WebWise: Web Sites of the Human Genome Project

Kim D. Pruitt

Genome Res. 1998 8: 1109-1111
Access the most recent version at doi:10.1101/gr.8.11.1109

WebWise: Web Sites of the Human Genome Project

Kim D. Pruitt[1]

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894 USA

In this last decade of the twentieth century the scientific community has witnessed the progression of a historic project—The Human Genome Project (HGP)—and a corresponding escalation in the rate of DNA sequence data accumulation. Managing, understanding, tracking, and storing this volume of data are not trivial tasks. The public data repositories have expanded their resources to receive, store, and retrieve sequence and map data. Furthermore, recognizing that the full extent of the data (including project descriptions and protocols) should be organized and publicly accessible, the HGP sequencing centers provide their own web sites. These sites allow for rapid dissemination of a wealth of data and information, including map and sequence data, protocols, software tools, and overviews of goals and progress. As such, it is useful to have an understanding of the general organization and content of these web sites.

The HGP Web Sites

Over the past year, this WebWise series has reviewed a dozen web sites hosted by the larger HGP sequencing centers. These sites and the URLs to their Home pages are listed, in order of review date, in Table 1. Each review provided an in-

Table 1. HGP Sequencing Center Web Sites Included in the WebWise Series

Sequencing centers		URL	Genome Research
GSC	Washington U. Genome Sequencing Center	http://genome.wustl.edu/gsc/	7: 1118–1121
SC	Sanger Center	http://www.sanger.ac.uk/	8: 4–8
SHGC	Stanford Human Genome Center	http://www.shgc.stanford.edu/	8: 86–90
W/MIT	Whitehead/MIT Genome Center	http://www.genome.wi.mit.edu/	8: 86–90
BCM	Baylor College of Medicine Human Genome Sequencing Center	http://gc.bcm.tmc.edu:8088/	8: 170–174
JENA	Institute of Molecular Biology—Jena	http://genome.imb-jena.de/	8: 334–338
UTSW	U. of Texas Southwest Genome Science and Technology Center (McDermott Center)	http://gestec.swmed.edu/	8: 422–426
UWGC	U. of Washington Genome Center	http://www.genome.washington.edu/UWGC/	8: 572–575
CGM	Washington U. Center for Genetics in Medicine	http://www.ibc.wustl.edu/cgm/	8: 686–689
ACGT	U. of Oklahoma Advanced Center for Genome Technology	http://www.genome.ou.edu/	8: 763–767
JGI	Joint Genome Institute	http://www.jgi.doe.gov/	8: 864–869
TIGR	The Institute for Genome Research	http://www.tigr.gov/	8: 1000–1004

The Genome Research column includes the volume and page number for each sequencing center WebWise review.

[1]Corresponding author.
E-MAIL pruitt@ncbi.nlm.nih.gov; FAX (301) 435-2433.

dication of the contents and main features, as well as an organizational diagram, of the center's web site. As web sites are often subject to revision, these sites were revisited for this final review to determine whether any large changes have been made since the original review.

Naturally, there has been an increase in the quantity of data available, and several centers have expanded the overall long-term goals for their center. Table 2 provides an indication of the centers that are currently engaged in sequencing each chromosome. Please note that the table only includes the reviewed groups, but there are additional HGP sequencing centers. Only three centers have made alterations that impact the general organization of the web site (web sites were not scrutinized at length to identify small changes). The Washington University's Genome Sequencing Center (GSC), the Joint Genome Institute (JGI), and the Stanford Human Genome Center (SHGC) have reorganized some part of their web sites. The GSC and JGI have redesigned their Home pages, resulting in less clutter; some links to closely related pages have been condensed. These new Home pages are an improvement; they are more streamlined and, more importantly, data access is more straightforward.

Beyond the Home page changes, the GSC reorganization primarily impacts some of the more general information pages where links have been removed from the Home page but are still available via navigation links on internal pages. Search services are also pulled together into a single point of access, whereas previously they were accessed from two different pages linked to the Home page. The sequence data pages appear to be basically the same in terms of organization and layout.

The JGI web site underwent a major revision at approximately the same time as publication of the JGI WebWise review. Both the general design style and basic organization of the web site have changed; the end result is that the web site is better organized and data are more accessible. The redesigned web site removes pages offering limited information and some intermediate pages, and offers an improved organization of the sequence and map data into one main section. In addition, links to pages that are not ready yet have been removed.

Table 2. Chromosome-Based View of Sequencing Activity

Chr.	Sequencing center
1	ACGT, SC[a]
2	GSC[a]
3[b]	BCM,[a] CGM, GSC, SC, W/MIT
4	JGI, SC, SHGC[a]
5	GSC, JGI[a]
6[b]	SC,[a] W/MIT, UWGC
7	GSC,[a] JENA, UWGC[a]
8[b]	GSC, JENA[a]
9[b]	ACGT, SC,[a] W/MIT
10[b]	SC[a]
11	ACGT, JENA, SC, UTSW[a]
12	GSC/SC, BCM[a]
13	GSC, SC,[a] W/MIT
14[b]	GSC
15[b]	UTSW[a]
16[b]	GSC, JGI,[a] TIGR[a]
17	W/MIT[a]
18	W/MIT[a]
19	JGI[a]
20	JGI, SC[a]
21[b]	JENA
22[b]	GSC, SC,[a] W/MIT, ACGT[a]
X[b]	BCM,[a] CGM,[a] GSC, JENA, SC,[a] W/MIT
Y	GSC,[a] W/MIT[a]

Data for this table were derived from the sequencing center web sites and from the Human Genome Sequence Index (HGSI) and only include web sites reviewed in the WebWise series. (See Table 1 for definitions of center abbreviations.)

[a] Centers that consider a given chromosome to be one of their major sequencing efforts as indicated either on the center's web site or on the HGSI web site.

[b] Those chromosomes for which sequence data are also generated at additional sequencing centers not reviewed in the WebWise series. These centers include, but are not limited to, the Japan Science and Technology Corporation (JST), Genoscope, the Chromosome 21 Consortium, the University of Washington Multimegabase Sequencing Group, and Genome Therapeutics Corporation. For further information on these centers, please see the HGSI or Sanger web sites (http://www.ncbi.nlm.nih.gov/HUGO/; http://webace.sanger.ac.uk/HGP/).

The underlying data pages are basically the same as before; however, these pages are also slated to be redesigned as part of the effort to integrate the Department of Energy (DOE) sequencing effort into a single point of access (R. Sutherland, pers. comm.).

The SHGC web site has also undergone some revision; here the Home page is unchanged but the sequence data pages have been reorganized and some chromosome 4 map data are now available. Data are accessed through map data, unfinished sequence data, and finished sequence data links. Links are also provided to explanatory text concerning goals and methods.

Contents and Features—What's Important?

The primary reason why the sequencing center web sites serve such an essential role is that sequence data are most meaningful when placed into the context of location and/or function. Data context information may include map markers, gene names, cytogenetic band, chromosome identity, or identification of several neighboring clones with overlapping sequences (contigs). The sequencing center web sites provide the context necessary to formulate some understanding of the raw data. At a minimum, data are oriented in terms of general chromosomal location, and many of the sites provide a graphic map display, from which one can come to understand the bigger picture, including the relationship between clones. This type of information is currently only available at these web sites; sequence data submitted to NCBI, for instance, do not generally include the larger context.

Clearly the most important key to a web site's success is good organization. A well-designed web site groups related information together, minimizes redundant links and pages with little useful content, uses meaningful link nomenclature, and provides consistent navigation links throughout the site. Of course, there is a balance to be had—it is possible to organize excessively or provide too many navigation links which, in turn, can lead to some confusion and hinder navigation of the web site.

Data access should be straightforward, the data should be clearly labeled in a consistent manner, and links to a public database should be consistently supplied for all submitted sequence data. Consistent nomenclature use is also critical, as clone names are used frequently to label both map and sequence data. If the naming format changes even slightly between these two pages, it is no longer possible to determine unambigu-

ously which clone on a map graphic corresponds to a particular clone listed in a sequence data table or on an FTP site. Unambiguous clone names support maintaining the cross-reference between the sequencing center's data and that deposited in a public database. For the most part the sequencing center web sites are providing links to public data records and consistent clone nomenclature; however, there are a few inconsistencies and elimination of these will enhance the overall data utility.

BLAST Revisited

Although the context provided at the web sites is quite useful if you are browsing the data, it is also useful to have a mechanism whereby you can place your own sequence of interest into the larger context of the HGP. Therefore, it is important to be able to identify sequence identities and from that to identify flanking clones or contigs as well as information such as general chromosomal location. To do this, you must be able to carry out a BLAST analysis against both finished and unfinished human sequence data. The question remains whether this service should be available at the sequencing center web sites themselves. All HGP sequence data are released rapidly into the public domain so one can reasonably argue that providing separate BLAST servers is simply not necessary. Yet a handful of the centers do provide their own BLAST server even though this service is costly to develop and maintain in terms of both personnel and computer hardware. This underscores the importance of placing the data into context. It is essential that people have the ability to find sequence identities, but it is also useful to place those results into the larger context of chromosomal, map and/or clone location. The public databases do not currently provide all of this added information, and so it becomes necessary to visit the sequencing center web sites to obtain the larger context.

The sequencing centers supplying a BLAST service on their web sites are in effect providing a tighter linkage between the two resources needed to first determine a sequence identity and subsequently place it into some context. In my opinion, this illustrates the underlying problem—that the public database repositories do not provide this type of close link back to the sequencing center web sites. For example, if you carry out a BLAST query at the public databases you usually obtain a result quickly. But if you are interested in identifying other sequenced clones that map to the same general location, you must find the sequencing center's web site first and then determine the relevant information. As long as data generated at the sequencing centers continues to be submitted to GenBank or other public database repositories rapidly, one can formulate a logical argument against maintaining the individual center BLAST servers. However, in this case, it would be useful if the public databases could provide easily accessed links to the genome center's web sites. If following the data trail from the public database to the sequencing center web site represents a common protocol, then it follows that data retrieval at these web sites can be facilitated by use of search tools (e.g., to search by clone name or accession number). Such search tools greatly expedite navigation from the public database, where one has just identified a BLAST hit to an unfinished sequence record, directly to the sequencing center data about that clone.

Planning for the Future

The HGP is at an exciting turning point. Project planners recently set 2003 as the new target date for HGP completion—a full 2 years before originally scheduled. To achieve this goal the sequencing centers will shift their strategy from the current map-driven approach and increase the rate of sequence data production. The immediate benefit to the research community is that human genome sequence data will become available in the public domain at a more rapid pace; indeed, the sequencing community predicts that a working rough draft covering ~90% of the genome will be available as early as 2001.

The anticipated onslaught of sequence data will undoubtedly require additional tools to help us retrieve and make sense of these data. For instance, the strategy shift will impact on our ability to place the data into some context. Currently, the map data drive the sequencing effort; by the time sequence data are available, a lot of context information is already available. In contrast, most of the context for future HGP sequence data will likely have to be elucidated after the data are generated. Consequently, we may see some new approaches to organize, coordinate, and interpret the accumulating data in the future. Certainly we will see a tremendous growth in available data at the sequencing center web sites as the HGP pushes forward into the twenty-first century.

ACKNOWLEDGMENTS

I thank Mark Boguski for his support and editorial advice through the course of this series.