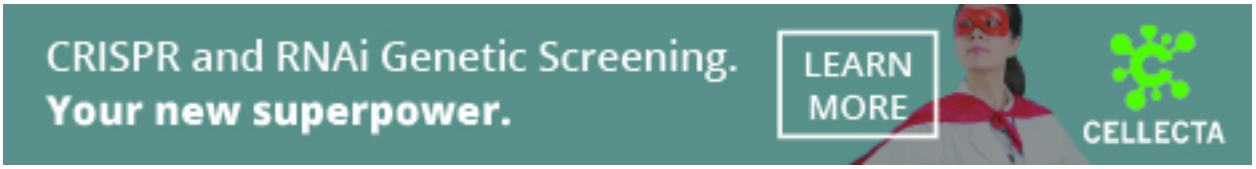
WebWise: Guide to the Institute of Molecular Biology Genome Sequencing Center—Jena Web Site

Kim D. Pruitt

Genome Res. 1998 8: 334-338
Access the most recent version at doi:10.1101/gr.8.4.334



WebWise: Guide to the Institute of Molecular Biology Genome Sequencing Center—Jena Web Site

Kim D. Pruitt[1]

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

This, the sixth installment of the WebWise series, reviews the Institute of Molecular Biology (IMB) Genome Sequencing Center Jena (JENA) web site. As with all of the WebWise reviews, you may find it useful to point your browser window to the JENA Home page while reading. When comparing this web site to the site map depicted in Figure 1, you may notice that some duplicate and minor links have been omitted. The main features of this web site, as well as the features of those sites already reviewed, are indicated in Table 1.

General Information

The JENA Home page (http://genome.imb-jena.de/) is aesthetically pleasant and well organized. Navigation links are provided in the left column of the Home page and on all of the General Information pages. The navigation links do not include a link to the sequence data (this is provided in text form on the Home page), and the sequence data pages do not include these general navigation links. To explore both the data and general information pages of this web site, one must return to the Home page to navigate between these different topics. Fortunately, most of the pages do include a link back to the Home page (in the form of a button or circular DNA icon).

The JENA web site includes an assortment of general information of use to the research community. The About Us page relates information about the JENA sequencing center, including a list of research interests (the List of Projects link). Contact information pertaining to the web site and the JENA bioinformatics program is included at the bottom of this page. A timely, thorough answer was received when questions were submitted to this contact point. Additional information about other research groups at the IMB Jena is available on the IMB Groups page.

An extensive collection of links can be accessed via the side column navigation links. For example, the Learning, Spotlight, and Biolinks pages include many useful links to bioinformatics, molecular biology, bibliography, database, and analysis resources. The links on these pages are organized into general categories and the pages are easy to browse. Although it is somewhat inconvenient to go to three separate pages to review all of the available links, a collection of this size does need added organization as effected by splitting into different web pages.

Click on the Learning button to access links to general informational resources concerning bioinformatics and molecular biology, including guides, tutorials, and manuals. The Spotlight page includes meeting reports, announcements of significant scientific results, and a few links to additional human genome-related web sites (e.g., links to genome sequencing statistics and NCBI's Gene Map of the Human Genome). The Biolinks page (http://genome.imb-jena.de/uselinks.html) presents links related to web-accessible sequence analysis tools, databases, journals and bibliographic collections, and general resources. A large list of sequence analysis resources is available by following the Sequence Analysis Form Sites and Sequence Analysis Workbenches links located at the top of this page. The Specials page includes several links (in German) targeted toward the local Jena community, a link to a job search engine, and links to computer and programming web resources (e.g., Perl Manual, HTML tutorial, Java documents). This section also includes a link to BioToolKit, a search engine for sequence analysis and other genome-related resources. Although the collection of links presented on these pages is not a complete representation of the biological and computational resources available on the web (indeed, that would be a challenging undertaking), there are many links to valuable resources included on these pages.

Data

Follow the Human link from the Home page to access the human sequence data (http://genome.imb-jena.de/printHum.html). This page includes several links to chromosome-specific data, to a BLAST server (see the Tools section), and a graphic overview of the human sequencing targets. The graphic overview is not very specific but does indicate the chromosome and general target areas. An indication of the size of the overall sequencing effort is available on the Home page, and the sizes of individual targets are indicated at the bottom of each target-specific data page. JENA has targeted the sequencing of ~11 Mb of chromosomes 7, 8, 11, 21, and X.

The various types of information relevant to the sequence data are organized into tables that can be reached by following one of the data links on the Human Sequence page (e.g., Xp11). These tables (the Target Status Tables in Fig. 1) include information on maps, project status reports, analysis results, number of contigs, vector, sequence length, status, availability in public database, and contact information. Links to associated

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

8:334–338 ©1998 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/98 $5.00; www.genome.org

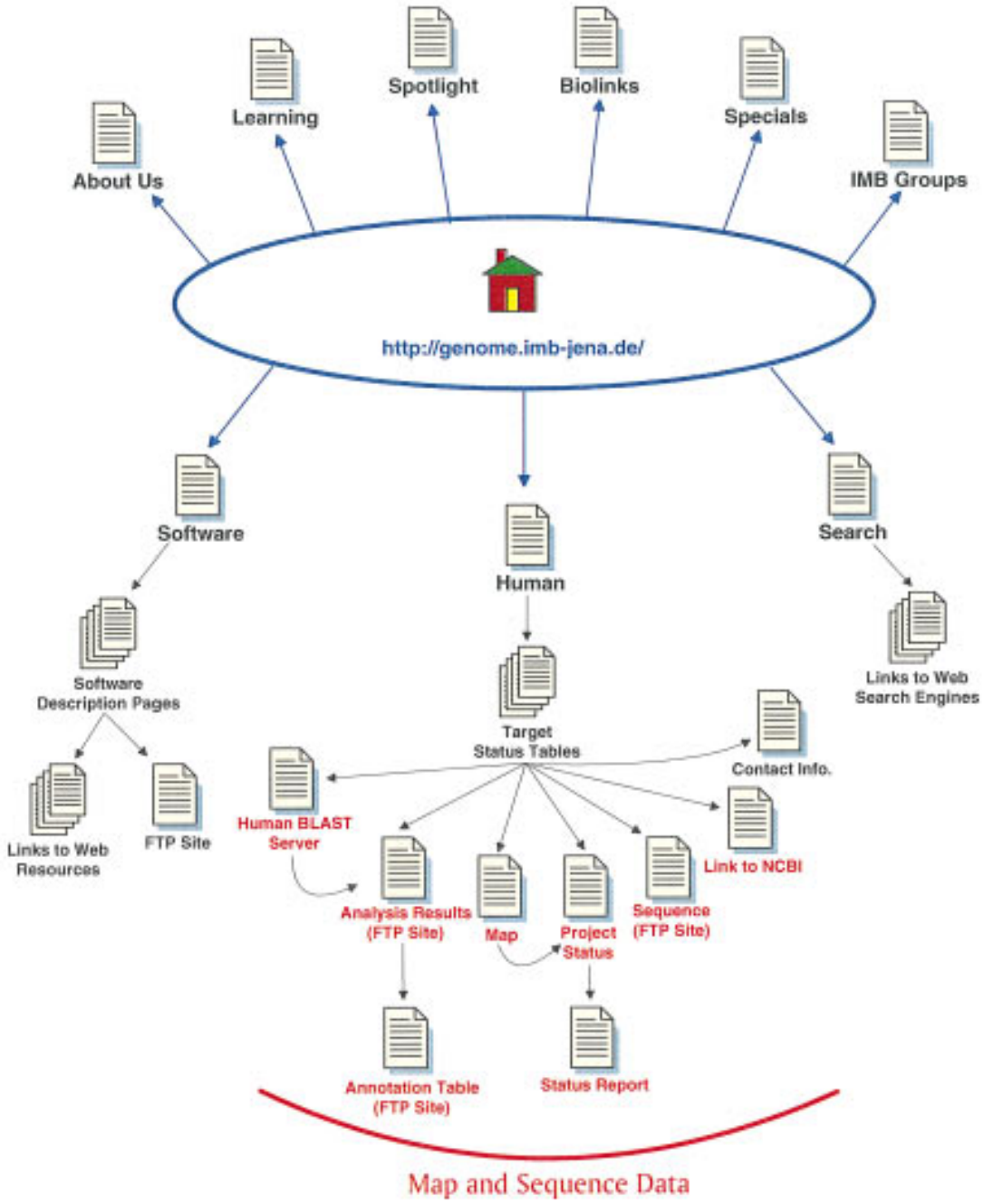


Figure 1 The IMB Genome Sequencing Center Jena Site Map. The main links to pages discussed in the text are illustrated here. Links at the *top* of the Home Page are to general informational resources; links located at the *bottom* are to the data and additional tools. Some links available on the web site are not indicated here.

data (when available) are provided in most of the columns. The Target Status pages present a table of data for the clones associated with a given target, and as contigs are formed a new table indicating the joined project(s) is included at the top of each page. One can access the map data (discussed below), sequence data, and analysis results (see below) from these tables. The DNA sequence data are maintained on the JENA FTP site and are available in two formats: as a FASTA file, and, once the sequence is submitted to GenBank, as a file reflecting the GenBank format style. In addition, links directly to a GenBank Genomes division graphical representation are available for some sequences (e.g., for submitted sequences). Unfortunately, at the time of this writing, ''anonymous'' access to the FTP site is not supported; in the course of reviewing this site, access to the FTP site was inconsistent and depended on the computer used. This likely reflects whether the accessing computer is in the Domain Name Server (DNS); this is an unfortunate policy that will severely limit the dissemination of this data.

Additional information about the

progression of a given project (either individual clones or contigs) is available by following the linked text in the Project Status column. The Project Status pages present a table of data pertaining to an individual clone or contig of interest. The Project Status table is quite similar to, but not identical to, the Target Status table. The Project Status tables include links to analysis results, to sequence files on the JENA FTP site, and, once submitted, to the sequence record in GenBank. The table column titled Sequence as GenBank (Vector) represents an effort to condense the table; the text in this column indicates the type of cloning vector and it is linked to the DNA sequence (on the FTP site) in GenBank file format. Below this table is a list of project checkpoints (e.g., Joined, Sequence finished, Database entry released) with the date at which each checkpoint was reached indicated. A quick survey of these status reports indicates that this is a good example of a sequencing center for which the data available on the web site are more up to date than that available in the public databases. For example, the status report for the Xp11–Join project indicates that the sequence was finished in July 1997; however, it is not yet available in the GenBank database.

Although the Target and Project Status pages provide many of the same links, the map data are only available from the Target Status pages (follow the linked text in the Map column). Maps are provided for each target, or in the case of the X chromosome, the Xq28 target map is broken down into smaller more digestible pieces. A single map style depicting the clone name, order, and status is used, and significant markers are indicated above the map. The maps indicate both single clones and joined projects; joined projects are represented by a string of + signs within the color-coded rectangle. Each map includes a legend key, so it is a straightforward matter to interpret the status of each clone or joined project. The clones and joined projects illustrated on each map are hot linked to the relevant Project Status page, from which you can access the DNA sequence on the FTP site in FASTA format. These maps successfully depict the critical information needed to integrate map and sequence data. An additional feature is the inclusion of the overall length of the depicted map, as well as an indication of the last update. The maps are kept very current as they are revised automatically following an update to the underlying database (B. Drescher, pers. comm.).

The JENA sequencing center imparts added value to its sequencing data by making available the results obtained from a series of computational analysis. DNA sequence data generated at JENA are automatically checked for features such as potential ORFs and homology to sequence data in the public domain. Follow the Analysis Results links from the Target or Project Status table to reach a summary of the analysis results. This table includes links to the detailed analysis data (for each contig associated with a given project) and indicates the sequence length, GC content, and any putative genes identified. Links are provided from the Putative Genes Found

Table 1. Features of the JENA Web Site

Center			GSC	SC	SHGC	W/MIT	BCM	JENA
Map Data	Static Map		●				●	
	Image Mapped		●	●	●			●
	Tabular List				●	●	●	
	Clones Linked to Sequence			●	●	●	●	
Sequence Data	Downloaded data from FTP Site		●	●	●	●	●	●
	Download a Database			●				
	Links to Public Databases			●		●	●	●
	Update Frequency	Daily	●	●	●	●	●	●
		Weekly						
		Unknown						
	Sequence Annotation	Graphic						●
		Text	●					●
		Not Available	●	●	●	●	●	
Search Services	Similarity Searches	Performance a	○	○	○	○	○	○
		b	○	○	○	○	○	○
		c	○	○	○	○	○	●
		d	●	●	○	○	○	○
		f	○	○	○	○	○	○
		Quality of Output a	○	○	○	○	○	○
		b	●	○	○	○	○	○
		c	○	○	○	○	○	○
		d	○	●	○	○	○	●
		f	○	○	○	○	○	○
		Not Available			●	●	●	
	Search the Maps			●				
	Search for Sequences		●	●			●	
	Search the Web Site		●		●	●	●	
Software	Documentation	a	○	●	○	○	○	○
		b	●	○	○	○	○	●
		c	○	○	○	●	○	○
		d	○	○	○	○	●	○
		f	○	○	○	○	○	○
	Available from:	FTP Site a	○	○	○	○	○	○
		b	○	●	○	●	○	○
		c	○	○	○	○	○	○
		d	●	○	○	○	○	●
		f	○	○	○	○	○	○
		Web Page Link a	○	○	○	○	○	○
		b	○	●	○	○	○	○
		c	○	○	○	○	○	●
		d	●	○	○	○	○	○
		f	○	○	○	○	○	○
		Contact the Site				●	●	

The red circles indicate features that are available at this web site or the quality of a given feature within a general range of ''better'' (*a*) to ''worse'' (*f*). Sequence data are assessed for their availability from an FTP site, availability in a database (such as ACEDB), whether archived sequences are linked directly to the public database records, the frequency of FTP site update, and whether any sequence annotation is provided in either a text or graphic format. Each web site is scored for the availability of various search services, including the ability to carry out similarity searches against the sequences in their database or perform a key word search of the map data, sequence data, or web site. Documentation and availability of software tools developed by the center are also indicated. (GSC) Washington University's Genome Sequencing Center; (SC) The Sanger Centre; (SHGC) Stanford Human Genome Center; (W/MIT) Whitehead/MIT Genome Center; (BCM) Baylor College of Medicine Human Genome Center; (JENA) The Institute of Molecular Biology Genome Sequencing Center Jena.

column to the GenBank record at NCBI. However, the link uses an unusual format that evokes a record display lacking the NCBI Entrez neighbors links (Schuler et al. 1996). The analysis summary and detailed analysis data also reside on the JENA FTP site, so access may be limited (see discussion above).

The Detailed Analysis page presents an enormous quantity of data; consequently, these pages are quite large and slow to download. Sequence data are subject to multiple analysis for ORFs, promoters, repeats, and homologies and all of the results for a given contig are made available on a single page. The date at which the analysis was carried out is indicated at the top of the page. These pages are not updated automatically; thus, the analysis might not reflect the most current sequence data. Furthermore, although the homology results certainly provide useful information, additional homologies might be identified if the search were repeated more frequently.

Analysis results are presented in both graphical and tabular format. A separate Results table is provided for each type of analysis performed, and links are provided at the top of the page to each Results table (located in the same page). These links do facilitate jumping directly to the particular analysis results that one might be interested in; however, this is an insufficient navigation tool given the large size of the page. The data provided on these pages would be more accessible if they were presented on more than one regular web page (e.g., rather than an FTP site page). The analysis results could quite easily be separated into general categories, such as ''Repeat regions'' or ''BLAST results,'' which could be accessed by linking to a separate page. In addition, including a set of consistent navigation links on these pages would facilitate use of the data. As a minimum, the analysis data should be organized into general categories and links should be provided to the top of each category between the large tables.

Tools

This web site supports homology searches of the JENA sequence data and also provides documentation concerning several sequencing-related software tools. Unfortunately, a tool is not provided to search the JENA site for either web pages, maps, or sequence data. One might assume that such a tool could be accessed by following the Search link from the Home page, but alas this Search page merely presents a collection of other search tools available on the web rather than a JENA web site-specific search tool. The extensive collection of links available through the general information pages also points to many sequence analysis resources. What is lacking is a general overview of the process of sequence generation, tracking, analysis methods, and tools used at JENA. For example, several software tools are used to generate the Analysis Results pages, but no documentation is provided describing the tools used (e.g., their action, source, and availability).

The JENA sequence data are available for homology comparisons via a BLAST server. To reach the JENA BLAST server (http://genome.imb-jena.de/cgi-bin/GDEWWW/blastserver.cgi), follow the Human Blast Server link toward the top of the Human Sequencing page. Because this web site includes sequence data that are not yet available in the public databases, the JENA BLAST server is an important mechanism to identify sequence homology. To submit a sequence, simply paste your FASTA formatted sequence into the box provided and click on the Select Program button. This refreshes the screen and adds a pull-down menu from which one can elect to search against either the entire JENA human sequence database or against individual chromosomes. Click on the Run BLAST button to begin the search. Search results are returned to the browser window in an expeditious manner as a tabular summary with links provided to the Analysis Results page (on the FTP site). This result provides an indication that a sequence of interest is homologous to a particular clone or contig, and by examining the analysis results for that contig one can extrapolate additional information such as possible ORFs or homologies. However, it is unfortunate that the BLAST result does not include an alignment, as that imparts considerable information about the homology significance. If one's sequence of interest is homologous to the central region of a large contig then it can be quite challenging to extrapolate any meaningful information from the Analysis Results page.

In addition to the BLAST server, the JENA web site does present some software documentation (via the Software button on the Home page; http://genome.imb-jena.de/software.html). Links to extensive documentation (much of it on the JENA web site) in the form of general descriptions, manuals, and/or tutorials are available for five software packages: GCG (the Genetics Computer Group), GDE (Genetic Data Environment), Prophet, the Staden Package, and Bass. The first three packages deal with sequence analysis, and evidently the Analysis Results pages are generated using a customized version of the GDE package (the JENA version of this package is available by anonymous FTP—ftp://genome.imb-jena.de/pub/software/GDE/). The latter two packages are related to sequence data tracking and interpretation (e.g., lane tracking, basecalling, etc.). Some indication of availability is included in the documentation of most of these packages.

Conclusions

The JENA web site provides a range of useful information, including links to additional resources, maps, sequence data, precomputed analysis, and software tools. The JENA BLAST server is an important feature of this site that is easy to utilize and provides a rapid result. The utility of these results could be enhanced by including alignments and by restructuring the Analysis Results pages to make them more accessible and more navigable. The current configuration of the FTP site does limit access and should be changed. Although the general organization of this site is robust, with the exception of the Analysis Results pages, the aesthetic quality of this site is somewhat uneven and the sequence data pages could be enhanced if more navigation links were added. The Home page and top-level General Information pages are easy on the eye, but some of the more internal pages resort to the first-generation web site design strategy of gray backgrounds and heavy table borders (Siegel 1997). A better strategy is to minimize the noise surrounding the data—this enhances the process of interpreting the data as well as the aesthetics (Tufte 1997).

In addition to convenient map displays and sequence data, the JENA web site presents a quantity of preliminary analysis results. Generating these data

does require some committed resources, and JENA places value on providing annotated data to the community at an early stage. Andre Rosenthal stressed the importance of annotation and strongly urged the sequencing community to annotate preliminary sequence data in his talk at the Ninth International Genome Sequencing and Analysis Conference (1997). In contrast, a recent commentary by Wheelan and Boguski (1998) argues for a minimalist approach to annotation of high-throughput sequence data submitted to the public databases. These investigators make the valid points that gene prediction software is still not sufficiently accurate and that predictions based on homology searches can rapidly become obsolete as new data become available. Clearly though, annotation can provide a perspective on the data that is quite valuable for a number of reasons. For instance, preliminary annotation, such as identification of potential ORFs, is a necessary part of ''gene hunting.'' The World Wide Web is an excellent forum for providing automatically computed, and frequently updated, annotation of Human Genome Project sequence data. However, this type of information will be most useful only if careful attention is given to accessibility, organization, and presentation style. Although the precomputed analysis results available at the JENA web site are a step in this direction, the organization and presentation style adopted here is not optimal.

REFERENCES

Schuler, G.D., J.A. Epstein, H. Ohkawa, and J.A. Kans. 1996. Entrez: Molecular biology database and retrieval system. *Methods Enzymol.* 226: 141–162.

Siegel, D.S. 1997. *Creating killer web sites*, 2nd ed. Hayden Books, Indianapolis, IN.

Tufte, E. 1997. *Visual explanations—Images and quantities, evidence and narrative.* Graphics Press. Cheshire, CT.

Wheelan, S.J. and M.S. Boguski. 1998. Late night thoughts on the sequence annotation problem. *Genome Res.* 8: 168–169.

Next Month: The McDermott Center for Human Growth and Development at the University of Texas Southwestern Medical Center