3-12-2001

Application of Spatial Concepts to Genome Data

Mary-Kate Beard-Tisdale
Principal Investigator; University of Maine, Orono, beard@spatial.maine.edu

Carol Bult
Co-Principal Investigator; University of Maine, Orono, carol.bult@jax.org

Max J. Egenhofer Editor
University of Maine, max@spatial.maine.edu

Final Report for Period: 08/1997 - 07/2000
Submitted on: 03/12/2001
Principal Investigator: Beard-Tisdale, Mary-Kate .
Award ID: 9723873
Organization: University of Maine
Title:
Application of Spatial Concepts to Genome Data

Project Participants

Senior Personnel

Name: Beard-Tisdale, Mary-Kate
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Bult, Carol
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Egenhofer, Max
Worked for more than 160 Hours: Yes
Contribution to Project:

Name: Wheeler, Tom
Worked for more than 160 Hours: Yes
Contribution to Project:
Tom particapted as a data modeling and database consultant on developing teh second round of the conceptual model

Post-doc

Graduate Student

Name: Holden, Connie
Worked for more than 160 Hours: Yes
Contribution to Project:
Connie has been involved in all aspects of the project from data modeling, visualization and display of genomes. In particular she worked on developing spatial analysis routines for statistical analysis of genomes.

Name: Holan, Mary
Worked for more than 160 Hours: Yes
Contribution to Project:
Mary was involved in analyzing the overall software engineering process and with the development of spatial analysis routines for statistical analysis of genomes.

Name: Rodriquez, Andrea
Worked for more than 160 Hours: Yes
Contribution to Project:
Andrea was involved with the design, layout and prototyping of a user interface for GenoSIS

Undergraduate Student

Name: Bethell, Amber
Worked for more than 160 Hours: Yes
Contribution to Project:

Amber attended all weekly project meetings. She was respsonsible for developing code for downloading and extracting Gen Bank files, becoming familiar with the Oracle database model and loading and querying the database.

Name: Hall, Suzannah

Worked for more than 160 Hours: Yes

Contribution to Project:

Suzannah participated in the weekly project meetings and worked on developing procedure for loading genome sequence into the Oracle database model.

Technician, Programmer

Name: Guerney, Nathan

Worked for more than 160 Hours: Yes

Contribution to Project:

Nathan was involved in writing initial software in Java to read and display genomes

Other Participant

Research Experience for Undergraduates

Organizational Partners

Jackson Laboratory

Drs. Jim Kadin and Janan Eppig of the Mouse Genome Informatics Group particpated in the inital workshop and in subsequent discussions on framing the biological questions of interes. Several members of the Jackson Lab staff participated in a survey on graphic representational forms of genomes.

Other Collaborators or Contacts

We held a workshop early in the project which included the following participants:

Workshop Participants

Dr. Kate Beard ? Workshop Co-organizer
National Center for Geographic Information and Analysis (NCGIA)
5711 Boardman Hall
University of Maine
Orono, ME 04469
beard@spatial.maine.edu
http://www.ncgia.maine.edu

Dr. Carol Bult ? Workshop Co-organizer
National Center for Geographic Information and Analysis (NCGIA)
5711 Boardman Hall
University of Maine
Orono, ME 04469
cbult@spatial.maine.edu
http://www.ncgia.maine.edu

Dr. Marc Armstrong
University of Iowa

Dept. Geography
316 Jessup Hall
Iowa City, IA 52242
marc_armstrong@uiowa.edu
http://www.uiowa.edu/~geog/faculty/armstrong.html

Dr. Judith Blake
Mouse Genome Informatics Group
The Jackson Laboratory
600 Main Street
Bar Harbor, ME 04609
jblake@informatics.jax.org
http://www.informatics.jax.org

Ms. Marisa Da Motta
Department of Spatial Information Science and Engineering
University of Maine
5711 Boardman Hall
Orono, Maine 04469
marisa@spatial.maine.edu

Mr. John Guidi
Department of Computer Science
Room 3228
A.V. Williams Building
University of Maryland
College Park, MD 20742
jng@cs.umd.edu
Guidi@mit.edu

Dr. Jim Kadin
Mouse Genome Informatics Group
The Jackson Laboratory
600 Main Street
Bar Harbor, ME 04609
jak@informatics.jax.org
http://www.informatics.jax.org

Dr. Dan Griffith
Department of Geography
144 Eggers Hall
Syracuse University
Syracuse, NY 13244-1090
tele: 315-443-5637
FAX: 315-443-4227
griffith@pop.maxwell.syr.edu

Dr. Dong-Guk Shin
Computer Science and Engineering
University of Connecticut
191 Auditorium Road, U-155
Storrs, CT 06269-3155
TEL: (860) 486-2783/3719
FAX: (860) 486-4817
shin@eng2.uconn.edu
http://www.ucc.uconn.edu/~wwwgeog/shintest.htm

Dr. Lei Liu
Computer Science and Engineering
University of Connecticut
191 Auditorium Road, U-155
Storrs, CT 06269-3155
leiliu@brc.uconn.edu

Dr. Fei Shi
Computer Science and Engineering
University of Connecticut
191 Auditorium Road, U-155
Storrs, CT 06269-3155
feishi@eng2.uconn.edu

Dr. Joseph Spitzner
TopoGEN, Inc.
1275 Kinnear Road
Columbus, OH 43212
614-890-7744
FAX 614-890-1865
joe.topogen@iwaynet.net
http://www.topogen.com/sbir/pubgraph.html

Ms. Aparna Yerragudi
Department of Spatial Information Science and Engineering
University of Maine
5711 Boardman Hall
Orono, Maine 04469
aparna@spatial.maine.edu

Activities and Findings

Research and Education Activities: (See PDF version submitted by PI at the end of the report)

Findings: (See PDF version submitted by PI at the end of the report)

Training and Development:

Training and development

The project contributed to research and training developments on several fronts but most particular I the promotion of interdisciplinary training. Undergraduate students in the Department of Spatial Information Science and Engineering are now exposed on a regular basis to genome informatics as one application area for geographic information systems. One presentation to the first year class in their required first year seminar is a presentation on spatial genomics by Carol Bult. This presentation has been very well received and more than one student has chosen this as the topic for their final class reports. Two undergraduate students were supported under an associated REU grant. These students attended the regular project meetings among faculty from spatial information science and engineering, bio-informatics and computer science. They were exposed

to view points of these other disciplines as well dynamics of interdisciplinary research. Several graduate students have been exposed to this interdisciplinary project by undertaking independent study projects. One graduate student from computer science found this area to be excellent niche for herself and is continuing this work under another related grant. All of the participating faculty have benefited from a new interdisciplinary perspective, by participating in regular meetings and attending conferences outside of their normal disciplinary expertise.

The project was unexpectedly cited at the 1999 UCGIS Summer Assembly in Minneapolis in the context of a panel convened on the NSF report 'Geographic Information Science: Critical Issues in an Emerging Cross-Disciplinary Research Domain' (http://www.geog.buffalo.edu/ncgia/workshopreport.html). One of the panelists Shashi Shekhar, from the Computer Science Department at the University of Minnesota. criticized the GIS community for not thinking big enough, and then spent 5 minutes praising the efforts of the researchers at UMaine for their investigation of the application of spatial concepts to genome research. He called this research an approach to demonstrate that the results from GIS research are 'main stream'.

Outreach Activities:

Outreach Activities

Outreach activities included visits to local schools, presentations at local meetings, an article in the Portland Press Herald, and articles in the Maine Alumni magazine and in Wired magazine. The project was routinely mentioned as an interesting interdisciplinary collaboration to prospective students visiting the Department of Spatial Information Science and Engineering. The project has been presented to the Expanding Your Horizons project, a group of 7th- 8th grade girls who visit campus to learn about science and engineering careers.

The PIs also worked collaboratively with Dr. Joseph Spitzner of Visual Genomics, Inc. (currently LabBook.com) to develop a visualization strategy based on the concepts of interactive sequence feature maps that emerged from the Workshop on Spatial Genomics. LabBook incorporates an XML-based browser that translates GenBank Feature Tables into a symbolic map of the annotations that users can dynamically scale using some of the paradigms developed as part of our prototyping efforts.

Journal Publications

D. Flewelling and M. Egenhofer

, "Using Digital Spatial Archives Effectively.

", International Journal of Geographical Information Science,
, p. 1, vol. 13, (1999). Published,

M. Egenhofer and A. R. Shariff

, "Metric Details for Natural-Language Spatial Relations
", ACM Transactions on Information Systems,
, p. 295, vol. 16 (4), (1998). Published,

A. R. Shariff, M. Egenhofer, and D. Mark.
, "Natural-Language Spatial Relations Between Linear and Areal Objects:
The Topology and Metric of English-Language Terms.
", International Journal of Geographical Information Science, p. 215, vol. 12 (3), (1998). Published,

A. RodrÝguez and M. Egenhofer
, "A Comparison of Inferences about Containers and Surfaces in
Small-Scale and Large-Scale Spaces.", Journal of Visual Languages and Computing, p. 639, vol. 11(6), (2000). Published,

Beard, M. K., "Propagation of Metric Uncertainty to Topological Uncertainty", International Jounral of Geographic Information Systems, p. , vol. , ( ). Submitted,

Books or Other One-time Publications

F. Fonseca, M. Egenhofer, and
C. Davis
, "Ontology-Driven Information
Integration.
", (2000). , Published
Editor(s): C. Bettini and A. Montanari
Collection: AAAI-2000 Workshop on Spatial
and Temporal Granularity,

Bibliography: Austin, TX
, August 2000

M. Bertolotto and M. Egenhofer
, "Progressive Vector Transmission.
", (1999). , Published
Editor(s): Bauzer Medeiros
Collection: 7th ACM Symposium on Advances
in Geographic Information Systems
Bibliography: pp. 152-157, November 1999.

K. Hornsby, M. Egenhofer, and
P. Hayes
, "Modeling Cyclic Change
", (1999). Book, Published
Editor(s): P. Chen, D. Embley, J.
Kouloumdjian, S. Liddle, and
J. Roddick

Collection: Lecture Notes in Computer
Science, Vol. 122

Bibliography: Springer-Verlag, pp. 98-109,
November 1999.
A. RodrÝguez and M. Egenhofer.
, "Putting Similarity Assessments
into Context: Matching
Functions with the User's

Intended Operations.
", (1999). Book, Published
Editor(s): P. Bouquet, L. Serafini, P.
Brezillon, and F. Castellani
(eds.),
Collection: Modeling and Using Context,
CONTEXT-99
Bibliography: Lecture Notes in Artificial
Intelligence, Vol. 1688,
Springer-Verlag, pp. 310-323,

K. Hornsby and M. Egenhofer
, "Shifts in Detail through
Temporal Zooming", (1999). Book, Published
Editor(s): A. M. Tjoa, A. Cammelli, and
R. Wagner

Collection: Tenth International Workshop
on Database and Expert Systems
Applications: Spatio-Temporal
Data Models and Languages,

Bibliography: IEEE Computer Society, pp.
487-491, August 1999.
K. Hornsby and M. Egenhofer
, "Identity-Based Change
Operations for Composite Objects
", (1998). Book, Published
Editor(s): T. Poiker and N. Chrisman (eds.)

Collection: Eighth International
Symposium on Spatial Data Handling

Bibliography: Vancouver, Canada
, pp. 202-213, July 1998.
K. Hornsby and M. Egenhofer
, "Qualitative Representation of
Change
", (1997). Book, Published
Editor(s): S. Hirtle and A. Frank (eds.)

Collection: COSIT '97
Bibliography: Laurel Highlands, PA Lecture
Notes in Computer Science,
Vol. 1329, Springer-Verlag,
pp. 15-33, October 1997.

Web/Internet Site

URL(s):

www.spatial.maine.edu/~cbult

Description:

This site was developed to track progress of this project and inform others about developments under the project. it inlcudes links to related projects and well as project specific information.

Other Specific Products

Product Type:

Data or databases

Product Description:

We developed an Oracle database. Input data for the database consists of GenBank formatted files that include the nucleotide sequence, feature names, coordinates, and some feature attributes. We assume that the genome information represented in GenBank represents the best current understanding of what features occur in an organism?s genome. Over a dozen complete genomes, including those of viruses, bacteria, archaea, and yeast have been loaded into the database.

Sharing Information:

We plan to make this database web accessible to other researchers

Product Type:

Software (or netware)

Product Description:

We developed an L-Scan and R-scan modules in Visual Basic, a 1D K function in Visual basic, a spatial query processor in Visual C

Sharing Information:

This software is freely availbe to anyone interested in working with it. The next implementation will be web based so it will be more easily shareable.

Contributions

Contributions within Discipline:

The contributions of our investigation include foundation components for a system that can improve understanding of the nature of spatial structure and spatial relationships among genome features that may ultimately drive discovery of biologically significant organizational and structural features encoded in genome data.

Contributions to Other Disciplines:

The contributions to spatial information science and computer science were the challenges of adapting concepts to a new problem domain. Geographic spatial problems have focused on 2D representations. Genome model development opened new insights on one dimensional spatial representations and relations including modeling the uncertainty of spatial relations. The complexity of genetic feature interactions also prompted new ideas for modeling partially ordered structures within an object ?relational database.

Contributions to Human Resource Development:

The contribution to human resource development was the exposure of several female students to this topic as an important and growing research area. Also the project was able to successfully foster several interdisciplinary interactions which would not have occurred without the support from the project.

Contributions to Resources for Research and Education:

An institutional contribution has been the closer collaboration between the University of Maine and The Jackson Laboratory. The initial collaboration of this project has formed a key basis for new collaborative research initiatives as well as broader interdisciplinary collaborations between the institutions. The work of

this project formed a foundation piece for a follow on research proposal to NIH.

Contributions Beyond Science and Engineering:

We are confident that the outcomes of this research can provide the basis for commercial spin offs particularly in the area of tools for drug development.

Categories for which nothing is reported:

Research Findings

The database of genomes and the suite of tools we have developed to analyze them are not sufficiently large and integrated to make significant biological findings as yet. However we are able to address findings on the utility of spatial concept similarity between geographic and biological structure that will support further developments. Specific findings of the project are organized and summarized under the following sections:

1. Spatial Modeling of the Genome

The goal of the modeling effort was to identify differences and commonalities between biological and geographic spatial structure and adapt geographic spatial modeling concepts where possible to the biological context. We identified several commonalities in the approach to spatial representation, visualization, and analysis. Like geographic entities, genomes are complex objects with internal spatial structure. Geographic entities exhibit some hierarchical spatial structures in which smaller entities nest cleanly within larger entities, but more often clear hierarchical structures are not maintained. Genome features of interest exhibit similar characteristics with some clear hierarchical structure but many exceptions that needed to be accommodated in the model. Geographic entities have various characteristics and behaviors that are influenced by their spatial context or situation. We can identify multiple overlapping but individually cohesive geographic neighborhoods defined by various underlying spatial processes. Similarly genome spatial structures appear to be influenced by and aligned with spatially bound biological processes. Like the symbolic abstractions used to represent geographic entities and interact with them, symbolic abstractions can be applied to represent genome features and support interaction with them. These commonalities allow us to directly or with minimal revision apply several existing spatial modeling concepts for genome representation. These include:

- support for partial hierarchical relations
- discovery of non-hierarchical relations through spatial overlay or segmentation
- support for multiple representations of entities at different scales or resolutions
- support for spatial zooming and panning
- support for representation of spatial relations

Some modification is required on this last point. Most geographic information systems consider spatial relations within a 2 dimensional context. Assuming a 1 dimensional abstraction of a genome, one dominant order applies and the set of applicable relations is better modeled by Allen's (1983) temporal intervals.

Our spatial data model supports abstractions of genomes at a high symbolic level. The model allows us to identify both qualitative and quantitative spatial relations among genome features, to support multiple spatial representations of genome features for different scale views and to use these structures to deduce currently unknown structures and relations.

A spatial reference framework can be imposed on a genome to locate and inter-relate genetic substructures. As in the geographic case we can use the spatial reference frame to identify shared spatial locations of features (overlaying bedrock, soils and land cover in the geographic case, overlaying genetic features in the biological case). As in the geographic case, when clear spatial relations are evident these can be explicitly incorporated in the model. The model further allows for deduction of other spatial structures (functional sub-systems) and the ability to explicitly record these as spatial objects if desired through a recursive feature – super-feature relationship. We have also built a foundation for recording uncertain spatial relations among genetic features by expressions of probabilities for a set of interval spatial relations (Beard – forthcoming).

Independently observed geographic variables typically require a common spatial reference frame in order to perform comparisons. When a common spatial framework does not exist, well-defined features or control points and a rubber sheet transformation are typically employed to obtain a common reference frame. Within our model each genome has an independent spatial reference framework and thus transformation to a common reference frame is not possible. The alternative approach for comparative analysis across genomes is to use qualitative spatial and semantic relations with metric refinements to determine patterns of similarity across genomes. This topic is discussed further in section 3 below.

Visualization

The user survey of graphic depictions of the genome described in the Activities Section, was not sufficiently large for statistical tests, however some trends were apparent in the results. The majority of respondents preferred horizontal to vertical depiction of the genome. A majority of respondents also depicted the genome figures in a consistent way as intervals scaled by the length (number of base pairs between start and end points). These preferences were, in general, implemented within the graphical genome display. There was much less consensus among respondents on the depiction of uncertainty in genome features. The display of uncertainty is not yet operational within the system.

Spatial analysis

One goal for the spatial analysis of genomes was to develop methods for the detection of spatial dependencies among genetic features. At one scale genomes can be conceptualized as spatial point patterns. That is each genome feature can be conceptualized as a discrete point. A spatial point pattern is a data set consisting of a series of point locations within some spatial region at which events of interest (genome features in this case) have occurred. This conceptualization allows us to apply statistical tests for spatial patterns in genome feature distributions. We can test whether patterns of genome features of a specific type, subtype or function have a significant spatial pattern or are no different than random. We implemented a 1 dimensional K function (see activities) which allows us to test for spatial dependence among genome features. We can test if there is spatial clustering among features in general or more specifically if there is clustering among features of a specific type or functional role. To apply spatial point process statistics requires converting the interval based feature representation to a point. The representative point can be the start point, end point or mid point of the interval. This

abstraction omits the spatial information pattern of the interval lengths. Scan statistics are another statistical approach for detecting spatial structure that takes into consideration the lengths of the intervals. We have implemented an L scan statistic and used it to discover spatial structure within the yeast genome. These statistics allows us to determine if the pattern of genetic features are random, or if they deviate from randomness in the direction of clustering or regularity. Once more functional annotations have been added to the database more extensive analysis will be possible. We will, for example, be able to generate hypotheses about why a particular set of features are clustered.

Another analysis component was to develop methods for the comparison of genomes across organisms. A spatial comparison of genomes poses some problems if quantitative metrics are used. A qualitative comparison becomes useful in that it allows an ability to compare and reason with incomplete or weak information and in this case across he independent (non-compatible) spatial reference frames of each genome.

In geographic space, various representations can be transformed (scaled, translated and rotated) to a common coordinate system based on references to an absolute (Earth based) coordinate system (lat, long) or based on a relative image to image mapping that uses a common set of landmarks appearing in each image. The analogy of landmarks in a genome is possible but not robust. Landmarks are domain dependent and scale dependent. Landmarks have been defined as uniquely distinguishable points that are visible in a scene. The alternative is to base a comparative analysis on a qualitative reasoning approach. This approach employs a similarity metric based on a set of qualitative spatial relations. A set of genetic features from a genome can be characterized according to numbers of objects, semantic similarities among these objects and pairwise qualitative spatial relations among these objects. This similarity metric allows genome comparisons at levels higher than a molecular level, (a need expressed by Ouzounis et al (1996)). This approach provides a solid basis for comparative analysis of genomes at many different levels of spatial detail above the sequence level. The approach does not require complete genome sequences.

Future Directions

The data model, database, spatial query, and spatial data analysis components that were developed during this funding period form the basis for an integrated analysis workbench for genome sequence feature maps that is currently underway. We anticipate that this integrated system will be useful for detecting and analyzing aspects of the spatial organization of genome features and will serve as a general tool for pattern detection and comparison down the road.

Allen, J. 1983. Maintaining Knowledge about Temporal Intervals. Communications of the ACM. 26(11) 832-843.

Ouzounis, C, G. Casari, C. Sander, J. Tamames and A. Valencia. 1996. Computational comparisons of model genomes. TIBTech Vol 14 pp 280-285.

Research Activities

The primary goal of this project was to develop a data model and computational representations of genome data that could enhance the ability of biologists to identify, compare, and test for biologically significant aspects of genome organization and function. Under the primary goal the project had three main objectives and four specific tasks as outlined below.

Objectives

1. To develop a conceptual spatial model of genomes
2. To develop visualization and query tools for genome analysis
3. To develop spatial statistical analytical approaches based on genome spatial representations

Tasks

Task 1. Investigate the properties of genomes as spatial objects and formalize them.
Task 2. Integrate spatial genome information in a queryable database architecture.
Task 3. Develop an interactive graphical user interface for analyzing, comparing, and displaying spatial and non-spatial genome data
Task 4. Develop the utility of geographic information science concepts and spatial data analysis methods to the analysis and interpretation of genome data.

A summary of research activities is organized by the three objectives given above.

1. Conceptual Model Development

Because this work is interdisciplinary we felt the modeling effort needed early interdisciplinary input to develop a common understanding of concepts and terms. A first step was to gather several researchers from relevant disciplines for a workshop. A "Spatial Genomics" workshop was held in Northeast Harbor, Maine in October 1997. Workshop presentations by participants covered the following topics:

- a summary of current genome mapping methods and data representation issues,
- an overview of how mapping data are represented in the Mouse Genome Database,
- reviews of previous work in map integration inference methods,
- genome sequence analysis and data representation issues,
- an overview of geographic information system methods and concepts,
- a description visualization standards efforts for genome data,
- ideas on how some spatial data analysis methods might apply to genome data, and
- ideas on spatial models and methods that might be appropriate for genome data.

Outcomes of the workshop included a functional overview of a Spatial Genomics System, a collection of types of questions such a system should address and specification of some analytical approaches Examples of representative questions to be supported by such a system included:

Show the locations of all of genes that contain inteins.

How many genes containing inteins have insertion elements within 500 bp of their start codon?
Show all features within 1000 bp upstream of the XYZ gene.
Show the locations of all genes that have names like "dehydrogenase"
Show the locations of all of the genes in this genome whose protein translations contain at least two predicted transmembrane regions.
Show the location of all genes that are longer that 500 bp and are annotated as being involved in amino acid biosynthesis.
Show only those genes that are expressed in this genome under the conditions described in experiment X and NOT under the conditions of experiment Y.

Statistical/Pattern Detection methods that a model should support were suggested to address questions such as the following:

Is the organization of the amino acid biosynthesis genes in this genome clustered or random?
Is the topology of amino acid biosynthesis genes in a genome similar to that found in other genomes?
Is the expression of gene XYZ significantly correlated with the up or down regulation of expression with another gene in the genome?

The development of a conceptual spatial model for genomes evolved from the workshop discussions and from subsequent weekly discussions among the PIs. A first conceptual model was developed and implemented in Oracle 7.3. Over the course of the project this model has been refined as a result of on going discussion among the PIs and with input from biologists, database specialists, and software engineers. The goals for the spatial model were that it provide 1) a spatial description of an individual genome such that sub-structures could be mapped, related and analyzed and 2) a spatial structure that supported the comparison of one genome and its sub-structures to another genome.

The newest version of the conceptual model is being implemented in the Oracle 8i Spatial database. The current model includes the entities: organism, genome, chromosome, feature, feature set, DNA sequence, transcript, protein, and the role of protein sequences. The organism is the creature from which a genomic sequence is obtained and annotated. The genome is the complete genetic information of an organism. The chromosome is a self-reproducing structure onto which an organism's genome is organized, packaged and replicated. A feature is a biologically significant region in a genome that can be predicted computationally or experimentally. A feature set is an aggregate of features that may or may not be spatially contiguous. The DNA sequence is the sequence of nucleotides. A transcript is a nucleotide sequence resulting from an actual or predicted transcription process. Protein is the amino acid sequence for a protein coding gene that is the result of an actual or predicted translation of a transcript. A role is the high level cellular job performed by a feature, transcript, or protein.

One of the benefits of the data model we have developed is that it generalizes the concept and representation of a genome "feature." The concept of feature allows the

representation of genes as well as other biologically significant features that can be represented as points or intervals. The concept of "feature set" means that features can be combined to form more complex features. For example introns and exons are features in and of themselves. They can also be combined to form a gene feature. Genes, in turn, can be combined to form the concept of an operon feature or a gene cluster feature.

The current form of the model is shown in Figure 1.

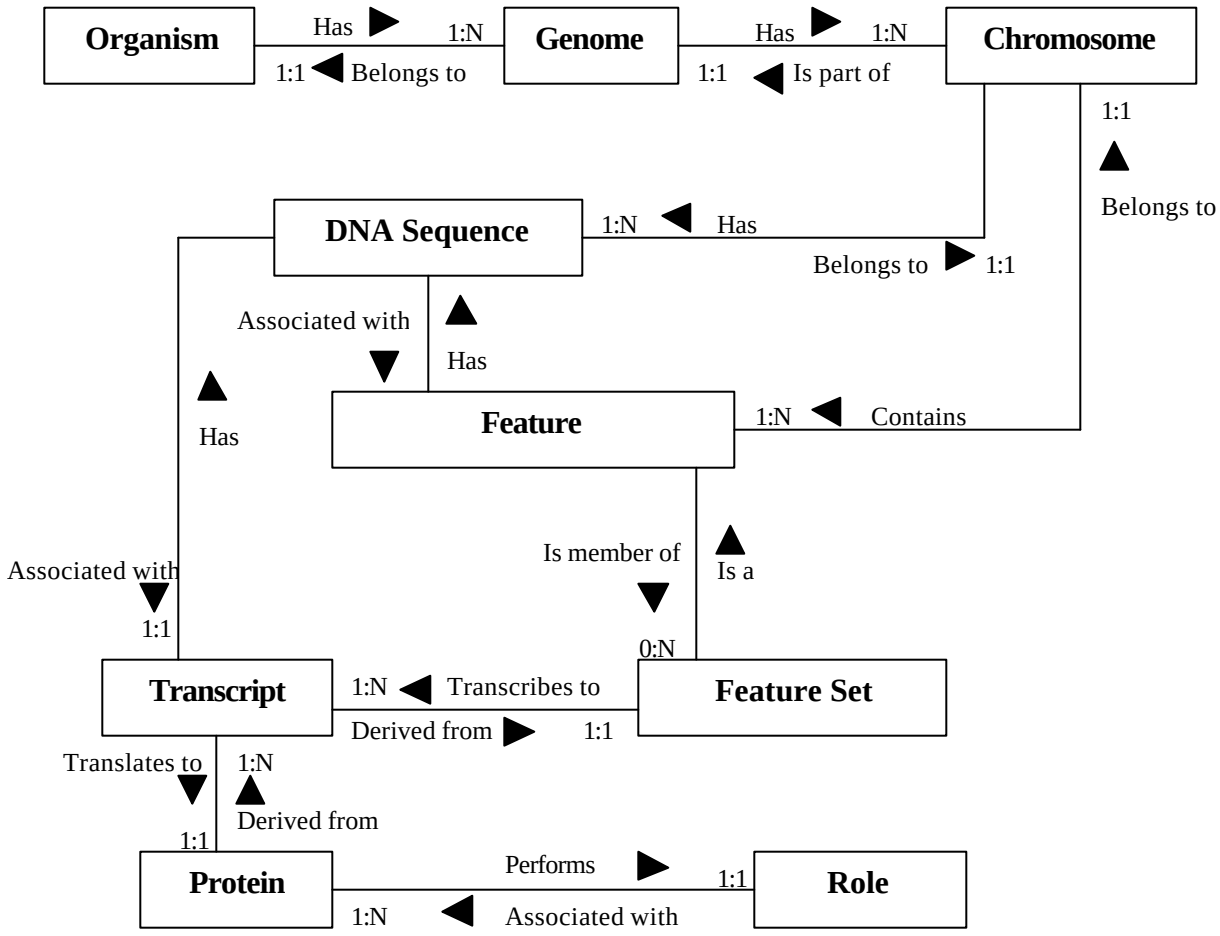


Figure 1. Entity Relation diagram of a genome and related entities

Of the entities in the model: genome, chromosome, feature, feature set, DNA sequence, transcript and protein are modeled as spatial objects. As spatial objects, these entities can be represented by any spatial primitive. The default representation is a spatial primitive called a chain. A chain is a directed sequence of non-intersecting line segments. The motivation for the chain primitive is that it supports interval logic for 1D spatial relations and since it has direction it supports modeling transcriptional direction. Under this model a single nucleotide is a zero length chain or equivalent to a point.

A chain has as attributes a unique identifier, and start and end nodes. As 1 dimensional spatial primitives, chains have the following spatial relationships as identified by Allen (1983). Genome features represented as chains are ascribed to have this set of qualitative spatial relations. In the biological context, genome features have functional and evolutionary relations. By modeling them as spatial objects we can investigate whether functional and evolutionary relations have spatial structures.

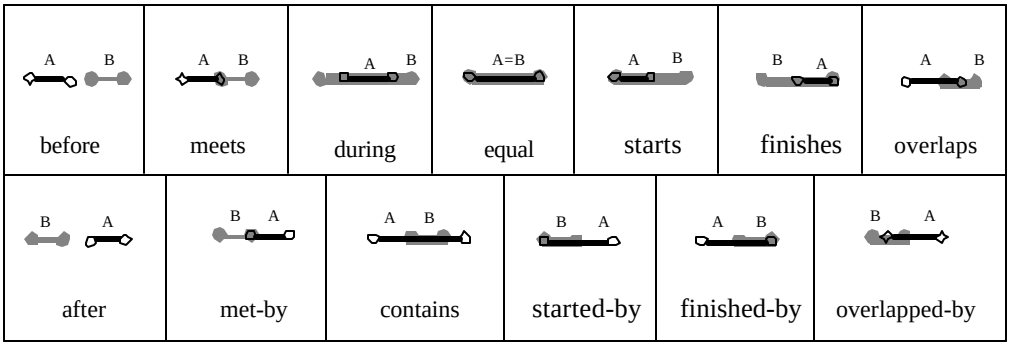


Figure 2. . Interval relations as defined by Allen (1983).

This data model has the advantage of allowing the representation of genome features at many different resolutions. For example in addition to allowing the representation of a traditional view of genes as features composed of exons and inteins the model can accommodate higher level features such as functional neighborhoods.

2. Visualization and Query of Genomes

A conceptual spatial model of a genome provides a foundation for query and display of genome data. Our second objective was to develop effective visualization and query methods for use within an interactive exploratory environment. As a first step in developing a visualization approach we investigated several depictions of genomes and their substructures.

Perception of Genome Features and Their Spatial Relations

We conducted a survey of biologists from The Jackson Laboratory and graduate students at NCGIA to solicit their input on graphic representations of genomes. Users were asked to create graphical representations of several simple genome sequence feature maps. Respondents had varying degrees of familiarity with genome data. The goal was to use the results to provide insight into how people perceive genome features and their spatial relationships to one another. Out of 32 surveys distributed, 24 were completed. These results were used to guide development of a graphical query interface. Key results from the survey are summarized below:

Genome Orientation: the majority of respondents oriented the genome feature and sequence feature maps horizontally.

Feature Depiction: Features and feature sets were graphically distinguished by symbolic representations with the preferred order being color, shape, and fill pattern

Distance representation – Respondents used tick marks to indicate distance units along a linear axis

Strand representation: The majority of respondents represented strandedness by drawing features above or below a single linear axis
Representation of relationship uncertainty- Respondents used various symbolic representations to indicate uncertainty in the relationships among features. Responses included use of an approximation symbol, question marks, and buffer areas at feature end points to indicate uncertainty.

Prototype graphical user interface.

From the results of the survey as well as a review of the literature on genome and sequence displays, we designed a prototype user interface. A genome is displayed in a manner similar to sheet music using a series of panels as shown below in Figure 3. A smaller panel at the bottom of the screen is an index panel that indicates how much of the genome is currently displayed. Some basic attribute information about the selected genome is included at the top of the display.

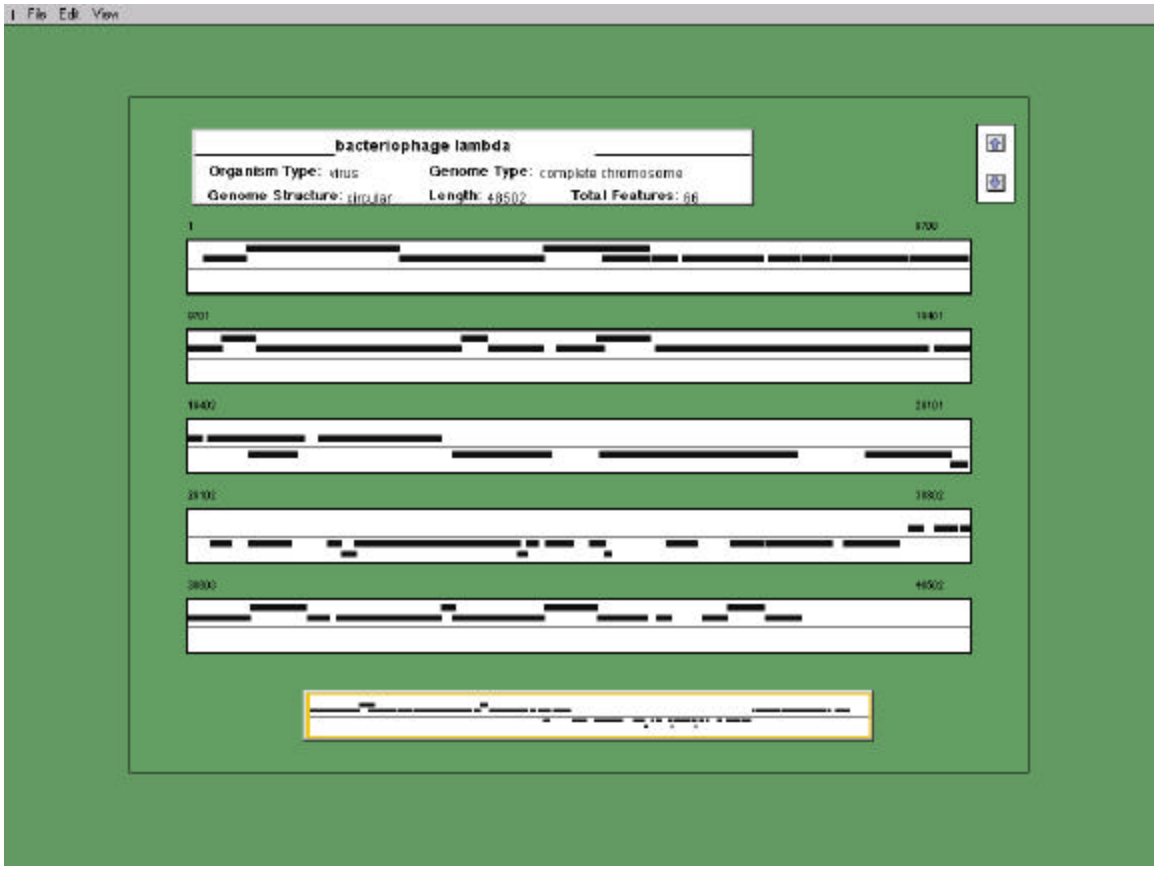


Figure 3. Display of a single genome across several display panels.

This format allows the display of a single genome across several panels or the display of one genome per panel to support visual comparison of multiple genomes. In the former case the format supports changes in scale or resolution. The up and down arrows in the

upper right corner allow an increase or decrease in the number of panels supporting the expansion or contraction of the display of the genome (e.g., from a display across five panels to two panels) (see Figure 4).

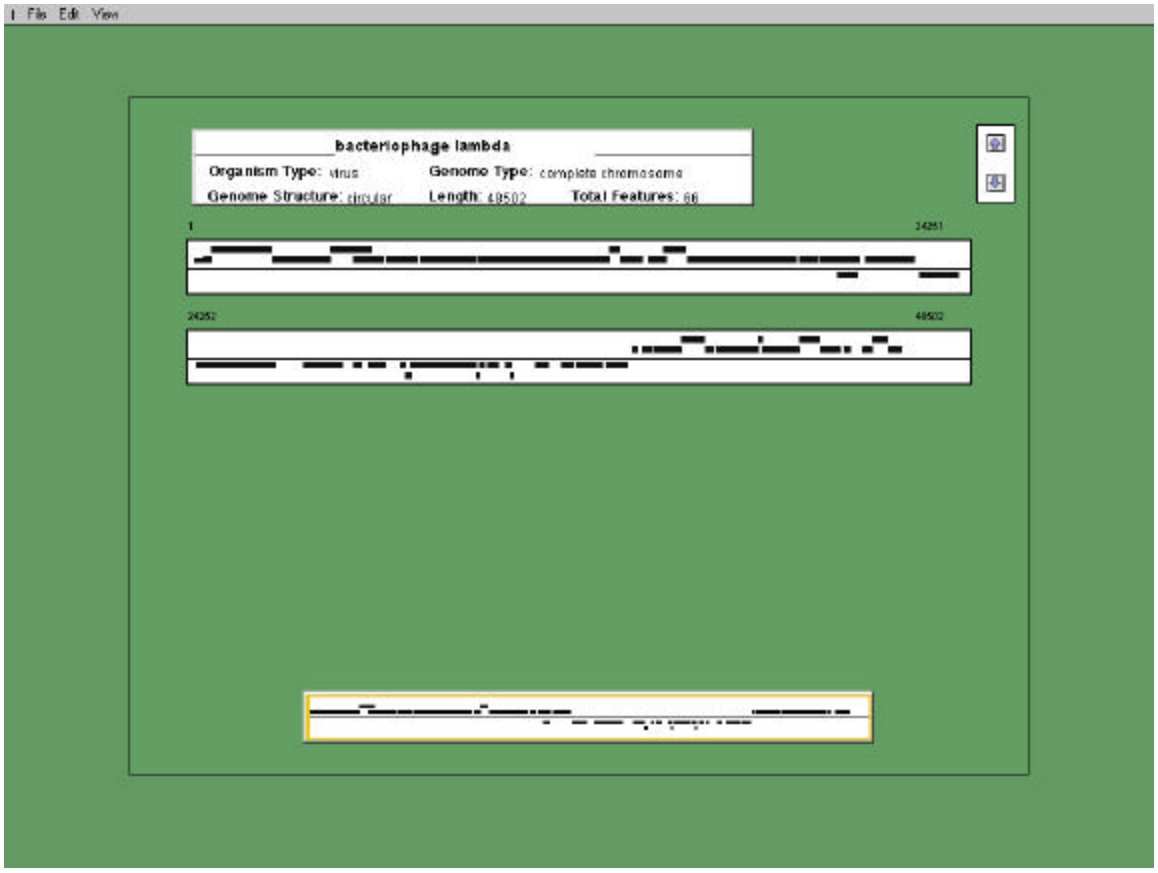


Figure 4. The number of display panels can be collapsed which is comparable to zooming out. The whole genome is now compressed into two panels.

The interface allows graphic display of features and feature sets in black and white that provides a general pattern of feature distribution across a genome. Features can also be color coded by any of the feature attributes stored in the database. In Figure 5 features are shown color coded by functional role. This dynamic interaction between the user and the sequence feature map is very similar to the creation of a choropleth map in geographic information systems. This is a simple form of pattern based data mining that allows users to search genomes visually for clusters of like elements on the map.

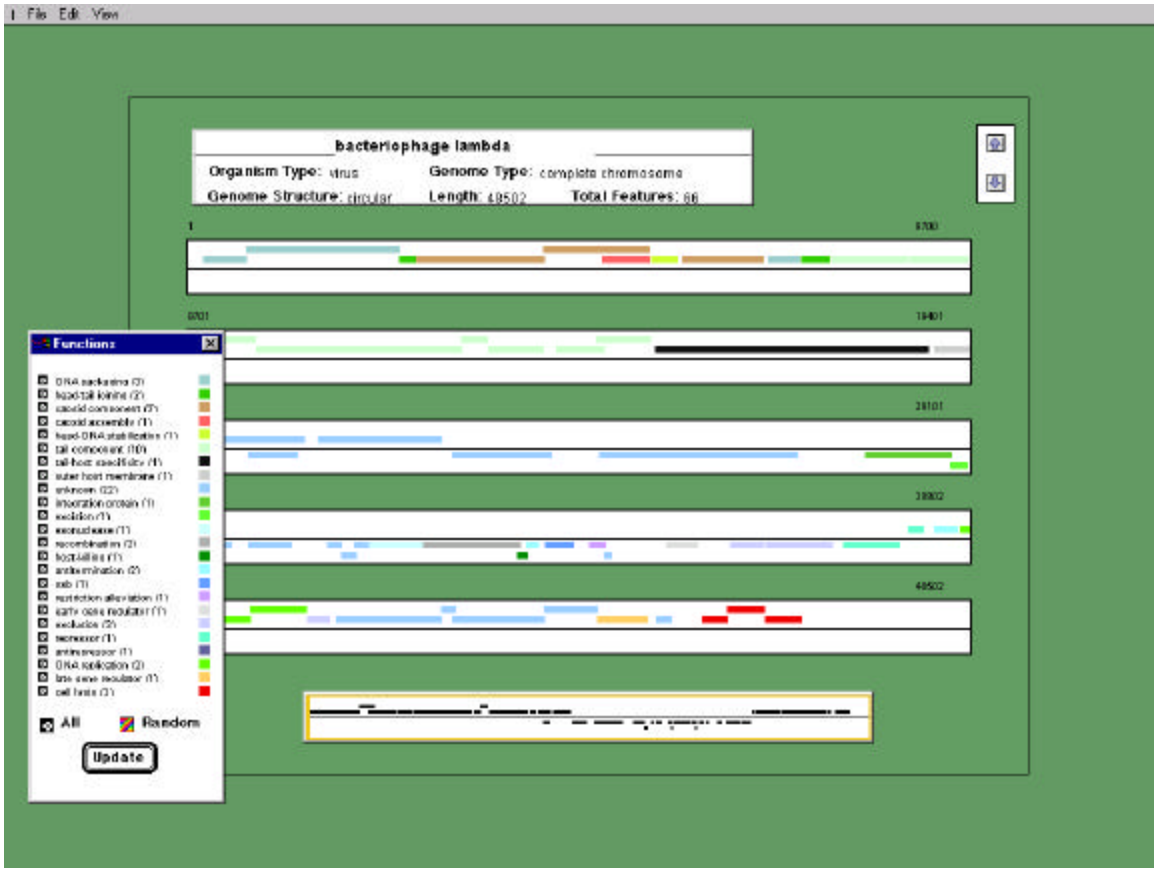


Figure 5. Users can toggle from black and white to color displays of genome features.

Every feature is an interactive object that can be directly queried for its attributes. Figure 6 illustrates use of the eyeglass tool. Clicking a feature with the eyeglass tool displays its attributes. Each feature can also be queried for its underlying DNA sequence.

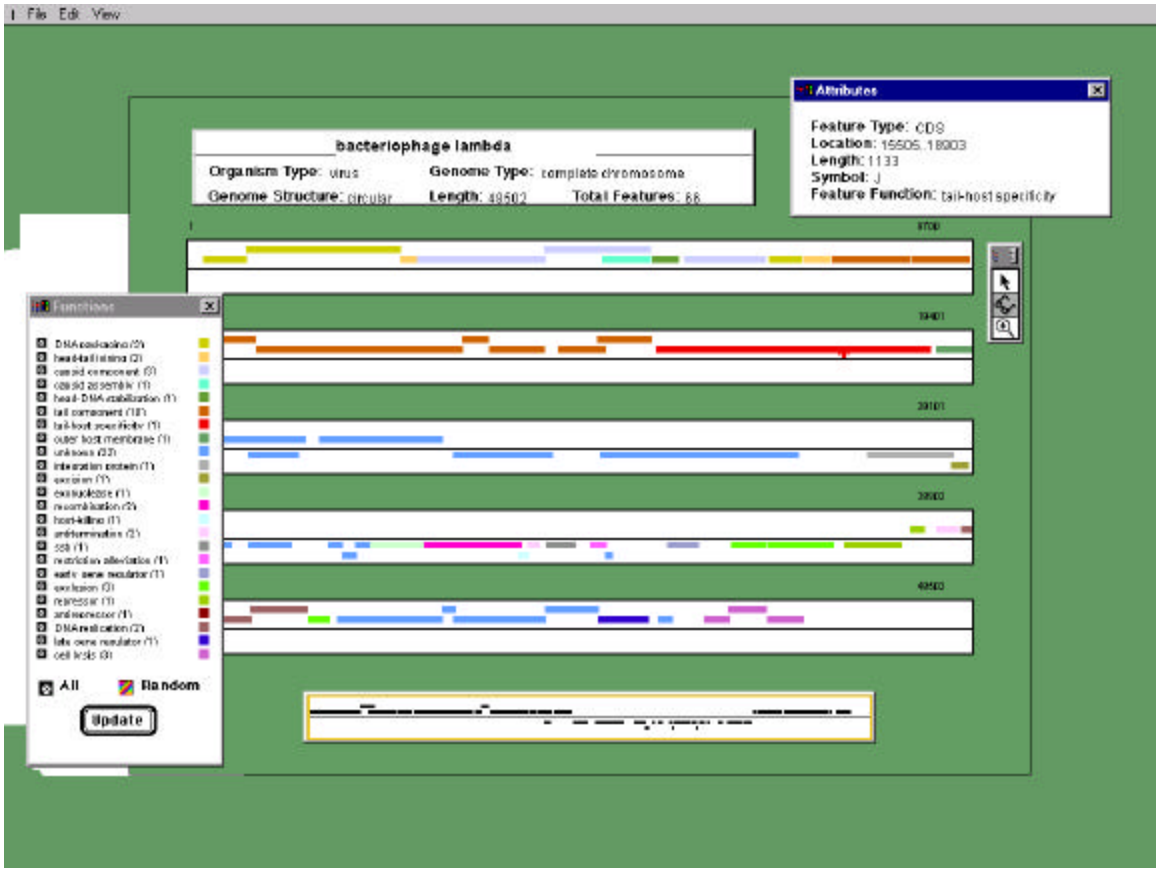


Figure 6. Clicking on features with the eyeglass tool displays the attributes of the selected feature.

Zoom capability is supported by the up/down arrows that collapse or expand the panels as well as through the index map at the bottom of the page. The index map supports both panning and zooming. Zooming is achieved by changing the size of the box, panning by changing the position of the box as shown in Figure 7.

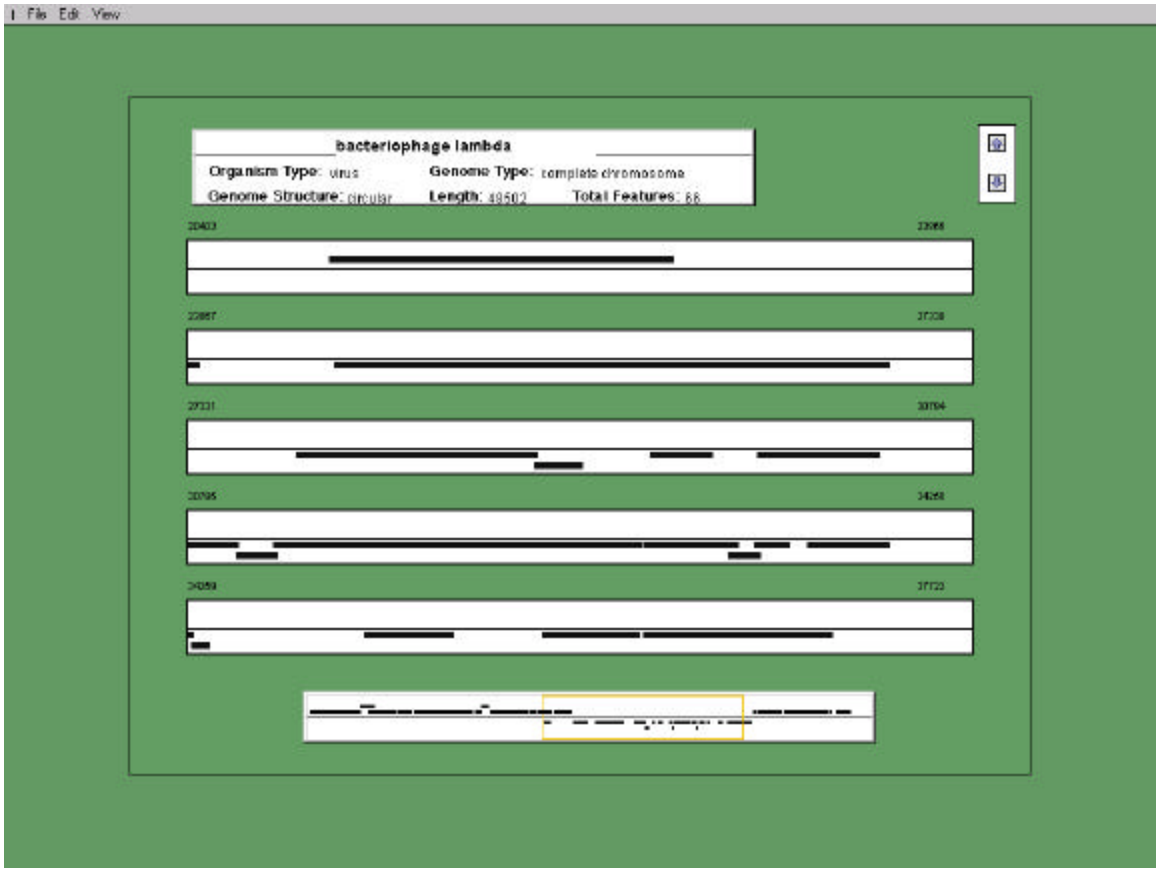


Figure 7. The index map at the bottom of the screen includes a box that can be moved or re-scaled to pan and zoom respectively.

Spatial Query of Genome Features

The ability to query a genome for spatial relations was considered to be an important function. Because of the one-dimensional nature of the conceptual genome representation, the thirteen interval relations defined by Allen (1981) (see Figure 2) were adopted. We developed a prototype spatial query logic, implemented in C++, based on these relations. The spatial relationships that can currently be queried include Meets, Equals, Overlaps, Overlapped By, Contains and Contained By. In the case of circular genomes the concepts of before and after become ambiguous. By defining a start site for a circular genome, coordinates can be assigned and the same reasoning that applies to linear genomes can apply to the circular form. These qualitative relationships can be refined metrically in some cases (e.g. Gene A before Gene B by 150 bp).

3. Spatial Analysis of Genomes

The third objective was to develop spatial statistical approaches for genome analysis. Spatial analyses of a genome that were considered to be of interest included questions such as:

Are selected features distributed in a regular pattern, clustered or random?
Are there spatial dependencies in the organization of features in a genome?
If spatial dependencies are present what are the explanations for this dependency?

Our efforts included development of a set of statistical analytical tools to investigate spatial dependencies

Scan Statistics in Genome Data Analysis

The L-scan statistic as described in Braun and Müller,[1] was implemented. as part of the project. The L-scan is a tool that can be used for interactive visual data mining and genome data interpretation. The underlying assumption is that there is information in a pattern if the pattern varies from a random arrangement.

The L-scan statistic involves smoothing or averaging point data to indicate whether the points are unusually clustered or whether or not there are segments of different composition in the sequence. The method consists of defining a weight function, g, mapping from an alphabet, Y, to the real numbers. The alphabet represents the possible values along the sequence.

We apply the method to a sequence of features along the genome. In this case, we have a two-letter alphabet, Y = {E, N }, where E represents the presence of a feature and N represents the absence of a feature. We define g:

$$g(E) = 1$$
$$g(N) = 0$$

The L-scan statistic is calculated as:

$$S_j = \sum_{i=j}^{j+L} g(Y_i)$$

that depends on the parameter L and the size of the subsequence over which the sum is calculated.

For the analysis, the L-scan for a given genome sequence is compared with the L-scan for a series of test patterns. We include in the implementation a series of test pattern files:

- a sequence of regularly spaced events,
- a sequence of randomly spaced events,
- a sequence of clustered events, and
- a sequence of events with a linear trend in the number of events.

These last two patterns were generated using a Poisson cluster process model and a heterogeneous Poisson process model respectively.

[1] J. V. Braun and Hans-Georg Müller, Statistical Methods for DNA Sequence Segmentation, *Statistical Science*, 13, 2, (1998), 142-162.

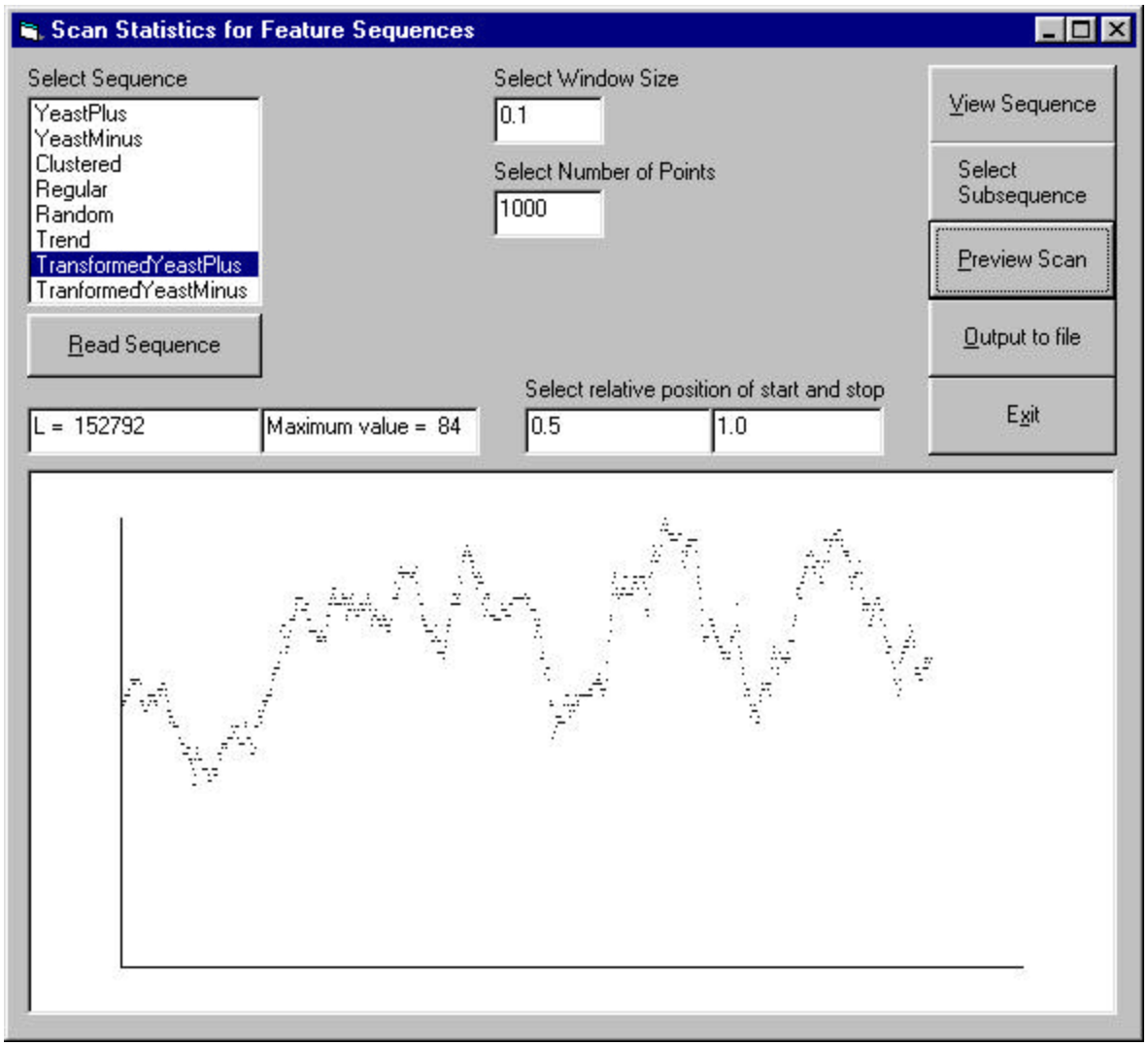
Further analysis can be used to test the significance of deviations from a random pattern, if detected. A number of sources[2] consider the distributions of several quantities that can be obtained from the L-scan. It is possible to consider:

- N_d, largest number of events in an interval of length d,
- D_n, the smallest interval containing n events,
- W (n, d), the wait interval for n events in an interval of length d.

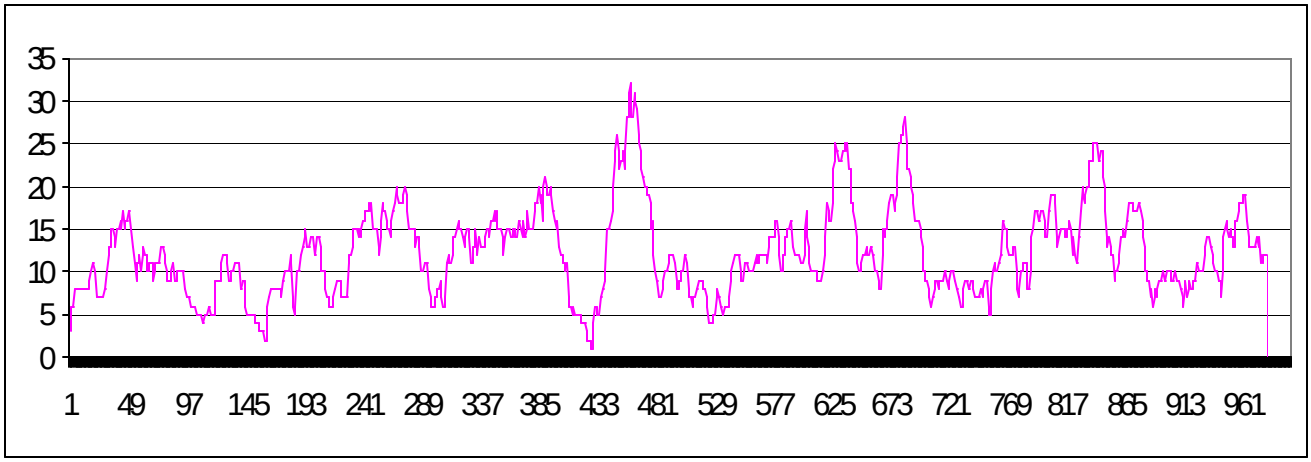
These distributions are related:[3]

$$\Pr(N_d \geq N) = \Pr(D_n \leq D) = \Pr(W(N, D) \leq D')$$

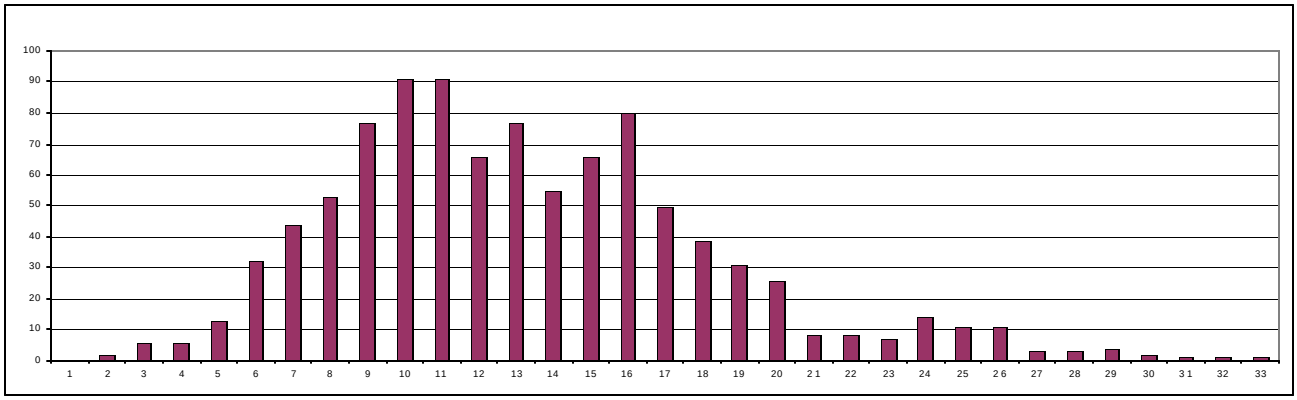
In the implementation, we generate the histogram of L-scan values, which can be used in this analysis. Several screen shots show the user interfaces. The user has the option of selecting a real feature sequence or a test pattern sequence. The user can view the feature sequence and perform the L-scan. Interactively, the user can adjust the window size, L, and the number of points at which the L-scan is calculated. The user has the option of selecting a sub-sequence from the entire sequence to view or to perform the L-scan.



Transformed Yeast – Plus strand
Scan values vs. position



Transformed Yeast – Plus strand
Scan values histogram



1D Kernel Estimate and K function

Because the genome is modeled as a one-dimensional structure, temporal analogs of the spatial statistics for point patterns were adapted. Code was developed to perform kernel estimates (Diggle, Liang and Zeger,1994) and the K function (Cressie, 1993, and Kernan, Mullenix, Kent, Hopper and Cressie, 1988). The kernel estimate is a weighted average of data points where the weights are specified using a kernel function K and a bandwidth h. The role of the kernel estimate is to investigate first order effects or spatial trends in the density distribution of features. The kernel function is normally a non-negative-valued function symmetric about zero. In this project, a Gaussian kernel ($K(u)=\exp(-0.5u^2)$)was chosen. The two-dimensional kernel, "visits" each of the events (gene features) in succession, and weights the area surrounding the event proportionately to its distance to from the event. The sum of these individual kernels is then calculated for the chromosome. The formulae used for the analyses were:

$$w_{ij}^{*}(t) = h^{-1} K\left(\frac{t_{ij} - t}{h}\right)$$

$$w_{ij}(t) = w_{ij}^{*}(t)\left[\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}^{*}(t)\right]^{-1}$$

$$u(t) = \sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}(t)\, y_{ij}$$

where t was the position along the chromosome, y is the number of features within the window, and h was the bandwidth. The user specifies the size of bandwidth. In general, the larger the bandwidth, the smoother the resulting graph.

The K function provides a summary of spatial dependence over a wide range of scales of pattern.. It is the expected number of events within a distance h of an arbitrary event. The formula used for the analysis was:

$$K_3(t) = \frac{\frac{T}{N}\sum_{i=1}^{N-1}\sum_{j=i+1} I(0 < |t_j - t_i| \le t, d_i > t}{\sum_{i=1}^{N-1} I(d_j > t)}$$

where T is the length of the transformed genome, N is the number of features, t_i is the transformed length of the i_{th} feature and d_i is the transformed length between t_i and T+1. I was an edge correction, and was evaluated as 0 or 1.

These methods were applied to the yeast chromosome that was selected because of its size and availability. Following the method of Cressie (1993), and Kernan et. al. (1988) the chromosome data was transformed, to show only the feature starts, and the length of the between feature segments; i.e. all but the first base of a feature was removed. These first bases of features became the events that were analyzed. The data consisted of the total number of bases on one strand of DNA, and the start and stop coordinates of the features identified along this strand. A comparison chromosome of the same length was generated in which features were randomly distributed along the chromosome. The number, average length and range of sizes were the same for both chromosomes. The results of both analyses suggested that features were not evenly distributed along the chromosome. There was a region of very low density (more than 2 standard deviations below the mean) near the beginning (3' end) of the chromosome, with clusters of features near the middle and the end (5' end) of the chromosome

T. C. Bailey and Anthony C. Gatrell, *Interactive Spatial Data Analysis*, Longman Scientific and Technical, UK, 1995.

Pierre Baldi and Soren Brunak, *Bioinformatics: The Machine Learning Approach*, MIT Press, Cambridge, 1998.

J. V. Braun and Hans-Georg Müller, Statistical Methods for DNA Sequence Segmentation, *Statistical Science*, **13**, 2, (1998), 142-162.

F. W. Huffer and Chien-Tai Lin, Approximating the Distribution of the Scan Statistic Using Moments of the Number of Clumps, *Journal of the American Statistical Association*, **92**, 440, (1997), 1466-1475.

C. Mack, An exact formula for $Q_k(n)$, the probable number of k-aggregates in a random distribution of n points, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **39**, (1948), 778-790.

J. I. Naus, Scan Statistics, in *Encyclopedia of Statistical Sciences*, N.L. Johnson and S. Kotz, eds., Wiley, New York, 1988, **8**, 281-284.

J. I. Naus, The Distribution of the size of the maximum cluster of points on a line, *Journal of the American Statistical Association*, **60**, 440, (1965), 532-538.

Spatial Correlation of Genome Features

We also implemented the calculation of spatial correlation of transcription factor binding sites and open reading frames (ORFs) in *Saccharomyces cerevisiae* (baker's yeast) based on r-scan statistics (Karlin and Brendl, 1992). The transcription factor binding sites and open reading frames were determined using the Saccharomyces Genome Database (SGD) at Stanford University[4] and the Transfac Database at Genomatix[5]. These were inserted

[4] http://genome-www.stanford.edu/Saccharomyces/
[5] http://genomatix.gs.de/

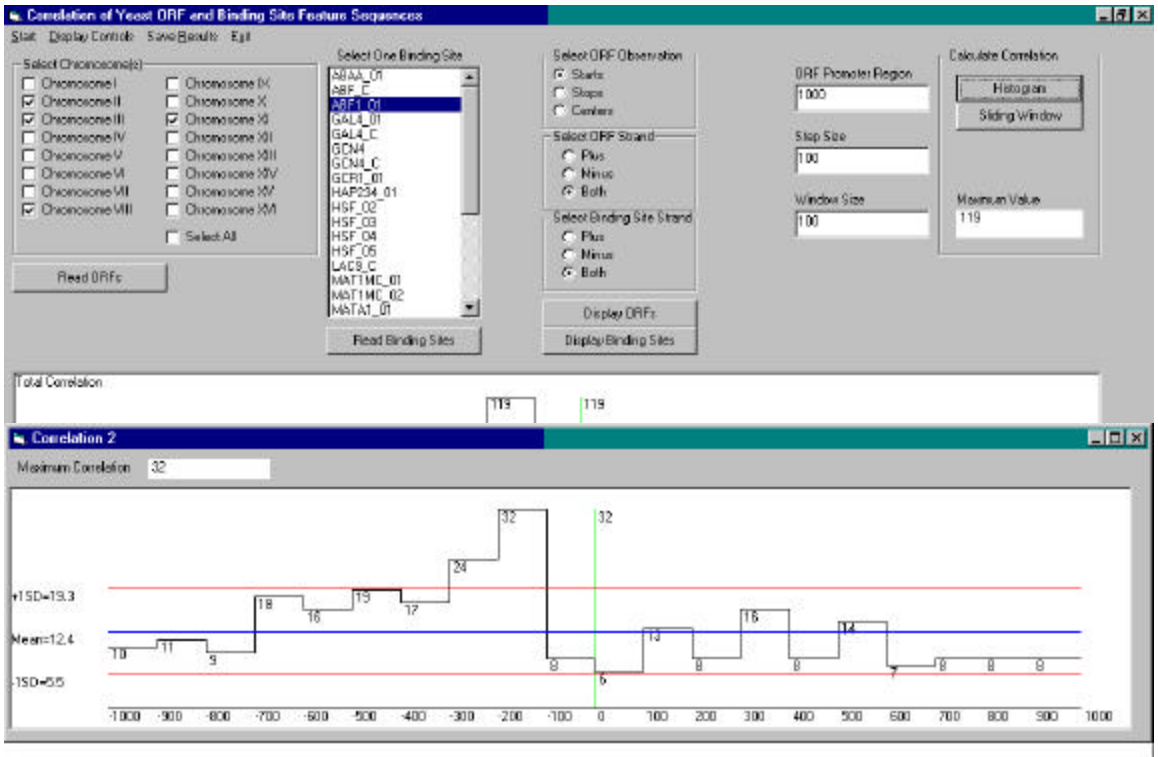
into a local database. The methods used here are completely general and can be applied to any organism or set of genome features.

We examined the promoter region upstream of ORFs for the presence of transcription factor binding sites using the methods described by Quandt et al (1996a,b). This method allows us to consider one or more types of binding sites to explore possible cooperative effects of different types of binding sites. In the database we have included the putative cellular role of the ORF products in order to investigate the association of significant correlation in the promoter region of the ORF and cellular role of the ORF. Our implementation of r-scan statistics confirmed the results of Quandt et al. (1996a,b).

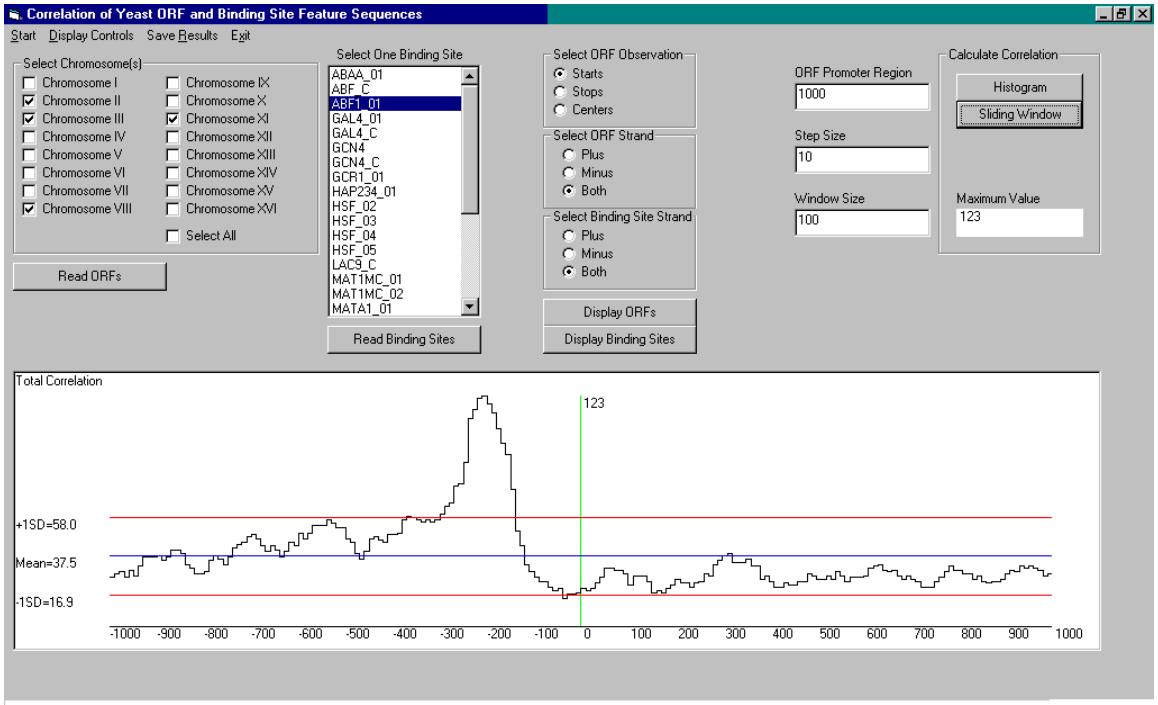
Our plans are to take the data mining potential of the r-scan method one-step further by generating "neighborhood profiles" of the regions that appear to have significant spatial dependencies. These neighborhood profiles, a form of associative data mining, will allow users to place the statistical results into biological context by showing the attributes of the features that are found to be statistically significant. In the case of the transcription factors in yeast, for example, we will display to the user the functional annotations of the ORFs that are significantly associated with specific transcription factor binding sites in order to help place the results in biological context.

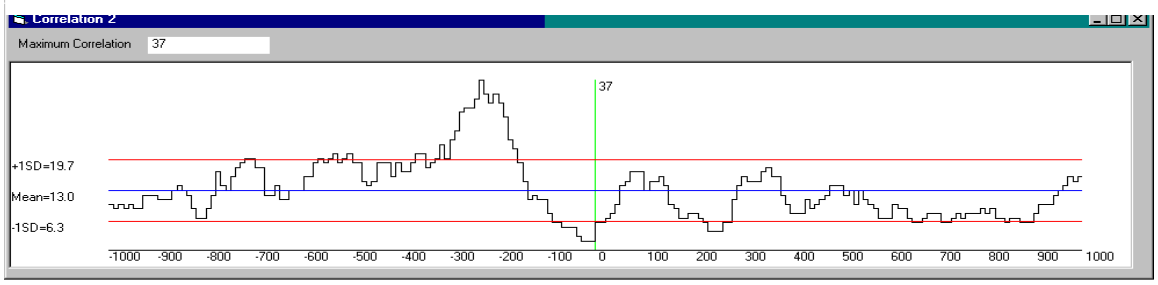
The screen shots below show the user interface. The user has the option of selecting the chromosome(s) to be analyzed, the type(s) of transcription factor binding sites to be examined, option to examine both plus and minus strands, and the size of the promoter region to be considered. For the analysis the user can choose a coarser-grained histogram or a finer-grained sliding window scan.

Histogram for all selected chromosomes, view of sequence, and histogram for chromosome 2:



Results of sliding window scan for all selected chromosomes and for chromosome 2.





Database development

Input data for the database consists of GenBank formatted files that include the nucleotide sequence, feature names, coordinates, and some feature attributes. We assume that the genome information represented in GenBank represents the best current understanding of what features occur in an organism's genome. A parser was written in Perl to extract data from GenBank files to a format that can be directly loaded into the Oracle database. Over a dozen complete genomes, including those of viruses, bacteria, archaea, and yeast have been loaded into the database. Functional annotations for 3 of the

genomes (Bacteriophage lambda, *Mycoplasma genitalium*, and *Saccharomyces cerevisiae*) were gathered from various sources and entered by hand.

Quandt K, Grote K, Werner T. 1996. GenomeInspector: a new approach to detect correlation patterns of elements on genomic sequences. Comput Appl Biosci 12(5):405-13

Quandt K, Grote K, Werner T.1996. GenomeInspector: basic software tools for analysis of spatial correlations between genomic structures within megabase sequences. Genomics 33(2):301-4

Karlin S, Brendel V. 1992. Chance and statistical significance in protein and DNA sequence analysis. Science 257(5066):39-49

Presentations

This research has been presented or featured at the following meetings:

- C. Bult. The Institute for Genomic Research First Annual Computational Genomics Meeting (1997)
- C. Bult. Frontiers in Genetics Research (University of Connecticut) (1997)
- C. Bult. Carl Von Linneaus Lecture (Upsala, Sweden) (1997)
- M.K. Beard *Propagation of Metric Uncertainty to Topological Uncertainty.* 3rd International Symposium on Spatial Data Accuracy Assessment in Natural Resources and Environmental Sciences. Quebec City, Quebec, Canada. May 1998.
- C, Bult. 13th Annual International Mouse Genome Conference (1999)
- M. Egenhofer. Spatial Data Models: State of the Art and Beyond, After the Genome V, Jackson Hole, WY, October 1999
- C. Bult. Maine GIS Users Group Annual Meeting (1999)
- C. Bult. University of Maine, Dept. Biochemistry, Microbiology and Molecular Biology. (1999)

Scheduled Presentations:

- M. K. Beard. Harvard GIS colloquium, Boston, MA. March 2001
- M. K. Beard. Genomics: Beyond the sequence, Society of General Microbiology, Edinburgh, Scotland. March 2001
- C. Bult. GIS and Bioinformatics, Virginia Tech University. May 2001