G-nome Surfer: a Tabletop Interface for Collaborative Exploration of Genomic Data

Orit Shaer[1], **Guy Kol**[2], **Megan Strait**[1], **Chloe Fan**[3], **Catherine Grevet**[4], and **Sarah Elfenbein**[1]

[1]Wellesley College, 106 Central St, Wellesley, MA, USA, 02482

[2]Babson College, 231 Forest Street, Babson Park, MA, USA, 02457

[3]HCII, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA, USA, 15213

[4]GVU Center, Georgia Institute of Technology, Atlanta, GA, USA, 30332

ABSTRACT

Molecular and computational biologists develop new insights by gathering heterogeneous data from genomic databases and leveraging bioinformatics tools. Through a qualitative study with 17 participants, we found that molecular and computational biologists experience difficulties interpreting, comparing, annotating, sharing, and relating this vast amount of biological information. We further observed that such interactions are critical for forming new scientific hypotheses. These observations motivated the creation of G-nome Surfer, a tabletop interface for collaborative exploration of genomic data that implements multi-touch and tangible interaction techniques. G-nome Surfer was developed in close collaboration with domain scientists and is aimed at lowering the threshold for using bioinformatics tools. A first-use study with 16 participants found that G-nome Surfer enables users to gain biological insights that are based on multiple forms of evidence with minimal overhead.

Author Keywords

Reality-Based Interaction, Tabletop interaction, Bioinformatics, Genome Browser

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: *User Interfaces*

General Terms

Design, Human Factors

INTRODUCTION

Over the past two decades, Human-Computer Interaction (HCI) research has generated a broad range of interaction styles that move beyond the desktop into new physical and social contexts. Key areas of innovation in this respect are tangible, tabletop, and embodied user interfaces. These interaction styles share an important commonality: leveraging users' existing knowledge and skills of interaction with the real non-digital world such as naïve physics, spatial, social and motor skills [11]. Drawing upon users' pre-existing knowledge and skills of interaction with the real non-digital world, these interaction styles are often unified under the umbrella of Reality-based Interfaces (RBIs) [11]. By basing interaction on pre-existing real world knowledge and skills, RBIs offer a more natural, intuitive, and accessible form of interaction that reduces the mental effort required to learn and operate a computational system and supports high-level cognition [11].

While these advances in Human-Computer Interaction have been applied to a broad range of application domains including problem-solving, education, and entertainment, little HCI research has been devoted to investigating RBI in the context of professional scientific research. However, it is important to study RBI in this context where reducing the mental workload associated with accessing information and supporting collaborative high-level reasoning could potentially lead to new scientific discoveries. Those RBIs that examined the possibilities of supporting scientific discovery in fields such as molecular biology [8], chemistry [4], and geophysics [6], highlight the potential contribution of RBIs to supporting scientific discovery, but focus on the representation and manipulation of information that has an inherent physical or spatial structure such as proteins, molecules, and maps. We are interested in a broader use case, investigating whether reality-based interaction techniques can enhance scientific discovery in areas where vast amount of abstract information is accessed and manipulated. Examples include molecular energy levels, aggregated physiological data, and genomic information.

Advances in genomic technologies have led to an explosive growth in the quantity and quality of biological information available to the scientific community. The ability to simultaneously collect detailed information about the structure and activity of multiple genes has fundamentally changed the way molecular biology research is conducted. Rather than focusing on small scale, lab-based experiments, researchers often conduct large scale experiments in which information from multiple genes is simultaneously measured, recorded,

CHI 2010, April 10 – 15, 2010, Atlanta, Georgia, USA

and stored in a database. The need to analyze such large and complex data sets has driven a change in the tools used in biological research: next to having a pipette and a pen, a web browser is currently the most widespread tool available for biologists as it provides access to powerful computational and statistical tools [21]. Web technologies have been massively adopted by molecular biology software developers to allow easier access for biologists who are not computer experts. However, existing web-based genomic tools show severe limitations in terms of persistence, usability, and support of high-level reasoning [3, 15, 23].

Through a study of molecular and computational biologists we observed that to develop insights, biologists gather a wealth of heterogeneous data from genomic databases and leverage a diverse set of bioinformatics tools. However, they experience difficulties interpreting, comparing, annotating, sharing, and relating biological information. We further observed that these manipulations are critical for developing deep insights and forming hypotheses. These observations motivated the creation of G-nome Surfer, a tabletop interface that provides collaborative and fluid interaction with heterogeneous genomic data. G-nome Surfer supports searching, comparing, annotating, and relating large amounts of genomic information.

The rest of the paper is organized as follows: the next section summarizes our study of molecular and computational biologists. Following that, we present the two other contributions of this paper. The first is G-nome Surfer, a tabletop interface for collaborative exploration of genomic information that was designed based on findings from the observational study. The second contribution is a first-use study of this system and the lessons we learned. The study demonstrates that G-nome Surfer enables users to gain biological insights that are based on the multiple forms of evidence it provides, with minimal overhead.

SEMI-STRUCTURED INTERVIEWS WITH BIOLOGISTS

To understand the current work practices of genomic researchers, we conducted a series of interviews with 17 molecular and computational biologists (eight female, nine male) from leading genomic research institutions, industry, and an undergraduate research institution. The title and research area of each participant are listed in table 1. Most interviews took place at the researcher's primary work place (except three that took place in our HCI laboratory). The interviews were semi-structured and lasted 45-60 minutes. During the interviews, we asked participants to educate us about their research goals, their work practices, and the computational tools they use. We asked each participant to walk us through a particular instance of research work. We collected data by audio-taping the interviews, taking pictures of partcipants' workspace, collecting relevant work samples, and saving screen captures of participants' computers as they were demonstrating how they perform various tasks. The second author of this paper, a bioinformatician with experience building commercial genomic analysis tools, aided our need-finding efforts and directed us towards issues most critical for genomic researchers. Two additional authors have background in biology. We analyzed this data by identifying common high-level tasks and themes. We then distilled design implications for genomic exploration tools. Following, we describe our findings.

ID	Title	Research Area
P1	Faculty	Developmental Genetics
P2	Faculty	Evolutionary Biology
P3	Faculty	Animal Physiology
P4	MSc Student	Metagenomics
P5	Faculty	Cell Biology and Genetics
P6	Industry Researcher	Cancer Drug Therapy
P7	Postdoc	Viral Infections
P8	Faculty	Neuroscience
P9	Faculty	Proteins Structure
P10	Industry Researcher	Next Generation Sequencing
P11	Research Assistant	Evolutionary Biology
P12	Postdoc	Genetic Networks
P13	Student Researcher	Animal Physiology
P14	Student Researcher	Protein Structure
P15	Faculty	Bacterial Systems
P16	Faculty	Proteins Structure
P17	Postdoc	Computational Genetics

Table 1. The background of study participants. All participants were interviewed individually except P3 and P13, and P9 and P14 who were interviewed in pairs.

Information Tasks and Bioinformatics Tools

We found that although our subjects are interested in finding answers to a diverse set of biological problems, they use similar techniques for accessing and analyzing genomic information. Specifically, we identified five basic information tasks that are commonly performed by our subjects:

- *literature searching:* presents what is already known about a gene, a condition, or a biological function.
- *locating a gene on a genome:* verifies the structure of a gene and its relationship with neighboring genes.
- *retrieving a genomic sequence:* provides the base-pair or amino-acid sequence of a gene.
- *searching for similarity between sequences:* highlights the similarity between the researched sequence and other sequences.
- *annotating genomic information:* adds finding and conclusions so that the researcher or other team members can further explore and query the information.

Databases

While some of our subjects maintain their own database for storing and managing molecular data and experiment results (mainly for confidentiality reasons), there exists a large number of public online databases for depositing and importing genomic data. Public databases are used by all of our subjects on a frequent basis. Such databases range from those that contain genomic information of a specific organism to large databases that contain DNA sequence data for many different organisms. Also useful in designing experiments and forming hypotheses are literature databases (i.e. PubMed) and ontology databases (e.g. GeneOntology). The National Center for Biotechnology Information (NCBI) web

site is typically a starting point for researchers looking for resources. All of these databases provide access through a web browser but implement different methods and tools for storing and retrieving information.

Genome Browser

A genome browser [5] is an online tool that visualizes the spatial relationships between different pieces of genomic data. In genomics, because spatial relationships often indicate functional relationships, a genome browser aims to help users form hypotheses about the function of different genomic elements. A genome browser displays several collections of data (i.e. tracks) that are aligned in respect to the genomic sequence. Often, the most important tracks are those that indicate genes, but tracks could also contain other information. Users can pan left and right on a genomic sequence as well as zoom in and out. For example, a user may request to view TP53, a human gene that is known to be related to cancer. A genome browser then represents the gene as a rectangle along a chromosome. The coordinates of the rectangle represent the location of the gene on the chromosome. The user may also pan left and right through the chromosome to view other genes in the region. Upon zooming, a gene is represented as a series of rectangles and arrows. Rectangles represent exons (areas of the gene that code for proteins), arrows represent introns (areas of a gene that are not translated to proteins), the direction of the arrows represents the direction in which the sequence of a gene is read (because the DNA is double-stranded, some genes are located on the leading strand and are read from left to right while others are located on the lagging strand and are read from right to left). Figure 1, shows a screen capture of the widely used UCSC Genome Browser. The center of the screen displays the gene track which shows the structure of the TP53 gene in terms of exons and introns. The top part of the screen shows a picture of the chromosome and highlights the area on the chromosome that is currently displayed. The bottom part of the screen displays the expression track, which uses color coding to convey the expression level of the TP53 gene in different tissues. Most current web-based genome browsers (e.g.

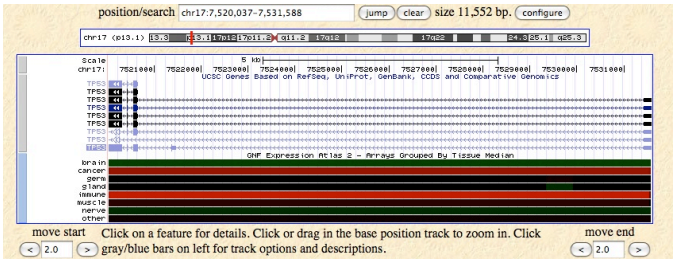

Figure 1. The UCSC Genome Browser. In the center, the gene track shows the structure of the TP53 gene. At the top, a figure of the chromosome highlights the viewed area. At the bottom, the expression tracks use color coding to convey the expression level of the TP53 in different tissues.

UCSC [22], Ensemble [7]) are implemented using HTML, allowing users to navigate genomic sequences using form-based controls. Thus, when users navigate through a region on the chromosome, they proceed through a series of static pages. These discontinuous page transitions often impair users' sense of location and context, leaving them wondering how the displayed data points are related [21]. Emerging web-based genome browsers (e.g. [1, 21]) attempt to address this problem by using Ajax technology to implement continuous zooming and panning. Also, a genome browser often serves as a gateway for a myriad of other web-based sources of information. Our subjects often access the literature database PubMed through a genome browser as well as search for ontology information. However, current genome browsers do not provide means for organizing and relating multiple forms of evidence. Thus, our subjects often develop ad-hoc techniques for relating the information they collect: three of our subjects maintain their own databases in which they link related publications to sequence information, two other subjects create visual networks of related genes where they annotate the connection between two genes, while others print out text files and publications and annotate them manually.

BLAST

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between genomic sequences. The BLAST tool is often accessed through a web browser. It compares a nucleotide (DNA or RNA) or protein (amino acid) sequence to sequences in different databases and calculates the statistical significance of the matches. All the matches that are above a certain threshold are displayed. Researchers use BLAST for a wide variety of tasks. For example, a researcher that plans to test a new cancer drug that targets the TP53 gene may use BLAST to compare the sequence of the human TP53 with the mouse and rat genomes in order to determine which of the animals is a better candidate for drug testing (one-to-many search). Or, a researcher that designs drugs that target particular areas on a gene may use BLAST to verify that this area has minimal similarity to other genes on the human genome so that the drug will only affect that particular area (one-to-one search). Finally, one of our subjects uses BLAST to find the extent to which the bacteria population in the human gut is genetically similar to the bacteria population on the human skin. To do so he uses the BLAST algorithm to compare all the genomic sequences extracted from a sample of the gut with all the genomic sequences extracted from a sample of the skin (many-to-many search). Existing web based BLAST tools present the results of one-to-one and one-to-many searches visually so that the matching sequences are aligned with the reference sequence sorted according to a similarity score (see Figure 2). Following the visual presentation are the details of the matching sequences (the matching score, location on the genome, and length). It is important to note that a BLAST search can return hundreds (or in some cases thousands) of results. Some of our subjects find the visual display of BLAST results confusing, while others ignore the visual display and only review the results in a text format.

Even though other bioinformatics tools are available for researchers, the above tools are fundamental to accessing and analyzing genomic sequences.

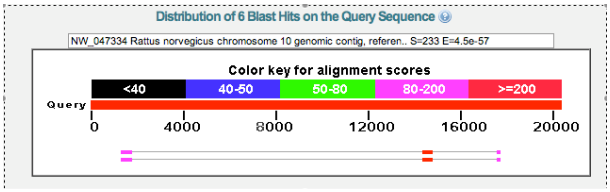

Figure 2. A visual display of similarity search results (by the NCBI BLAST). The human TP53 gene is represented by a red rectangle; below this rectangle two short sections of the rat genome that are similar to the human TP53 gene are represented by small red rectangles that are aligned with the the human TP53 gene.

Data Explosion

In the last decade, the cost per reaction of DNA sequencing has fallen with a Moore's Law precision [14]. In 2005, a person could sequence his whole genome for the cost of 350k, four years later several companies offer to sequence human genome for 60k, and soon the cost mark of 5k or even 1k will be reached. Although subsistence DNA sequencing is currently conducted at a single investigator, departmental, or university facility setting, high-throughput DNA sequencing is currently only performed in a handful of sites. However, the recent introduction of next-generation sequencing technology, capable of producing millions of DNA sequence reads in a single run, is rapidly changing the landscape of genomics. In the near future, such sequencing instruments will become available for more researchers, allowing a single lab to create in one year the same amount of data that was held in all the NIH sequence databases just 3 years ago. In the context of HCI, these advances present the challenge of providing researchers with means for sharing, searching, comparing, connecting, and organizing this vast amount of data.

One of our subjects (P4) is involved in the Human Microbiome Project that seeks to sequence the human genome together with the genome of the entire human micro-bacterial population. Every experiment that he conducts results in hundreds of millions of short genomic sequences that together describe the several thousands of bacteria species that populate the human body. The challenge he faces is how to organize these complex data sets into a form that will allow him to learn about the similarities and differences in the genomes of those micro organisms. In several cases, we observed that the ability to visualize large data sets is extremely helpful for biologists and can help them to understand and generalize complex phenomena they were not capable of understanding before the visualization existed.

Integrated Workflow

To gain insight into complex biological systems, our subjects often link together several data sets, each being handled with a special bioinformatics tool. We observed that subjects with background in biology often create a linear workflow by manually fetching data from one spot, reformatting the data, applying the next bioinformatic tool, parsing the results, reformatting the results, and so on. Often not comfortable with programming, they only rarely automate a workflow (typically by asking for help from a bioinformatician) and instead repeat required steps as necessary. As genomic sets grow larger, this method of operation becomes more and more time consuming. Thus, there is a clear need for providing means for easily linking both data and tools to create a workflow that can be repeated across experiments. In the words of one of our subjects:

> We searched for NOS3 in the Zebra fish genome, and when we got the DNA sequence we picked the largest exon. Taking the sequence for that exon, we pasted it into NCBI BLAST. Since it returned the NOS3 gene, we proceeded to the IDT site to input our sequence and get suggestions for possible primers for this exon. (P13)

Subjects with background in computational biology typically write scripts to execute a linear or parallel workflow. Our subjects work with shell scripts, Perl, Python, Matlab, and C. Some scripts are designed for single-use, other scripts are designed to be used in multiple experiments and sometimes by other users.

Multiple Forms of Evidence

Biological data is often noisy due to the complexity of living systems and the imperfection of measurement technologies. As one of our subjects describes:

> Sequencing DNA is not black and white, and therefore it is not so simple that the sequence we receive is 100 percent correct. Depending on the quality of the sequence, we need to adjust the acceptable level of agreements. (P2)

To address this uncertainty, researchers often combine multiple forms of evidence, as well as examine evidence from multiple resources. For example, because genes are detected through experimental evidence, no data indicates 'the genes' unambiguously [5]. Thus researchers often combine evidence from multiple gene data sources and view it in parallel. Our subjects often combine multiple forms of evidence not only to overcome uncertainty but also to discover connections and casual relationships. For example, some of our subjects often compare the genomes of multiple organisms to infer evolutionary relationships. To do so, they display the genome from each organism in a single genome browser track (i.e. row). This sometimes results in displaying more than a dozen tracks in parallel. Once inferring relationships between multiple genomes, users are often interested in aggregating the information and presenting their findings in a single track (i.e. row). Some subjects often examine information in different levels of granularity - moving back and forth between viewing a large chromosome area containing multiple genes and the base-pair level showing only a small area of a single gene:

> Since we don't know the exact location sometimes, we need to zoom in to the base-pair level, and zoom out to the level of viewing 10 or so genes to see their context. (P15)

We found that existing tools often overwhelm users with the amount of data presented on the screen, making it difficult for the user to organize the information in a way that highlights the connections between multiple forms of evidence. This observation is also supported in the literature [3, 15]. We thereby identify a clear need to support the display of multiple forms of evidence while allowing users to compare pieces of evidence, highlight the similarities and differences, and aggregate comparison results.

From Novice to Expert

Our subjects differ in their level of expertise both in terms of domain knowledge and of computer experience. Student researchers typically have 2-4 years of experience in their field of study, while researchers and faculty typically have more than 10 years of experience in their field. Although all of our subjects are comfortable using bioinformatics applications that have a GUI, subjects with background in molecular biology typically use only limited functionality of bioinformatics tools and are often cautious of trying new features. Computational biologists on the other hand, are trained in computer science and are expert users. When needed, they develop new computational tools to solve a particular biological problem. We observed that bioinformatics tools in general and genome browsers in particular have a high threshold: they are powerful, but in order for one to really take advantage of their power, they require both a broad domain knowledge and an extensive training. To support a wider range of researchers and empower users to examine new forms of evidence, there is a need for tools with lower threshold that encourage exploration of advanced features.

Collaboration

In leading research institutions and in industry, biological research is often conducted in multidisciplinary groups that are made of biologists, computational biologists and bioinformaticians. We learned from our subjects that in such teams biologists often come up with a biological question, and computational biologists translate the question into a process in which data is collected and analyzed. Communication in such teams is often the key to success. Our subjects indicated that collaborative work is typically based on emails and research meetings during which one of the researchers presents slides or distributes hard copy of results. In addition, researchers store their results in a shared database so that other group members can access the information. In smaller labs that consist of faculty and student researchers, several researchers often work together on the same computer exploring, analyzing, and discussing biological data. We found that despite the importance of collaboration in biological research, bioinformatics tools do not provide support for collaborative exploration.

THE G-NOME SURFER SYSTEM

Our user study helped us define a set of user-driven design goals for a visual computing tool for genomic exploration. These goals include: 1) facilitating collaborative, immediate and fluid interaction with large amounts of heterogeneous genomic information, 2) lowering the threshold for using advanced bioinformatics tools, 3) reducing mental workload associated with accessing and manipulating genomic information, and 4) improving current information workflow processes in genomic research. Based on existing research of tabletop user interfaces indicating that tabletop interfaces support collaboration through visibility of actions and egalitarian input [10], facilitate active reading [16], as well as afford distributed cognition (that could potentially lower mental workload) [17], we decided to utilize *tabletop* interaction in our system design. We thereby created G-nome Surfer, a tabletop interface for genomic research with the intention of meeting our user-driven goals. G-nome Surfer supports the five information tasks identified in the study: searching literature, locating a gene on a genome, retrieving genomic sequences, searching for similarity between sequences, and annotating genomic information.

Seven of our study participants participated in the iterative development process of G-nome Surfer by providing feedback to a series of prototypes in increasing fidelity. Their feedback informed the current design of G-nome Surfer that is implemented on top of the Microsoft Surface platform. Following, we describe G-nome Surfer‘s primary interaction techniques and implementation.

Navigating Genomic Maps

G-nome Surfer supports the navigation of genomic maps in multiple zooming levels. To access a genomic map, a user first selects a clade, an organism, and an assembly, and then specifies a chromosome, a range, or a name of a particular gene. The gene is displayed in the center of a chromosome track. The default zooming level captures approximately five adjacent genes. An indicator that is shaped as a box below the chromosome track shows what portion of the chromosome is displayed. Each gene is represented as an arrow upon one of the two DNA strands. The direction of the arrow represents the direction in which a gene is read. The coordinates of a gene represent its location upon the chromosome in terms of base-pairs (see Figure 3). Users can then pan right and left through the chromosome simply by using flick gestures. Continuous visual feedback maintains users‘ sense of location. When the user taps on a gene, its structure in terms of exons and introns is displayed in a structure track below the chromosome. A polygon connects the gene and its structure track to support the user‘s sense of location. Exons are represented as rectangles upon the structure track. Introns are represented by the empty space between exons (see Figure 3). To display the DNA, RNA, or amino acid sequence of a gene, the user taps on the structure track. If the user taps on an exon, she can choose either to display the sequence of the entire gene or only the sequence of that particular exon. The sequence is then displayed in a separate window that is connected to its source gene. Users can display multiple sequences of the same gene (e.g. a user may display the RNA and amino acid sequences of a gene), each sequence opens in a new window that is connected to its source exon. Users can move, orient, resize, and arrange the windows as well as annotate genomic sequences. To compare two sequences, a user can overlay two windows and align the sequences. Figure 3 shows the aligned RNA and amino acid sequences of TP53.

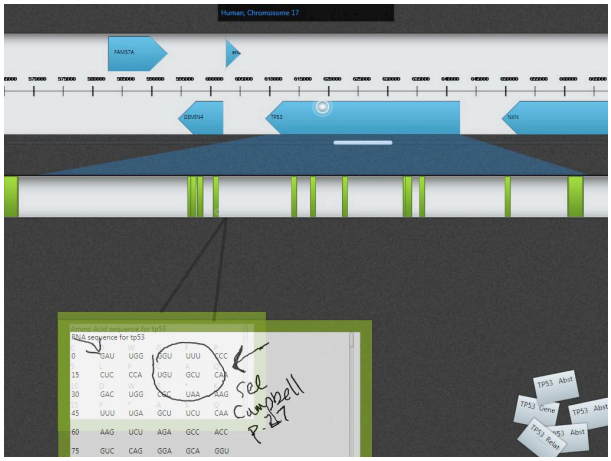

Figure 3. A genomic map of TP53. A chromosome track (top) shows TP53 chromosomal environment. A structure track (center) displays its structure in terms of exons and introns. Aligned RNA and amino acid sequences of TP53 are displayed at the bottom.

Heterogeneous Information Upon Request

G-nome Surfer enables researchers to access and relate heterogeneous information. When holding a finger upon a gene, the gene star-shaped context menu appears, allowing the user to choose between ontology, publications, and gene expression information. Ontology displays a summary of the publicly known information about the gene from the Entrez Gene database. Publications displays the titles of publications related to this gene from the PubMed database. When tapping on a publication, the abstract (with a link to the full paper) opens in a separate window. Gene expression presents expression levels in different tissues. While the gene expression information is taken from the UCSC Genome Browser [22], we created a new visualization that uses the accepted color coding scheme (red for high expression, green for low expression) but displays the information in hierarchical structure upon an organism's body. Figure 4 shows the gene context menu and the different pieces of information that relate to the human gene TP53. Each piece of information is displayed in a separate window so that users can move, orient, resize, and spatially arrange the information upon the surface.

To allow users to save all the information related to a particular gene, we decided to use a tangible test tube container (see figure 5) that is associated with a particular gene. When placed upon the surface, the tube attracts all the pieces of evidence related to that particular gene, then when removed, that information disappears. When placed upon the surface again the information reappears on the surface. We selected to represent storage with a physical tube because it is a familiar object - tubes are often used as a portable storage object in the lab. In the future, we plan to allow users to place multiple tubes on the surface to explore genes interaction, as well as to move information between surfaces. We believe that the immediacy and persistence of tangible interaction will enhance these tasks. A recent paper by Kirk et al. [13] informed our design by highlighting design considerations and tradeoffs of choosing between physical and digital representations. Currently, we pre-assigned a limited number of tubes to represent a pre-defined set of human genes. However, we are developing a mechanism for dynamically coupling tube objects to genes.

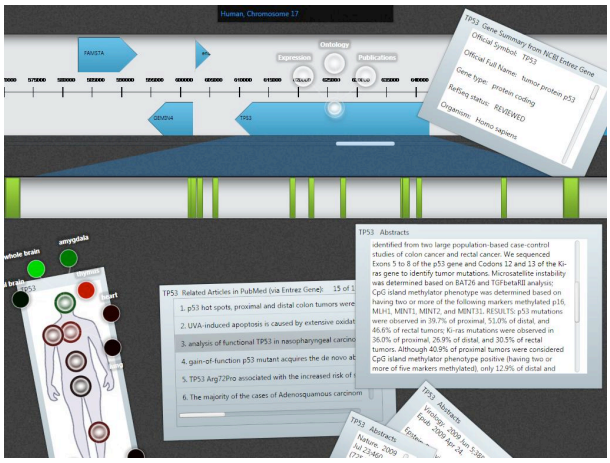

Figure 4. The gene context menu and heterogeneous information related to TP53 (structure, publications list, expression, ontology, and various papers).

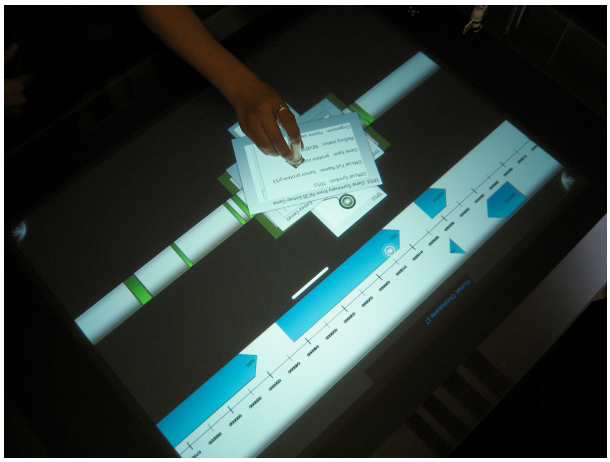

Figure 5. A tangible test tube container stores all the heterogeneous information related to a particular gene associated with the tube.

Similarity Search

After a particular sequence is displayed on the surface, a user can perform a BLAST search to find regions of local similarity between that sequence and the genomes of other organisms. To do so, the user places a tangible BLAST tool upon the sequence (see Figure 6). A semi-transparent layer then covers the surface (i.e. BLAST layer) and a BLAST context menu is presented. We chose to represent the BLAST tool using a playful tangible object to make this mode change immediate, visible, and easily reversed. After the user selects from the menu one or more organisms, the BLAST search is invoked and search results are displayed. Because such

search can yield numerous results, and considering the limitations of current representations of BLAST results, we created a new visualization for BLAST results. Figure 6 shows the BLAST results of TP53 against the mouse, rat, and monkey genomes. Each Blast result (i.e. an alignment to a genomic region with a similarity score that crosses a certain threshold) is represented as a rectangle in a shade of green. The color of the rectangle represents the similarity score of that result - the brighter the color the higher the similarity score. Each rectangle contains further details about that particular BLAST result including the target organism, the exact similarity score, the length of that genomic region, and its location upon the genome. The results are organized in a flower-like structure around a target organism so that target organisms with more results are displayed closer to the bottom of the surface.

In order to align a BLAST result with the source gene, a user taps a result. The result is then automatically aligned with the source gene. The alignment is presented at the bottom of the BLAST later. A user can align multiple results with the source gene. Finally, upon locating a set of BLAST results of interest, a user can save those results to the main G-nome Surfer layer. When the user removes the tangible BLAST tool, the BLAST layer disappears and the selected results appear on the main layer.

Figure 6. Interacting with BLAST results of the human TP53 gene against the mouse, rat, and monkey genomes. A physical BLAST tool (shaped as a spacecraft) presents a star-shaped search menu when placed upon the surface. Search results are organized in a flower-like structure around a target organism, with bright results represent high similarity score.

Reality-Based Interaction

The design of G-nome Surfer draws on users' existing knowledge and skills to provide a tabletop reality-based interface [11]. Specifically, G-nome Surfer uses naïve physics metaphors such as inertia, transparency, and mass in the layout of chromosomes and genes and in the representation of BLAST results. The interface also leverage users' spatial skills, allowing them to spatially organize information upon the surface to express relationships between multiple forms of evidence. Like tabletop interfaces in general [10], G-nome Surfer draws upon users' social skills and existing social protocols to afford collaborative interaction: the system provides multiple points of entry (through multiple forms of evidence that can be simultaneously manipulated), and makes modes visible to all users through the use of visual and physical physical objects (e.g. the BLAST tool).

Implementation

G-nome Surfer is written in C# using the Microsoft Surface SDK. It uses web services to draw genomic information from various databases including UCSC, Pub Med, and Entrez Gene. The BLAST search is implemented using the Washington University BLAST (WU-BLAST) web service. In order to improve the performance of G-nome Surfer, we are currently working on implementing a local database and BLAST search. Tangible objects are tagged with the Microsoft Surface fiducial tags.

Extensibility

G-nome Surfer was designed with extensibility in mind. For example, additional forms of evidence (e.g. protein structure and gene family information) could be added through web services, the star-shaped context menus could be extended using hierarchical organization to support additional functionality, and the layer metaphor that is used for displaying BLAST results could be reused to support additional services. We are currently extending the system to support the display of multiple genomes in parallel (e.g. human next to other organisms), and the display of isoforms (i.e. multiple RNA sequence sections from the same gene). In addition, we are currently developing a gene interaction layer that will support the investigation of gene interaction.

SYSTEM EVALUATION

To evaluate the usability of our design, we conducted a first-use study of G-nome Surfer. In particular, we were interested in validating that: 1) G-nome Surfer enables users to gain biological insights based on the multiple forms of evidence it provides, with minimal overhead, 2) G-nome Surfer presents BLAST results in a way that facilitates rapid identification of relevant results, and 3) G-nome Surfer supports smooth transition between the chromosome level and the base-pair/protein levels.

To define an insight, we draw on Saraiya et al. [19] that view insight as "an individual observation about the data by the participant, a unit of discovery" (p. 444). They group bioinformatics insights into four categories: overview (overall distribution), patterns (identification or comparison across data attributes), groups (identification of comparison of groups of entities), and details (focused information about a specific entity). In our study, we asked subjects to find answers to biological questions that require *surface* biological insights from all four categories. Surface insights are based on information findings and typically answer "what" questions. We define overhead as training time combined with the time users spend on tasks that are not directly related to biological research such as reformatting data, importing and exporting files, and sorting data.

Sessions were held at the HCI lab. In each 40-minutes session we asked two participants to work together to complete a task. The 16 participants (all female) included undergraduate biology students with background in genomics and some research experience. Some participants had limited previous experience with bioinformatics tools. We chose to test the system with students because they represent an important user population (i.e. student researchers) and were most available for first-use testing. Following a five minutes introduction to the Microsoft Surface and to the basic functionality of G-nome Surfer, we asked the participants to work together to complete a task on their own: provide evidence that the human gene TP53 is a candidate for the development of cancer gene therapy treatment and select a model organism for testing the treatment. We modeled this task to mimic a real scenario that we observed in our background study. To accomplish this task, subjects needed to complete four subtasks that include: gene location, evidence search, sequence retrieval, and similarity search. We asked our subjects to follow steps and record answers to biological questions that reflect surface insights as they proceed toward accomplishing the task. Following the task, we asked the participants to complete a questionnaire with their opinion on the task and G-nome Surfer. We also conducted an informal debriefing with participants and examined their answers to the biological questions. We observed the participants and took notes during the session as well as videotaped each of the sessions.

Results

Following a brief training, all participants were able to complete the task (and found *correct* answers to the biological questions) within a reasonable time. Table 2, shows mean times for completion, and mean confidence score (on a 7-points scale where 7 is most confident) for the overall task and for each of the four subtasks. In general, participants reported that the overall task was moderately difficult (mean score of 5.3, where 7 is most difficult), that they were fairly confident in their overall findings (mean score of 5.8, where 7 is most confident), and that the task had a relatively low mental workload (mean score of 2.5, where 7 is highest workload).

Description	Mean	SD	Confidence
Locating a gene and examining its struture and environment	6.2	2.3	6.4
Gathering heterogeneous information (expression, publications, and ontology)	5	0.7	5.9
Retrieving and exmaining genomic sequences	3.3	1.2	6.1
Conducting and interperting similarity search	3	0.5	6.2
Overall task- investigation of TP53	17.5	4	5.8

Table 2. Time for completion (in minutes) of the study sub tasks.

Identifying BLAST Results

All users successfully conducted a similarity search (BLAST). The times for completion of the BLAST task (see Table 2) indicate that users were able to rapidly identify relevant BLAST results with high confidence level. In the debrief interview participants mentioned that the "Blaster as an external tool was great" and that "it is easy to compare different organisms". However, some participants reported that due to the color contrast we used in the BLAST visualization, it was hard to read the details of BLAST results with low scores. Because some biologists are interested in further investigating results with low similarity scores, we plan to address this issue promptly.

Transition Between Zooming Levels

The subtask of locating a gene and examining its structure and environment required users to transition from the chromosome level to the base-pair level and back several times. All users successfully completed this task with high confidence level (see Table 2). This includes correctly answering biological questions about the location, structure, and the chromosomal environment of the TP53 gene. In the debrief interview, participants mentioned that transitions are "easy to follow" and that "it is easy to overlay amino acid / RNA sequences".

Collaboration

As expected, participants collaborated during the task. We asked each participant to what extent she and her partner were working together, and responses had a mean score of 5.9 where 7 is 'we highly collaborated when exploring the information'. We further observed that in all teams, both participants were actively interacting with the information. Some teams collaborated in a turn-taking manner, while others worked in parallel. We also observed that all teams were frequently gesturing and actively discussing the task, results, and interaction techniques aloud as advised by the study instructions.

Summary

Overall, participants were excited about working with G-nome Surfer, with a mean enjoyment score of 6.3 where 7 is most enjoyable. They described the experience as "fun", "hands on", "easy to use, smooth, good response time" and "visually stimulating". We allot some of this excitement to the novelty of the Microsoft Surface. More specifically, participants liked the integration of information, flexibility of moving and resizing screens, and searching for specific genes and linking them to pertinent articles. During the debrief interview, one participant described G-nome Surfer as "designed for multiple users but simple enough that one user can also accomplish the task". Another participant mentioned that G-nome Surfer "helped visualize gene location and synthesized a lot of information in one place".

Unfortunately, at the time of the study our mechanism for deleting information from the surface was not stable, so multiple participants reported that their surface became too cluttered with information that they wished to remove. Participants also suggested to add more text-based information (e.g. adding coordinates on the structure track, exon length) that will increase confidence level in results, as well as additional functionality for searching within publications (e.g. for keywords), sorting and annotating publications. We plan to address these issues as we expand G-nome Surfer.

Discussion

This study was intended as a first-use evaluation of our design. Its results show that G-nome Surfer supports the five information tasks identified by our user study and sets a relatively low threshold for using bioinformatics tools. This study also demonstrates that G-nome Surfer provides a collaborative, immediate and fluid interaction with heterogeneous genomic information. However, an additional investigation is required to evaluate whether G-nome Surfer reduces the mental workload associated with accessing and manipulating genomic information in comparison to current GUI tools. Also, to distinguish the strengths and limitations of G-nome Surfer in comparison to a *collaborative* GUI, a *comparative study* is required: specifically, we are interested in separating the limitations of current tools from limitations that are inherent to conventional GUIs by identifying which of the G-nome Surfer's advantages are products of its visual design and integrated workflow, and which are products of a horizontal tabletop interface. To truly assess whether G-nome Surfer improves current workflow processes, a longitudinal study within a research lab is required. Finally, it is important to note that all of our subjects in this study were female. This eliminated a confounding variable of gender, but does not represent most research settings of mixed gender. In the future, we will look at male and mixed gender groups.

RELATED WORK

Evaluation of Bioinformatics Tools

Several studies indicate that current web-based bioinformatics tools show severe limitations in supporting users to find answers to complex biological questions [23, 3, 15, 12]. To better understand the requirement for supporting deep causal insights, Mirel performed a longitudinal field study of biomedical researchers [15]. Her findings indicate that using the tested tool, scientists could successfully pose and answer "what" questions that are based on information findings, but could not easily develop explanatory insights that require answering "how" and "why" questions. Her study results further suggest that design choices of such tools should support contextualizing relationships, highlighting biologically meaningful concepts, and flexible interactivity. These findings informed our design.

Reality-Based Interfaces for Scientists

A number of systems illustrate the vast possibilities of supporting scientific discovery through reality-based interaction. Brooks et al. [4] developed the first haptic display for scientific visualization. This display improved the perception and understanding of force fields and was used by chemists to investigate docking positions for drugs. Gillet et al. [8] presented a tangible user interface for molecular biology that used augmented reality technology to augment 3D molecular models. While users collaboratively manipulate the physical object, the system superimposes graphical information upon the physical model. The system was only preliminarily evaluated with professional researchers. Schkolne et al. [20] developed an immersive interface for the design of DNA molecules that uses tangible objects to create and edit digital representations of DNA molecules. A user study with scientists finds this system to be more satisfying for users than a corresponding 2D system. While these systems focus on the representation and manipulation of objects that have an inherent physical structure. We are interested in a broader use case, where *abstract* information is represented and manipulated. Labscape [2], is a smart environment for the cell biology laboratory that helps biologists to produce more complete records of their work. The system allows biologists to easily record, relate, and share heterogeneous information about their lab work. ButterflyNet [25], is a mobile capture and access system for field biologists that integrates paper notes with digital photographs captured during field research. While addressing different needs and requirements, both systems share our challenge of organizing and relating heterogeneous information.

To date, a few systems have been developed to facilitate collaboration among scientists across large displays and multi-touch tables. WeSpace [24], integrates a large data wall with a multi-user multi-touch table and personal laptops to provide group members equal access to touch manipulation as well as live rendering and interactive visualization. TeamTag [18] is a tabletop interface that allows biodiversity researchers to collaboratively search, label, and browse digital photos. Finally, Involv [9] is a tabletop application that uses the Voronoi treemap algorithm to create an interactive visualization for the Encyclopedia of Life. Involv is the closest to our work as it shares the challenge of creating effective tabletop interaction for exploring massive data spaces.

CONCLUSIONS AND FUTURE WORK

This paper makes three contributions. First, we described a study of molecular and computational biologists that identifes design requirements for supporting scientists searching through and organizing vast amounts of heterogeneous information in order to gain biological insights. Second, we described G-nome Surfer, a tabletop user interface for collaborative exploration of genomic data that employs multi-touch and tangible interaction techniques. G-nome Surfer enables users to gain biological insights based on the multiple forms of evidence it provides with minimal overhead. It also facilitates rapid identification of BLAST results, and supports smooth transition between the chromosome level and the base-pairs level. As a result, it lowers the threshold for using bioinformatics tools and encourages exploration of multiple forms of evidence. Finally, we presented results from a first-use usability study of G-nome Surfer.

We plan to expand G-nome Surfer so it will be suitable for use in scientific and educational (college level) settings. This includes improving G-nome Surfer's performance through local database and search, as well as supporting the display of multiple genomes in parallel (e.g. human next to other organisms), and the display of isoforms (i.e. multiple forms of the same gene). We intend to further evaluate G-nome Surfer's strengths and limitations in comparison to current state-of-the-art GUI based tools, and to collaborative GUI

with multiple mice. We also plan to conduct a longitudinal study in both research and educational settings.

While the domain of bioinformatics provides the frame for this work, this research contributes to tabletop interaction in general by demonstrating its application for supporting domain experts interacting with large amount of abstract, heterogeneous and complex information.

ACKNOWLEDGEMENTS

We would like to thank the biologists that participated in our study. We also thank the Brachman Hoffman foundation for supporting this work.

REFERENCES

1. Argo Genome Browser, http://www.broadinstitute.org/annotation/argo/.
2. Arnstein, L., Hung, C.-Y., Franza, R., Zhou, Q.H., Borriello, G,Consolvo, S., Su, J., Labscape: A Smart Environment for the Cell Biology Laboratory. In *IEEE Pervasive Computing Magazine*, 1(3). pp.13-2, 2002.
3. Bolchini, D., Finkelstein, A., Perrone, V., Nagl, S. Better bioinformatics through usability analysis, *Bioinformatics*, 25(3), 406-12, February 2009.
4. Brooks, F. P., Ouh-Young, M., Batter, J. J., and Jerome Kilpatrick, P. Project GROPEHaptic displays for scientific visualization. In *Proceedings of the 17th Annual Conference on Computer Graphics and interactive Techniques*, Dallas, TX, USA, 1990.
5. Cline, M.S. and Kent, J.W., Understanding genome browsing, *Nature Biotechnology*, 27(2):153155.
6. Coutere, N., Riviere, G., Reuter, P., GeoTUI: A Tangible User Interface for Geoscience, In *Proceedings of the Second International Conference on Tangible and Embedded Interaction*, ACM Press, 2008.
7. Ensembl, http://www.ensembl.org/index.html.
8. Gillet, A., Sanner, M., Stoffler, D., Goodsell, D., and Olson, A. 2004. Augmented Reality with Tangible Auto-Fabricated Models for Molecular Biology Applications. In *Proceedings of the Conference on Visualization '04*. IEEE Visualization, October 2004.
9. Horn, M.S., Tobiasz, M., Shen, C., Visualizing Biodiversity with Voronoi Treemaps. In *Proc. Sixth International Symposium on Voronoi Diagrams in Science and Engineering*, Copenhagen, Denmark. June 23-26, 2009.
10. Hornecker, E., Marshall, P., Dalton, N.S., Rogers, Y., Collaboration and Interference: Awareness with Mice or Touch Input. *Proc. of ACM CSCW Conference*, 2008.
11. Jacob, R.J.K. , Girouard,A., Hirshfield, L.M., Horn, M.S., Shaer, O., Solovey, E.T., and Zigelbaum, J., Reality-Based Interaction: A Framework for Post-WIMP Interfaces, In *Proc. ACM CHI 2008 Human Factors in Computing Systems Conference*, pp. 201-210, ACM Press, 2008.
12. Javahery, H. and Seffah, A., Refining the Usability Engineering Toolbox: Lessons Learned from a User Study on a Visualization Tool, *HCI and Usability for Medicine and Health Care*,185–198, 2007.
13. Kirk, D., A. Sellen, S. Taylor, N. Villar, and S. Izadi. 2009. Putting the physical into the digital: Issues in designing hybrid interactive surfaces. In *HCI 2009*, pp. 35-44. British Computer Society.
14. Mardis, E., The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3), 133-141, March 2008.
15. Mirel, B. Supporting cognition in systems biology analysis: findings on users' processes and design implications, *J Biomed Discov Collab*. 27(2),153155, February 2009.
16. Morris, M.R., Brush, A.J.B., and Meyers, B., Reading Revisited: Evaluating the Usability of Digital Display Surfaces for Active Reading Tasks. *Proceedings of IEEE Tabletop*, 79-86, 2007.
17. Patten, J., and Ishii, H., A Comparison of Spatial Organization Strategies in Graphical and Tangible User Interfaces, in *Proceedings of Designing Augmented Reality Environments* , pp. 41-50, 2000.
18. Ryall, K., Forlines, C., Shen, C., Morris, M.R., Everitt, K., Experiences with and Observations of Direct-Touch Tabletops. In *Proc. IEEE Tabletop'06*, 2006.
19. Saraiya, P. and North, C. and Duca, K., An insight-based methodology for evaluating bioinformatics visualizations, *Visualization and Computer Graphics*, 11(4), 443-456, 2005.
20. Schkolne, S., Ishii, H., Schroder, P., Immersive design of DMA molecules with a tangible interface, *Visualization*, IEEE, pp. 227- 234, 2004.
21. Skinner, M.E., Uzilov, A. V., Stein, L. D., Mungall, C.J., and Holmes, I.H., JBrowse: A next-generation genome browser, *Genome Research*, 19(9), 1630-1638, September 2009.
22. UCSC Genome Browser, http://genome.ucsc.edu/.
23. Veretnik, S. and Fink, J. and Bourne, P., Computational biology resources lack persistence and usability, *PLoS computational biology*, 7(4), July 2008.
24. Wigdor, D., Jiang, H., Forlines, C., Borkin, M., Shen, C. 2009. The WeSpace: The Design, Development and Deployment of a Walk-Up and Share Multi-Surface Visual Collaboration System. Proceedings of CHI 2009, Boston, MA.
25. Yeh, R., Liao, C., Klemmer, S., Guimbretière, F., Lee, B., Kakaradov, B., Stamberger, J. and Paepcke, A., ButterflyNet: a mobile capture and access system for field biology research, in *Proceedings of the SIGCHI conference on Human Factors in computing systems*,571–580, 2006.