Visualization Challenges for a New Cyberpharmaceutical Computing Paradigm

Russell J. Turner Kabir Chaturvedi Nathan J. Edwards Daniel Fasulo Aaron L. Halpern
Daniel H. Huson Oliver Kohlbacher Jason R. Miller Knut Reinert Karin A. Remington
Russell Schwartz Brian Walenz Shibu Yooseph Sorin Istrail

Celera Genomics Corporation
45 W Gude Drive, Rockville, MD 20850, USA
{Russell.Turner,. . . ,Sorin.Istrail}@Celera.com

Abstract

In recent years, an explosion in data has been profoundly changing the field of biology and creating the need for new areas of expertise, particularly in the handling of data. One vital area that has so far received insufficient attention is how to communicate the large quantities of diverse and complex information that is being generated. Celera has encountered a number of visualization problems in the course of developing tools for bioinformatics research, applying them to our data generation efforts, and making that data available to our customers. This paper presents several examples from Celera's experience. In the area of genomics, challenging visualization problems have come up in assembling genomes, studying variations between individuals, and comparing different genomes to one another. The emerging area of proteomics has created new visualization challenges in interpreting protein expression data, studying protein regulatory networks, and examining protein structure. These examples illustrate how the field of bioinformatics is posing new challenges concerning the communication of data that are often very different from those that have heretofore dominated scientific computing. Addressing the level of detail, the degree of complexity, and the interdisciplinary barriers that characterize bioinformatic problems can be expected to be a sizable but rewarding task for the field of scientific visualization.

1 Introduction

The process by which new pharmaceutical agents are discovered and developed is undergoing radical change. The combination of gene sequencing technology with advanced algorithms and sufficiently powerful hardware has now made it possible to rapidly determine the entire genetic code of a higher organism. The existence of complete reference nucleotide sequences for human, mouse, and other model organisms has revolutionized our approach to the study of biology and brought the new field of bioinformatics to the forefront of medical research.

Only a small fraction of the human and mouse genome sequence represents protein-coding genes. However, much of the variation we see in human populations is probably due to differences in gene sequences — or their regulatory region sequences — that result in changes in the protein product or the abundance or expression pattern of the protein. These protein products are the target of most drugs on the market so a more comprehensive understanding of human proteins and protein variation is expected to accelerate and improve the discovery of novel drug targets. By comparing these reference genomes against comprehensive protein databases, we can distinguish protein-coding regions that correspond to specific genes, including those associated with specific diseases. This information becomes the basis for identifying and annotating the complete set of genes in a genome.

By comparing the genome sequences of closely related species, we can discern those regions, adjacent to the protein-coding regions, which are responsible for regulating a given gene's expression of proteins. We can also build syntenic maps of closely related chromosomal segments, from which we can infer evolutionary relationships between species and between individual homologous genes. This information can be used to determine gene function and correlations of genotype to phenotype, and to identify specific genes. Since the reference human genomes are based on samples from a small number of diverse individuals, it is possible to extract information about genetic variation in humans at the level of individual nucleotides, including single-nucleotide polymorphisms in coding or regulatory regions. These polymorphisms could be responsible for congenital diseases. This information can be supplemented with targeted sequence data from larger sample populations to directly study the genetic components of variation in human populations.

Despite the tremendous amount of information about the genome itself, we are still far from being able to understand the complex interactions of protein products and gene regulatory mechanisms that make up the workings of individual living cells. To undertake this task it is necessary to turn to the emerging field of proteomics, the study of all the proteins in all the cells of the body.

State-of-the-art mass spectrometers allow high-throughput sampling and analysis of large varieties of cell samples. Liquid chromatography in combination with mass spectrometers allow scientists to isolate individual protein peptides in these samples and compare them with similar samples to detect differential protein expression. Candidate peptides can then be selected and analyzed through a secondary mass spectrometry process to determine its exact amino acid sequence.

New technologies for rapidly gathering mRNA and protein expression data will allow scientists to infer temporally-ordered networks of expressed genes which can then be correlated with protein pathway datasets to study the efficacy and possible toxicity of potential drug targets. Protein interaction datasets, derived empirically or computationally through literature mining, can be used to speed new drug design and reduce the need for empirical drug screen-

ing. Three-dimensional protein structure, which is necessary for structure-based drug discovery, will eventually be computationally determined directly from peptide amino acid sequence information.

The common thread and underlying force behind all of these innovations in biological science is high-performance computing, which is so thoroughly transforming the drug discovery process that we believe a new cyberpharmaceutical paradigm [6] is now taking shape. From a computational point of view, this paradigm poses a variety of problems characterized by large and often incomplete datasets, high experimental error rates and variances, computationally and memory-intense algorithms, and discrete, often string-oriented data types. Visualization techniques play a variety of roles in addressing these problems such as aiding in the development of algorithms, quality control, interactive human annotation of final datasets, and presentation of specific information to drug discovery scientists. In this paper, we will review some of the major types of visualization techniques we are using in our drug discovery research, and discuss the issues and challenges that they pose for the development of cyberpharmaceutics.

2 Genomics

The human genome can be thought of as a 3.1 billion-letter string over a four-letter alphabet, a relatively simple concept which belies the significance of the information contained within the sequence. In February 2001, two groups simultaneously published the near-complete sequences of the human genome [26, 17], a milestone for biomedical research comparable to chemistry's Periodic Table of the Elements. That achievement capped more than a century of breakthroughs in understanding heredity at the molecular level. Mendel described dominant and recessive genes in 1866. Morgan associated mutation with chromosomes in 1910, and Hershey and Chase implicated the DNA in particular in 1952. Watson and Crick described DNA's double helix structure in 1953, and Sanger devised the first DNA sequencing method in 1977. The 1990s saw the publication of the entire genomic sequences for several viral and bacterial organisms. Since 2000, several large genomes have been sequenced to near completion, including *Drosophila melanogaster* (fruit fly) [20], *Caenorhabditis elegans* (worm) [25], *Arabidopsis thaliana* (plant) [24], *Mus musculus* (mouse) (unpublished data), and of course, *Homo sapiens* (human).

Celera entered the genomics arena upon its formation in 1998, championing the technique of whole genome shotgun sequencing (WGS) as the most practical, affordable, and time efficient method for decoding large scale genomes, setting an ambitious three year timetable for its own completion the human genome. The key obstacle to sequencing on the scale of such a genome is the limitation of current DNA sequencing machines, which are able to produce only very short contiguous sequences (approximately $500 - 600$ nucleotides at a time). WGS was pioneered at The Institute for Genomic Research with the successful sequencing of the *Haemophilus influenzae* genome in 1995 [10]. The method requires random sampling of fragments across the entire target genome, and leveraging the ability of current sequencing machines to generate paired fragment reads of known distance and orientation (mate-pairs). With sufficient oversampling and appropriate mate-pair accuracy and distance distributions, accurate reconstruction of the complete genome sequence can be obtained *in silico*. This reconstruction process is known as sequence assembly.

The scale of the human genome posed significant challenges in terms of the sheer volume of sequencing required to assure assembly, algorithms and data management capabilities needed for the assembly process itself, and, ultimately, mining and curation of the biological information encoded within the assembled sequence. A variety of specialized visualization tools were created to support the genomics effort and remain critical as further genomes are sequenced and comparative studies across genomes are undertaken.

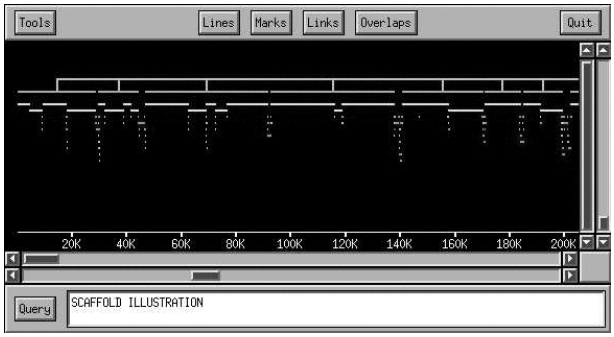

Figure 1: An assembly viewer.

2.1 Assembly Visualization

The assembly process starts from the individual fragment level and builds a hierarchical representation of the genomic sequence. Driven by fragment-to-fragment sequence comparisons and mate-pair distances, clusters of fragments are pieced together to form contiguous stretches of sequence (contigs), which are in turn ordered and oriented with respect to one another and fixed into position by means of a scaffold construct. These scaffolds comprise an assembly, and are assigned a chromosomal location during a subsequent mapping process.

2.1.1 Celamy

As development of the Celera Assembler began, before the sequencing factory was in operation, simulated data was the only testbed for the new algorithms being devised. Even then, the ability to visualize the assembly process and its data agglomeration became an imperative to enable analysis of the algorithms and debugging of the process pipeline. The Celamy viewer was developed to display linearly ordered line segments and linkage information among such segments, while allowing a user to create custom line categories (e.g., fragment, contig) and link types (e.g., scaffold, mate-pair). Controls for navigation via zooming and scrolling were required to make the tool useful on the scale of a three gigabase genome. Further, a sophisticated query language was built into the viewer, providing the ability to rapidly obtain quantitative analysis of the viewed data without requiring the formulation of a specific inquiry prior to runtime. With the development of this tool, it soon became standard practice for each assembly software component to output diagnostic snapshots of the assembly which could be analyzed offline and compared from run to run to measure the effect of changes to the assembly codebase. Due to its generic design and flexibility, Celamy has been adopted in a variety of settings outside the scope of its original WGS assembly design point. Figure 1 shows an exploded view of a $200kb$ region of a genome assembly. This snapshot was taken after construction of all unique contigs (shown as medium-length horizontal lines), but before incorporation of sequence that lies in repeat regions (short isolated lines). In this example, all contigs are contained in one scaffold, which is indicated by the top line that connects vertically to all contained contigs.

2.1.2 Annograph

Celera published its historic human genome assembly in February 2001. But even before that, Celera was providing its customers

approximations of large subsections of the genome. The approximations consisted of DNA segments from the public databases, enhanced by Celera data and expertise. The public data consisted of genome sections called BACs that had been partially shotgun sequenced, and in some cases, reassembled, forming either a partially complete sequence, or a complete (finished) sequence for that genomic region. The Celera data consisted of short DNA reads produced by its whole genome shotgun (WGS) method, including a distance measure between each mate pair. At an early stage of Celera's sequencing effort (while Celera's human scale WGS assembler was still under construction), the Combining Assembly Problem was to use Celera data to (1) verify, correct, order and orient the BACs into linear layouts called scaffolds, (2) order and orient components within unfinished BACs, and (3) fill gaps within BACs [16]. Data problems precluded automatic scaffolding, even with heuristics. At that time, the public data was largely unfinished and contained many errors such as mislabeled database entries, chimeric sequences (distinct regions accidentally joined), and over-sampling of some genomic regions (indistinguishable from ordinary sampling of genomic repeats).

Annograph is a software suite developed at Celera to permit human curation of regional scaffolds. Annograph represents scaffold data as graphs. Nodes in the graph represent known DNA sequences (BACs and Celera reads). Edges in the graph represent predicted overlaps or distances between BACs, based on sharing of Celera reads and mate-pairs. The graphs were intractable, due to false edges introduced by low-copy repeats in the genome and errors in the data. Annograph helped Celera's curators to manually identify, analyze, and remove the false edges in the graphs.

2.2 Annotation

Once the assembly process is complete, the next step is to attach meaning to the nucleotide sequence data through structural annotations. This is accomplished by loading the entire sequence into a relational database and adding annotation information through a collection of automated and interactive tools. First, the sequence is searched against a collection of protein sequence databases using standard search algorithms such as BLAST [2] and SIM4 [11]. The resulting "hits" are expressed as regions on the genome that closely match known proteins and are therefore are likely to correspond to the coding regions of genes. Comparisons are also made against the genomes of other species. The data generated in this manner are referred to as pre-computes.

Next, a set of automated gene predicting tools such as Genscan [7] and Celera's human genome annotation pipe-line *Otto* [26] are run using these pre-computes to generate annotations. While the raw sequence data for a human genome is on the order of 10GB, the additional pre-compute information can approach a total of one terabyte in size.

2.2.1 Genome Browser

At this stage, the pre-compute and annotation data can be viewed in detail using an interactive graphical tool called the Genome Browser 3. This is a three-tiered application consisting of a Java-based interactive graphical client connecting to the database via an EJB-based middle tier running in a commercial application server; see Figure 4. The Genome Browser is analogous to a geographic information system viewing a one-dimensional nucleotide terrain. The sequence itself is displayed at the bottom of the view, while the various types of precomputed data are displayed at the top in a series of tiers, distinguished by a standard color scheme. The tiers immediately above the sequence data show the annotation information in magenta.

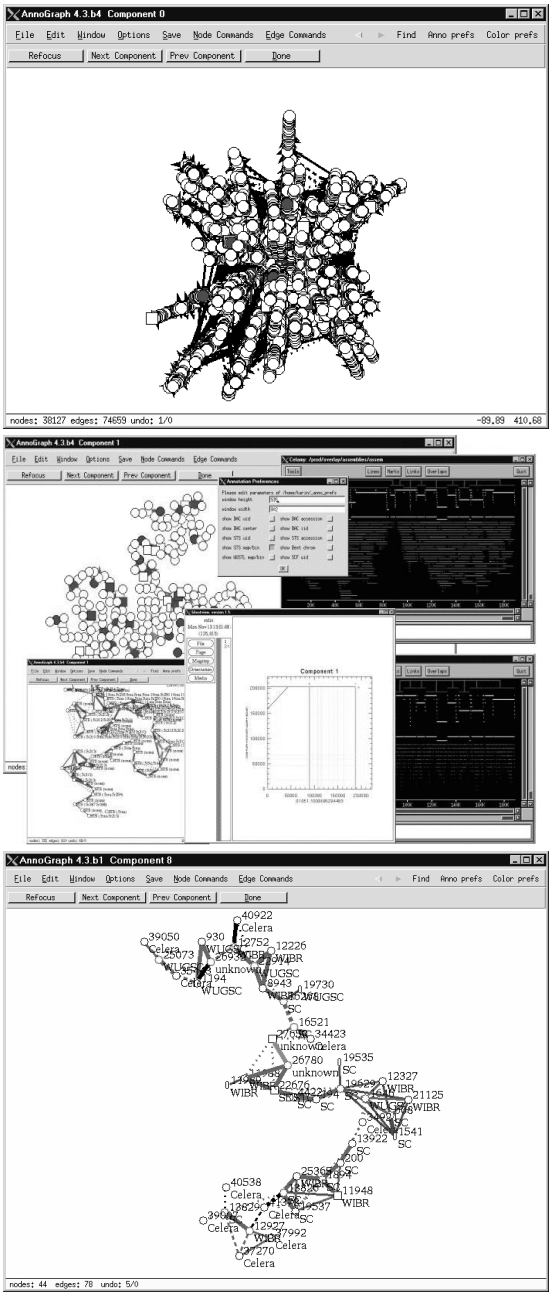

Figure 2: (a) Before curation, the graphs were intractable. False edges predominated, obscuring any true paths. (b) Bioinformatics experts interacted with an editor and a variety of pop-up visualization tools. A pipeline was built around collaborators whose edits were merged nightly. (c) After curation, the edge count was vastly reduced, and potential paths became apparent.

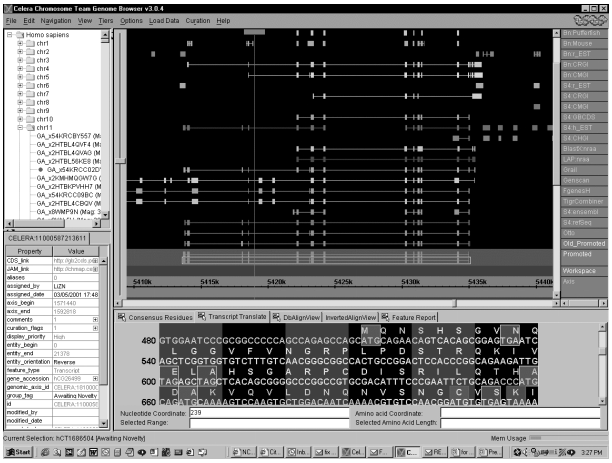

Figure 3: Celera's Genome Browser.

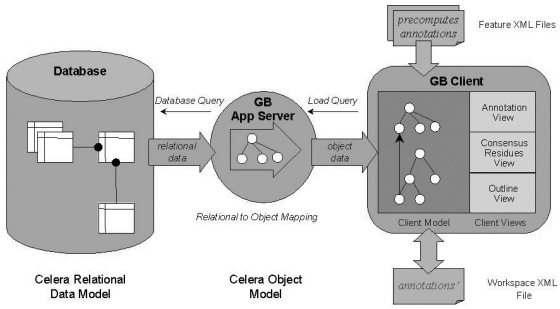

Figure 4: The three-tiered design of Celera's Genome Browser.

The data in each tier consists of individual features, which are usually displayed in groups of adjacent bars corresponding to a range along the sequence. In annotation features, these groups represent individual genes or transcripts, while the bars represent exons, which are the coding regions of the gene. The regions in between the exons represent the non-coding (or "junk") DNA. Features can be selected by mouse click, causing the feature to be highlighted and more detailed information about it to be displayed in the property inspector and subeditor panel at the bottom of the window.

The data in the main window can be panned along the sequence and zoomed over more than six orders of magnitude of scale, ranging from viewing an entire chromosome (on the order of 200 MB nucleotides) to a couple of hundred nucleotides at a time. We are able to maintain interactive frame rates (5 Hz) for reasonable numbers of visible features (30,000) during the zooming process by storing the features in local memory. Using a navigational approach to accessing data, we initially only load in the minimal amount of geometric information necessary to display each feature. Additional database queries are made to load more detailed data about individual features on demand, for example, when an individual feature is selected or when a subeditor needs to display nucleotide information.

Since we are viewing an approximately terabyte-sized database, it is obviously necessary to load only small portions of data at a time. This is accomplished by either interactively selecting a relatively small (< 1 million nucleotide) region of the sequence directly using the mouse, or searching for a specific feature by name, which loads in a small range around the feature. This process can be repeated until the client reaches the memory limits of the machine (typically it requires 70 bytes per feature), at which point features can be dynamically unloaded to make more memory available.

Data is loaded from the middle tier via the server API. Queries from the client result in SQL select statements being made to the relational database. The resulting data is converted to Java objects in a relational-to-object mapping operation. These objects are then serialized and sent over the network to the client, where they are instantiated into the client's object model. The client uses a standard model-view-controller (MVC) software architecture. The server uses an object caching mechanism which enables data loading rates from the middle tier on the order of 2500 features per second.

Users can integrate their own data with the database data by loading XML files into the client application. Annotation is performed via interactive "drag and drop" creation and editing of annotation features. Typical annotation operations include addition or deletion of exons, adjustment of exon boundaries, and adding comments.

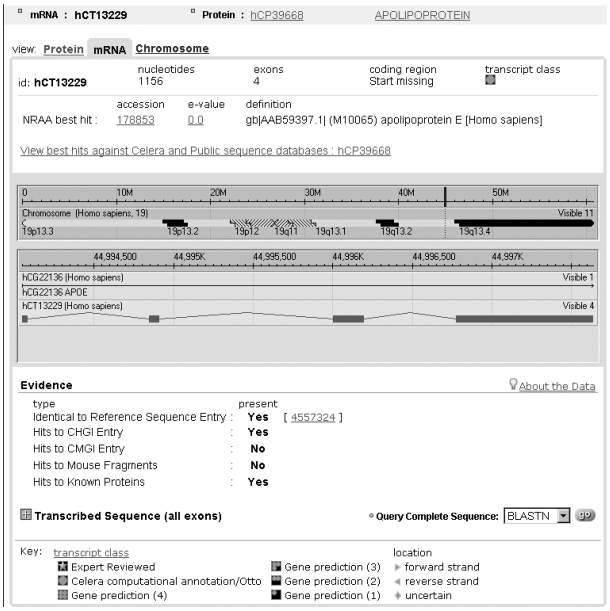

Figure 5: Screen-shot from CDS for the gene APOE.

These operations take place in a local client workspace with infinite level undo capability, and the resulting workspace file can be saved in an XML file for later promotion into the database, via a separate promotion utility.

2.2.2 Biomolecule Pages

The Celera Discovery System (CDS) is a web application hosted by Celera for its subscribers. CDS uses a tabbed report called the Biomolecule Report (BMR); see Figure 5. The BMR coalesces data from three different realms of biology. For every gene, the BMR accumulates genomic information such as chromosomal mapping and genomic neighbors, transcript information such as RNA splice variants, and protein information such as predicted function and homologs in other species. Some of the BMR information is presented within a Java applet that permits scroll, zoom, mouse-over, and drill-down. The BMR's extensive value-added clustering of information is unique among online bioinformatics resources.

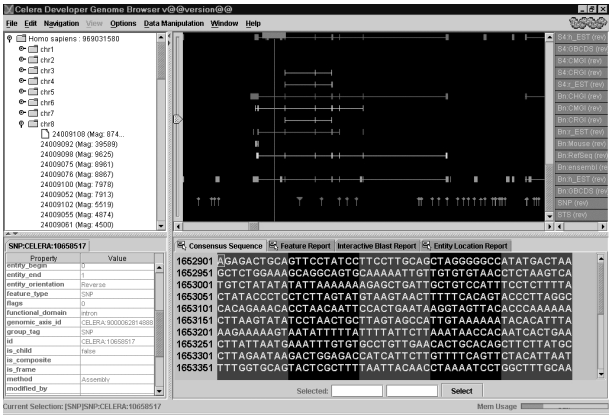

Figure 6: Snapshot of the Celera Genome Browser displaying a segment of DNA containing polymorphisms.

2.3 Genome Polymorphisms

Although the genome of one random individual is nearly identical to that of another, locating the relatively rare sites of difference is crucial to such problems as tracing ancestry and finding genetic predictors of disease. These differences, or polymorphisms, occur largely in the form of SNPs, sites at which a single DNA base may take on different values in different individuals. Both Celera [5] and the public Human Genome effort [8] considered developing maps of sequence variations important goals of sequencing the human genome. Celera [26] and the public SNP Consortium [3, 19, 22] have both since located millions of polymorphisms, focusing specifically on SNPs. As of this writing, the public dbSNP database contains just under 3 million SNPs and Celera's human database just over 3.3 million SNPs within the approximately three billion bases of human DNA. The total number of known SNPs is likely to continue growing for the foreseeable future.

Understanding the sequence context of individual polymorphisms is likely to be crucial to judging which are most likely to have functional significance, suggesting the need for tools to allow researchers to study polymorphisms in parallel with other sequence features. For example, it has been suggested in particular that SNPs affecting protein coding are most likely to be significant (see, for example, [21]). Furthermore, SNPs must be studied at multiple levels of detail alongside other sequence features. For instance, SNP density on levels of hundreds of kilobases can be revealing about the relative functional importance of different gene regions, while the individual bases in the vicinity of a polymorphic site may be essential to assessing its likely functional effect, if any.

Celera presently provides a means for both internal users and subscribers to visualize the locations of SNP polymorphisms via the Genome Browser. Figure 6 provides a snapshot of a Genome Browser view showing a polymorphic segment of DNA. The "lollypops" in the bottom of the black panel show different types of SNPs in relationship to the gene structure specified above them. Below, the consensus sequence shows the sequence context of the SNPs. At low zoom levels, such visualization capabilities are useful in finding the approximate locations of individual SNPs as well as distinguishing regions of low and high SNP density over genomic spans of a few hundred kilobases, while at finer resolutions, the Genome Browser can show individual bases at which variations are known to exist.

Providing informative but comprehensible visualizations of genomic polymorphisms remains, however, a largely unsolved problem. One crucial piece of information about a SNP is its correlation with data about the health of the donors who have it, a difficult concept to communicate visually when many potential forms of illness are considered. Another important feature of SNPs is the relative frequencies of the different variants in the population. These frequencies may themselves vary by sub-populations, for example with one variant being prevalent in one part of the world, another elsewhere. We have not even considered here the problem of visualizing non-SNP polymorphisms, such as variable-length tandem repeats, at which a short pattern may be repeated a variable number of times, and how those polymorphisms may complicate visualization of genomes in general. Furthermore, we are often interested not in single SNPs but in haplotypes, sets of all genetic variants found on one chromosome of one individual in some region. There may be many known haplotypes on a region potentially overlapping with one another, making the task of usefully communicating them in relation to one another and to other properties of an organism daunting.

2.4 Comparative Genomics

Comparative genomics — the comparison of organisms at a genomic level — reveals both the striking unity and also complex diversity of life on earth. Genome comparison is a key technology for determining *in silico* which features of a genome are important and for discovering what their function might be. Visualization plays a crucial role is this field [15].

A common characteristic of large genomes is the presence of duplications of large stretches of sequence both within the same and also between different chromosomes. These duplications give rise to genes found within a given organism that share a common history. A deeper understanding of the "redundancy" provided by such "paralogs" is necessary in the context of developing new drugs. For example, genes with high redundancy may be more difficult to target. In Figure 7, we depict segmental duplications between chromosomes in the human genome for chromosomes 1, 2, 5, 6, 9, 10, 17 and 18 (each represented by a heavy horizontal line). Each segment contains at least 3 paralogs on each of the chromosomes on which they appear. The inset displays a close-up of a one duplication between chromosomes 18 and 20 and 12 of the 64 duplicated genes are labeled by their names.

Similar techniques can be used to visualize sequence segments that are highly conserved between chromosomes of different species. Finding genes that share a common history in two or more organisms (orthologs) offers a quick way of learning what the function and biology of a protein might be. In Figure 8 we show a comparison of (Celera's assemblies of) the human and mouse genomes, particularly a region of human chromosome 3 which is highly similar to a region of mouse chromosome 13. The matches where computed using BLAST with subsequent elimination of non-unique matches. In a section of over 20 mega-basepairs (nearly 1% of the entire genome), we see two large blocks of highly conserved sequence; the first block (dark grey lines running from top left to bottom right) is in the same orientation in the two genomes, while the second block (light grey lines running from top right to bottom left) is oppositely oriented in the two genomes, indicated by the block's hourglass shape, and transposed with respect to the first block. Since the two chromosomes were assembled completely independently of one another, the agreement also acts as a simultaneous validation of both assemblies within the blocks. Some smaller differences within blocks are visible; for instance, blank wedges (tapered at one end relative to the other end) within the blocks correspond to regions in which one genome has an insertion of additional sequence relative to the other genome.

Given different assemblies of the same target chromosome, visual and analytical tools are needed to compare the assemblies, judge their quality and also to map and track features from one assembly to another. There are a number of properties that are de-

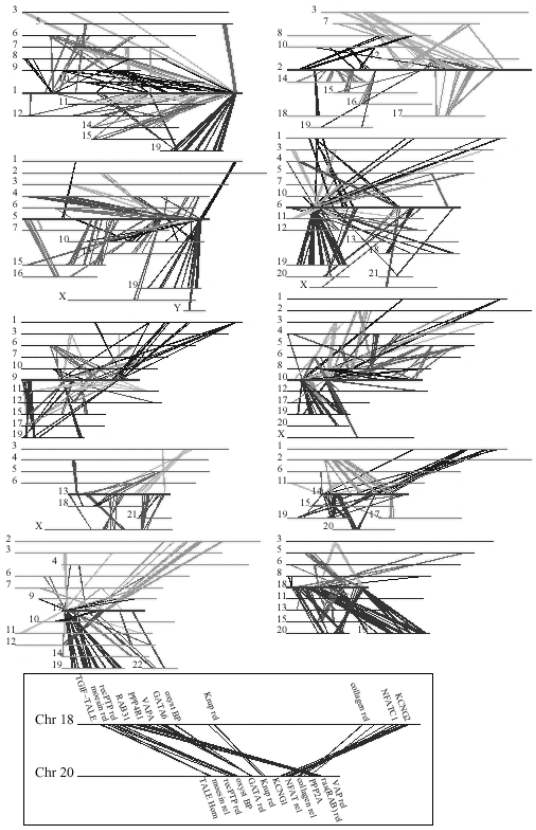

Figure 7: Significant duplications between human chromosomes.

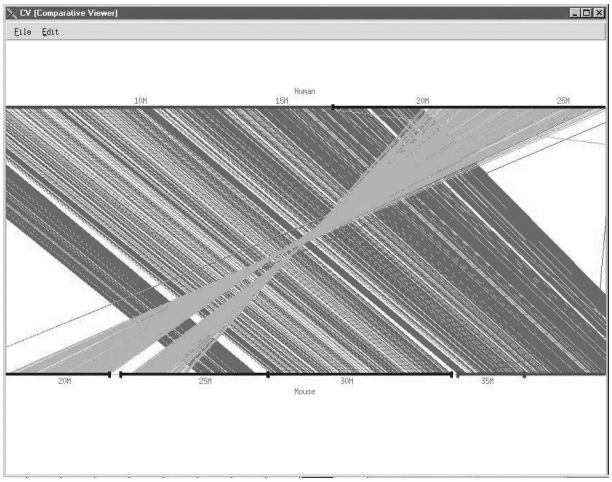

Figure 8: Comparison of similar regions of human chromosome 3 (top) and mouse chromosome 14 (bottom).

sirable for such a comparison viewer:

- interactivity, allowing inspection at any desired level of detail;
- flexibility, allowing easy integration of data from many different sources;
- scalability, running both on small and very large data sets;
- and, as always, speed.

In Figure 14 we compare a local region of two different assemblies of human chromosome 19.

The two assemblies are represented in the bottom and top thirds of the picture. In the center, a mapping is shown between both as discussed in Figure 8. In the bottom and top black panels, the location of "recruited mate-pairs" is shown and these are used to determine "breakpoints" (positions of probable misassembly) in the given assembly which are indicated in blue.

3 Proteomics

Proteomics, or the direct analysis of the expressed protein components of a cell, is critical to our understanding of cellular biological processes. Building on a foundation of genomic sequence, proteomics answers questions about the structure, function and control of biological processes and pathways by a systematic and comprehensive analysis of the proteins expressed in a cell or tissue. Unlike genomic sequence, the proteome, or the expressed protein complement of a genome, is highly dynamic; the types of expressed proteins, their abundance, state of modification, subcellular location are all dependent on the physiological state of the cell or tissue. Therefore, proteomics studies the cellular state or the external conditions encountered by a cell, in order to differentiate and study cellular states and to determine the molecular mechanisms that control them. This is a daunting task, as the proteome is estimated to consist of hundreds of thousands of different proteins with an estimated dynamic range of expression of at least five orders of magnitude. Mass spectrometry provides a tool to study the proteins present in a sample and their relative quantities in a high throughput setting. For an overview we refer the reader to a recent survey by Aebersold and Goodlett [1].

3.1 Mass Spectrometry

In mass spectrometry, a sample containing many compounds is ionized, providing many of the compounds in the sample with a charge. The charged compounds are placed in an electromagnetic field to measure their mass-to-charge ratio (M/Z). The observed ratios of all the ionized compounds in the sample form a mass spectrum, in which a peak at a particular M/Z value indicates the number of compounds (also called the intensity) observed at that value. The basic mass spectrometry machinery can be used in a variety of ways to explore the proteome. To find the relative quantities of a large number of proteins simultaneously, which then enables differential expression analysis, a complex mixture is first digested with a restriction enzyme, such as trypsin, to break each protein into pieces of mass suitable for mass spectrometry. The peptide mixture is then fed into a liquid chromatography column that separates the peptides in the mixture according to some physiochemical property, typically hydrophobicity. As the peptide mixture elutes from the liquid chromatography column over a period of one to two hours, the instrument produces a mass spectrum (called a scan) of the material leaving the column every two seconds.

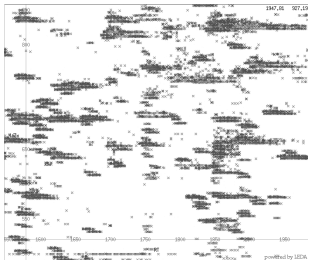

Figure 9: Part of a peptide map.

To identify the proteins present in a complex protein mixture, a different experimental protocol is used. Again, the sample is digested with a restriction enzyme, breaking each protein into peptides, and a mass spectrum taken of the mixture. The mass spectrum is analyzed on the fly, and one of the intense peaks is selected. In real time, the mass spectrometry instrument selects the peptides with this M/Z value, subjects them to secondary fragmentation, and measures the M/Z values of the resulting fragments. The goal is to identify the sequence of the peptide, given the spectrum produced by its fragments.

In the following two subsections we describe the visualization needs for differential analysis and peptide identification.

3.1.1 Differential Analysis

To analyze the relative expression and hence the differential expression of proteins, numerous scans are collected at fixed time intervals, usually every two seconds. This experiment yields a set of points in three-dimensions, called a map (see Figure 9 for a two-dimensional projection of this map). A peptide usually elutes over about 15 scans (time, the X axis) and shows a isotope pattern (M/Z, the Y axis) typical for its amino acid content. The intensity (Z axis) is coded by color in Figure 9. The goal is to identify two-dimensional regions corresponding to a single peptide, and then to identify differentially expressed peptides in normal and diseased samples. Viewing the raw spectral data proves to be extremely helpful to the researcher. The visualization tool needs to be able to browse (zoom, translate) a two-dimensional projection (normally the RT and M/Z dimensions). It should be possible to mark subsets and project the points in these sets onto different two-dimensional sub spaces (M/Z and IT, RT and IT), in order to evaluate isotope patterns and elution profiles.

In order to conduct the differential analysis the raw data corresponding to a single peptide is further reduced to a single, representative three-dimensional point. This typically results in a 500-fold data reduction. The user's data viewing requirements change at this point. In addition to browsing the reduced data, we must now show differential expression, and annotate the reduced data with additional information computed independently. Figure 15 shows a plot displaying the ratio of expression levels of identical peptides in two different data sets. The left picture shows a three-dimensional representation of the intensity ratios, the right picture shows the two-dimensional projection. The color coding and height of bars indicate different levels of expression. The views are interactive and allow the researcher to query additional information such as the peptide sequence or the quality of the measurements and predictions.

3.1.2 Peptide Identification

In the previous section we described how to identify differentially expressed peptides. Unfortunately this method does not yield the actual sequence of the peptide. In order to get this information a second (tandem) mass spectrum of the peptide fragments is taken.

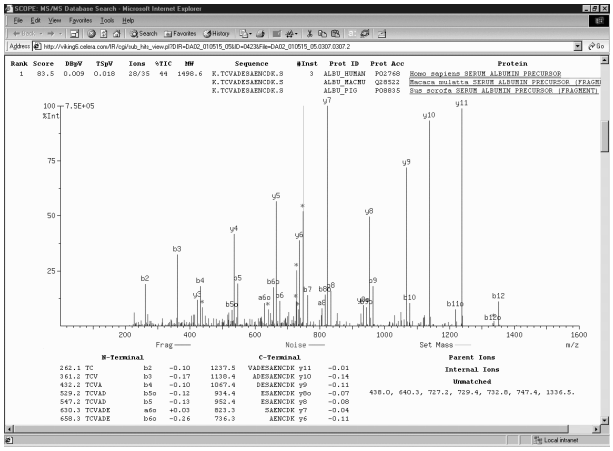

Figure 10: SCOPE, Celera's peptide identification tool.

Tandem mass spectra provide information about the position and composition of the amino-acid backbone that forms a peptide. The peptide identification problem takes as input the M/Z of the ionized "parent" peptide selected for secondary fragmentation and the tandem mass spectrum, and typically outputs a ranked list of peptides that may be responsible for the observed spectra. To date, the most successful high throughput approach to peptide identification uses a protein or genomic sequence database to provide candidate peptides satisfying various properties, including the parent M/Z, which can then be scored and ranked. Matching putative peptide fragments in the tandem mass spectrum is used for scoring candidate peptides, where many fragment matches indicate a good score.

Visualization challenges in peptide identification revolve around presenting the proposed peptide identification to the user in a way that helps them decide whether or not the identification is correct. This typically involves a tandem mass spectrum with peaks annotated by the corresponding fragment, a list of backbone bond cleavages represented in the spectrum, an indication of the difference between theoretical and observed fragment M/Z values, and a list of the unmatched fragments and spectra peaks to assist the user in interpreting spectral peaks with no annotation.

As is the case with most algorithms that score and rank candidate solutions, the user must also be presented with the top scoring candidates in a way that allows them to decide if the top candidate is the unambiguously correct identification or a fluke. In addition, the user must be able to interpret the proposed identification in the context of the other peptides in the original scan or even in the entire mixture, as many identified peptides from a single protein lends support to the hypothesis that a protein was present in the original sample. Figure 10 shows a typical result of a peptide identification query in an interactive web-based tool.

3.2 Molecular Networks

3.2.1 Pathways

Visualization and browsing of metabolic and signal transduction pathways allows one to ascertain relationships between either genes or metabolites involved in a common cellular process. The availability of data from comparative genomics (ortholog tables for different genes) and microarray experiments (clusters of co-regulated genes) is driving the development of a web-based platform that allows interactive browsing of individual genes in a pathway and inter-pathway comparison. In the current model, a pathway record

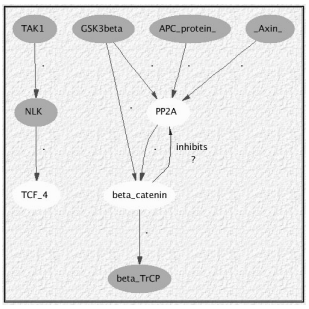

Figure 11: Partial view of a canonical Wnt-signaling pathway.

is represented as a set of binary relations which are then modeled as a directed multigraph. The edges of the graph can assume different values corresponding to the different interaction types (inhibitory, stimulatory, or putative), while the nodes represent the genes or metabolites and their subcellular location, if specified. It is possible to search all pathway records by sets of common interactions and then visualize the result through a composite representation (a snapshot of "crosstalk" between pathways). Facilities for annotation/curation are available: a scientist may revise the directionality and links in a pathway diagram, annotate them with evidence from literature, create entirely new pathway records or merge existing ones. Currently a small number of public pathway databases have been integrated into a single data framework; work is in progress on an algorithm to bound the set of all protein-protein interactions by process ontologies "on the fly" and present them as pathway records.

A typical metabolic pathway diagram contains 20 to 60 metabolites or genes; approximately 150 major metabolic pathways are described in the literature. The number of distinct signal transduction pathways is yet not known, and computational approaches will be needed to fit experimentally determined protein-protein interactions into existing networks as well as to highlight redundancy between different pathways. An updated version of the software aims to replace the simple graphical model with a colored petri net: transition arcs will contain functions and data that govern feasibility of pathway interactions and will lay out optimal paths through a pathway based on a given set of initial conditions, thereby allowing simulation. The initial conditions could represent microarray expression values for genes in a particular experiment which have been mapped to a pathway or data from kinetic studies (which would allow modeling of pathway flux).

3.2.2 Expression Patterns

Oligonucleotide arrays and cDNA microarrays [18, 27, 9, 23] allow for the measurement of the mRNA expression levels of thousands of genes at a time. The massive parallelism offered by these genome chips looks ready to revolutionize areas like drug discovery and diagnostics. For several organisms, the expression patterns of the entire gene set can now be monitored simultaneously. For instance, the entire complement of $6000+$ genes of the yeast *S. cerevisiae* has been put on a single microarray. There have been many types of problems addressed using these chips, including the identification of co-regulated genes (i.e., genes involved in a common regulatory process), the identification of gene networks, and the identification/classification of tissues based on gene expression levels. The visualization problems in this area share similarities with those encountered in the clustering and classification domains, in particular the visualization of high dimensional data points. We describe briefly two of the application areas.

In the search for co-regulated genes, the mRNA expression levels

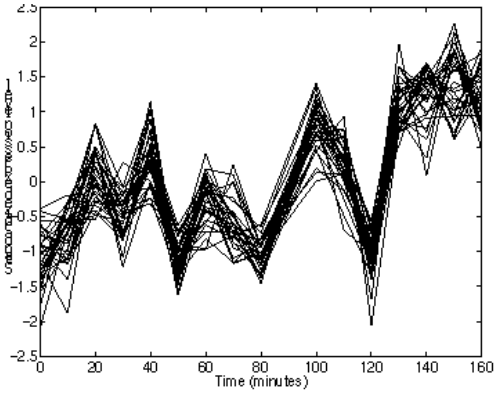

Figure 12: Clustered expression profiles.

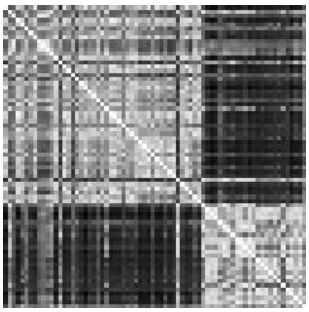

Figure 13: ALL/AML sample similarity.

of thousands of genes are monitored across the set of conditions of interest. The genes are then clustered and each cluster represents a potential set of co-regulated genes. Figure 12 shows the expression profiles of a group of genes that have been clustered together based on their expression levels as yeast goes through various phases of its cell cycle. Simple plots such as this can give a better understanding of what the expression patterns within clusters look like, and also what the average values and variances are within a cluster.

In the tissue identification problem, the goal is to distinguish different types of tissues by their mRNA content as measured by microarray experiments. This is typically accomplished by clustering the microarray data to reveal or confirm the tissue labels, and then finding differentially expressed genes among the tissues. Visualization can help to both confirm the cluster structure and the quality of the differentially expressed genes.

As an example, consider a publicly-available 72 sample leukemia data [12] with 47 acute lymphoblastic leukemia (ALL) samples and 25 acute myeloid leukemia (AML) samples. Figure 13 shows the 72 by 72 similarity matrix between all pairs of the 72 samples. The degree of similarity between samples i and j is proportional to the intensity at block (i, j). The ordering of the rows and columns is based on the known tissue types, and quickly reveals that approximately two tissue types are present, although some samples are potentially bad or misclassified.

Figure 17 shows the relative expression levels of 25 genes (rows) across the 72 tissues (columns). A gene in a particular tissue is red if it is underexpressed relative to the other samples, or green if it is overexpressed. These genes were chosen because statistically they appear to be differentially expressed between ALL and AML; that is, their expression levels are significantly different between the two tissue types. Hence, one expects the first 47 columns to be of mainly a single color, and the second 25 columns to be of the

other color. The figure makes this easy to verify visually, and also indicates some samples that may either be misclassified or may not be well identified by the set of genes.

3.3 Protein Structure

Visualizing three-dimensional structure has a long tradition in chemistry. Ever since chemists started thinking about the structure of molecules, they used models, mostly made out of wood, metal, or polymers, to visualize the structural information revealed by their methods. Even with those very crude models, important insights concerning molecular structure could be derived. With the advent of biochemistry and the availability of large structural data sets due to the development of X-ray crystallography those material models reached their limits. Although important insights were gained at the very beginning by constructing material models and comparing them to X-ray data (e.g., the famous double helix structure), virtual computer-generated models have completely taken over that role since becoming widely available in the 1980s. Since then, the interactive visualization and interpretation of the three-dimensional structure of biomolecules (mainly proteins) became an irreplaceable tool in structural biology.

3.3.1 3D Structure

There are a number of standard graphical representations employed in the visualization of molecular 3D structure. Stick models mimic the common two-dimensional structural formulas drawn by chemists, while ball and stick models were long common in material models. Besides visualizing the structure of a protein at the atomistic level, other models have been developed to illustrate important structural features of proteins. Secondary structure is usually displayed as cylinders (for helices), bands (for pleated sheets), and tubes or ribbons (for loops or the backbone in general). By reducing the level of detail, those models present the major structural features of proteins at a glance. Surface-based models (e.g., the solvent-excluded or solvent-accessible surface) are used to get an impression of the overall shape of a protein.

Figure 18 shows the structure of the AIDS drug Nevirapine [13] (red, center) in the binding site of its target, the enzyme Reverse Transcriptase. The drug is shown as a ball and stick model with a superimposed translucent surface, the receptor is displayed with tubes, cylinders, and bands to indicate the protein secondary structure, and the amino acid side chains of the receptor are drawn in a simple stick model. The figure was produced using VMD [14] and PovRay.

3.3.2 Structure-Based Drug Design

Visualization of proteins has become an important tool in the rational design of drugs, which became feasible with the availability of high-resolution X-ray structures of receptor-ligand complexes.

From the visual inspection of those crystal structures alone, it is possible to gain valuable insights about the mode of binding and the nature of the receptor-ligand interaction. For example, the geometric vicinity of hydrogen bond donors and acceptors or the geometric complementarity of a ligand and a binding pocket becomes obvious just by looking at a structure of the complex.

That information in itself is helpful for the (*de novo*) design of ligands or the optimization of lead structures, but visualization of physicochemical properties in combination with the structure has proven to be an even more powerful tool. Those properties, e.g., hydrophobicity, flexibility, or electrostatic properties, can be used immediately to develop strategies for the optimization of drug affinities. The integration of that information is achieved by a number of different visualization techniques, ranging from simple color coding of the structural models, over mapping of properties to molecular surfaces, to the generation of iso-surfaces and vector fields from grid-based scalar and vectorial properties.

Another important aspect of visualizing molecular structures is the integration of their dynamics. Most biochemical processes can be understood from their dynamics only, a fact that is often hidden by the snapshot character of X-ray structures. Computer simulations can be used to understand that dynamics and visualization is crucial to the analysis of those simulations.

3.3.3 Structure Algorithms

Since 1972, the Protein Data Bank [4] has provided a database of protein structures. Its growth has been explosive: in 1982, there were 200 structures; in 1992, 1,000 structures; in 1999, 10,000 structures; as of May 2001, over 15,000 structures are available. As each structure contains, on average, several hundred amino-acid residues, this database contains roughly 100 million atomic coordinates. Additionally, residues are usually annotated with the type of secondary structure (alpha-helix, beta-sheet, turn) they are within.

Relationships between these structures are at the heart of structural genomics, with applications to drug design, protein fold prediction and evolutionary studies. To assist in understanding algorithms for these applications, an interactive development framework has been developed. The framework is entirely object-oriented, and is fully extensible in both the types of data and algorithms it uses. The algorithms work with the data via an event-driven shared-workspace — any change to the workspace causes an event to be generated. Algorithms monitor the events, performing their work when feasible.

For example, the user could load a protein sequence into the framework, allowing an algorithm to predict the secondary structure, in turn, allowing a folding algorithm to predict tertiary structure, which would allow several localized refinement algorithms to operate. Once the structure has been predicted, a structure comparison algorithm could then classify the structure against a database of existing structures. Figure 19 shows an algorithmically folded protein model (upper left), an experimentally derived protein structure (upper right), a self-assembled hydrophobic core (lower left), and a raw PDB text file (lower right).

4 Conclusions

Despite their diversity, the examples presented in this paper illustrate some common themes about the nature of difficult visualization problems in biology. We have shown a selection of very different biologically-inspired visualization problems in such areas as genome sequencing, comparative genomics, proteomics, and protein structure analysis that required the development of novel tools and techniques. Some of these problems are not hard to solve. Some of the hard problems, very likely more than we are aware of, have already been solved by visualization work in entirely different domains. Of greatest interest, though, are the hard problems that are distinct to biological applications involving large data sets, which are most likely to be remote from the past experience of the visualization field. We will therefore conclude this paper by considering three lessons we can draw from the Celera experience about how our biologically-inspired visualization problems tend to differ from those encountered in other areas of scientific visualization and what new kinds of challenges they might pose to the visualization field.

Lesson one is that biological systems are messy. The preceding examples give some hint of the many details involved in considering even simple biological problems. Even a seemingly straightforward problem, such as expressing the string of three billion bases that is

the human genetic code, quickly gets bogged down in the details of the underlying biology and the specific experimental methodologies used to generate and analyze the data. As we move to more complicated problems, such as analyzing protein expression, the amount of detail increases dramatically. Even so, the above presentation substantially understates the level of detail an expert in the field would like to access as he or she examines such data. Furthermore, we can be nearly certain that new instruments and experimental protocols will be developed generating wholly new kinds of information that will need to be integrated into our databases and communicated effectively. Biological systems, unlike, for example, those typically dealt with by chemistry and physics, are not reducible to simple fundamental laws. They reflect not only the laws of physics but also the legacy of evolution, leaving them generally specialized, diverse, and extremely ad hoc. This level of detail implies tradeoffs will be needed in the design of visualization tools to avoid both information overload and oversimplification.

Lesson two is that biological systems are highly interconnected. The many aspects of a system and the many different kinds of data that may be generated about it all relate to one another in often unpredictable ways. Although this paper presented Celera's visualization problems as a series of isolated tasks, that is a vast oversimplification of the real challenge: providing all of the information available to us in ways that allow researchers to explore how the different bits of data relate to one another. Consider a single gene in an organism. It has a sequence, which may have several variants among individuals. The sequence may be parsed different ways to produce one or more proteins. Each of those proteins has a pattern of expression throughout an organism tied into a regulatory network possibly involving many other genes and various metabolites. The gene likely also has paralogs within the organism being examined and orthologs in other organisms. It has a structure, possibly several under different conditions or when complexed with different other molecules. These structures may themselves have structural homologs beyond the sequence homologs of the genes. And every aspect of this web of data may be tied into a literature of diverse bits of information such as diseases associated with a particular property of the gene, drugs affecting it, and physiological processes with which it is involved, which themselves likely involve other genes. Creating the tools that will facilitate drilling down into biological data and following these convoluted chains of inference is likely to remain a formidable problem for the foreseeable future.

Lesson three is that interdisciplinary boundaries can be a major obstacle to progress in biology. The examples presented above draw on expertise from highly divergent disciplines. A single project may require input from biochemists, geneticists, computer scientists, software engineers, mathematicians, statisticians, or a host of other specialists. Such diverse interdisciplinary collaborations, which were rare only a few years ago, are central to the work of Celera and others in bioinformatics. The distances between these disciplines in background knowledge and even styles of thought are typically far broader than those that have traditionally been encountered by scientific computing. Biology, for example, is traditionally a reductionist, qualitative discipline, creating difficult communication barriers with the high-level, quantitative disciplines of mathematics and computer science. Much of the difficulty in the aforementioned tasks comes from communicating detailed biological knowledge to computational disciplines and representing computational solutions to biological problems in ways biologists can understand and assess.

We hope that the challenges of our field will prove attractive to the visualization community. The examples presented above show solutions to a few dilemmas, but the visualization of bioinformatic data is largely an open problem. As our three lessons indicate, bioinformatic visualization problems are likely to be demanding, particularly in the detail and complexity of information that solutions must accommodate. But while the work will be difficult, it will also likely be exciting and rewarding. Lesson three in particular suggests both an obstacle and an opportunity: the sizable interdisciplinary boundaries make the communication problem challenging, but they also indicate a great need for improvements in visualization techniques due to their potential to provide a common ground that specialists in many disciplines can understand. The nature of the problems and of the solutions we have seen to date suggest that successful solutions to the open bioinformatic visualization problems will frequently prove to be complicated, ad hoc, and highly problem specific, much like the biological systems themselves. This should not mean, though, that we have to start from scratch with each visualization project we encounter. Working out a common language that will allow us to communicate these diverse types of data and facilitate reuse of techniques across wide areas of biological visualization remains a vital challenge. Bioinformatics should be able to occupy interested visualization specialists for a long time to come.

Acknowledgments

This article reports on work undertaken at Celera Genomics in the course of the past three years. This work involved many more researchers than can be listed as authors of this paper. In particular, we would like to mention the team of scientists and software engineers that put together Celera's Genome Browser. Gene Myers designed and built some of the fundamental tools for visualizing sequence assemblies. Figure 7 was produced by Steven Salzberg of TIGR. Moreover, we would like to thank our colleagues at Celera for many useful discussions, including in particular: Vineet Bafna, Ellen Beasley, Deepali Bhandari, Peter Davies, and Richard Mural.

References

[1] AEBERSOLD, R., AND GOODLETT, D. R. Mass spectrometry in proteomics. *Chemical Reviews 101* (2001), 269–295.

[2] ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W., AND LIPMAN, D. J. Basic local alignment search tool. *Journal of Molecular Biology 215* (1990), 403–410.

[3] ALTSHULER, D., POLLARA, V. J., COWLES, C. R., VAN ETTEN, W. J., BALDWIN, J., LINTON, L., AND LANDER, E. S. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature 407* (2000), 513–519.

[4] BERMAN, H., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T., WEISSIG, H., SHINDYALOV, I., AND BOURNE, P. The protein data bank. *Nucleic Acids Research 28* (2000).

[5] BRODER, S., AND VENTER, J. C. Sequencing the entire genomes of free-living organisms: the foundation of pharmacology in the new millennium. *Annu. Rev. Pharamacol. Toxicol. 40* (2000), 97–132.

[6] BRODER, S., AND VENTER, J. C. Sequencing the entire human genome: Towards a new cyberpharmaceutical paradigm. *Journal of Commercial Biotechnology 6*, 4 (2000), 270–283.

[7] BURGE, C., AND KARLIN, S. Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology 268* (1997), 78–94.

[8] COLLINS, F. S., PATRINOS, A., JORDAN, E., CHAKRAVARTI, A., GESTELAND, R., AND WALTERS, L. New goals for the U. S. human genome project: 1998-2003. *Science 282* (1998), 682–289.

[9] DERISI, J., IYER, V., AND BROWN, P. Exploring the metabolic and genetic control of gene expression on a genome scale. *Science 278* (1997), 680 – 686.

[10] FLEISCHMANN, R., ADAMS, M., WHITE, O., CLAYTON, R., KIRKNESS, E., KERLAVAGE, A., BULT, C., TOMB, J.-F., BOUGHERTY, B., MERRICK, J., MCKENNEY, K., SUTTON, G., FITZHUGH, W., FIELDS, C., JOCAYNE, J., SCOTT, J., SHIRLEY, R., LIU, L.-I., GLODEK, A., KELLEY, J., WEIDMAN, J., PHILLIPS, C., SPRIGGS, T., HEDBLOM, E., COTTON, M., UTTERBACK, T., HANNA, M., NGUYEN, D., SAUDEK, D., BRANDON, R., FINE, L., FRITCHMAN, J., FUHRMANN, J., GEOGHAGEN, N., GNEHM, C., MCDONALD, L., SMALL, K., FRASER, C., SMITH, H., AND VENTER, J. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science 269* (1995), 496–512.

[11] FLOREA, L., HARTELL, G., ZHANG, Z., RUBIN, G. M., AND MILLER, W. A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Research 8* (1998), 967–974.

[12] GOLUB, T., SLONIM, D., TAMAYO, P., HUARD, C., CAASENBEEK, J., COLLER, H., LOH, M., DOWNING, J., CALIGIURI, M., BLOOMFIELD, C., AND LANDER, E. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science 286* (1999), 531–537.

[13] GUSSIO, R., PATTABIRAMAN, N., ZAHAREVITZ, D. W., KELLOGG, G. E., TOPOL, I. A., RICE, W. G., SCHAEFFER, C. A., ERICKSON, J. W., AND BURT, S. K. All-atom models for the non-nucleoside binding site of HIV-1 reverse transcriptase complexed with inhibitors: a 3D QSAR approach. *J. Med. Chem. 39* (1996).

[14] HUMPHREY, W., DALKE, A., AND SCHULTEN, K. VMD – visual molecular dynamics. *J. Molecular Graphics 14* (1996).

[15] HUSON, D. H., HALPERN, A. L., LAI, Z., REINERT, K., MYERS, E. W., AND SUTTON, G. G. Comparing assemblies using fragments and mate-pairs. To appear at: "Workshop on Algorithms in Bioinformatics" (WABI-01), 2001.

[16] HUSON, D. H., REINERT, K., AND MYERS, E. W. The greedy path-merging algorithm for sequence assembly. In *Proceedings of the Fifth Annual International Conference on Computational Molecular Biology (RECOMB-01)* (2001), pp. 157–163.

[17] INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM. Initial sequencing and analysis of the human genome. *Nature 409*, 6822 (2001), 860–921.

[18] LOCKHART, D., DONG, H., BYRNE, M., FOLLETTIE, M., GALLO, M., CHEE, M., MITTMANN, M., WANG, C., KOBAYASHI, M., HORTON, H., AND BROWN, E. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol. 14*, 13 (1996), 1675 – 1680.

[19] MULLIKIN, J. C., HUNT, S. E., COLE, C. G., MORTIMORE, B. J., RICE, C. M., BURTON, J., MATTHEWS, L. H., PAVITT, R., PLUMB, R. W., SIMS, S. K., AINSCOUGH, R. M., ATTWOOD, J., BAILEY, J. M., BARLOW, K., BRUSKIEWICH, R. M., BUTCHER, P. N., CARTER, N. P., CHEN, Y., CLEE, C. M., COGGILL, P. C., DAVIES, J., DAVIES, R. M., DAWSON, E., FRANCIS, M. D., JOY, A. A., LAMBLE, R. G., LANGFORD, C. F., MACARTHY, J., MAL, L. V., MORELAND, A., OVERTON-LARTY, E. K., ROSS, M. T., SMITH, L. C., STEWARD, C. A., SULSTON, J. E., TINSLEY, E. J., TURNEY, K. J., WILLEY, D. L., WILSON, G. D., MCMURRAY, A. A., DUNHAM, I., ROGERS, J., AND BENTLEY, D. R. An SNP map of human chromosome 22. *Nature 407* (2000), 516–520.

[20] MYERS, E. W., SUTTON, G. G., DELCHER, A. L., DEW, I. M., FASULO, D. P., FLANIGAN, M. J., KRAVITZ, S. A., MOBARRY, C. M., REINERT, K. H. J., REMINGTON, K. A., ANSON, E. L., BOLANOS, R. A., CHOU, H.-H., JORDAN, C. M., HALPERN, A. L., LONARDI, S., BEASLEY, E. M., BRANDON, R. C., CHEN, L., DUNN, P. J., LAI, Z., LIANG, Y., NUSSKERN, D. R., ZHAN, M., ZHANG, Q., ZHENG, X., RUBIN, G. M., ADAMS, M. D., AND VENTER, J. C. A whole-genome assembly of Drosophila. *Science 287* (2000), 2196–2204.

[21] RISCH, N. J. Searching for genetic determinants in the new millennium. *Nature 405* (2000), 847–856.

[22] SACHIDANANDAM, R., WEISSMAN, D., SCHMIDT, S. C., KAKOL, J. M., STEIN, L. D., MARTH, G., SHERRY, S., MULLIKIN, J. C., MORTIMORE, B. J., WILLEY, D. L., HUNT, S. E., COLE, C. G., COGGILL, P. C., RICE, C. M., NING, Z., ROGERS, J., BENTLEY, D. R., KWOK, P. Y., MARDIS, E. R., YEH, R. T., SCHULTZ, B., COOK, L., DAVENPORT, R., DANTE, M., FULTON, L., HILLIER, L., WATERSTON, R. H., MCPHERSO, N. J. D., GILMAN, B., SCHAFFNER, S., VAN ETTEN, W. J., REICH, D., HIGGINS, J., DALY, M. J., BLUMENSTIEL, B., BALDWIN, J., STANGE-THOMANN, N., ZODY, M. C., LINTON, L., LANDER, E. S., AND ATSHULER, D. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature 409* (2001), 928–933.

[23] SCHENA, M., HELLER, R., THERIAULT, T., KONRAD, K., LACHENMEIER, E., AND DAVIS, R. Microarrays: biotechnology's discovery platform for functional genomics. *Trends Biotechnol. 16* (1998), 301 – 306.

[24] THE ARABIDOPSIS GENOME INITATIVE. Analysis of the genome sequence of the flowering plant *arabidopsis thaliana*. *Nature 408* (2000), 796–815.

[25] THE C.ELEGANS SEQUENCING CONSORTIUM. Sequence and analysis of the genome of *c. elegans*. *Science 282* (1998), 2012–2018.

[26] VENTER, J. C., ADAMS, M. D., MYERS, E. W., ET AL. The sequence of the human genome. *Science 291* (2001), 1145–1434.

[27] WODICKA, L., DONG, H., MITTMANN, M., HO, M., AND LOCKHART, D. Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nat. Biotechnol. 15* (1997), 1359 – 1367.

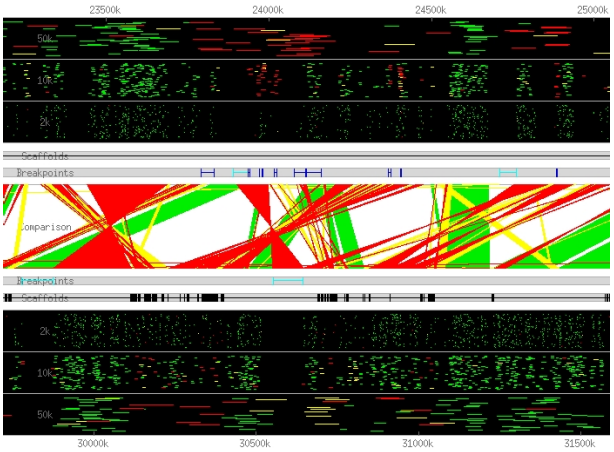

Figure 14: Assembly comparison tool.

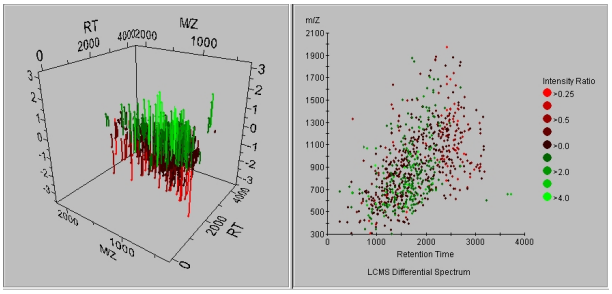

Figure 15: Visualization of differential expression.

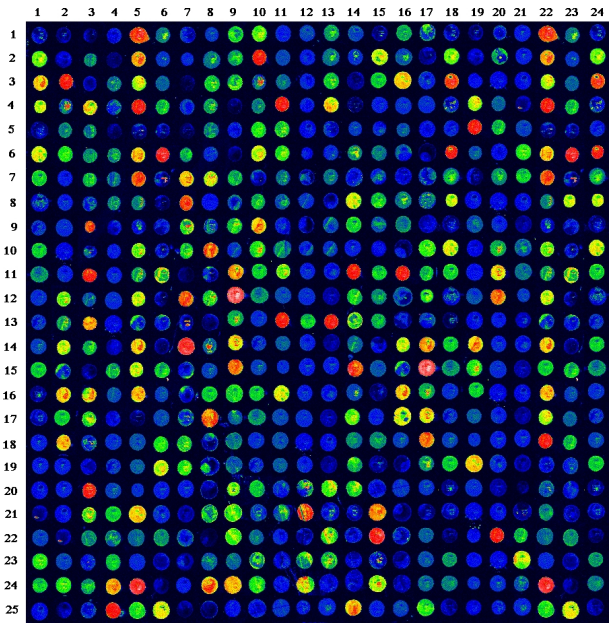

Figure 16: Gene expression levels.

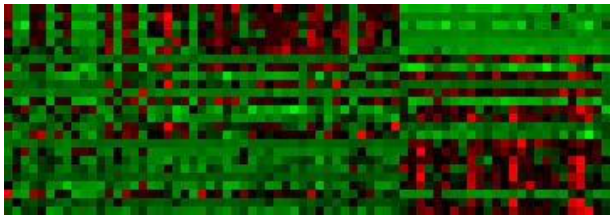

Figure 17: 25 differentially expressed genes.

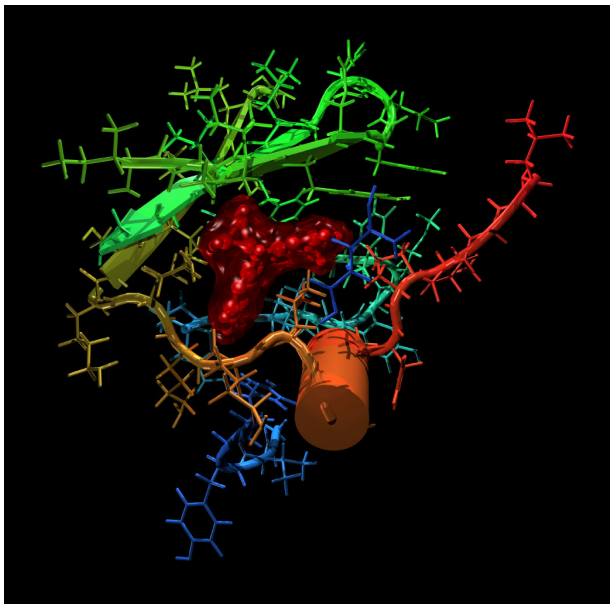

Figure 18: The AIDS drug Nevirapine in the binding site of Reverse Transcriptase.

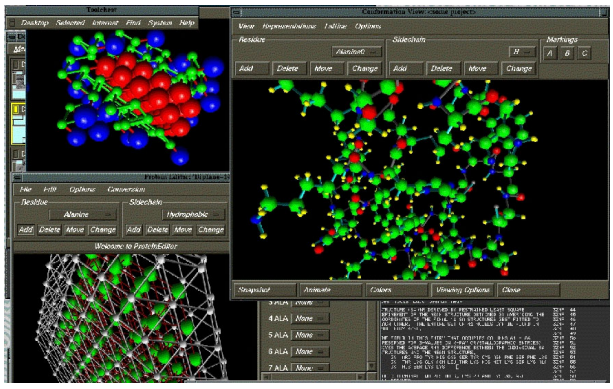

Figure 19: The protein structure algorithm framework interface.