were used in an object-oriented approach. Parrish's example is a frame-based central repository, which appeared to be able to handle object relationships, generally the hardest task in data integration.

David Hansen (Lion Biosciences, Heidelberg, Germany) discussed integrating extensible markup language (XML) data with the sequence retrieval system (SRS) and relational databases. XML is the web language for data interchange. XML data is structured and is provided with a description and the rules for data structure are stored in a document called the document type declaration (DTD). SRS stores sequence data in a text-file indexing system which allows the data to be queried by field. It uses powerful web-based tools for interrogating databases and abstracting information. Application of XML using a meta-data approach is rapidly becoming the method of choice for exchanging chemical and biological data. Features of the meta-data approach to data integration include its widespread use and its flexibility in dealing with a variety of different data sources. XML, coupled with appropriate meta-data, provides consistency and transparency about the structure of data within an individual data source, as well as a set of exchange standards. DTDs are used to map XML data sources to SRS libraries, using meta-data and metaphors to map XML object attributes with SRS fields and loading instructions or scripts. A metaphor allows conditional indexing of one field to be dependent on the content of another and contains micro-syntax parsing rules to deal with data inconsistencies. To integrate relational databases, Hansen and his colleagues defined conceptual objects on top of the schema(s) and used object-relation mapping for integration; this is essentially a variation of database federation. By mapping XML data into SRS concepts they have been able to overcome inconsistencies in vocabularies and ontologies, map between XML data standards, and integrate XML and non-XML data sources into a common environment. Hansen presented an example of an integrated database using an XML application in which sequence information was tied to protein structure, to signaling or metabolic pathway, and to function.

The consensus of the conference was that there is no simple solution to database integration. Federated databases using three-layer architecture with web-based query tools were popular because of their easy implementation. There was a good deal of discussion on developing common ontologies. Many of the participants felt that this approach might not provide the optimal solution. Major challenges to developing common ontologies were discussed, including difficulties in capturing all information from biological systems because our understanding of biological systems keeps changing, and the disparate technical domains crossed by genomics and bioinformatics.