BioMedia: Multimedia Information Systems for Biology Research, Education and Collaboration


Computer Science Department
* Biology Department
San Francisco State University
1600 Holloway Avenue, San Francisco, CA, 94132

Abstract
The long-term goals of the recently started Biomedia project at SFSU are to provide multimedia information systems and applications for the research and education needs of several projects in the SFSU Biology Department. These applications involve a considerable amount of images and image sequence data, in addition to traditional text, genomic, and experimental measurement data. Our systems will allow biology researchers and students to store, index, annotate, search, visualize, analyze, collaborate, and share a large amount of heterogeneous biomedical data. Our initial focus in BioMedia is the creation of collaborative WWW site for the Hedgehog gene pathway. The Hedgehog (Hh) protein super family constitutes a group of closely related secreted proteins that control many crucial processes during the embryogenesis of tissues. The overall goals of the Hh WWW Site project are as follows: a) to provide a WWW site to be used by researchers and students studying the Hedgehog gene pathway and made available to broad community, and b) to provide advanced and innovative functionality enabling easy usage and management, community based content submission and updates, and asynchronous collaboration between researchers and students. In this paper we present the status and first results in building and researching technologies necessary for this WWW site.

1. Introduction
The Computer Science Department and the Biology Department at San Francisco State University have embarked on a collaborative activity, BioMedia. The long-term goals of this collaboration are to provide multimedia information systems and applications for the research and education needs of several projects in the SFSU Biology Department. Our application candidates are from the areas of cell and molecular biology, ecology and systematic biology, insect neurology and others. These applications involve a considerable amount of images and image sequence data, in addition to traditional text, genomic, and experimental measurement data. Our systems will allow biology researchers and students to store, index, annotate, search, visualize, analyze, collaborate, and share a large amount of heterogeneous biomedical data. The first chosen application relates to the Hedgehog gene pathway.

While this paper focuses specifically on a multimedia information repository for the hedgehog gene pathway, the technology we develop and the research we perform into usability and user centered design for biology applications will have, it is hoped, a much broader scope. In particular, the effective design of applications for one domain will give us an understanding of the technology needed by biologists in other domains. We envision this application area as a first step toward a broader understanding of the computational requirements of real-world researchers working in the Life Sciences.

While the size of the bioinformatics data is enormous, advances in hardware and basic system software (OS, databases, multimedia support, interfaces and the Internet) and networking will make the cost of capturing, storing, processing, transmitting and rendering these data feasible. (Under data here we really mean bits vs. the semantic meaning of the data). However, the key obstacles that will remain in spite of these advances are as follows:

- There are no systems that enable effective iterations and collaboration in a typical biology research and education process.
- High cost, low usability, and low efficacy of data and WWW site management, content submissions and manual metadata entry are prohibitive for the large bio-information systems we envision.
There is a lack of adequate fully automated algorithms for data analysis, interpretation, and processing, requiring considerable expert and manual user involvement.

Our initial focus in BioMedia is the creation of a collaborative WWW site for the Hedgehog gene pathway. The Hedgehog (Hh) protein super family constitutes a group of closely related secreted proteins that control many crucial processes during the embryogenesis of tissues. Originally discovered in Drosophila (fruit fly), this protein super family and the cell pathway it activates in cells has now been well characterized in a variety of organisms from flies to humans (Ingham, and McMahon, 2001). This pathway has now become to be known as the "Hedgehog Gene Pathway". There are many protein components of this gene pathway and current research has focused on the function and inter-relationships of these genes. For example, the Hedgehog gene pathway's fundamental role in such diverse processes as: bone development, spinal and brain development, endodermal tissue formation, skin development and axis formation has led researchers to recognize its importance not only in normal development but also in a variety of human diseases (Ingham, and McMahon, 2001). More specifically, Hedgehog's role in several cancers such as basal cell carcinoma, small-cell lung cancer, pancreatic cancers, and digestive tumors, is testament to the pathway's ever increasing clinical importance (Ingham and McMahon, 2001; Watkins et. al., 2003; Thayer et. al. 2003; Berman et. al. 2003;).

As a result of this intense effort to understand this pathway much data has been generated. This bioinformatics data is complex and comes in many different forms from sequence gene homology (text and images), to molecular interaction analysis (text, images, video), to more recently microarray profiles of cancerous cells and emerging therapeutic agents. Surprisingly, no organized repository of all this research on the Hh pathway exists even though new information is generated weekly by researchers around the world.

Motivation for this project in part comes from a very successful model for a repository of another important gene pathway in embryogenesis and human diseases, the Wnt/Wingless pathway, established and maintained by Dr. Roel Nusse and researchers at Stanford Medical School (http://www.stanford.edu/~rnusse/wntwindow.html). This site is an important tool used by researchers around the world as a place where all that is known about this gene pathway is stored and/or linked. It is also a site where researchers can add and share information on this gene pathway.

The overall goals of the Hh WWW Site project are as follows:

- To provide a WWW site to be used by researchers and students studying the Hedgehog gene pathway and to make it available to the broader community.
- To provide advanced and innovative functionality enabling easy usage and management, community based content submission and updates, and asynchronous collaboration between researchers and students.

The focus of our initial work, which will be presented in this paper, is based on a User Centered Design methodology (Torres, 2002). All the major decisions are driven from the user perspective (instead of from the technology perspective). Our design process first involved the analysis of key use cases, benefits and user needs, followed by usability design (specifically a socio-technical analysis of how key requirements play out in the system), and UI design and analysis. Only after we have analyzed and considered the key usability requirements do we consider technology and architectural issues.

In Section 2 we present a socio-technical analysis of the requirements of the multimedia information repository. In section 3, we discuss our approach to the design of the repository, focusing on asynchronous "annotations" as a metaphor for content submissions and collaboration. In section 4, we describe the current design of the system and include screenshots of the prototypes of the repository. Finally, in section 5, we conclude with a description of our contributions to date and planned future work.

2. Hedgehog (Hh) Gene Pathway WWW site: Requirements and Usage cases

To design a system around our users, we adopt a socio-technical analysis of the usage characteristics of our system. This analysis is based around the concept of “stakeholders”, individuals and groups that have an interest in the design and implementation of the project. In our project, we consider the design criteria from the perspective of the following stakeholders:

- The Primary Stakeholder is the website owner, namely a researcher in the SFSU Biology Department. Let us call this user the owner. There could conceivably be more than one owner.
• Secondary Stakeholders are members of the Hedgehog community who contribute to this site, and include researchers inside and outside of the SFSU Biology Department. We call these users *contributors*.

• Tertiary Stakeholders are those who visit this site to become informed about the Hedgehog gene pathway or who can submit some comments. We call these users *observers*. They can be active (submitting comments) and passive (only reading comments).

• *Regular* users are normal WWW users accessing the site. They may not be able to see all the information of Hh site (such as comments).

• Finally, technical stakeholders are the website designers, called *designers*.

Owners, contributors and observers are registered users and they form the Hh *community*. The Hh WWW site, in its design, has the property that it is in essence populated and updated by the Hh community, with owner(s) having the ability to “police” the site content. The process of content update is asynchronous, meaning any approved user can do it at any time.

To analyze the owner’s concerns, we held extensive discussions on the design and implementation of the website with the site owner. His primary goal in creating the website is to inform him and his research community of information on and advances in the Hedgehog gene pathway. The site owner also has some idea of the goals and content of the website. He describes the typical user of the site as a biology researcher or student (both basic and biomedical) at a university with access to a high speed internet connection and an understanding of database searching on the web.

Expectations of the website are:

• A comprehensive catalog of all research on Hh that is easy to search, browse and visualize.

• Fast download of information.

• The ability to easily (via graphical and textual means) add to content using a web page interface.

• The ability to communicate and share data with other researchers in other parts of the world.

The site owner notes that there are several main data types that the Hh web site will contain, including sequence information, associated genes in the pathway, and current experimental visual data. This data, representing an expression of the gene, can take on various forms including text, images, spreadsheets, video and graphics. The sizes of files are from a few bytes to over 1GB. Content should include, minimally:

• A Database of all known members of the Hh super family in various organisms, including DNA and protein sequence data for sequence comparisons between Hh subtypes. The DNA sequencing data should include coding and non-coding regions to aid researchers in searching for conserved regulatory motifs. Links to various gene/protein bank databases should also be provided.

• A Database of all genes known to be involved in the Hh pathway in various organisms. This should include DNA and protein sequence data, protein sequence motifs, molecular interaction data, and associated mutational or disease information.

• A Database of all known gene targets of the Hh pathway derived from microarray experiments involving the Hh pathway.

• A Database of all known human diseases associated with the Hh pathway. This will also include experimental data on therapeutic agents and their molecular targets.

• A Database of protocols, tissue cell lines, and reagents used by Hh pathway researchers.

• A Database of all researchers working on the Hh pathway in various organisms.

• A Newsgroup area for researchers in the Hh pathway to use as a means to reach others to discuss current research and problems in the field.
Note that some of the data will be represented as links to original databases, due to practicality of incorporating and maintaining very large DB, as well as other legal issues.

Of paramount importance will be the continual updating of the data in the multimedia repository by researchers around the world. One constraint on the construction of the website is that the owner cannot contribute all data. Even beyond contribution, it will be impossible for the owner to view all content. In section 3, we describe how we keep owner-control without time-consuming gate keeping.

To enhance the educational and collaborative aspects of our repository, it is essential to have a mechanism to support informal content as well. Asynchronous annotations (submitted at any time and order) are one means of associating informal meta-data with core information. As a result, our observer users can be viewed as two groups. There is an active browser of the website, who may wish to post commentary either for him or her self or for a broader audience. There are also observers who simply wish to read the material (data and annotations) on the website. We support both the notion of “active participation” and “passive participation” through the optional use of informal annotations.

Typical use cases from the perspective of contributor would be:

- A researcher has a large data set that she wishes to post on the website. If the contribution becomes a large burden, she will be reluctant to contribute information to the website. She prefers a graphical means of user interface, possibly with some templates. In some cases she contributes raw content with its metadata, in some cases she submits links to a large volume of content located elsewhere.

- A researcher publishes a paper that extends other work currently posted on the website, but also makes a connection with other concepts in the Hh gene pathway ontology. The researcher submits the paper to the site and also adds linkages (with optional annotations) among the newly observed connections.

- A researcher notices inconsistencies in a paper on the website. She highlights the inconsistent text and posts annotation with her comments addressed to the author of the paper. The author of the paper with the alleged inconsistencies, on seeing the comments, responds by providing information to clear up any perceived inconsistencies.

Beyond contributors, the website is also a repository for learning about the Hedgehog gene pathway. As such, people studying the site (i.e. observers) have a variety of behaviors that we wish to support:

- A student studying the website to gain knowledge notices a piece of information that she does not understand. She highlights the confusing data and poses a question. She can chose to address her annotation to all (make it public) or to the author, if she knows who he/she is. She can optionally use integrated e-mail to alert the author to the question. The author (or anybody from the Hh community upon seeing the question) provides the answer.

- A professor using the website as a resource during course teaching creates a series of questions embedded in contributions on the website, perhaps a digital video of a certain aspect of the pathway. His students browse the website and see the questions while browsing the site. The students can then answer the questions on the website. The professor can verify their answers to ensure that they understand the material.

- A group of researchers has a large data set that must be annotated with metadata to be useful. They post this data set on the website and invite others to contribute comments and metadata to make the data set more useful. Over the next two or three weeks, other website users see and annotate small portions of the data set. By exposing this to Hh community, a large amount of data is annotated and categorized, which would otherwise be impractical by a single group or organization.

- A research project obtains some experimental data that requires advanced computer vision algorithms. They, however, have no such expertise. They post the data or an explanation of the problem on the website. Another research group, noting the problem, contributes a computer vision algorithm that effectively analyzes the data.

3. A Conceptual Model for Hh WWW site: Annotations as Content

In our design, we envisage the Hh WWW Site as a framework or ontology for information about the genetic pathway. In other words, the informational organization of the Hh WWW site follows the well known Hh gene pathway ontology, i.e. the organization and objects, and the concepts and relationships related to the Hh gene. A graphical representation of this
ontology is shown in Figure 1. We can then view content submissions and updates as, in fact, the process of annotation of this ontology originally provided (with some content) by the Hh WWW site owner. This view of our website as a basic ontology with annotation as a main contribution method provides us with both a cognitive model of our website and of the contribution process.

Contributions or annotations to Hh WWW site can therefore occur at three levels:

- **Content annotations:** Contributors may create new pieces of information on the website by attaching content (raw data with metadata, or links to raw data, with metadata) to a particular object or concepts in the ontology (allowed for owners and collaborators).

- **Linkage annotations:** Contributors may wish to enhance information already on the website with new linkages/relationships pointing to new observed relationships among objects or concepts (allowed for owners and collaborators). These will be combined with implicit linkages (obtained by analyzing the metadata of submitted content) and further available for filtering, searching and browsing.

- **Collaborative annotations:** Overall, the notion of annotation has been one of associating value-added metadata to information already available. In order to support this, some free-form notion of annotation to support commentary, questions and interpretations necessary to provide collaborative functionality. This enhances the “teaching/learning” aspect as opposed to the “informing” aspect of the website (allowed for owners, collaborators and observers). Collaborative annotations, which are critical for collaborative aspects of Hh WWW Site, can be further categorized by:

  - **Target** denoting the precision or “resolution” by which an annotation can be “attached” or targeted to a particular data item. In general, annotations can be targeted to: the whole data item (collection, full video, book, paper); a data element such as a page, keyframe, gene segment, particular image, or video segment; and finally to data items like words, letters in the gene sequence, image region, video keyframe region, etc.

  - **Type** denoting the type of data contained in the annotation. It can be a combination of: user identification, timestamp, title, keywords, free text, graphical outlines (sequence of X, Y coordinates), time segment in video or gene sequence markers, and recorded voice. Finally, it can contain URLs to any type of resource including image, video, executables (algorithms, alert e-mail), WWW sites, and references to other annotations.

  - **Functionality** denoting the types of operations allowed on annotations. They include searching, browsing, submission, attaching response (other annotation), removal (by administrators). It also refers to the target audience (public, private, or addressed to specific persons). Searching, particularly, should depend on the application, and can include search by annotation author, title/topic/keywords, time, etc.
The above annotations can be made by any authorized user, at any time and in any order. Of course, the whole content including collaborative and linkage annotations can be searched, browsed and examined by members of the Hh community. We also note that annotations resemble the metaphor of e-mail, but with a precisely defined target to which they refer.

3.1 Related work on Annotations

Initially, research work on annotations focused on digital text documents, followed by textual WWW documents, and then by multimedia documents. For examples of reviews of related work we refer the reader to (A. J. Bernheim Brush et al., 2000, and J. J. Cadiz et al., 2000). Collaborative aspects of annotations have also been studied (see for example J. J. Cadiz et al., 2000, N. Bouvin et al., 2002, T. Sannomya et al., 2000, D. Karger et al., 2003). They all point to the value of annotations in general collaborative applications. Recently, annotations for multimedia documents and data have also been studied, mainly to enable better anchoring and synchronization of those annotations, such as in applications when video is used for training or learning (e.g. D. Bergeron et all, 1999). Results of user studies are positive in that they show that much better collaboration can be obtained when annotations can be viewed synchronously with video. The IMarkup (www.imarkup.com) product offers users some ability to markup WWW pages with text, free ink and even voice data. Users may also share the annotations via a dedicated server. SGI IRIS Annotator (www.sgi.com/software/annotator) offers users the ability to attach a variety of annotation types (text, pointers, media, URLs) to visual documents like 3D models, and share them. Recently, we have seen the emergence of successful WWW communities where any user can edit the pages, see for example Wiki Web pages (e.g. http://c2.com/cgi/wiki?WikiWikiWeb). Finally, recent personal productivity products from Microsoft (like MS Word) and others include collaborative functions.

In terms of architecture, annotations can be personal or shareable. Personal ones are usually stored locally and used by the annotation author only. Shareable annotations are usually stored in annotation or collaboration servers indexed by documents or URLs, such that they are displayed with the page the user is viewing. Naturally, a means of managing access must also be provided. Some authors advocate local storage of annotations for better scalability (L. Denoue, L. Vignollet, 2002) but the problem is that in those cases it is hard to provide access to annotations by others.

Of relevance to our project, we can make several observations regarding collaborating annotations:

- None of the related work analyses the problems of annotation in the context of specific biology applications.
- Well-known problems of annotations being attached to dynamically created WWW pages or removed pages are minimized in our context since a great majority of our key pages are static data pages (i.e. research content) that is not being modified or deleted.
- We want annotations that are in general shared with all community users, thus requiring a centralized (in a logical sense) architecture of storing the annotation data.
- We want to minimize the burden on the user in terms of deviating from normal WWW usage metaphors in the UI, as well as in requiring complex downloads and installation of client SW applications related to collaborative annotations.

4. Current Status and Implementation

Our prototype repository is currently located at http://hedgehog.sfsu.edu/. The website information architecture represents the known ontology of the Hedgehog gene pathway, and contains an initial data set representing some of the information available on the Hh pathway. The website is implemented using a standard SQL database (IBM DB2) and the standard apache webserver. We make extensive use of

![Figure 2 Hedgehog website search page.](image)
php to design queries on the database and to coordinate data from our multiple annotation servers.

In Figure 2, we show the search page for our website. Note a visual map of the Hh gene pathway as a visual interface for the search (similar to Stanford’s Wnt WWW site), augmented by text entries on the left. The user can search for data related to a certain object or concept in the ontology by clicking on the map and/or combining this with text queries.

In Figure 3, we see a results page, containing some information related to a query posed by an observer of the website. Note, as well, the informal annotations the observer has made on the website. Our currently envisioned annotation interface has many similarities to email, allowing an observer to choose whether the annotations she submits are personal annotations (available only to her), directed annotations (available to a select group of individuals) or public annotations (available for view by anyone). It is also possible, using the same interface, to control which annotations are visible on the page. Does she want to see only her own annotations, only her annotations and those directed to her, or all annotations available?

Finally, Figure 4 shows the submission page for our website. Recall that all data on the website, whether content, linkage, or collaborative annotation, is conceptually viewed as an annotation. In the case of a content annotation, the target is the ontology described by the website architecture. To support this notion of content annotation, we allow multiple forms of submission. Contributors can submit based on a textual description of site architecture (on the left side of the screen) and/or based on a graphical depiction of the Hh pathway (in the center of the screen) using the same image map as the search-based image map. Clicking on either the textual site architecture or on the image map brings up a submission panel already partially populated with relevant data (shown on the right side of the page). A contributor can select the type of their submission (whole WWW project, raw data, text, images, video data, or a link to another URL are currently supported) and add the data to the repository with the pre-existing metadata/keywords populated from the image map. A contributor may add additional metadata/keywords from a pre-defined list. We are also investigating what controls should be given to contributing users in terms of content presentation and placement.
5. Conclusions

We describe the design and on-going implementation of a multimedia information repository to support the research community studying the Hedgehog Gene Pathway. Challenges involved in this project include:

1. The design of an information repository supporting multiple data types with data sizes ranging from a web link to a data file in excess of 1Gb.

2. The design of a collaborative community site where overall architectural control of the site is maintained by the website owner, but submissions to the site are allowed by members of a broader community, accomplished using content annotations to populate the website.

3. The design of a repository that supports high-level collaboration both between specific individuals and within the community as a whole, accomplished through the integration of informal annotations.

4. The design of a repository where information can be extended over time due to new research. We support this through the use of linkage annotations, which allow for the creation of a two-way hyperlink between two pieces of data on the website, or between data on the website and a new submission. Linkages, themselves, can be annotated with content annotations, permitting a description of how the linkage was observed. These linkages, together with implicit linkages (obtained by keywords and metadata) will have to be searched, browsed and filtered properly to ensure effective usage by the Hh community.

5. Usability and ease of use have to be provided at all levels of design and implementation, including in maintenance, usage, annotation, and the submission process.

In addition to building a useful WWW site for the Hh gene pathway, our on-going research is in two veins. First, we are focusing on the content annotations (submission of main data) and collaborative annotations from a UI and a usability standpoint. Understanding the ease of use and burden annotations place on the contributors will allow us to design more accessible interfaces for the submission of annotations. As well, as we collect a significant data set of these collaborative annotations, we wish to analyze their usefulness as a source of metadata for the content annotations that they modify. Understanding the relative role of content annotations, linkage annotations and collaborative annotations within the Hh community will allow us to tailor search tools more appropriately for our user community.

Second, we are examining system architecture issues associated with collaborative annotations. Our goals are to provide absolutely minimal burden on the client user in terms of additional SW tools needed, as well as to design scalable system with standard components as much as possible. For now, we are focusing on centralized server for linkage and collaborative annotations, since they have to be perused in general by the whole Hh community.

Ultimately, we plan to have the Hh WW Site be available and well supported for the benefit of the whole research and educational community.

Acknowledgments: We thank Professors Murphy and Yoon for valuable suggestions and G. Papilion (SFSU IT) for his help and support. We also thank Jamie Lemus for helping with our usability experiments.

References


Wnt web page: [http://www.stanford.edu/~rnusse/wntwindow.html](http://www.stanford.edu/~rnusse/wntwindow.html)

Hh WWW site: [http://hedgehog.sfsu.edu](http://hedgehog.sfsu.edu)


IMarkup [www.imarkup.com](http://www.imarkup.com)


L. Denoue, L. Vignollet: “Annotations in the wild”, SAAKM workshop at ECAI 2002, Lyon, France