LBL-37808
UC-408
CONF-960158--2

Lawrence Berkeley Laboratory
UNIVERSITY OF CALIFORNIA

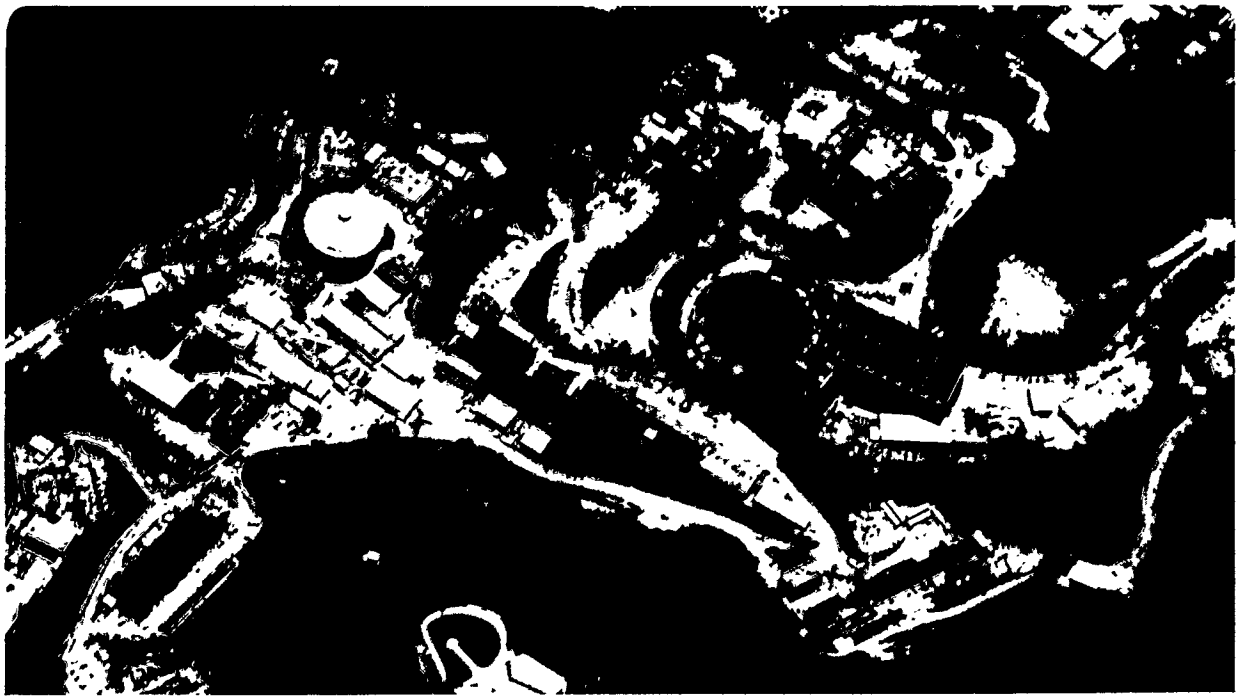
Information and Computing Sciences Division

To be presented at the Pacific Symposium on Bio Computing, Hawaii, HI, January 3–6, 1996, and to be published in the Proceedings

An Electronic Laboratory Notebook Based on the World Wide Web

J.E. Marstaller and M.D. Zorn

October 1995



Prepared for the U.S. Department of Energy under Contract Number DE-AC03-76SF00098

LBL-37808
UC-408

An Electronic Laboratory Notebook Based on the World Wide Web

J.E. Marstaller and M.D. Zorn

Information and Computing Sciences Division
Lawrence Berkeley Laboratory
University of California
Berkeley, California 94720

October 1995

This work was supported by the Director, Office of Energy Research, Office of Health and Environmental Research, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

LBL-37808

AN ELECTRONIC LABORATORY NOTEBOOK BASED ON THE WORLD WIDE WEB

J. E. MARSTALLER, M. D. ZORN

Ernest Orlando Lawrence Berkeley National Laboratory
University of California, Berkeley, CA 94720

The LBNL/UCSF Resource for Molecular Cytogenetics has been created to facilitate the application of molecular cytogenetics in clinical and biological studies. One of the primary tasks is the selection of probes optimized for use in fluorescence in situ hybridization (FISH). Our group provides data management support for all the activities in the Resource. In this paper we describe an electronic laboratory notebook based on the World Wide Web. The data are located in a central database. The user interface consists of a set of HTML forms that handle data input and retrieval from a database from two locations several miles apart. A WWW client allows users to formulate retrieval and edit operations that are sent to the database. Results are filtered through Perl scripts which generate HTML documents with Hypertext links that are sent back to the client. Besides tracking laboratory information through the various stages in the biology laboratory, the system also feeds into a public web server that makes the data available to the community.

1 Introduction

The Resource for Molecular Cytogenetics is a joint project at the Ernest Orlando Lawrence Berkeley National Laboratory and at the University of California, San Francisco. Funded by the U.S. Department of Energy, it has been created to facilitate the application of results in genomics and molecular cytogenetics to clinical applications in, for example, cancer biology. Work is being pursued in three main areas: the development and application of improved hybridization technology, the selection of probes optimized for use in fluorescence in situ hybridization (FISH), and the development of digital imaging microscopy.

One of the goals of the Resource is the development of an evenly spaced set of probes along each chromosome that are mapped cytogenetically and tie in with genetic and physical maps that are being generated by the Human Genome project.

The electronic laboratory notebook described in this paper handles the data management for this probe mapping effort.

The probe mapping effort starts with the selection of an appropriate target locus that is well defined on genetic and physical maps. The selection is largely based on public information found in the Genome Data Base[1,2]. For each target locus a set of primers is defined to select a clone out of a human P1 library and to develop this clone into a FISH probe. The clone is mapped onto metaphase chromosomes to determine the position along the axis in fractional length from the p terminus of the chromosome.

A number of reasons, including the volume of data that needs to be tracked, the physical separation between the two main participating laboratories, and not the least to make the results available to the community as quickly as possible, led to the decision to develop an electronic laboratory notebook. It would also provide reasonable documentation for future commercialization of the probes.

2 Laboratory Notebooks in the Human Genome Project

In the Human Genome project several attempts have been made to provide a laboratory notebook to support the biology efforts. These notebooks should replace manual record-keeping and provide a history of the reagents and protocols used in the laboratory. While most of the genome centers focussed on various mapping efforts, there are relatively few laboratory notebooks that support mapping projects. Mapping is very labor intensive, not very automated, and the methods change frequently. We want to mention two examples of successful development of mapping notebooks. The **Livermore Contig Browser**[3] which features a graphical user interface based on a custom widget set and an underlying Sybase database management system. It is based on Unix workstations and nudges the biologists to learn Sybase SQL to interact fully with the database. Although the details change, the Livermore genome center focussed on one major strategy, cosmid fingerprinting, to provide the backbone of the system, i.e., the protocols and reagents are hardwired into the system. Other biological methodologies were folded into the system during the mainte-

nance phase. Both the user interface and the database are unique pieces that cannot easily be ported to other uses. An example at the other end of the spectrum is the Laboratory Information Management System created for the **MIT/Whitehead genome center**[4]. A Workflow manager uses a simple state/transition model to represent laboratory protocols and connects to LabBase, an object-oriented database. For user interaction it takes advantage of the capabilities of Microsoft Excel spreadsheets running on Macintosh computers. Forms created in Excel interface via electronic mail with the LabBase data manager application. Much of the actual biological processes are captured in Perl scripts that interact with laboratory instruments so that the users are alleviated from the tedious tasks of feeding the computer. The workflow manager offers an interface to Perl[13] that facilitates changes to the protocols and thus provides the flexibility of the lab notebook. Separating the lab protocols from the database schema allows each of them to evolve independently.

Notebooks for sequencing efforts are more abundant and date back to programs developed by **R. Staden**[5,6] in the late seventies which have been continually updated and are still in use in major genome centers around the world. The **LBNL Human Genome Center**[7] generating large amounts of human and fly sequences relies on the Unix file system to store and manage the sequence traces. Scripts in various languages and task-specific software modules in Smalltalk and C move the data from automatic sequencing machines to the final end sequence released to public sequence databases. Although the Unix-based system is fairly easy to implement, it does not adapt well to protocol changes and provides little data access security, no metainformation, e.g., who did what when, and no history. A similar strategy is employed by the **Sanger Centre**[8] in the United Kingdom. Probably the most sophisticated laboratory notebook is under development at the **University of Utah Genome Center**[9]**.** Hand-held Newton devices communicate via a radio controlled network with a Sybase laboratory database. Bar code readers feed important information directly into the computers.

Without spending a lot of money on hardware and software that would be involved in the transfer of any of these solutions by other genome centers, we wanted to propose a compromise solution that would allow our biologists to store their data in an electronic lab notebook that appears to be running on their personal computers, connect the two main laboratories that are a few miles apart on either

side of the San Francisco Bay, and be done with modest development efforts. We recognized the ease of use of the Web browsers[11,12] distributing Hypertext documents over the World Wide Web[10] and decided to use that to implement our lab notebook.

The World Wide Web browsers with HTML-based forms provide a fast and easy mechanism to create forms-based user interfaces. The developer can sit down with the biologist and rapidly make changes in responses to the comments of the user. Furthermore the HTML forms work almost equally well on a number of different hardware platforms. Thus the biologists may continue using their Macintosh computers and find a familiar interface once they do work on a Unix workstation. The web browser can be run from any machine connected to the Internet, thus the users are free to enter and view information even away from their labs at home or while on travel. Access can be restricted by passwords and other means to secure the confidentiality of the data. A bonus that is hard to implement otherwise is the facile connection to outside resources. Linking local information to data in public databases is only a hypertext link away with little or no additional programming effort.

3 World Wide Web Server as an Electronic Laboratory Notebook

The biological protocol involves defining an appropriate target locus and a set of primers that identifies the locus. Then the primer conditions are optimized and the P1 library is screened using PCR reactions. Potential candidate probes are sent to the mapping lab where the cytogenetic position is determined with FISH.

We identified four key reagents in the protocol: target, primer pair, picked probe, and mapped probe. Each element has a one to many relationship with the one following in the protocol, e.g., a target may have many associated primer pairs. The lab protocol and the representations in the HTML forms and the database are shown in Figure 1. Each of these elements is represented in the database. The database has been implemented using the Unix dbm libraries which support simple dictionary-type databases. Dbm stores data as tag - value pairs in a data file and provides an index on the tags in an index or directory file. Each of the four reagents is stored in a

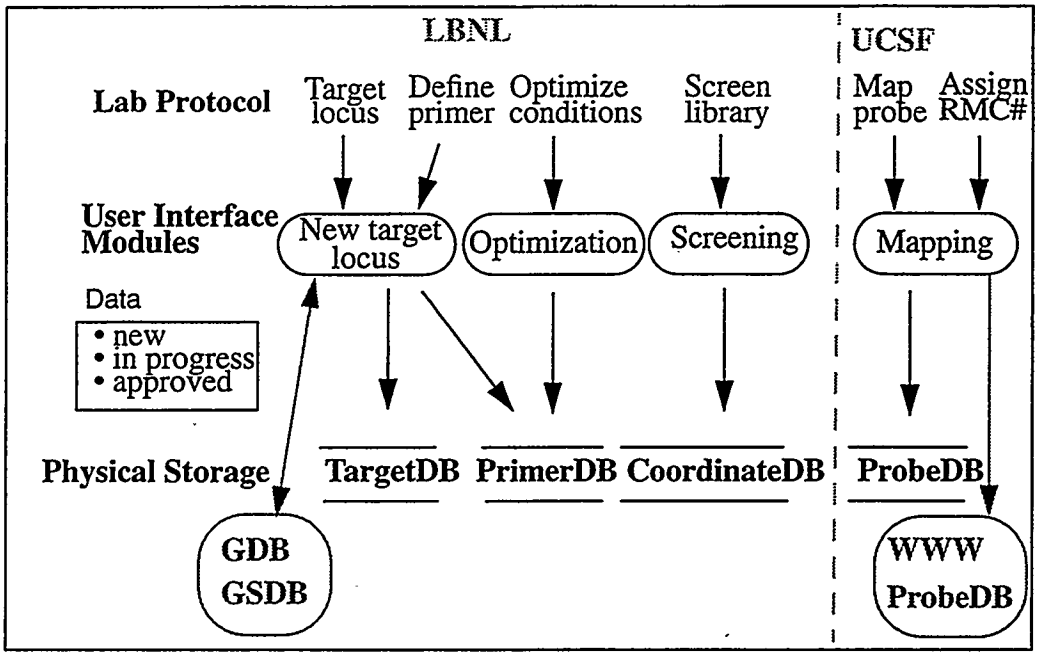


FIGURE 1. Data Flow in the Probe Mapping effort
A target locus is selected based on information in public databases and a pair of primers is defined by their sequences. The PCR conditions for this primer pair are then optimized to yield a good signal on test DNA. Optimized primer pairs are then used to select candidate probes out of a library of clones. Once the probes have been identified they are sent to UCSF to be mapped by fluorescent in situ hybridization onto human metaphase chromosomes. To determine the positions alongside the chromosomal axis a number of experiments are performed. After a candidate probe is proved to be successful it is assigned a unique RMC number. After a certain period the probe is made publicly available by both a public web server at the Resource and by submitting them to the GDB database. The dashed line shows the division of labor between LBNL and UCSF.

separate pair of dbm files. The tag is an automatically generated identifier, all attributes for the reagents are concatenated with a separation character and stored in the value portion. Relationships among the reagents are stored in additional dbm files. The structure resembles an entity relationship model where entitysets and relationshipsets are physically represented by tables or, as in this case, dbm dictionaries. This structure allows us to switch to a relational data management system if the need arises without having to invest prematurely. At present the data volume can easily be handled by the dbm data files.

The graphical user interface is implemented using HTML forms. The forms have been designed in close collaboration and over a series of iterations with the biologists who would handle the respective part of the lab protocol. A web client, NetScape[12], is used to display the forms on their computer, in most cases, a Macin-

tosh. Data entry forms prompt the user to enter all the relevant information. The forms are sent to the Web server where they are processed by Perl[13] scripts and the data stored in the corresponding dbm files. All entries are time-stamped and have administrative information added to each entry. For example, particular stages corresponding to processing steps in the lab protocol are maintained as states within each reagent. Stages are dependent on the information provided by the user and provide some kind of management control over the lab process, i.e., explicit approval is required to move an element to the next protocol step. For example, an optimized primer pair needs authorization to become available for the screening process. In some cases, the processing includes connecting to outside web servers to retrieve relevant background information, e.g., GDB identifiers for targets that have GDB Locus names.

Various browser interfaces show the contents of the database and allow retrieval of stored information. The HTML documents are generated automatically by Perl scripts that access the database and build either document tables or text documents. External identifiers are translated into hypertext links to the respective database. Essentially, the purpose of this part is to show the progress of the mapping process. Data may be viewed based on several criteria. Users may access the data using a range on a chromosome, or search for a specific probe.

Hypertext linked on-line help is available to the user to facilitate the use of the laboratory notebook.

Data access security is maintained through the web server access selection. Users are prompted for a password and have to be in a group that is granted write access to the database and connect from a machine on a trusted network. Read access is less restricted. Although the given security measures are not very strict and can easily be circumvented by a dedicated hacker, they are deemed sufficient to protect against accidental damage of the data. Only scientists involved in the research at each stage have access to that stage via a password.

Figure 2 shows the data entry steps that reflect the laboratory protocol and a sample form in Figure 3.

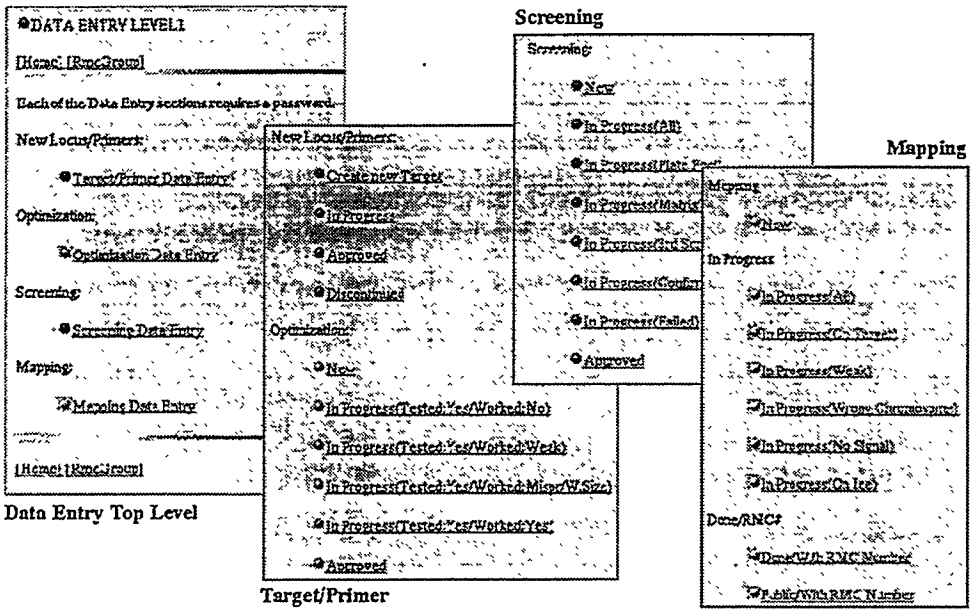


FIGURE 2. Data Entry steps
Shown are the top level browsers and data entry documents. Each row in a browser refers to one locus target and the collected information from each stage of the mapping process. A search engine based on the criteria for each stage retrieves the data. Documents are selected for updating by clicking on the appropriate link.

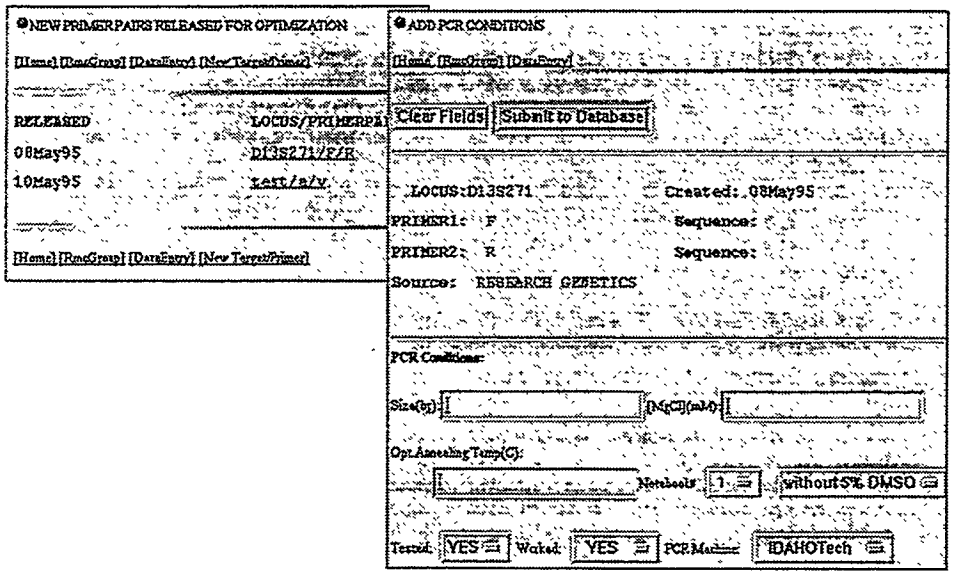


FIGURE 3. Example of a data entry form: Primer optimization data entry form

4 Public World Wide Web Server

In order to distribute information generated at the Resource, we built a public web interface[14] that allows outside researchers to view our data. The data can be accessed through a search engine that returns a list of probes. Each probe is a hypertext link to detailed information about the probe, the target, primer pairs used, etc. (Figure 4). Whenever a pointer to a public database, e.g., GDB, is present in the database, a link into the public database is created. The detailed probe form features a request button that forwards a probe request to the Resource. This request link brings up an e-mail form where the reader can enter name, address, and a short statement about the proposed use of the probe. The mail gets forwarded to the appropriate staff in the Resource which fulfill the order and send out the physical probe clone. The public RMC web server integrates local data with FISH mapping results from collaborators. Request forms are mailed directly to the persons responsible for physically distributing the probes. In addition to probe mapping information, the web server also features sample images for users of Resource imaging software, pages about policies, and procedures employed in the Resource. A summary usage statistic presents a graphical view of the history of access to the server. For security reasons, the public web server runs on a separate workstation on its own data set that is physically distinct from the data used in the laboratory notebook.

5 Conclusion

While it is quite common now to distribute information via the World-Wide Web, we used the technology to create a group-internal communication and information management tool. We developed an electronic laboratory notebook that tracks the progress of the probe mapping effort and its key reagents within the Resource for Molecular Cytogenetics. The lab notebook is implemented as a set of HTML forms that are displayed by a forms-capable web browser such as Mosaic[11] or Netscape[12].

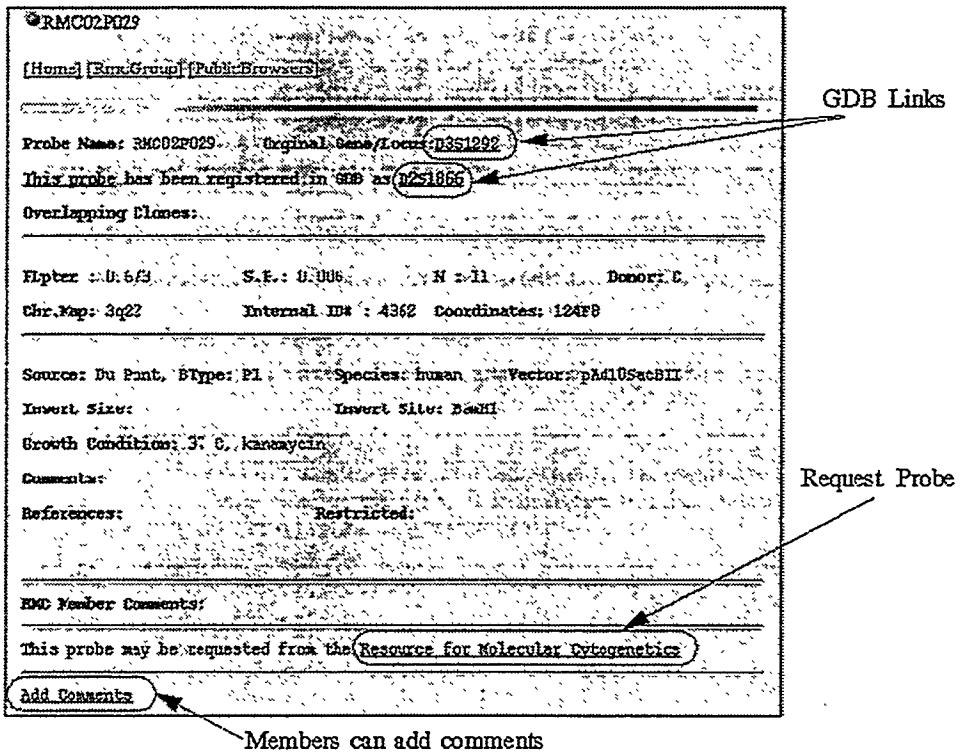


FIGURE 4. Public Display of Probe Information
The display shows links to the GDB database and an email probe request button.

The forms are generated and manipulated by scripts developed in the Perl language. The data are stored in a number of files using the Unix dbm facilities.

In using the web as the basis for the lab notebook we utilize not only the growing familiarity of biologists with this type of user interface, but also the hardware platform independence of the interface. By separating the user interface from the underlying database both ends may evolve more independently from each other. While many of the scripts and forms are peculiar to our implementation, the concepts and the lessons learned can be easily transferred to other lab notebooks with modest data volumes. For larger applications both the dbm data files and the interpreted Perl language may become bottlenecks.

Controlling access to the server provides reasonable prevention to misuse and misappropriation of internal data. Securing the data like banking statements is not warranted and not feasible for daily operations.

A feature our biologists found especially useful was provided by adding direct hypertext links to related information in public databases. Although such links do not invalidate the need for closer integration of biological information, they give users a first step into more of the publicly available data.

With comparably modest effort we were able to develop a laboratory information management system that fulfills the basic tracking and recording needs for the biologists. It also shows that hypertext-linked documents are not only useful for global sharing of information, but equally powerful as groupware within an organization to exchange and manage information.

Acknowledgments

This work was supported by the Director, Office of Energy Research, Office of Health and Environmental Research, UC-408, of the U.S. Department of Energy under Contract No. DE-AC03-76SF00098.

References

1. Fasman KH, Cuticchia AJ, and Kingsbury DT, The GDB™ Human Genome Data Base Anno 1994. *Nucleic Acids Research*, Vol. **22** No. 17: 3462-3469, (1994).
2. Genome Data Base, GDB™. The Human Genome Data Base Project, Johns Hopkins University, Baltimore, MD. World Wide Web <URL: http://gdbwww.gdb.org/gdb/browser/docs/topq.html>, (1995)
3. Livermore Contig Browser <URL: http://www-bio.llnl.gov/bbrp/genome/informatics.html>
4. MIT/Whitehead Genome Center <URL: http://www-genome.wi.mit.edu/genome_software/genome_software_index.html>

5. Staden, R., Computer Handling of DNA sequencing projects, in *Nucleic acid and protein sequence analysis, A practical approach*, 173-217, eds. M. J. Bishop and C. J. Rawlings, IRL press 1987.
6. Dear S; Staden R.: A sequence assembly and editing program for efficient management of large projects. *Nucleic Acids Research*, Jul 25, **19(14)**:3907-11 (1991)
7. LBNL Human Genome Home Page
<URL: http://www-hgc.lbl.gov/GenomeHome.html>
8. Sanger Centre Home Page
<URL: >http://www.sanger.ac.uk/
9. University of Utah Genome Center
<URL: http://www.genetics.utah.edu/genome/index.html>
10. The World Wide Web
<URL: http://www.w3.org/hypertext/WWW/TheProject.html>
11. Mosaic Browser
<URL: http://www.ncsa.uiuc.edu/SDG/Software/Mosaic>
12. Netscape Browser
<URL: http://home.netscape.com/>
13. Wall, L., Schwartz, R.L., Programming Perl, O'Reilly & Associates, Inc. 1991.
14. Resource for Molecular Cytogenetics Public Home Page
<URL: http://rmc-www.lbl.gov/>

LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
TECHNICAL INFORMATION DEPARTMENT
BERKELEY, CALIFORNIA 94720