

# **USER INTERFACE EVALUATION OF ONLINE MOLECULAR BIOLOGY DATABASES**

by

**Rock E. Turk**

A Capstone Project

Presented to the Department of Medical Informatics & Clinical Epidemiology

and the Oregon Health Sciences University

School of Medicine

in partial fulfillment of

the requirements for the degree of

Master of Biomedical Informatics

May 2004



School of Medicine  
Oregon Health & Science University

Master of Biomedical Informatics

Certificate of Approval

This is to certify that the Capstone Project of

Rock E. Turk

“USER INTERFACE EVALUATION OF ONLINE MOLECULAR BIOLOGY  
DATABASES”

Has been approved

---

Professor in charge of capstone project

---

Date



## Table of Contents

Acknowledgments.....	v
Abstract .....	vii
1. Introduction.....	1
2. NIH Databases .....	3
a. GenBank .....	3
b. UniGene .....	6
c. Blast .....	9
d. CGAP.....	12
e. Pubmed .....	13
f. OMIM .....	14
3. ENSEMBL.....	17
4. KEGG .....	19
5. EMBL .....	20
6. Analysis.....	23
a. Cross web interface analysis.....	24
b. Critical Journal Review.....	25
7. Conclusion .....	27
References.....	29
Appendix A.....	33
Interviewees .....	33
Appendix B.....	35
NCBI main web page/GenBank search engine.....	35

---

GenBank primary results page.....	36
GenBank information page.....	37
Appendix C.....	39
UniGene main web site.....	39
UniGene primary results page.....	40
UniGene information page.....	41
Appendix D.....	43
Blast main page.....	43
Blast queue page.....	44
Blast results page.....	45
Appendix E.....	47
CGAP main page.....	47
Appendix F.....	49
Pubmed main page.....	49
Pubmed results page.....	50
Appendix G.....	51
OMIM results page.....	51
Appendix H.....	53
Ensembl main page.....	53
Ensemble tool page.....	54
Appendix I.....	55
KEGG main page.....	55
Appendix J.....	57

EMBL main page.....	57
EMBL tools page.....	58

---





## **Acknowledgments**

I would like to thank my project advisor Christopher Dubay and our program administrator Andrea Ilg for their support and help. I would like to also thank the departmental staff and faculty. And finally, I would like to thank my sister Elif Eroglu, my parents Ayhan and Yucel Eroglu for their support over the years.



## **Abstract**

This capstone project will evaluate and analyze user interfaces for online molecular biology databases and corresponding journal reference databases. This project will assess the available features and the usability of the web interfaces to NIH (National Institutes of Health) sponsored molecular biology databases, and journal reference databases. The project will also assess the web interfaces to Ensembl, KEGG and EMBL databases. Evaluation will be done by interviewing scientists on their assessment of the above said databases. And furthermore, evaluation of the web interfaces will be supported by journal article references from studies and reviews conducted by the scientific community. Finally, this project will conclude with a cross web interface analysis to assess differences and similarities of the web interfaces and, a critical journal review to evaluate the substance of the included journal article references.



## 1. Introduction

Explosion of data is one of the direct results of leaps of advances made in the information age. Mass amounts of data are available with ease of access in various fields. Especially, in areas such as biomedical research, large projects like the various genome research projects have led to large scale data productions. Ability to access this data and analyze it properly will determine our capacity for expansion of our knowledge. Access being the keyword, requires us to develop proper tools for mass data analysis. User interfaces come into foreground when proper access to masses of data needs to be granted. Deployment of user interfaces is a conundrum in itself because the design of any user interface is a very subjective task. What might be a successful user interface to some might be an unusable interface to others. As such, in the age of expanding data and technology, successful deployment of user interfaces has become of crucial importance.

In this capstone project, I will evaluate and analyze user interfaces for online molecular biology databases and corresponding journal reference databases. I have chosen this project topic because of the tremendous expansion of data in the molecular biological sciences. Especially over the last few decades, our knowledge of molecular biological sciences has expanded exponentially. The advent of various technological advances such as micro-array technology development and the accomplishment of major projects such as the decoding of the human genome have led to the accumulation of tremendous amounts of data. Resultant data needs to be accessed and analyzed with the

use of tools to further advance science and technology. This is where online user interfaces step in. A user interface is what makes it easy or hard to access biological data in online databases. Therefore in this capstone project, I will assess the value of 9 web interfaces to molecular biology information by interviewing scientists (Appendix A, pg [33](#)) in the corresponding field and by reviewing journal articles. I will attempt to analyze what makes the web interfaces successful or unusable.

## **2. NIH Databases**

NIH is one of the foremost institutions where biological data is stored. Human and other organisms' genome sequencing products are stored at NIH databases. Hence, some of the molecular biology databases under evaluation by this capstone project are within NIH's main web site. They include GenBank, UniGene, Blast tool and CGAP. Another 2 web interfaces to be evaluated within NIH web site are OMIM and Pubmed. Pubmed is a journal reference database and OMIM is an online book with journal references.

### **a. GenBank**

GenBank is the name of the main nucleotide sequence database at NIH. It is the main molecular biology database that stores sequences submitted by various resources such as molecular biology labs and genome sequencing projects. There is not a specific web interface to GenBank. Instead, we will review the simple interface to Entrez search engine used by various molecular biology databases at NIH one of which happens to be GenBank.

There is a snapshot of the NIH simple Entrez interface within the appendix B on page [35](#). The NIH web site is relatively well designed and well balanced. There are links to internal major web sites on the right and left frames. The designers have avoided loading the web site with overwhelming amount of links and kept it simple for utility. The search engine is at the top and users can click the drop down menu to select which database to search.

Many of the NIH databases included in this capstone project start off with this interface and they follow to their own interfaces in the results windows.

The GenBank results window can be reached by choosing “Nucleotide” in the drop down menu and typing in either a gene name or accession number such as Hs.2 (Hs stands for the homo sapiens collection within the UniGene database). After a very quick search, results interface appears which can be found in the appendix B on page [36](#). The NIH results windows are relatively similar within various NIH search engines and are made up of a list of identifiers (such as accession numbers) followed by a brief description of the gene or disease or other category for which the user utilizes the search engine.

The identifier on the results page is also a link that takes the user to the web site where the user can find the information being sought. An example of this results page is in appendix B on page [37](#). For the GenBank example, this results page contains identifier (accession) number, name of sequence and other such identity information. Journal references for the sequence in question can also be located within this page. One point of importance that is found in most sequence databases is the inclusion of the sequence in either nucleotide or amino-acid form at the bottom of the page. In our example, a short sequence is at the bottom of the results page.

Interviewed scientists had many insights about the GenBank search functionality as an overall representation of the NIH web interface. To begin with, many complained about the help links located on the primary web site



(Appendix B, pg [35](#)). Some users were very unhappy with the content of the help links because they thought that their content were too cryptic and did not include any illustrations on how to navigate the web site. One scientist's major complaint about the main search engine was the lack of examples on what to put into the search engine. He felt that a link explaining what to put into the textbox or an example text right next to the textbox would prove very useful for novice users.

Most of the interviewed scientists liked the format of the primary results page (Appendix B, pg [36](#)). Some thought that the simple descriptions in a list form allowed scientists to sift through results quickly and reach the particular identifier that they were seeking. One user however thought that the primary results page could use reorganization. She said that if the top five secondary results pages (Appendix B, pg [37](#)) were to be put in one after the other in the fashion of blast results (Appendix C, pg [45](#)), it would help her save some time navigating through results pages. She indicated that a link can be included for results in a list form in case users did not want the top 5 results. There were no further comments about the secondary results page from the interviewees.

On the last additions of Nucleic acids research molecular biology database review [1](#), the authors of NIH databases overview paper indicated that sequences were still being entered at a regular pace into the databases at NIH [2](#) (pg 13). GenBank in particular continues to grow at an exponential rate [3](#) (pg 17). Over a period of 12 months between 2001 and 2002 4.6 million

new sequences have been added. In 2002, Genbank established “daily data exchange with the EMBL Data Library in the UK and the DNA Data Bank of Japan” in order to provide worldwide coverage of nucleotide data <sup>8</sup> (pg D23). At this stage Genbank had nucleotide sequences from 119,000 species and that expansion was at a rate of 1100 species per month. In 2003, number of species went up to 140,000. During the last year, most of the sequences contributed came from completed genome sequencing projects <sup>9</sup> (pg D23). This expansion makes it that much more important to have a good interface to properly access the data. To that affect NIH has provided the simple interface to Entrez search engine and the blast method to explore the Genbank database. I just discussed the use of Genbank through simple Entrez interface. I will explore the blast interface in section c.

## **b. UniGene**

One of the problems faced by online molecular biology databases such as GenBank is redundancy. Submitted sequences are not monitored against existing sequences for duplication. Also, overlapping sequences or similar sequences count for a large percentage of duplication anomalies within the databases. One attempt to correct this situation and provide more robust and better-screened sequences is the UniGene project. UniGene is an attempt at producing non-redundant collection of genomic sequences. UniGene group goes about producing these sequences via pattern matching algorithms. As a result their sequences are more likely to be gene representations than the bits and pieces available at databases such as GenBank.

Due to its higher level of annotation, UniGene sequences are utilized more in the molecular biology community within studies such as micro-arrays. Matching researched sequences against UniGene clusters enable scientists to better fit their research to their models. For the above stated reasons, I will explore the interface to UniGene's database on the NIH site.

Main UniGene web site (Appendix C, pg [39](#)) is very similar to the NIH main web site. There is a search tool available at the top of the web page. Users can enter search queries such as gene names and accession numbers to this search box. Similar to other NIH search engines, there is no advance search options. In addition to the search tool, the body of the main UniGene page contains some informational links about the database and how it is produced.

Primary results page is a summary list similar to the results of a GenBank search (Appendix C, pg [40](#)). However, unlike GenBank results, results list is shorter than the GenBank hits. This is due to fewer entries within the UniGene Database.

Clicking the favorite identifier link on the results page leads to the information page (Appendix C, pg [41](#)). The information page contains mapping, expression, and organism data. It also lists all the short sequences that contributed pieces to the UniGene cluster. With all the available information and links on the results page, UniGene acts as a gateway one can use to jump from one database interface to another to gather information about the particular sequence one is interested in.

The interviewed scientists had couple of comments about UniGene. As I mentioned above, the scientists have started to use UniGene more and more over the last couple of years due to its minimized-redundancy, robustness and gateway features. They indicated that UniGene accession numbers are sought out for any research that has to do with sequencing or expression work. However, couple of the scientists showed concerns about lack of good search tools to use on UniGene. They mainly complained about the fact that one cannot compare a raw sequence against UniGene. This has to do with the fact that there are no Blast tools available to compare raw sequences against the UniGene set on the web. To compare raw sequences against UniGene, scientists end up blasting against available databases such as GenBank and then using the GenBank accession numbers, they seek out UniGene identifiers via a UniGene search. This introduces problems such as the fact that GenBank matches are not perfect, especially against small sequences. And, a larger matched sequence in GenBank might not correctly identify the right UniGene cluster that the interested scientist is seeking for their research sequence.

Just like GenBank the UniGene database continues to expand. 6 new organisms have been added to the UniGene database in 2002 <sup>2</sup> (pg 14). UniGene database is updated bimonthly and background search algorithm is adjusted to accommodate the addition of this new data into the database. As of the end of 2003, UniGene contains clusters from “16 animals and 13 plants” <sup>10</sup> (pg D37). “The UniGene collection has been used as a source of

unique sequences for the fabrication of microarrays for the large-scale study of gene expression” [10](#) (pg D37). UniGene web interface is an important tool for scientists because it provides attractive features such as organized information and easy access to information through available Genbank accession numbers.

### **c. Blast**

So far, I have only discussed identifier based search tools. They are powerful tools that bring important functionality to users. However, they lack in certain features such as raw sequence comparisons. There is another tool to do just that: Blast. Blast is a sequence comparison tool originally developed based on the Smit-Wasserman sequence-matching algorithm. Modern version runs off the server farm at NIH and is used by scientists worldwide. The theory is simple enough. One takes a sequence of interest and compares against other sequences (generally against sequence databases) to find similarity. With an efficient algorithm and mass computing power, hopefully, this search completes in a finite amount of time. Of course with the computing power at NIH and the efficient algorithm, the web version of blast takes as little as 15 seconds to compare a small sequence against entire genomes and databases in gigabyte sizes.

The blast interface is markedly different compared to simple search engines we have reviewed so far. User accesses the utilities of the web interface by uploading a text sequence of nucleotides DNA search or amino acids for protein search. The main blast page contains a text box for pasting

of the sequence. Another option is for the user to upload a simple text file to the blast server (Appendix D, pg [43](#)). Following insertion of the sequence user can decide which database to blast against by picking one from the drop down menu. The choices are many and include main ones such as GenBank and dbEST (the EST database). Next, the user can modify blast parameters by changing the options within the option box. Various algorithm options can be manipulated here to optimize the blast process. The format box allows the user to change the format of the output. The user can opt to limit the displayed number of matches, to show various useful links and to include or not include various other features.

When the user is ready to proceed, the submit click is followed by a format window (Appendix D, pg [44](#)). In this window, the user gets another chance to change the format of the output. This window also serves as queue status display. In case of heavy server load, the window indicates the time left in the queue before the user is entitled to proceed with the blast process.

After a certain period (sometimes lengthy), the blast process concludes and the browser displays a results page (Appendix D, pg [45](#)). The results page starts with a graphical picture of the match of the small sequence against the database sequence. The bars in the picture are links to the database sequence's information page. The table below the picture shows identifiers as links to information pages along with short gene name and statistical scores for the particular match. Sequence matches are displayed

following this table by lining up the short sequence above the database sequence.

The scientists whom I interviewed utilize the blast tool often. According to them, it is one of the fastest ways to get from research data to analyzed data provided by the genomic databases. However, there is couple of issues in the user interface. All the scientists complained about the hard to understand language used in the help links of the options box. They indicated that difficulty in comprehending the help features led to none of them manipulating the blast algorithm options to fine-tune their processes. However, they all said that they use format options often and it is pretty straightforward to manipulate the format function. One scientist criticized the format window (Appendix D, pg [44](#)) that pops up after the blast button is clicked. She said that an extra click to reach the results leads to lost processes sometimes during multiple blasts. She suggested that maybe a pop-up window could be used to show queue information instead of interactive web form.

“The NCBI BLAST interface has been re-designed and offers several new search options including the specification of an Expectation Value range, rather than a threshold, for reporting alignments, and the specifications of a residue range to limit searches to a portion of the query sequence, XML output is now supported”<sup>2</sup> (pg 14). These added features enable users to better adjust search parameters to produce desired output. XML output also enables users to design their own tools to analyze results. The web BLAST

interface also “allows searches to be restricted to a database subset using standard Entrez search queries; the same restrictions may be used to screen the output of an initially unrestricted search” [11](#) (pg 29). This feature enables users to create their own customized search database.

#### **d. CGAP**

CGAP is another project web site within the NIH similar to UniGene (Appendix E, pg [47](#)). CGAP contains multiple sub-projects within its web site including sage, which is an expression database made using tags (short sequences following primer). Another sub-project worth mentioning is the cell libraries present in the CGAP database. The sequences for cancerous and none-cancerous cell lines from various organs are available for download thru CGAP. Chromatogram files for these sequences can be downloaded from trace file repositories (<ftp://genome.wustl.edu>).

I will not get into individual tools available from the CGAP site because there are quiet few of them. Instead, now I will delve into the scientists’ use of this site. Couple of the scientists whom I interviewed, did not know that this site existed. However, the others stated a reluctance to live without it. One scientist contributed the success of this site to its expression SAGE libraries. He said that he used these libraries against his own research to confirm and distinguish discrepancies with researched cell libraries. He continued by stating that he further investigates differences for possible discoveries. Another scientist used the sequences of the cell libraries and utilized the trace files to judge the quality of her own sequence and analyze sequence clusters



based on their quality. Both scientists indicated that the web site is not optimized for fast browsing and that usually it takes multiple clicks to get from one results page to another.

#### **e. Pubmed**

Pubmed is the web interface to the Medline database. It contains references to medical journals. Its interface is somewhat different than the search interfaces to the online genomic databases (Appendix F, pg [49](#)). Pubmed has links for advance searches, which can be accomplished by inputting author names, journal titles, subject names and others. Pubmed interface also allows MESH terms, which can make a search more accurate.

Primary results page for Pubmed is similar to other NIH search engines, which display results in a list form with an identifier and a short summary for each row in the list. In Pubmed's case the identifier is author names. As before the identifier link leads to the information page (Appendix F, pg [50](#)). Information page includes article information such as journal name, date, author name and location. Following the article information is the abstract of the article. Nowadays, most articles have a link to the journal website where they are available electronically. Unfortunately, most journals require subscriptions with substantial fees for access.

The interviewees utilized Pubmed extensively. However, most scientists did not use advance search options. Only one scientist was aware of MESH terms. She did not use them in her searches. Another scientist was not happy with the Pubmed interface at all. He indicated problems with

connection stability, disconnected links and lack of categorizations by journal volume. However, he did like to use the related articles feature of Pubmed searches.

Pubmed continues to serve as an interface to the 11 million references at Medline <sup>2</sup> (pg 13). The abundance of references makes it crucial to have a proper interface providing to access the data. Pubmed also provides multimedia data with some of its entries. It provides users with “streaming video and high resolution media” <sup>12</sup> (pg 2)

In a recent study, scientists tested the effectiveness of Pubmed in producing sought out results for tracking down systematic reviews. They found out that with personalized strategies, user can retrieve up to 90% of listed resources. Cochrane library and ACP journal club was used as the baseline for this study <sup>13</sup> (pg 161).

#### **f. OMIM**

OMIM is an online book of inheritable diseases. It used to have its own search engine but at present, it is incorporated in Entrez (Search engine for online journals and molecular biology databases). Therefore, it can be searched using the similar Pubmed interface. A query in OMIM brings up the list form of NIH search engine results. From the list results one can go to the information page of interest by clicking on the OMIM identifier link. The information page (Appendix G, pg [51](#)) is one of the longest ones within the NIH results pages. It contains text about the disease and links to relative journal articles about the disease.

Only one scientist I interviewed used OMIM before. He indicated that he used OMIM as a journal reference tool and that he countered many examples of out of date disease entries.

“OMIM is an easy and straightforward portal to the burgeoning information in human genetics” <sup>4</sup> (pg 52). OMIM has added statistics links to its main web page to give users a feel of the size and functionality of the OMIM database <sup>4</sup> (pg 53) OMIM is continuing to expand its functionality by offering help to researchers to unravel complex relationships between genes and diseases <sup>4</sup> (pg 54). To enable this functionality an easy to use and robust interface is of absolute necessity.

In a recent study by the Washington University School of Medicine, medical students were tested on capabilities to find sources on the internet for a given disease <sup>14</sup> (pg 852). In this study OMIM web interface proved particularly useful by providing the students with various genetic sources about the disease in question. OMIM provided information about genetic mutations causing the disease. Source of each mutation was cited with a brief explanation of molecular consequences of the mutations <sup>14</sup> (pg 854).

OMIM also provides links to related literature at Pubmed. User has the ability to move from a possibly outdated OMIM article to fresh articles about the subject at Pubmed by clicking the yellow light-bulb next to the article <sup>15</sup> (pg 184).



### 3. ENSEMBL

Ensembl is a venture among European molecular biology institutions to produce tools to annotate eukaryotic genomes. In a way it attempts to do the same thing UniGene is trying to accomplish which is to produce robust sequence sets from the available bits and pieces and to annotate those sets. Ensembl also added a very good web design to the mix. Use of buttons instead of links provides a feeling of an application tool to the web site (Appendix H, pg [53](#)).

The search page has some interesting tools (Appendix H, pg [54](#)). One of them is the graphical genome browser. The user can click on a part of a picture of a chromosome and zoom in to the map coordinate of the gene they are interested in. Ensembl also allows the user to download data in the form of maps or genomic dumps such as sequence data files.

Only one interviewed scientist had used Ensembl before and she did not use it extensively. She remembers the experience of browsing in the web site as being too complicated.

“The Ensembl web site provides a variety of alternate views of the data” [5](#) (pg 39). This is done to allow alternative views to provide a better understanding of the data. For this purpose, “ensemble can be accessed in a variety of ways apart from web pages” [5](#) (pg 39). Tools such as Apollo Java viewer can be used to organize data to fit the users’ specifications.

In 2002, Ensembl added new capabilities to its features. Ensembl now provides comparative genomic information between human and mouse genomes. The available information includes: “fine grained DNA–DNA

alignments; orthologous protein information and large scale synteny data” [16](#) (pg 38). Ensembl also predicts the addition of “drosophila and mosquito comparison information and a three way comparison of fugu human and mouse” [16](#) (pg 39). The increase in available features will enable scientists to use resource of effectively and efficiently.

In 2003, Ensembl added regular updates into its repertoire. A monthly release cycle has been adapted since February of 2003 [17](#) (pg D468). Other added features include a pre-ensembl web site which includes information about genome assemblies to be published [17](#) (pg D468). The new ‘Otter’ tool provides “an extended Ensembl schema for gene curation” [17](#) (pg D468). Further enhancements include added annotations for the comparative genome analysis tool, an overhauling of the Ensembl web-site and, the addition of Ensembl-Mart which is a tool exclusively written for data-mining genomes [17](#) (pg D469). With a mission statement of “continuously trying to improve all aspects of its work, from software engineering through to data analysis” [17](#) (pg D469), Ensembl web interface continues to refine its process of presenting scientists with necessary tools to analyze data.

Ensembl also has participated in studies to expand its reach. Ensembl has provided an XML interface to its data for a Distributed Annotation System research. This study attempted to access various annotation resources and provide centralized data [18](#) (pg 5-7).

#### 4. KEGG

KEGG is a pathway tool that shows metabolic or regulatory pathways for the sequence of interest. The main web site (Appendix I, pg [55](#)) contains various links to reach the pathway map. One way to reach pathways is the chromosome map, which can be browsed to go to a particular gene. KEGG can also be browsed with an OMIM map. The web site also offers categorized and browsable list to navigate in order to search for pathways.

The only scientist that used this interface was very disappointed by it. He indicated that the web site was very cryptic, hard to navigate and not very well explained. He found himself running around in circles and not reaching the needed pathway.

The authors of a KEGG related article discuss various ways of accessing the GENES databases at the KEGG site [6](#) (pg 43). They state “the GENES database can be accessed by three methods, although there are numerous links leading to this database in the KEGG system” [6](#) (pg 43). The multiplicity of routes to reach results was one the major problems with the KEGG interface. The sheer amount of links usually led the user to unwanted paths. If KEGG ever plans to have a robust interface, they will need to restrict the available number of pathways to reach results.

## 5. EMBL

EMBL is the European analog of the NIH's molecular biology section: NCBI. Some worldwide used genomic databases are produced by EMBL such as the Swiss-prot protein database. The main web page (Appendix J, pg [57](#)) is the entry point to the web site. The user has to navigate thru it in order to get to tools web page (Appendix J, pg [58](#)). The tools page has links to many different tools to access the data provided by EMBL. There is also a simple search engine to find genes via accession numbers.

None of the scientists I interviewed used the EMBL site before. The data from EMBL is available thru the NIH web sites.

“The EBI provides a comprehensive set of sequence similarity algorithms that can be accessed both interactively from the EMBL-EBI World Wide Web site (<http://www.ebi.ac.uk/Tools/>) or by email” <sup>7</sup> (pg 25). The abundance of tools increase EMBL's potential as a well developed device to analyze molecular data. However some of these tools need to be presented on the main page in order to allow users easy access.

In 2002, EMBL added new features to its functionality. The new con(struct) division provides “sequences of chromosomes, genomes and other long sequences constructed from segment entries” <sup>19</sup> (pg 20). During this period, EMBL also started providing whole genome shotgun (WGS) sequences <sup>19</sup> (pg 20). While expanding its information base, EMBL also looked into expanding its reach. To that affect, EMBL collaborated with Genbank and DDBJ to create a “Third Party Annotation” dataset <sup>19</sup> (pg 21). Daily builds of dataset ensure



constant updates. And, ftp access to dataset enable scientists to analyze annotation information from three major sequence repositories with ease.

In 2003, EMBL has adopted the XML standard for data exchange [20](#) (pg D29). Formatting of its data with this standard will make it easy for third parties to access EMBL information and adopt it quickly.



## 6. Analysis

An issue that came up repeatedly during the interviews is the number of links that users needed to go thru in order to reach the page of interest. Most NIH pages used an intermediary list page prior to access to information pages. Two clicks is relatively short when seeking information, however it is one more than absolute minimum. Especially during long lengths of multiple searches, users going thru an intermediary step would further lengthen the amount of time researchers invest in browsing. One way to minimize the intermediary step is displaying top 5 results and providing a link to the intermediary list results. Unfortunately, since different scientists seek different results sometimes for the same data, interface designers need to add easy to manipulate and intuitive parameters to the main search site in order to give scientist the ability to seek the correct top 5 entries for their searches. One way of providing these parameters is to do studies of the users and to categorize the most used parameters of the search engine. Then, provide these parameters as categories in the search site. However, I repeat that these parameters need to be intuitive and easy to use. Otherwise, it will lead to abandoning of the search tools.

Another issue that came up during interviews was the difficulty in using help sites. The major problem was the inability to understand what was said in the help sites. In most cases, help site failure is due to non-professional preparation of the help site. My suggestion in this situation is to hire a technical writer to prepare the help sites. I also suggest that illustrations must be used in order to prevent users from reading pages upon pages of text manuals.

## **a. Cross web interface analysis**

The main NIH database web interfaces start with similar search pages. Genbank and, UniGene provide a simple search page made from the same template. This template includes a drop down menu to choose from set of databases and a textbox to enter search query. Search query can be of various formats including accession numbers. Help on how to conduct searches is available through help links on all search pages. This form of simple search is pretty standard allowing users to conduct simple searches without too much hassle.

The results pages for NIH sequence databases' search engines differ from each other. Results provide database specific information formatted to fit user's needs. Genbank provides a list form of matching sequences with short descriptions which can be further explored by clicking provided links. UniGene's results page is annotation specific and provides list of categories where the sequence belongs. Both results page formats provide links to relevant publications.

CGAP is also oriented in an annotation specific structure. However, unlike UniGene, CGAP is structured more rigidly and provides access to clumps of data at a time in the form of libraries. For that reason CGAP is oriented more towards higher level analysis and organization.

Blast doesn't compare to the other web interfaces very well. It's a specific tool provided to search NIH databases by sequences. As such, the initial search user interface is completely different from prior query based search interfaces.

The only commonality between other search engines is the links provided to relevant publications.

Pubmed and OMIM provide interface to medical article databases. Pubmed has extensive configuration options for its searches. This extensive configurability allows users to filter their searches to their desire. OMIM provides access to genetic disease articles in a book format. It excels from a simple paperback book in that, being a web interface; it has the capability to go from one document to another by following provided links.

Ensembl also provides extensive annotation information like UniGene. It reaches further than UniGene in user accommodation by continuous upgrades to its features. Ensembl continues to expend its capabilities by adding new tools and, providing access to new data.

KEGG lacks behind in usability as compared to other web-interfaces. Its initial web page is not very user friendly. It does not provide simple descriptions for novice users thus making it hard to learn how to use.

EMBL provides another web interface that is on the rise. EMBL's collaboration with other databases and suit of tools for access and analysis makes EMBL web-site and attractive choice for scientists.

## **b. Critical Journal Review**

Evaluation of web interfaces for their utility is a subjective matter. Complicating this fact is the youthful nature of the web medium used to provide access and analysis features to molecular biology databases. Due to these

natures of molecular biology database web interfaces, it is difficult to locate quantitative analytical papers on the subject.

Most of the papers I have used as references in this project provide qualitative analytical results. Genbank series of articles provided me with information on overall qualitative progress made in the Genbank web interface [3](#), [8](#), [9](#). Similarly, the NIH database update papers from Nucleic Acids Research provide progress information for UniGene and Blast web-sites [2](#), [10](#), [11](#). Ensembl and EMBL have their own set of articles that track their progress over the years [16](#), [17](#), [19](#), [20](#).

Pubmed and OMIM have their own articles that track progress, show utilization studies and provide information on cross database collaborations [12](#), [14](#), [15](#).

One article on Pubmed does a quantitative study on search findings. The study is based on whether users can reproduce results found by Cochrane library and ACP journal club given certain search criteria. Authors go on to do statistical analysis of their results and draw conclusions based on findings [13](#).

The referenced articles proved useful because they showed the progress made by the web interfaces in trying to accommodate their user in providing useful access and analytical tools to the particular molecular biology databases.

## **7. Conclusion**

In conclusion, an optimal web interface for online databases should employ single click or double click at most to reach results from the start of search query. The interface should also employ easy to use and intuitive parameters, such as those described in this paper, for scientists to access search engines in order to enhance the features of the search. To complement the web interface, a help site to navigate the various features should be written in simple language and illustrations should be included. Length of the help pages should also be limited in order to provide users with fast and simple help. If deployed, such web interfaces will allow users to more thoroughly utilize the tremendous data available from various online databases.





## References

1. Andreas D. Baxevais; The Molecular Biology Database Collection: 2002 Update; *Nucleic acids research*, 2002 Vol. 30, No 1 (1-12)
2. David L. Wheeler\*, Deanna M. Church, Alex E. Lash, Detlef D. Leipe, Thomas L. Madden, Joan U. Pontius, Gregory D. Schuler, Lynn M. Schriml, Tatiana A. Tatusova, Lukas Wagner and Barbara A. Rapp; Database resources of the National Center for Biotechnology Information: 2002 update; *Nucleic acids research*, 2002 Vol. 30, No. 1 (13-16)
3. Dennis A. Benson\*, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, Barbara A. Rapp and David L. Wheeler; GenBank; *Nucleic acids research*, 2002 Vol. 30, No. 1 (17-20)
4. Ada Hamosh\*, Alan F. Scott, Joanna Amberger, Carol Bocchini, David Valle, and Victor A. McKusick; Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders; *Nucleic acids research*, 2002 Vol. 30, No. 1 (52-55)
5. T. Hubbard, D. Barker, E. Birney\*, G. Cameron<sup>1</sup>, Y. Che<sup>&</sup>, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyraas, J. Gilbert, M. Hammond<sup>1</sup>, L. Huminieck<sup>1</sup>, A. Kasprzyk<sup>1</sup>, H. Lehvaslaiho<sup>1</sup>, R Lijnzaad<sup>1</sup>, C. Melsopp<sup>1</sup>, E. Mongin<sup>1</sup>, R. Pettett, M. Pocock, S. Potter, A. Rust<sup>1</sup>, E. Schmidt<sup>1</sup>, S. Searle, G. Slater<sup>1</sup>, J. Smith, W. Spooner, A. Stabenau<sup>1</sup>, J. Stalker, E. Stupka<sup>1</sup>, A. Ureta-Vida<sup>P</sup>, I. Vastrik<sup>1</sup> and M. Clamp; The Ensembl genome database project; *Nucleic acids research*, 2002 Vol. 30, No. 1 (38-41)
6. Minoru Kanehisa\*, Susumu Goto, Shuichi Kawashima and Akihiro Nakaya; *Nucleic acids research*, 2002 Vol. 30, No. 1 (42-46)
7. Guenter Stoesser\*, Wendy Baker, Alexandra van den Broek, Evelyn Camon, Maria Garcia-Pastor, Carola Kanz, Tamara Kulikova, Rasko Leinonen, Ouan Lin, Vincent Lombard, Rodrigo Lopez, Nicole Redaschi, Peter Stoehr, Mary Ann Tuli, Katerina Tzouvara and Robert Vaughan; The EMBL Nucleotide Sequence Database; *Nucleic acids research*, 2002 Vol. 30, No. 1 (21-26)
8. Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell and David L. Wheeler\*; GenBank; *Nucleic Acids Research*, 2003, Vol. 31, No. 1 (23-27)

9. Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell and David L. Wheeler\*; GenBank: update; Nucleic Acids Research, 2004, Vol. 32, Database issue (D23-D26)
10. David L. Wheeler\*, Deanna M. Church, Ron Edgar, Scott Federhen, Wolfgang Helmberg, Thomas L. Madden, Joan U. Pontius, Gregory D. Schuler, Lynn M. Schriml, Edwin Sequeira, Tugba O. Suzek, Tatiana A. Tatusova and Lukas Wagner; Database resources of the National Center for Biotechnology Information: update; Nucleic Acids Research, 2004, Vol. 32, Database issue (D35-D40)
11. David L. Wheeler\*, Deanna M. Church, Scott Federhen, Alex E. Lash, Thomas L. Madden, Joan U. Pontius, Gregory D. Schuler, Lynn M. Schriml, Edwin Sequeira, Tatiana A. Tatusova and Lukas Wagner; Database resources of the National Center for Biotechnology; Nucleic Acids Research, 2003, Vol. 31, No. 1 (28-33)
12. Donald A.B. Lindberg; Internet Access to the National Library of Medicine; Effective Clinical Practice, September/October 2000
13. KAVEH G. SHOJANIA, MD, LISA A. BERO, PhD; Taking Advantage of the Explosion of Systematic Reviews: An Efficient MEDLINE Search Strategy; Eff Clin Pract. 2001; 4: 157-162.
14. Jeffrey Magee, Jeffrey I. Gordon, MD, and Alison Whelan, MD; Bringing the Human Genome and the Revolution in Bioinformatics to the Medical School Classroom: A Case Report from Washington University School of Medicine; AC ADEMIC MEDICINE, VOL. 76, NO. 8/AUGUST 2001 (852-855)
15. Ouellette F; Internet resources for the clinical geneticist; Clin Genet 1999; 56: 179–185
16. M. Clamp, D. Andrews, D. Barker, P. Bevan, G. Cameron, Y. Chen, L. Clark, T. Cox, J. Cuff, V. Curwen, T. Down, R. Durbin, E. Eyra, J. Gilbert, M. Hammond, T. Hubbard\*, A. Kasprzyk, D. Keefe, H. Lehvaslaiho, V. Iyer, C. Melsopp, E. Mongin, R. Pettett, S. Potter, A. Rust, E. Schmidt, S. Searle, G. Slater, J. Smith, W. Spooner, A. Stabenau, J. Stalker, E. Stupka, A. Ureta-Vidal, I. Vastrik and E. Birney; Ensembl 2002: accommodating comparative genomics; Nucleic Acids Research, 2003, Vol. 31, No. 1 (38-42)
17. E. Birney, D. Andrews, P. Bevan, M. Caccamo, G. Cameron, Y. Chen, L. Clarke, G. Coates, T. Cox, J. Cuff, V. Curwen, T. Cutts, T. Down, R. Durbin, E. Eyra, X. M. Fernandez-Suarez, P. Gane, B. Gibbins, J. Gilbert, M. Hammond, H. Hotz, V. Iyer, A. Kahari, K. Jekosch, A. Kasprzyk, D.

- Keefe, S. Keenan, H. Lehvaslaiho, G. McVicker, C. Melsopp, P. Meidl, E. Mongin, R. Pettett, S. Potter, G. Proctor, M. Rae, S. Searle, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, A. Ureta-Vidal, C. Woodwark, M. Clamp and T. Hubbard\*; Ensembl 2004; Nucleic Acids Research, 2004, Vol. 32, Database issue (D468-D470)
18. Robin D Dowell, Rodney M Jokerst, Allen Day, Sean R Eddy and Lincoln Stein\*; The Distributed Annotation System; BMC Bioinformatics, 2001, 2:7
19. Guenter Stoesser\*, Wendy Baker, Alexandra van den Broek, Maria Garcia-Pastor, Carola Kanz, Tamara Kulikova, Rasko Leinonen, Quan Lin, Vincent Lombard, Rodrigo Lopez, Renato Mancuso, Francesco Nardone, Peter Stoehr, Mary Ann Tuli, Katerina Tzouvara and Robert Vaughan; The EMBL Nucleotide Sequence Database: major new developments; Nucleic Acids Research, 2003, Vol. 31, No. 1 (17–22)
20. Tamara Kulikova\*, Philippe Aldebert, Nicola Althorpe, Wendy Baker, Kirsty Bates, Paul Browne, Alexandra van den Broek, Guy Cochrane, Karyn Duggan, Ruth Eberhardt, Nadeem Faruque, Maria Garcia-Pastor, Nicola Harte, Carola Kanz, Rasko Leinonen, Quan Lin, Vincent Lombard, Rodrigo Lopez, Renato Mancuso, Michelle McHale, Francesco Nardone, Ville Silventoinen, Peter Stoehr, Guenter Stoesser, Mary Ann Tuli, Katerina Tzouvara, Robert Vaughan, Dan Wu, Weimin Zhu and Rolf Apweiler; The EMBL Nucleotide Sequence Database; Nucleic Acids Research, 2004, Vol. 32, Database issue (D27-D30)
21. GenBank; <http://www.ncbi.nlm.nih.gov/>
22. UniGene; <http://www.ncbi.nlm.nih.gov/UniGene/>
23. Blast; <http://www.ncbi.nlm.nih.gov/BLAST/>
24. CGAP (Cancer Genome Anatomy Project);  
<http://www.ncbi.nlm.nih.gov/ncicgap/>
25. Pubmed; <http://www.ncbi.nih.gov/entrez/query.fcgi>
26. OMIM (Online Mendelian Inheritance in Man);  
<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM>
27. Ensembl; <http://www.ensembl.org/>
28. KEGG; <http://www.genome.ad.jp/kegg/kegg2.html>
29. EMBL (European Molecular Biology Laboratory);  
<http://www.embl-heidelberg.de/>



## Appendix A

### Interviewees

Nigel Clegg, PhD	D4-100 1100 Fairview Ave N, Seattle, WA 98109	<a href="mailto:nclegg@fhcrc.org">nclegg@fhcrc.org</a>
Camari Ferguson	D4-395 1100 Fairview Ave N, Seattle, WA 98109	<a href="mailto:cferguso@fhcrc.org">cferguso@fhcrc.org</a>
Stacy Moore, PhD	D4-395 1100 Fairview Ave N, Seattle, WA 98109	<a href="mailto:smoore@fhcrc.org">smoore@fhcrc.org</a>
Denise Abbott	D4-395 1100 Fairview Ave N, Seattle, WA 98109	<a href="mailto:dabbott@fhcrc.org">dabbott@fhcrc.org</a>



## Appendix B

### NCBI main web page/GenBank search engine [21](#)

The image shows a screenshot of the National Center for Biotechnology Information (NCBI) website. At the top, the NCBI logo is on the left, and the text "National Center for Biotechnology Information" is centered, with "National Library of Medicine" and "National Institutes of Health" below it. A navigation bar contains links for PubMed, Entrez, BLAST, OMIM, Books, TaxBrowser, and Structure. Below this is a search bar with a dropdown menu set to "Nucleotide" and a text input field containing "Hs.2". A "Go" button is to the right of the search bar. On the left side, there is a vertical menu with links for "SITE MAP", "Guide to NCBI resources", "About NCBI", "GenBank", "Molecular databases", "Literature databases", and "Genomic biology". The main content area features a "What does NCBI do?" section, a "Hot Spots" section with a list of resources, a "Mouse Genome" section with a mouse icon and links to "Map Viewer", "Sequencing Progress", and "Human-Mouse Homology", and a "BLink" section. At the bottom, a "Rat Genome Resource" section is partially visible.

**NCBI**  
National Library of Medicine National Institutes of Health

PubMed Entrez BLAST OMIM Books TaxBrowser Structure

Search Nucleotide for Hs.2 Go

**What does NCBI do?**  
Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. [More...](#)

**Hot Spots**

- ▶ Cancer genome anatomy project
- ▶ Clusters of orthologous groups
- ▶ Coffee Break
- ▶ Electronic PCR
- ▶ Gene expression omnibus
- ▶ Genes and disease
- ▶ Human genome resources
- ▶ Human/mouse homology maps
- ▶ LocusLink
- ▶ Malaria genetics & genomics
- ▶ Map Viewer
- ▶ Mouse genome resources

**Mouse Genome**  
*Resources: explore tools for manipulating the mouse genome.*

**Try these:** Map Viewer Sequencing Progress Human-Mouse Homology

**BLink** and get results fast!  
Use BLink to view a graphical alignment of protein sequence similarities, taxonomic trees, 3D structures, and more. BLink provides quick results based on precomputed BLASTp searches against the non-redundant (nr) protein database. [More...](#)

**Rat Genome Resource**

## GenBank primary results page [21](#)

The screenshot shows the NCBI Entrez Nucleotide search interface. At the top, the NCBI logo is on the left, and a decorative banner with the word 'Entrez Nucleotide' and a DNA sequence is on the right. Below the banner, a navigation bar includes tabs for PubMed, Nucleotide, Protein, Genome, Structure, PopSet, Taxonomy, OMIM, and Books. The search bar contains 'Nucleotide' as the search type and 'Hs.2' as the query. Action buttons for 'Go' and 'Clear' are present. Below the search bar, there are links for 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. A display control section shows 'Display Summary', 'Show: 20', and 'Send to File'. The results are listed as 'Items 1-20 of 23' on 'Page 1 of 2'. The first six results are:

- 1:** [AF521594](#) Cotton leaf curl virus isolate HS-2 coat protein (AV1) gene, complete cds  
gi|21745345|gb|AF521594.1|[21745345]
- 2:** [C24714](#) JC virus gene for VP1 and large T antigen, partial cds, VT-intergenic region, isolate:HS-2  
gi|13383027|dbj|C24714.1|[13383027]
- 3:** [U14032](#) Human cell line HD53 T cell receptor beta chain, V-D-J junctional region (TCR V beta 8.1, J beta 2.7) mRNA, partial cds  
gi|565073|gb|U14032.1|HSU14032[565073]
- 4:** [U14039](#) Human T cell receptor alpha chain, V-J junctional region (TCR V alpha 8.1, J alpha 39) mRNA, partial cds  
gi|557345|gb|U14039.1|HSU14039[557345]
- 5:** [U14038](#) Human T cell receptor alpha chain, V-J junctional region (TCR V alpha 8.1, J alpha 25) mRNA, partial cds  
gi|557343|gb|U14038.1|HSU14038[557343]
- 6:** [U14037](#) Human cell line HD59 T cell receptor beta chain, V-D-J junctional region (TCR V beta 6.2, J beta 2.7) mRNA, partial cds  
gi|565071|gb|U14037.1|HSU14037[565071]

Each result includes a checkbox, the accession number, a description, and a 'Links' button. On the left side of the page, there is a blue sidebar with navigation links: 'About Entrez', 'Search for Genes', 'Entrez Nucleotide', 'Batch Entrez', 'Check sequence revision history', 'How to create WWW links to Entrez', 'LinkOut', 'Cubby', 'Related resources', 'Reference sequence project', and 'Submit to GenBank'.



## GenBank information page [21](#)

1: U14032. Human cell line H...[gi:565073] Links

LOCUS HSU14032 45 bp mRNA linear PRI 28-OCT-1995  
DEFINITION Human cell line HD53 T cell receptor beta chain, V-D-J junctional region (TCR V beta 8.1, J beta 2.7) mRNA, partial cds.  
ACCESSION U14032  
VERSION U14032.1 GI:565073  
KEYWORDS .  
SOURCE Homo sapiens (human)  
ORGANISM [Homo sapiens](#)  
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.  
REFERENCE 1 (bases 1 to 45)  
AUTHORS Desai-Mehta,A., Mao,C., Rajagopalan,S., Robinson,T. and Datta,S.K.  
TITLE Structure and specificity of T cell receptors expressed by potentially pathogenic anti-DNA autoantibody-inducing T cells in human lupus  
JOURNAL J. Clin. Invest. 95 (2), 531-541 (1995)  
MEDLINE [95164681](#)  
PUBMED [7860735](#)  
REFERENCE 2 (bases 1 to 45)  
AUTHORS Desai-Mehta,A.  
TITLE Direct Submission  
JOURNAL Submitted (26-AUG-1994) Ami Desai-Mehta, Medicine/Arthritis, Northwestern University, Medical School, 303 East Chicago Avenue/Ward 3-315, Chicago, IL 60611-3008, USA  
COMMENT On Nov 8, 1994 this sequence version replaced gi:[557331](#).  
FEATURES  
Location/Qualifiers  
source 1..45  
/organism="Homo sapiens"  
/isolate="lupus patient HS-2"  
/db\_xref="taxon:9606"  
/cell\_line="HD53"  
/cell\_type="T-lymphocyte"  
[gene](#) 1..45  
/gene="TCR V beta 8.1, J beta 2.7"  
[CDS](#) <1..>45  
/gene="TCR V beta 8.1, J beta 2.7"  
/note="coding region sequence reported spans from nucleotide positions 274 to 318. 5' end of V beta sequence couldn't be defined due to position of primer, hence V gene is either 8.1 or 8.2 allele"  
/codon\_start=1  
/product="T cell receptor beta chain, V-D-J junctional region"  
/protein\_id="AAA80036.1"  
/db\_xref="GI:565074"  
/translation="CASSPGGSYEQYFGP"  
BASE COUNT 6 a 16 c 15 g 8 t  
ORIGIN  
1 tgtgccagca gccccggtgg ctctacgag cagtacttcg ggccg  
//



## Appendix C

### UniGene main web site <sup>22</sup>

NCBI






UniGene

Search UniGene for Hs.2

PubMed Nucleotide Protein Genome Structure Popset Taxonomy

Limits Preview/Index History Clipboard Details

UniGene is an experimental system for automatically partitioning GenBank sequences into a non-redundant set of gene-oriented clusters. Each UniGene cluster contains sequences that represent a unique gene, as well as related information such as the tissue types in which the gene has been expressed and map location.

	Chordata Mammalia	<p><b>Bos taurus</b> (cow) 10,703 entries</p> <p><b>Homo sapiens</b> (human) 115,523 entries</p> <p><b>Mus musculus</b> (mouse) 88,647 entries</p> <p><b>Rattus norvegicus</b> (rat) 65,804 entries</p> <p><b>Sus scrofa</b> (pig) 14,271 entries</p>
	Amphibia	<b>Xenopus laevis</b> (frog) 19,143 entries
	Actinopterygii	<b>Danio rerio</b> (zebrafish) 15,961 entries
	Arthropoda Insecta	<p><b>Anopheles gambiae</b> (malaria mosquito) 2,661 entries</p> <p><b>Drosophila melanogaster</b> (fruit fly) 11,189 entries</p>
	Nematoda Chromadorea	<b>Caenorhabditis elegans</b> 19,800 entries
	Embryophyta Eudicotyledons	<p><b>Arabidopsis thaliana</b> (thale cress) 27,159 entries</p> <p><b>Glycine max</b> (soybean) 8,664 entries</p> <p><b>Lycopersicon esculentum</b> (tomato) 3,693 entries</p> <p><b>Medicago truncatula</b> (barrel medic) 5,792 entries</p>
	Liliopsida	<p><b>Hordeum vulgare</b> (barley) 7,915 entries</p> <p><b>Oryza sativa</b> (rice) 16,842 entries</p> <p><b>Triticum aestivum</b> (wheat) 22,188 entries</p> <p><b>Zea mays</b> (maize) 12,536 entries</p>
	Chlorophyta Chlorophyceae	<b>Chlamydomonas reinhardtii</b> 6,517 entries

In addition to sequences of well-characterized genes, hundreds of thousands

## UniGene primary results page [22](#)

The screenshot shows the NCBI UniGene search interface. At the top left is the NCBI logo. The main header features the UniGene logo and a navigation bar with tabs for PubMed, Nucleotide, Protein, Genome, Structure, Popset, and Taxonomy. Below this is a search bar with 'UniGene' selected in the dropdown and 'Hs.2' entered in the text field. 'Go' and 'Clear' buttons are to the right. Underneath the search bar are links for Limits, Preview/Index, History, Clipboard, and Details. A control bar includes a 'Display' dropdown set to 'Summary', a 'Show' dropdown set to '20', a 'Sort' dropdown, and a 'Send to' dropdown set to 'File'. The search results section shows one result:  **1:** [Hs.2](#)  
**NAT2** N-acetyltransferase 2 (arylamine N-acetyltransferase)  
*Homo sapiens* , 24 sequence(s)

At the bottom of the page, there are links for [Restrictions on Use](#), [Write to the Help Desk](#), [NCBI](#), [NLM](#), and [NIH](#).


## UniGene information page [22](#)

<p style="text-align: center; margin: 0;">NCBI</p> <hr style="border: 0; border-top: 1px solid white; margin: 5px 0;"/> <p style="margin: 0;">UniGene Query Tips FAQ DDD Download UniGene</p> <p style="margin: 0;">Related Resources LocusLink HomoloGene dbEST Trace Archive CGAP</p>	<p><b>UniGene Cluster Hs.2 <i>Homo sapiens</i></b></p> <p><b>NAT2 N-acetyltransferase 2 (arylamine N-acetyltransferase)</b></p> <p><b>SEE ALSO</b> <b>LocusLink:</b> <a href="#">10</a> <b>OMIM:</b> <a href="#">243400</a> <b>HomoloGene:</b> <a href="#">Hs.2</a></p> <p><b>SELECTED MODEL ORGANISM PROTEIN SIMILARITIES</b> <b>organism, protein and percent identity and length of aligned region</b></p> <table border="0"><tr><td style="vertical-align: top;"><b><i>H.sapiens:</i></b></td><td style="vertical-align: top;"><a href="#">pir:B34585</a> - B34585 arylamine N-acetyltransferase (EC 2.3.1.5) 2 - human</td><td style="vertical-align: top; text-align: right;"><b>100 % / 289 aa</b> (see <a href="#">ProtEST</a>)</td></tr><tr><td style="vertical-align: top;"><b><i>M.musculus:</i></b></td><td style="vertical-align: top;"><a href="#">sp:P50295</a> - ARY2_MOUSE Arylamine N-acetyltransferase 2 (Arylamide acetylase 2) (N-acetyltransferase type 2) (NAT)</td><td style="vertical-align: top; text-align: right;"><b>74 % / 289 aa</b> (see <a href="#">ProtEST</a>)</td></tr><tr><td style="vertical-align: top;"><b><i>R.norvegicus:</i></b></td><td style="vertical-align: top;"><a href="#">ref.NP_446306.1</a> - N-acetyltransferase 2 (arylamine N-acetyltransferase) [Rattus norvegicus]</td><td style="vertical-align: top; text-align: right;"><b>73 % / 289 aa</b> (see <a href="#">ProtEST</a>)</td></tr></table> <p><b>MAPPING INFORMATION</b> <b>Chromosome:</b> 8 <b>Genome View:</b> <a href="#">Chromosome 8</a> <b>OMIM Gene Map:</b> <a href="#">8p22</a> <b>UniSTS entries:</b> <a href="#">GDB:386004</a> Genomic Context: <a href="#">Map View</a> <b>UniSTS entries:</b> <a href="#">stSG40</a> Genomic Context: <a href="#">Map View</a> <b>UniSTS entries:</b> <a href="#">SHGC-130680</a> Genomic Context: <a href="#">Map View</a> <b>UniSTS entries:</b> <a href="#">WI-7224</a> Genomic Context: <a href="#">Map View</a> <b>UniSTS entries:</b> <a href="#">GDB:310612</a> Genomic Context: <a href="#">Map View</a></p> <p><b>EXPRESSION INFORMATION</b> <b>cDNA sources:</b> liver ;hepatocellular carcinoma ;adenocarcinoma ;corresponding non cancerous liver tissue ;colon ;Cell lines <b>SAGE :</b> <a href="#">Gene to Tag mapping</a></p> <p><b>mRNA SEQUENCES (4)</b></p> <table border="0"><tr><td style="vertical-align: top;"><a href="#">NM_000015</a></td><td style="vertical-align: top;">Homo sapiens N-acetyltransferase 2 (arylamine N-acetyltransferase) (NAT2), mRNA</td><td style="vertical-align: top; text-align: right;"><b>P</b></td></tr><tr><td style="vertical-align: top;"><a href="#">D90040</a></td><td style="vertical-align: top;">Human mRNA for arylamine N-acetyltransferase (EC 2.3.1.5)</td><td style="vertical-align: top; text-align: right;"><b>P</b></td></tr><tr><td style="vertical-align: top;"><a href="#">BC015878</a></td><td style="vertical-align: top;">Homo sapiens, N-acetyltransferase 2 (arylamine N-acetyltransferase), clone MGC:27492 IMAGE:4716636, mRNA, complete cds</td><td style="vertical-align: top; text-align: right;"><b>P A</b></td></tr><tr><td style="vertical-align: top;"><a href="#">D90042</a></td><td style="vertical-align: top;">Human liver arylamine N-acetyltransferase (EC 2.3.1.5) gene</td><td style="vertical-align: top; text-align: right;"><b>P</b></td></tr></table> <p><b>EST SEQUENCES (20)</b></p> <table border="0"><tr><td style="vertical-align: top;"><a href="#">A1792606</a></td><td style="vertical-align: top;">cDNA clone IMAGE:1870937</td><td style="vertical-align: top;">adenocarcinoma</td><td style="vertical-align: top;">5' read 1.2 kb</td><td style="vertical-align: top;"><b>P</b></td></tr><tr><td style="vertical-align: top;"><a href="#">A1733799</a></td><td style="vertical-align: top;">cDNA clone IMAGE:1870937</td><td style="vertical-align: top;">adenocarcinoma</td><td style="vertical-align: top;">3' read 1.2 kb</td><td style="vertical-align: top;"><b>P A</b></td></tr></table>	<b><i>H.sapiens:</i></b>	<a href="#">pir:B34585</a> - B34585 arylamine N-acetyltransferase (EC 2.3.1.5) 2 - human	<b>100 % / 289 aa</b> (see <a href="#">ProtEST</a> )	<b><i>M.musculus:</i></b>	<a href="#">sp:P50295</a> - ARY2_MOUSE Arylamine N-acetyltransferase 2 (Arylamide acetylase 2) (N-acetyltransferase type 2) (NAT)	<b>74 % / 289 aa</b> (see <a href="#">ProtEST</a> )	<b><i>R.norvegicus:</i></b>	<a href="#">ref.NP_446306.1</a> - N-acetyltransferase 2 (arylamine N-acetyltransferase) [Rattus norvegicus]	<b>73 % / 289 aa</b> (see <a href="#">ProtEST</a> )	<a href="#">NM_000015</a>	Homo sapiens N-acetyltransferase 2 (arylamine N-acetyltransferase) (NAT2), mRNA	<b>P</b>	<a href="#">D90040</a>	Human mRNA for arylamine N-acetyltransferase (EC 2.3.1.5)	<b>P</b>	<a href="#">BC015878</a>	Homo sapiens, N-acetyltransferase 2 (arylamine N-acetyltransferase), clone MGC:27492 IMAGE:4716636, mRNA, complete cds	<b>P A</b>	<a href="#">D90042</a>	Human liver arylamine N-acetyltransferase (EC 2.3.1.5) gene	<b>P</b>	<a href="#">A1792606</a>	cDNA clone IMAGE:1870937	adenocarcinoma	5' read 1.2 kb	<b>P</b>	<a href="#">A1733799</a>	cDNA clone IMAGE:1870937	adenocarcinoma	3' read 1.2 kb	<b>P A</b>
<b><i>H.sapiens:</i></b>	<a href="#">pir:B34585</a> - B34585 arylamine N-acetyltransferase (EC 2.3.1.5) 2 - human	<b>100 % / 289 aa</b> (see <a href="#">ProtEST</a> )																														
<b><i>M.musculus:</i></b>	<a href="#">sp:P50295</a> - ARY2_MOUSE Arylamine N-acetyltransferase 2 (Arylamide acetylase 2) (N-acetyltransferase type 2) (NAT)	<b>74 % / 289 aa</b> (see <a href="#">ProtEST</a> )																														
<b><i>R.norvegicus:</i></b>	<a href="#">ref.NP_446306.1</a> - N-acetyltransferase 2 (arylamine N-acetyltransferase) [Rattus norvegicus]	<b>73 % / 289 aa</b> (see <a href="#">ProtEST</a> )																														
<a href="#">NM_000015</a>	Homo sapiens N-acetyltransferase 2 (arylamine N-acetyltransferase) (NAT2), mRNA	<b>P</b>																														
<a href="#">D90040</a>	Human mRNA for arylamine N-acetyltransferase (EC 2.3.1.5)	<b>P</b>																														
<a href="#">BC015878</a>	Homo sapiens, N-acetyltransferase 2 (arylamine N-acetyltransferase), clone MGC:27492 IMAGE:4716636, mRNA, complete cds	<b>P A</b>																														
<a href="#">D90042</a>	Human liver arylamine N-acetyltransferase (EC 2.3.1.5) gene	<b>P</b>																														
<a href="#">A1792606</a>	cDNA clone IMAGE:1870937	adenocarcinoma	5' read 1.2 kb	<b>P</b>																												
<a href="#">A1733799</a>	cDNA clone IMAGE:1870937	adenocarcinoma	3' read 1.2 kb	<b>P A</b>																												



## Appendix D

### Blast main page <sup>23</sup>

 **nucleotide-nucleotide BLAST**

Nucleotide Protein Translations Retrieve results for an RID

[Search](#)

```
aaatgttctt ccagttcttc atataacaaa tactgccaca acagtgtctg
ttgatgaatt
tggcgttggg ccactttgca aaggtgacaa ctgtatttg tcagctgttg
atgtttgtgg
aatgtttact aacagatctg gttcccagca gtggagagga ctgtccagat
```

[Set subsequence](#) From:  To:

[Choose database](#)

Now: **BLAST!** or [Reset query](#) [Reset all](#)

**Options** for advanced blasting

[Limit by entrez query](#)  or select from:

[Choose filter](#)  Low complexity  Human repeats  Mask for lookup table only  Mask lower case

[Expect](#)

[Word Size](#)

[Other advanced](#)

**Format**

Show  [Graphical Overview](#)  [Linkout](#)  [Sequence Retrieval](#)  [NCBI-gi](#)  in  [format](#)

Number of: [Descriptions](#)  [Alignments](#)

[Alignment view](#)

[Limit results by entrez query](#)  or select from:

[Expect value range:](#)

## Blast queue page [23](#)



Your request has been successfully submitted and put into the Blast Queue.

**Query** = (610 letters)

The request ID is

or

The results are estimated to be ready in 38 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.

**Format**

Show  [Graphical Overview](#)  [Linkout](#)  [Sequence Retrieval](#)  [NCBI-gi](#)  in  [format](#)

Number of: [Descriptions](#)  [Alignments](#)

[Alignment view](#)

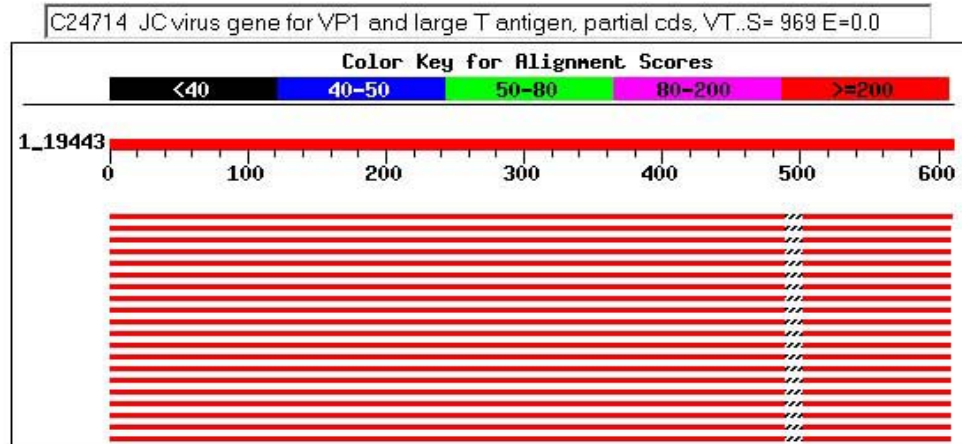
[Limit results by  
entrez query](#)  or select from:

[Expect value  
range:](#)



## Blast results page [23](#)

### Distribution of 40 Blast Hits on the Query Sequence



Sequences producing significant alignments:	Score (bits)	E Value
<a href="#">gi 13383027 dbj C24714.1</a> JC virus gene for VP1 and large T...	<a href="#">969</a>	0.0
<a href="#">gi 24636471 dbj AB081029.1</a> JC virus DNA, complete genome, ...	<a href="#">938</a>	0.0
<a href="#">gi 22830652 dbj AB081606.1</a> JC virus DNA, complete genome, ...	<a href="#">938</a>	0.0
<a href="#">gi 3452694 dbj AB008767.1</a> JC virus genes for VP1 and large...	<a href="#">938</a>	0.0
<a href="#">gi 761704 dbj D49941.1 JCVVTTIRL</a> JC virus (isolate:M-04) gen...	<a href="#">938</a>	0.0
<a href="#">gi 13383129 dbj C24748.1</a> JC virus gene for VP1 and large T...	<a href="#">938</a>	0.0
<a href="#">gi 13383126 dbj C24747.1</a> JC virus gene for VP1 and large T...	<a href="#">938</a>	0.0
<a href="#">gi 13383120 dbj C24745.1</a> JC virus gene for VP1 and large T...	<a href="#">938</a>	0.0
<a href="#">gi 13383102 dbj C24739.1</a> JC virus gene for VP1 and large T...	<a href="#">938</a>	0.0
<a href="#">gi 13383087 dbj C24734.1</a> JC virus gene for VP1 and large T...	<a href="#">938</a>	0.0

#### Alignments

Get selected sequences

Select all

Deselect all

[>gi|13383027|dbj|C24714.1](#) JC virus gene for VP1 and large T antigen, partial cds,  
VT-intergenic region, isolate:HS-2  
Length = 610

Score = 969 bits (489), Expect = 0.0  
Identities = 489/489 (100%)  
Strand = Plus / Plus

```
Query: 1 aaatgttctccagttcttcatataacaaactgcccacaacagtgctgcttgatgaatt 60
          |||
Sbjct: 1 aaatgttctccagttcttcatataacaaactgcccacaacagtgctgcttgatgaatt 60
```



## Appendix E

### CGAP main page [24](#)

**CGAP HOW TO** | **Genes** | **Chromosomes** | **Tissues** | **SAGE Genie** | **Pathways** | **Tools**

**CGAP Info**

- [Educational Resources](#)
- [Slide Tour](#)
- [Team Members](#)
- [References](#)

**CGAP Data**

- [Download](#)

**Quick Links:**

- [NCI Home](#)
- [NCICB Home](#)
- [NCBI Home](#)

**NATIONAL CANCER INSTITUTE**







### The CANCER GENOME ANATOMY PROJECT

**CGAP's Goals**

The goal of the NCI's Cancer Genome Anatomy Project is to determine the gene expression profiles of normal, precancer, and cancer cells, leading eventually to improved detection, diagnosis, and treatment for the patient. By collaborating with scientists worldwide, such as the [Ludwig Institute for Cancer Research](#) and [Lund University](#), CGAP seeks to increase its scientific expertise and expand its databases for the benefit of all cancer researchers.

**The Web Site**

The information is organized in a "biological sense" as follows:

 <p><b>Genes</b> Information on specific genes and collections of genes.</p>	 <p><b>Chromosomes</b> Gene mapping, BAC clones, and Mitelman database of chromosome aberrations.</p>
 <p><b>Tissues</b> Information on CGAP and other cDNA libraries, gene expression, and SNPs.</p>	 <p><b>SAGE Genie</b> Analysis of gene expression using SAGE data.</p>
 <p><b>Pathways</b> Diagrams of biological pathways and protein complexes, with links to genetic resources for each known protein/enzyme.</p>	 <p><b>Tools</b> Direct access to all the analytic and data mining tools developed for the project.</p>

The CGAP web site provides researchers with access to all CGAP data and biological resources. Briefly, you can find:

- Genomic data for human and mouse, including expressed sequence tags (ESTs), gene expression patterns, single nucleotide polymorphisms (SNPs), cluster assemblies, and cytogenetic information.
- Informatics tools to query and analyze the data.
- Information on methods and resources for reagents developed by the project.

If you have any questions, comments, or need information about CGAP, please contact the [NCI CGAP Help Desk](#).



## Appendix F

### Pubmed main page [25](#)

The screenshot shows the PubMed main page interface. At the top, there are logos for NCBI, PubMed, and the National Library of Medicine (NLM). Below the logos is a navigation bar with tabs for PubMed, Nucleotide, Protein, Genome, Structure, PopSet, Taxonomy, OMIM, and Books. A search bar contains the text 'nelson, p' with 'PubMed' selected in a dropdown menu. To the right of the search bar are 'Go' and 'Clear' buttons. Below the search bar are tabs for 'Limits', 'Preview/Index', 'History', 'Clipboard', and 'Details'. On the left side, there is a blue sidebar with links for 'About Entrez', 'Text Version', 'Entrez PubMed Overview', 'Help | FAQ', 'Tutorial', 'New/Noteworthy', 'E-Utilities', 'PubMed Services', 'Journals Database', 'MeSH Browser', 'Single Citation Matcher', 'Batch Citation Matcher', 'Clinical Queries', 'LinkOut', 'Cubby', 'Related Resources', 'Order Documents', 'NLM Gateway', 'TOXNET', 'Consumer Health', 'Clinical Alerts', 'ClinicalTrials.gov', 'PubMed Central', and 'Privacy Policy'. The main content area features a list of search tips, a yellow box with information about PubMed's citation history, and two highlighted sections: 'Bookshelf Additions' and 'PubMed Changes'. The 'Bookshelf Additions' section includes three book covers and a list of three new books available for interactive searches. The 'PubMed Changes' section describes updates to the search interface and MEDLINE citations. At the bottom, there are links to 'Write to the Help Desk', 'NCBI | NLM | NIH', 'Department of Health & Human Services', 'Freedom of Information Act', and 'Disclaimer'.

Search PubMed for

Limits Preview/Index History Clipboard Details

- Enter one or more search terms, or click [Preview/Index](#) for advanced searching.
- Enter [author names](#) as smith jc. Initials are optional.
- Enter [journal titles](#) in full or as MEDLINE abbreviations. Use the [Journals Database](#) to find journal titles.

PubMed, a service of the National Library of Medicine, provides access to over 12 million MEDLINE citations back to the mid-1960's and additional life science journals. PubMed includes links to many sites providing full text articles and other related resources.

#### Bookshelf Additions

There Are Three New Books Now Available for Interactive Searches on **Bookshelf** :

1. Basic Neurochemistry.
2. The NCBI Handbook.
3. The Human ATP-Binding Cassette (ABC) Transporter Superfamily.

#### PubMed Changes

The Order, Clip Add, Text, and Save buttons in PubMed have been replaced by selections under a new Send to menu. A new select page function includes links for Previous and Next page display.

Also, MEDLINE citations in PubMed have been modified to reflect 2003 MeSH vocabulary.

See [New/Noteworthy](#) for additional information.

[Write to the Help Desk](#)  
[NCBI](#) | [NLM](#) | [NIH](#)  
[Department of Health & Human Services](#)  
[Freedom of Information Act](#) | [Disclaimer](#)

## Pubmed results page <sup>25</sup>



PubMed | Nucleotide | Protein | Genome | Structure | PopSet | Taxonomy | OMIM | Books

Search PubMed for [ ] Go Clear

Limits Preview/Index History Clipboard Details

Display Abstract Show: 20 Sort Send to File

1: J Natl Cancer Inst 2002 Nov 6;94(21):1641-7 [Related Articles, Links](#)

### Effects of the herbal extract PC-SPES on microtubule dynamics and paclitaxel-mediated prostate tumor growth inhibition.

**Bonham MJ, Galkin A, Montgomery B, Stahl WL, Agus D, Nelson PS.**

Division of Human Biology, Fred Hutchinson Cancer Research Center, Seattle, WA 98109-1024, USA.


**BACKGROUND:** PC-SPES is a botanical preparation shown to have efficacy in patients with androgen-dependent and androgen-independent prostate carcinoma. Several herbal constituents in PC-SPES inhibit tumor growth through cell cycle arrest and apoptosis, although the mechanisms of these activities are poorly defined. We sought to identify PC-SPES-induced changes in gene expression, specifically in those genes encoding cytoskeletal proteins that could be associated with PC-SPES-induced cytotoxicity. **METHODS:** LNCaP prostate carcinoma cells were treated with PC-SPES, and changes in gene expression were determined by complementary DNA (cDNA) microarray hybridization and northern blot analyses. PC-SPES and paclitaxel, a microtubule-stabilizing drug, effects on microtubules were assessed by immunofluorescence of treated cells and by in vitro tubulin polymerization assays. In vivo effects of PC-SPES and paclitaxel were assessed using CWR22R androgen-independent prostate cancer xenografts. All statistical tests were two-sided. **RESULTS:** PC-SPES treatment of LNCaP cells for 24 hours altered the expression of 17 cytoskeletal genes. mRNA levels of alpha-tubulin decreased sevenfold. Although paclitaxel stabilized and PC-SPES treatment disrupted microtubule architecture in LNCaP cells, the combination of both agents had an intermediate effect. PC-SPES inhibited tubulin polymerization in vitro, even in the presence of paclitaxel. Compared with tumors in control mice (mean tumor volume = 2983 mm<sup>3</sup>, 95% confidence interval [CI] = 2380 to 3586 mm<sup>3</sup>), tumors were statistically significantly smaller in mice that received PC-SPES (mean tumor volume = 2018 mm<sup>3</sup>, 95% CI = 1450 to 2568 mm<sup>3</sup>; P = .028), paclitaxel (mean tumor volume = 1340 mm<sup>3</sup>, 95% CI = 697 to 1983 mm<sup>3</sup>; P < .001), or the combination of PC-SPES and paclitaxel (mean tumor volume = 1955 mm<sup>3</sup>, 95% CI = 1260 to 2650 mm<sup>3</sup>; P = .034). **CONCLUSION:** PC-SPES may interfere with microtubule polymerization. This activity has implications for the clinical management of patients with advanced prostate cancer who may be taking PC-SPES concurrently with microtubule-modulating chemotherapeutic agents, such as paclitaxel.

PMID: 12419791 [PubMed - indexed for MEDLINE]

Display Abstract Show: 20 Sort Send to File

## Appendix G



### OMIM results page [26](#)



OMIM #600116  
Text  
References  
Contributors  
Creation Date  
Edit History

- Clinical Synopsis
- Gene map

LinkOut



OMIM  
Online Mendelian Inheritance in Man

PubMed	Nucleotide	Protein	Genome	Structure	Popset	Taxonomy	OMIM
--------	------------	---------	--------	-----------	--------	----------	------

Search  for

Limits Preview/Index History Clipboard Details

Display  Show:  Send to

[#600116](#) Links


**PARKINSON DISEASE, JUVENILE, AUTOSOMAL RECESSIVE; PDJ**


*Alternative titles; symbols*

**PARKINSON DISEASE, AUTOSOMAL RECESSIVE EARLY-ONSET  
PARKINSONISM, JUVENILE; JP  
PARKINSONISM, EARLY-ONSET, WITH DIURNAL FLUCTUATION; EPDF  
PARKINSON DISEASE 2, INCLUDED; PARK2, INCLUDED**

Gene map locus [6q25.2-q27](#)

**TEXT**

A number sign (#) is used with this entry because of evidence that a form of autosomal recessive juvenile Parkinson disease is caused by mutations in the parkin gene (PARK2; [602544](#)). Other loci for autosomal recessive early-onset Parkinson disease include PARK6 ([605909](#)) and PARK7 ([606324](#)), both of which map to 1p. 


An autosomal recessive form of familial juvenile parkinsonism was described in a Japanese family by [Takahashi et al. \(1994\)](#). Juvenile parkinsonism (JP) is defined as parkinsonian manifestations with onset before age 40 years. JP is symptomatically different in several aspects from Parkinson disease (PD; [168600](#)), although classic symptoms of PD, such as bradykinesia, rigidity, and tremor, are present. [Takahashi et al. \(1994\)](#) commented that the familial occurrence in Japanese JP cases was approximately 40 to 50% and that the inheritance pattern appeared to be mostly autosomal recessive. They reported a family in which 4 of 5 sibs were affected and the parents were first cousins. A full pathologic examination of 1 of the sibs, a 67-year-old woman, was presented. Changes were confined to the substantia nigra pars compacta and locus ceruleus. The former region showed obvious neuronal loss and gliosis in the medial and ventrolateral regions. In the remainder of that region and in the locus ceruleus, the population of neurons was reduced and there was low melanin content in most of the neurons but no detectable gliosis or extraneuronal free melanin pigment suggestive of a neurodegenerative process. There were no Lewy bodies. The entire pathologic picture was different from that of Lewy body Parkinson disease ([168601](#)). This patient had been well until about the age of 10 years when gait disturbance appeared. By age 14, she was unable to walk long distances. By her forties, she was unable to walk without assistance. There was no evidence of dementia. She had been slow-moving and had shown frozen gait and tremor, more evident on motion, in the head and upper and lower limbs. She showed improvement of the movement disorder after waking up in the morning. When she was young, the improvement lasted until evening, but as she aged it became progressively shorter, eventually lasting only about 10 minutes. In the other sibs, gait disturbance began at the age of 8 or 9 years. One sister had died at age 42 years and a brother at the age of 27 years, both in a bedridden state. 






## Appendix H


### Ensembl main page <sup>27</sup>



project **Ensembl**



The Wellcome Trust  
Sanger Institute




EMBL

---

**Ensembl Genome Browser**

#### About Ensembl



Ensembl is a joint project between [EMBL - EBI](#) and the [Sanger Institute](#) to develop a software system which produces and maintains automatic annotation on eukaryotic genomes. Ensembl is primarily funded by the [Wellcome Trust](#). Access to all the data produced by the project, and to the software used to analyse and present it, is provided free and without constraints.

Ensembl presents up-to-date sequence data and the best possible automatic annotation for eukaryotic genomes. Available now are [human](#), [mouse](#), [zebrafish](#), and [mosquito](#). Others will be added soon.

For an introduction to the Ensembl project, take the [Ensembl tour](#), and then go through a step-by-step [worked example](#) which introduces Ensembl's main functions. For more information read this short [paper](#) in Nucleic Acids Research.

For all enquiries, please contact the Ensembl [HelpDesk](#) ([helpdesk@ensembl.org](mailto:helpdesk@ensembl.org)).

#### Ensembl provides ....

- ▶ Easy access to sequence data
- ▶ For known genes, predicted structure and location in the genome sequence
- ▶ Prediction of novel genes, all with supporting evidence
- ▶ Annotation of other features of the genome
- ▶ Targetted connections to other genome resources worldwide

#### Easy access to the data via ....

- ▶ A web-based genome browser (which can be customized as required)
- ▶ A web-based system for data export and data mining
- ▶ 'Dumps' of sequence and other data sets for you to download
- ▶ Direct access to the databases
- ▶ A Perl-based object layer

#### Ensembl Species

<b>Human</b>	v. 9.30a.1	2 Dec 2002
<b>Mouse</b>	v. 9.3a.1	2 Dec 2002
<b>Rat</b>	v. 9.1.1	25 Nov 2002
<b>Zebrafish</b>	v. 9.08.1	18 Nov 2002
<b>Fugu</b>	v. 9.1.1	18 Nov 2002
<b>Mosquito</b>	v. 9.1a.1	2 Dec 2002

Access to whole genome shotgun data (includes additional species) [Trace Server](#)

#### Help and documentation

- ▶ Species-specific documentation is available via the species home pages above.
- ▶ Take the [Ensembl tour](#), go through a step-by-step [worked example](#), or read this short [paper](#) in Nucleic Acids Research.
- ▶ For context-sensitive help on any web page click: Help
- ▶ There is also an [index](#) of context-sensitive help pages, and a set of guided [How do I...?](#) trails.

Recent Ensembl news [News](#)

Multi-species data retrieval [Data Mining](#)

Display your own data in Ensembl [DAS](#)

Apollo genome browser [Apollo](#)

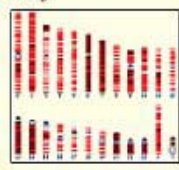
Questions or suggestions? Try the [Help Desk](#) [Help Desk](#)

Documentation (includes tutorial on direct data access & instructions for installing Ensembl on your own site) [Documentation](#)

#### Have you tried ....?

**Human Assembly History**

View the changes in genome coverage between the human assembly releases



[Click for more information](#)

# Ensembl tool page [27](#)

**Human Genome Browser**

**Ensembl Entry Points**

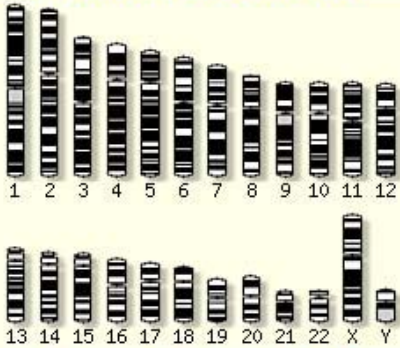
Search for  with

Display Chr  From  To

Retrieve a sequence

BLAST your sequence

**Browse a Chromosome**



1 2 3 4 5 6 7 8 9 10 11 12  
13 14 15 16 17 18 19 20 21 22 X Y

**Documentation & Help**

About Ensembl

For context-sensitive help on any web page click

Questions or suggestions? Try

Documentation (includes tutorial on direct data access & instructions for installing Ensembl on your own site)


**Current Release 9.30a.1**

This release is based on the NCBI 30 assembly of the human genome.

View the [status history](#) of the human assemblies.  
Last Update: 2-12-2002

Ensembl gene predictions: 22980  
GenScan gene predictions: 73128  
Ensembl gene exons: 204094  
Ensembl gene transcripts: 27628  
Contigs: 59936  
Clones: 26407  
Base Pairs: 3342501213  
Golden Path Length: 2815843060


**Ensembl Links and Site Map**



**Other Species**

## Appendix I

### KEGG main page [28](#)



## KEGG - Table of Contents

---

[ [PATHWAY](#) | [GENES](#) | [SSDB](#) | [EXPRESSION](#) | [BRITE](#) | [LIGAND](#) | [DBGET](#) ]

Search  for

### 1. Pathway Information

#### 1.1. Pathway Maps and Ortholog Tables -- [Update Status](#)

Category	Pathway Map Ortholog Table	Search & Compute	DBGET Search
Pathway	<a href="#">Metabolic pathways</a>	Search objects in pathway maps Color objects in pathway maps Search or color genes in ortholog tables	PATHWAY
	<a href="#">Regulatory pathways</a>	Search similar sequences in pathway maps Search similar sequences in ortholog tables Generate possible reaction pathways	

#### 1.2. Hierarchical Classifications

Category	Classification	DBGET Search	LIGAND	Chemscape Search
Disease	ICD disease classification	OMIM gene map OMIM morbid map		
Ortholog	KEGG Orthology (KO)	Gene Ontology (GO) COG		
Enzyme	EC number classification	PIR superfamilies SCOP 3D-folds PROSITE motifs	ENZYME	
Reaction			REACTION	REACTION
Compound	Compound classification			
Element	Periodic table		COMPOUND	COMPOUND

### 2. Genomic Information

#### 2.1. Genomes in KEGG

Category	Genome	Search & Analyze	DBGET Search
Organism	<a href="#">Complete genomes in KEGG</a>	Compare two genome maps Identify gene clusters in two genomes	GENOME



## Appendix J

### EMBL main page [29](#)

GRENOBLE · HAMBURG · HEIDELBERG · HINXTON · MONTEROTONDO



[info@embl.de](mailto:info@embl.de)

**EMBL at a glance**

**EMBL staff**

**Job opportunities**

**Alumni association**

**Visitors programme**

**International PhD Programme**

**Contact us**

EMBL  
Meyerohofstrasse 1  
D-69117 Heidelberg  
Germany  
Tel. + 49 6221 3870  
Fax. + 49 6221 387306  
E-mail [info@embl.de](mailto:info@embl.de)

**Related organisations**

EMBO  
EIROforum

# EMBL

European Molecular Biology Laboratory

## RESEARCH IN MOLECULAR BIOLOGY



Research	Services
<b>Research groups</b>	<b>EMBL services</b>
<b>Research programmes</b>	<b>Biological databases at EBI</b>
<b>Browse keywords</b>	<b>Computational services</b>
<b>EMBL Main Laboratory</b>	<b>Core facilities</b>
<b>EMBL outstations</b>	<b>Szilárd Library</b>
<b>Interdisciplinary research</b>	<b>Technology Transfer</b>

**Upcoming Events**

**Today at EMBL**

**Seminars**

**Courses, conferences and workshops**

**Science and Society at EMBL**

**Search**

**Current News**

12 December 2002

**Talking genes:  
New database shares  
DNA chip results**

 [Download PDF](#)



 [DNA chips teaching kit](#)

 [Surfing genomes teaching kit](#)

4 December 2002

**The power of two:  
Draft mouse genetic code provides a powerful tool for understanding our own genome**

 [Download PDF](#)

**News archives**

**EMBL Publications**

EMBL Internal  
Local Information  
EMBL Services

Updated: 03-12-2002

**RESEARCH IN MOLECULAR BIOLOGY**

Design Jason Soffe© EMBL 2002[Webmaster@embl.de](mailto:Webmaster@embl.de)



## List of Services

[computing](#) - [database](#) - [DNA translation](#) - [gene prediction](#) - [mail](#) - [protein colourer](#) - [protein motif discovery](#) - [protein searches](#) - [search](#) - [sequence alignment](#) - [sequence analysis](#) - [sequence retrieval](#) - [sequence search](#) - [sequence statistics](#) - [similarity searches](#) - [structure prediction](#) - [web](#)

### computing

<a href="#">database applications</a>	web-aware database applications (these pages, conference registrations...)	<a href="#">Computer &amp; Networking Group</a>
<a href="#">SSH applet</a>	Access to the gateway machine ssh.embl-heidelberg.de via a java applet.	<a href="#">Computer &amp; Networking Group</a>

### database

<a href="#">reptile database</a>	An online information resource on reptile taxonomy with a focus on the species level	
----------------------------------	--	--

### DNA translation

<a href="#">Protein Engine</a>	Translates DNA to Protein	<a href="#">Stoehr Group</a>
--------------------------------	---------------------------	------------------------------

### gene prediction

<a href="#">Genemark (at ebi)</a>	Gene prediction service external: <a href="#">? HELP</a>	<a href="#">Stoehr Group</a>
-----------------------------------	--	------------------------------

### mail

<a href="#">MailService</a>	EMBL account owners can set an automatic mail reply message (vacation mail), check their mail quota, set a mail forwarding and manage mailinglists.	<a href="#">Computer &amp; Networking Group</a>
<a href="#">webmail</a>	Service for EMBL account owners to check the incoming mailbox using a web-front end. This service can be used while being outside EMBL (e.g. on a conference...) or from within EMBL.	<a href="#">Computer &amp; Networking Group</a>

### protein colourer

<a href="#">Protein-Colourer</a>	Tool for colouring your amino acid sequences	<a href="#">Stoehr Group</a>
----------------------------------	--	------------------------------

### protein motif discovery

<a href="#">Pratt (at ebi)</a>	Protein motif discovery external: <a href="#">? HELP</a>	<a href="#">Stoehr Group</a>
--------------------------------	--	------------------------------

### protein searches

<a href="#">Fingerprintscaan (at ebi)</a>	PRINTS Protein fingerprint searches external: <a href="#">? HELP</a>	<a href="#">Stoehr Group</a>
<a href="#">PPsearch</a>	Protein motifs searches external: <a href="#">? HELP</a>	<a href="#">Stoehr Group</a>

### search

<a href="#">BioWURLd</a>	Database of resources in bioinformatics external: <a href="#">? HELP</a>	<a href="#">Stoehr Group</a>
--------------------------	--	------------------------------

### sequence alignment

<a href="#">Alignment tools</a>		<a href="#">Bork Group</a>
<a href="#">ClustalW</a>	Multiple Sequence Alignment external: <a href="#">? HELP</a>	<a href="#">Stoehr Group</a>