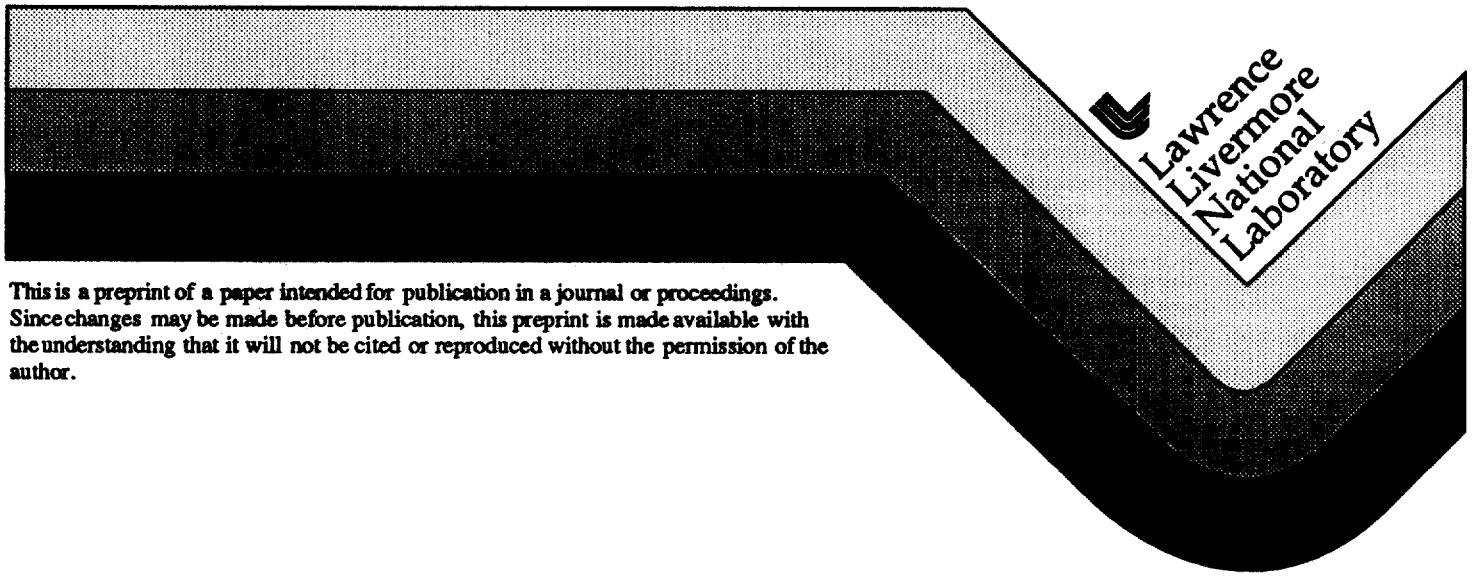
UCRL-JC-127338
PREPRINT

Large-Scale Data Mining Pilot Project in Human Genome

R. Musick
K. Fidelis
T. Slezak

This paper was prepared for submittal to the
Workshop on Research and Development Opportunities in Federal Information Systems
Arlington, VA
May 13-14, 1997

May 1997



Large-Scale Data Mining Pilot Project in Human Genome

Ron Musick (Principal Investigator)
Center for Advanced Scientific Computing
Lawrence Livermore National Lab

Krzysztof Fidelis
Structural Biology

Tom Slezak
Biology and Biotechnology Research Program

Introduction

We are at a point in time where science, business, and government must face many common technical challenges in their struggle to keep moving forward. The challenges stem from our growing abilities to create and collect complex data at tremendous rates, followed naturally by our need to manage and analyze that same data. To maximize the value of this data, we first must be able to access and manipulate it as structured, coherent, integrated information. Scientific data management, the discipline of improving the scientist's interactions with data, addresses this need in the scientific arena (the domain of this pilot project). Data mining and knowledge discovery is a cornerstone of scientific data management, and of advanced management of data from all sectors in general. Data mining has the narrower focus of helping automate the exploration and characterization of the data being generated. This leaves the analyst/scientist free to concentrate on the actual interpretation and use of the data.

This white paper briefly describes a new, aggressive effort in large-scale data mining at Lawrence Livermore National Labs (LLNL). The implications of "large-scale" will be clarified in the Barriers Section. In the short term, this effort will focus on several mission-critical questions in the Human Genome project. We will adapt current data mining techniques to the genome domain, extend them to quantify the accuracy of inference results, and lay the groundwork for a more extensive R&D effort in large-scale data mining. A major aspect of the approach is that we will be leveraging a fully-staffed data warehousing effort in the human genome area. The long term goal is to build a strong applications-oriented research program in large-scale data mining. The tools, experience and skill set gained will be directly applicable to a wide spectrum of tasks involving advanced analytics for large spatial and multidimensional data. This includes applications in ensuring Global Security (non-proliferation, stockpile stewardship), enabling Global Ecology (Materials Database for Industrial Ecology), advancing the Biosciences (Human Genome Project), and supporting work for others (Battlefield Management, Health Care).

Barriers

There are many challenges on the path to effective large-scale data mining. The issues are broad, and their solutions require expertise from several domains in computer science, mathematics and statistics. We introduce them by laying out several axes along which current technology must be scaled to match the demands of a data-intensive domain:

- *Large datasets* - Astrophysics currently has terabytes of data available for processing; earth science data will soon be measured in the petabytes. One must deal with data storage, and data transport across networks and through the storage hierarchy from tertiary storage to disk to main memory.
- *Highly complex data* - Data mining algorithms typically handle "simple" data consisting of rows of feature vectors. Native data models (before being transformed into feature vectors)

are often highly structured 2D or 3D objects with complex, inter-related components. Mapping from native data into feature vectors can explode the number of features of each object, and potentially lose information in the process.

- *High algorithmic complexity* - Algorithms in this field vary dramatically. Simple naive Bayes, a fast but limited predictor, can be linear in the number of features + training examples. Decision trees, which are excellent classification tools, can range from O(nlogn) to O(n^2) or worse depending on the type of pruning done. Model induction algorithms (e.g. belief networks) with built-in assumptions that reduce complexity start at O(n^4 D^2) where n is the number of variables, and D is the average domain size for the variables. More complex algorithms are exponential.
- *Oppressive data management demands* - There are significant difficulties in mining "living data", or data that is used on a daily basis. This includes (1) different repositories hold critical pieces of information, (2) non-standard nomenclatures, (3) radically different data types and models, (4) data may be duplicative and erroneous, and (5) the models for representing data content may change at a high rate.
- *Qualification of results* - This last issue is important, but rarely mentioned. Data mining tools are decision making aids. As such, if the tool is poorly understood, then expensive and incorrect decisions could easily be made based on the output (e.g. treatment decisions in the medical world). Data mining tools generate models that have built in biases, and are often the product of heuristic, non-exhaustive searches through a problem space. Unfortunately the quality of these models (or inferences) are rarely measured, leaving the user to rely on intuition as to how far to trust the results. If these methods are to be trusted tools of any profession where the (real and opportunity) cost of a mistake is high, then the qualification of the inference results must be addressed.

Approach

There are four main strategies for overcoming the barriers facing large-scale data mining. Each addresses a different aspect of scalability.

1. Design environments and tools that support the entire data mining process from data collection, data integration and cleaning, data selection, to visualization of results.

2. Build scalable I/O architectures and interfaces. For example, build intelligent interfaces between DBMS systems and tertiary storage.

3. Develop algorithms that work with non-traditional data types such as time sequences, protein sequences, or 3D structures.

4. Scale algorithms to work with larger data sets. This involves reducing algorithm complexity, effective parallelization, and controlled sampling or filtering.

The first approach is geared towards moving data mining technologies to large scale real world applications; the data warehousing project described in the Application Domain Section below addresses the challenges raised there. The second and third approaches are important as well, and we plan to leverage results in from these areas at some point in the future. In particular, there are several major efforts at LLNL that focus on scalable I/O architectures and interfaces, including the High Performance Storage System (HPSS), the Message Passing Interface, and the Scalable I/O Facility. Other work revolving around scalable interfaces includes the Interface Data Repository, and the Conquest/OASIS projects [2, 9]. However, in our judgment the highest payoff research area for genome, and for large-scale data mining in general is the fourth approach, scaling algorithms to work with large data. The key enablers will come from a focus on parallel codes, and controlled sampling techniques.

For the pilot, we will adapt existing parallel codes like SPRINT [8] and MLC++ [3] to the human genome domain. Where advantageous we will extend other algorithms with the more obvious parallel optimizations such as numerical and graph-search parallel techniques. Our primary focus, however, will be to build a controlled random sampling framework based on the expected utility of additional sampling.

The key concept behind random sampling is the realization that, often, the economically rational decision is to use a small portion of the available data to answer a question. For interactive data mining tasks, a user might want only a rough idea of the predicted value of a variable, and so would want to cut off search when the expected accuracy reaches, say, +/-10%. Or similarly, stop when it is 95% certain that the current answer will not change no matter how much of the remaining data is seen. The methods that can support these abilities can also be used to answer questions like: "What sample size will probably be large enough to produce inferences within a given expected error," and "What is the estimated accuracy of this prediction given the data that was used to produce it."

The theoretical basis for these techniques stems from the idea that the utility of additional sampling is a measure of the expected marginal gain in inference quality [4, 5]. Determining marginal gain is a difficult problem, both from the aspect of creating an accurate mathematical model of the learning algorithm at hand, and tying the model to the sampling decision process. Initial results for decision trees can be found in Musick et. al. [4] (http://www.llnl.gov/CASC/people/musick/papers/dt.ps). The research is strongly related to ideas from learning from sparse data [6], sampling theory [1], game theory [7], and anytime algorithms [10].

There are two clear benefits of providing a controlled sampling framework. First, this research will work a critical shortcoming in many state of the art mining algorithms: the lack of a well-grounded metric by which to judge the quality of inference results. The marginal gain in inference quality is the key to providing this metric. The second major benefit is that we can control the large-scale aspect by taking samples of the data large enough to return high quality answers, but much smaller than the full-blown corpus of data.

Application Domain

The technical insights from this pilot project should be readily transferrable to nearly any application of data mining to large data in the scientific, business or government arenas. Our particular focus for the short term is on applying these ideas to domain-specific questions from the human genome project. We also believe that data mining will become a central technology in data warehousing and intelligent integration techniques, aiding in schema and row identity analysis. We are pursuing that direction as well.

The genetics domain is stimulating and challenging, and the staff in that area in LLNL is interested and already contributing to a highly leveragable R&D project: "Data Warehousing and Integration for Scientific Data Management". The warehousing project marks a significant push to collect, organize and integrate the corpus of genetic map data, sequence data, protein structure, taxonomy, and other information that exists at LLNL and other sites around the world. The data infrastructure is currently under construction, and should be in place by FY98. The warehousing effort addresses "Oppressive Data Management Demands" described in the Barriers Section, and will establish an excellent testbed environment for the data mining effort.

The total data collected so far in the genome effort is not as large as in other domains (on the order of tens to hundreds of gigabytes). However, the data is growing rapidly, it is extremely complex in nature, and misinterpretation of analysis results in this domain could lead to expensive errors in deciding, for example, which gene to sequence next. The domain is a natural fit for large-scale data mining. Data mining is a vital enabling technology in this area that will help address several important research questions in the human genome effort. Furthermore, the ability to characterize

the quality of the data mining results will be critical to the general acceptance of the techniques. Some examples of where these techniques are needed:

- Characterize tumor suppressing genes, and apply that model to all the genes in chromosome 19 to see if there are others that match. Note that this process can be repeated for any particular cluster of genes and chromosomes in any organism in which the scientist is interested.
- Discover new methods to find and compare protein homologs by incorporating knowledge from many non-traditional sources into the classification and modeling techniques. For example, take protein from any species where the sequence may or may not be known. Digest the protein, and develop a mass spectrometry footprint of it. Compare with other footprints. Include functional information when available, and attempt to link the unknown protein to a particular sequence in the genome.

Deployment

The first usable prototype of the warehouse will be deployed by the end of FY97, and will provide the testbed for the data mining research. From the research standpoint, we will be productive within a few months of project initiation. We expect that with two scientists committed to the pilot project, strong practical results will be achievable within half a year of the start date. Based on research results achieved by one of the PI's in this area [4, 5, 6], we believe the chance of success to be high.

Leverage

There are significant resources that this effort will be able to leverage. The largest impact will come from the data warehousing R&D effort between the Center for Applied Scientific Computing (CASC) and the Biology and Biotechnology Research Program (BBRP), as discussed above.

Other resources include:

- *Significant high performance computing competence* - CASC and the Computation Directorate has a history of excellence in large-scale parallel computing. Many of the issues faced in this domain are similar to the key challenges facing large-scale data mining. Computation is also heavily involved in the data management aspect of SDM, examples include the data warehousing project, the Intelligent Archive, work done by the Nuclear Weapons and Information Group, and storage-related projects such as HPSS, MPI, and SIOF. The expertise is local, well within reach of this project.
- *Computing resources* - We have access to a 256 processor Cray T3D, a massively parallel cluster of Dec 8400's, and limited access to HPSS platforms and the teraflop-capable ASCI machines.
- *Collaborations* - There is an existing, effective collaboration between BBRP and CASC upon which this effort would build. The principal investigators on this proposal are already working together on the data warehousing project.
- *Large-scale applications* - LLNL houses efforts visible at the national level in human genome, stockpile stewardship, astrophysics, energy, materials science, environmental sciences and more. All of these domains have similar footprints regarding their need to analyze large, complex data.

LLNL is strongly positioned to make a significant impact on the Human Genome effort in the short term, and on both the theoretic and practical aspects of this new field of large-scale data mining. The combined knowledge, tools, collaborations and other resources available to this effort at LLNL will enable significant progress in a short time frame.

References

[1] R. Bechhofer and D. Goldsman; "Sequential Identification and Ranking Procedures", The University of Chicago Press, Chicago, IL, 1968.

[2] P. Brown, R. Troy, D. Fisher, S. Louis, J. McGraw, R. Musick; "Metadata for Balanced Performance", Proc. of the 1st IEEE Metadata Conference, 1996.

[3] R. Kohavi, G. John, R. Long, D. Manley, K. Pfleger; "MLC++: A Machine Learning Library in C++", Tools with Artificial Intelligence, 1994.

[4] R. Musick, J. Catlett, S. Russell; "An Efficient Method for Constructing Approximate Decision Trees for Large Databases", Proc. of the 10th Int'l Conference in Machine Learning, 1993.

[5] R. Musick; "Belief Network Induction", Ph. D. Dissertation, UCB Tech Report CSD-95-863. University of California, Berkeley, 1994.

[6] R. Musick; "Rethinking the Learning of Belief Network Probabilities", Proc. of the 2nd Int'l Conference on Knowledge Discovery and Data Mining, 1996.

[7] S. Russell and E. Wefald; "Do the Right Thing: Studies in Limited Rationality", MIT Press, Cambridge, MA, 1991.

[8] J. Shafer, R. Agrawal, M. Mehta; "Fast Serial and Parallel Classification of Very Large Databases", Proc. of the 22nd Int'l Conference on Very Large Database, 1996.

[9] P. Stolorz, H. Nakamura, E. Mesrobian, R. Muntz, E. Shek, J. Santos, J. Yi, K. Ng, S. Chien, C. Mechoso, J. Farrara; "Fast Spatio-Temporal Data Mining of Large Geophysical Datasets", Proc. of the 1st Int'l Conference on Knowledge Discovery and Data Mining, 1995.

[10] S. Zilberstein; "Using Anytime Algorithms in Intelligent Systems", AI Magazine, 17(3), 1996.

Administrative Information

Qualifications of Principle Investigators

The principle investigators on this project all have significant experience and visibility in their respective fields.

Dr. Ron Musick (Center for Applied Scientific Computing) earned his Ph.D. in Computer Science at the University of California, Berkeley. At LLNL Dr. Musick has been deeply involved in data and information management, and is the P.I. of the above mentioned data warehousing R&D project. He is the Program Chair for the IEEE Metadata Conference, is on the Program Committees for the International Conference on Machine Learning, the IEEE Advanced Digital Libraries Conference, and several others involving scientific data management. He is a member of IEEE, the Association of Computing Machinery, and the American Association for Artificial Intelligence. He is the Co-Editor of a special issue on Scalable High-Performance Computing for KDD to appear in the Journal of Data Mining and Knowledge Discovery. Dr. Musick has published several papers regarding the expected quality of results for belief networks and decision trees, as well as several others in the area of scientific data management.

Dr. Krzysztof Fidelis (Structural Biology) is leading the LLNL effort in protein structure prediction and assessment. He is an organizer and co-founder of international conferences on Critical Assessment of Techniques for Protein Structure Prediction and Director of the Livermore based Center for Critical Assessment of Protein Structure Prediction. His current research is devoted to the development of protein fold recognition with applications to structural and functional characterization of the genomic and expressed sequence data. Ph.D. Biophysics/Phys.Chem.

Mr. Tom Slezak (Human Genome) has spent more than 18 years at LLNL, most involving computational support of BBRP research (nine of those years on the Genome project). He is the architect of all of BBRP's database, analysis, and graphical tools. He designed and implemented the physical map assembly and integration solutions and all database abstractions. He is an acknowledged expert in genome informatics, has been invited to speak at major genome labs in England, France, Germany, and Japan and has consulted at many major biotech and pharmaceutical companies. BS/MS CS.

Prior Results Achieved

The data warehousing R&D effort was begun in FY97, and the pilot project described in this white paper has not yet begun. Most of our related prior results are research-oriented. Publications directly relevant to the research problems in large-scale data mining include:

P. Brown, R. Troy, D. Fisher, S. Louis, J. McGraw, R. Musick; "Metadata for Balance Performance", Proc. of the 1st IEEE Metadata Conference, 1996.

R. Musick, J. Catlett, S. Russell; "An Efficient Method for Constructing Approximate Decision Trees for Large Databases", Proc. of the 10th Int'l Conference in Machine Learning, 1993.

R. Musick; "Belief Network Induction", Ph. D. Dissertation, UCB Tech Report CSD-95-863. University of California, Berkeley, 1994.

R. Musick; "Rethinking the Learning of Belief Network Probabilities", Proc. of the 2nd Int'l Conference on Knowledge Discovery and Data Mining, 1996.

R. Musick; "Minimal Assumption Distribution Propagation in Belief Networks", Proc. of Int'l Conference on Uncertainty in Artificial Intelligence, 1993.

Large Scale Data Mining Pilot Project

Ron Musick
Center for Applied Scientific Computing

Krzysztof Fidelis
Biology and Biotechnology Research Program

Tom Slezak
Computations / Human Genome Center

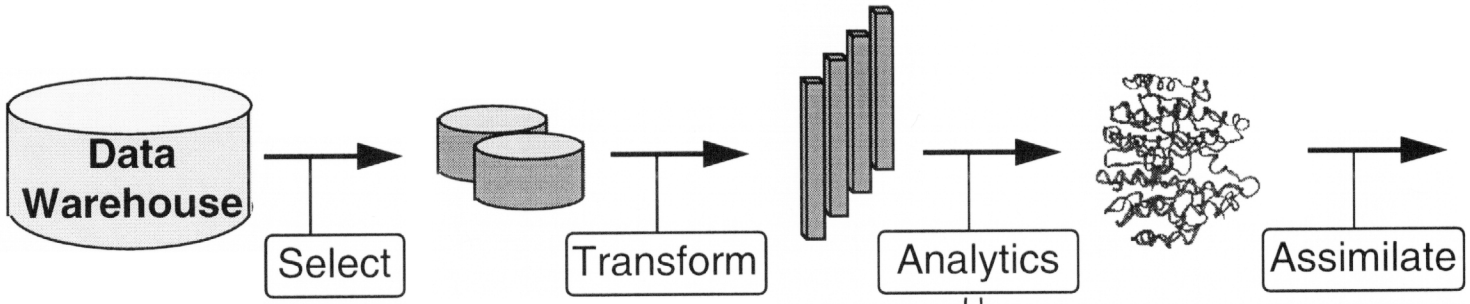
Lawrence Livermore National Laboratory

May, 1997

Talk outline

- **Background**
  - — **Scalable data mining - motivation, challenges, approaches**
- **Random Sampling Framework**
- **Mediated Data Warehouse**

Advanced Analytics - a step in the Knowledge Discovery Process



Data Mining - help automate the exploration and characterization of data

- Build models of data, and
- Find and flag data fitting those models

Interactive exploration is limited to small data, or "simple" algorithms

Model construction is expensive:

- Protein Homology
  - Neural networks => 1-10MB
- Insurance
  - Belief networks => 10-100MB
- Financial
  - Rule finding => 1-10GB
- Astronomy
  - FOA + decision trees => 10-100GB

Currently,

Time:
Data size:

Imagine..

Time:
Data size:

Barriers to scaling up are numerous and difficult

- Transport of large data
- High algorithmic complexity
- Highly complex data
- Oppressive data management demands
- Limited qualification of results

Strategies for large scale data mining

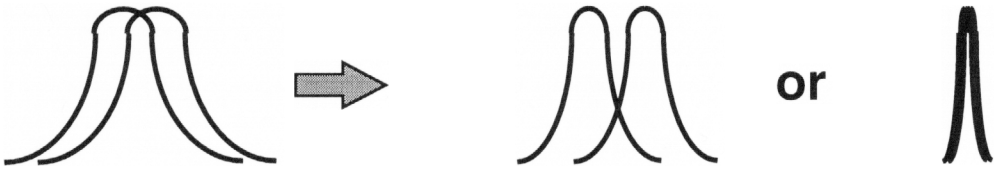
- Manage practical aspects of data collection and integration
- Build scalable I/O architectures and interfaces
- Develop algorithms for non-feature vector data
- Scale algorithms to work with larger data sets

Random sampling framework for large data and complex algorithms

Economically rational decision -
Use only a small portion of available data to answer a question

Measure the marginal expected utility of a sample

- Reduce result distribution variance
- Distinguish choices, or
- Render choices indistinguishable



Significant challenges, promising early results

- Research issues
  - Build mathematical model of algorithm
  - Tie model to sampling decision process
- Early results
  - Sequential sampling
  - Belief networks - Dirichlet vs. point probability
  - Decision trees - expected loss as stopping criterion

Sampling framework lends itself to interactive mining

- Methodology can be used to
  - Cut off search when E(*error*) < 10%
  - Stop when 95% certain current answer will not change
  - Estimate sample size needed for a "good" answer
- Barriers addressed
  - High algorithm complexity
  - Large data
  - Result qualification

Data warehousing and integration for scientific data management

- Transport of large data
- High algorithmic complexity
- Highly complex data
- **Oppressive data management demands**
- Limited qualification of results

Managing "living" data

— **Distributed, autonomous, heterogeneous repositories**

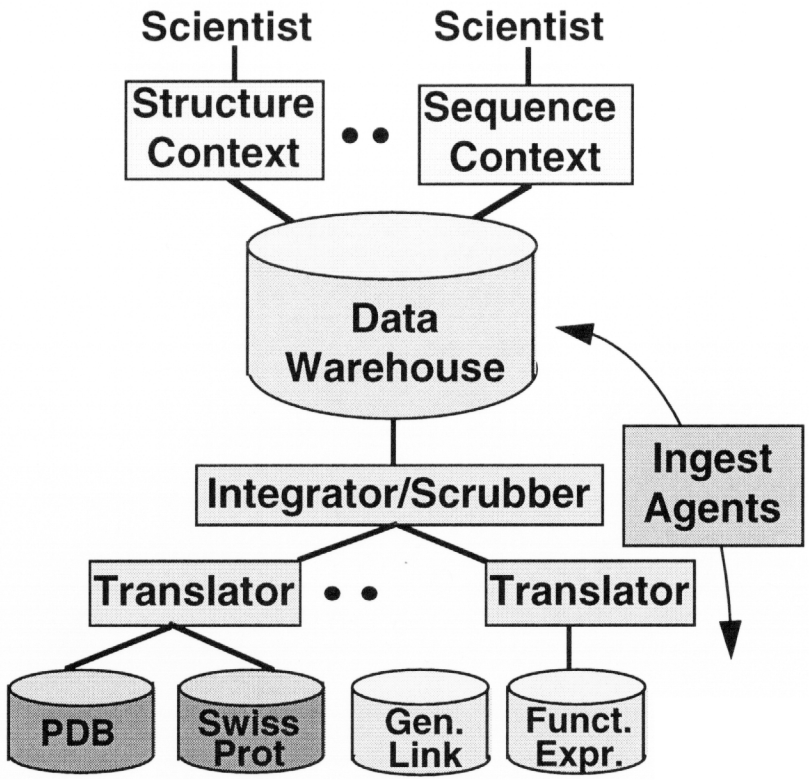
— **Non-standard nomenclatures**

— **Radically different data schemas and models**

— **Duplicative, erroneous, incomplete data**

— **High rates of change**

DataFoundry - a mediated data warehouse



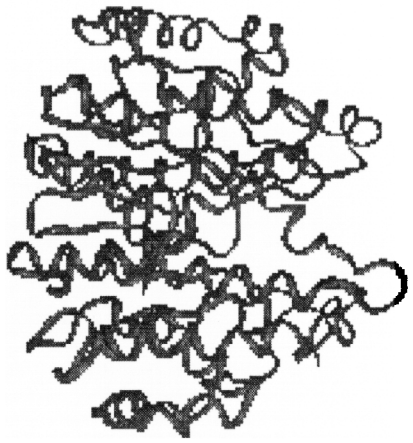
- **Mediated Data Warehouse**
  - Independent data sources
  - Many data types and models
  - Ad-hoc queries
- **Primitive Model, Data Mining**
  - Nomenclature
  - High rates of change
  - Erroneous, missing data
- **Partially Materialized Views**
- **Incremental Data Ingest**
  - Storage and computational efficiency

Planned information infrastructure for the LLNL component of the DOE Joint Genome Institute

Data warehousing will enable breakthrough functional genomics

Sequence of
Human Rad51
MAMQMLEANAD
TSVEEESFGPQP
ISRLEQCGINAN
DVKKLEEAGFHT
VEAVAYAPKKEL
INIKGISEAKADKI
LAEAAKLVPMGF
TTATEFHQRR...

~100,000 sequences



~1000 structures

- **Experimental techniques not keeping pace**
- **Human genome project - 100k genes in 5 years**
- **Next phase -**
  - **3D modeling**
  - **function identification**

For computational analysis,

Scientists must be able to access and analyze data from multiple data sources.

DataFoundry is a synergistic approach to large scale data mining

- **Advanced data warehouse**
  - Provides excellent testbed for mining
  - Scales to real environments
- **Random sampling research**
  - Provides metric for judging quality of results
  - Scales up to large data

- **Significant practical impact**
- **Widely applicable results**

Technical Information Department • **Lawrence Livermore National Laboratory**
University of California • Livermore, California 94551