Searching NCBI Databases Using Entrez

Gretchen Gibney[1] and Andreas D. Baxevanis[1]

[1]Bethesda, Maryland

ABSTRACT

One of the most widely used interfaces for the retrieval of information from biological databases is the NCBI Entrez system. Entrez capitalizes on the fact that there are pre-existing, logical relationships between the individual entries found in numerous public databases. The existence of such natural connections, mostly biological in nature, argued for the development of a method through which all the information about a particular biological entity could be found without having to sequentially visit and query disparate databases. Two basic protocols describe simple, text-based searches, illustrating the types of information that can be retrieved through the Entrez system. An alternate protocol builds upon the first basic protocol, using additional, built-in features of the Entrez system, and providing alternative ways to issue the initial query. The support protocol reviews how to save frequently issued queries. Finally, Cn3D, a structure visualization tool, is also discussed. *Curr. Protoc. Bioinform.* 34:1.3.1-1.3.25. © 2011 by John Wiley & Sons, Inc.

Keywords: Entrez • NCBI databases • biological databases • integrated information retrieval

INTRODUCTION

One of the most widely used interfaces for the retrieval of information from biological databases is the NCBI Entrez system. Entrez capitalizes on the fact that there are pre-existing, logical relationships between the individual entries found in numerous public databases. For example, a paper in MEDLINE (or, more properly, PubMed) may describe the sequencing of a gene whose sequence appears in GenBank. The nucleotide sequence, in turn, may code for a protein product whose sequence is stored in the protein databases. The three-dimensional structure of that protein may be known, and the coordinates for that structure may appear in the structure database. Finally, the gene may have been mapped to a specific region of a given chromosome, with that information being stored in a mapping database. The existence of such natural connections, mostly biological in nature, argued for the development of a method through which all the information about a particular biological entity could be found without having to sequentially visit and query disparate databases.

Basic Protocols 1 and 2 describe simple, text-based searches, illustrating the types of information that can be retrieved through the Entrez system; Basic Protocol 2 also illustrates the use of Cn3D, a viewer that is used to visualize three-dimensional structures. The Alternate Protocol builds upon Basic Protocol 1, using additional, built-in features of the Entrez system, as well as alternative ways of issuing the initial query. The Support Protocol reviews how to save frequently issued queries.

QUERYING ENTREZ

BASIC PROTOCOL 1

The Entrez Web interface is located at *http://www.ncbi.nlm.nih.gov/Entrez*. In addition, Entrez search boxes are provided at the top of many of the pages comprising the NCBI Web site, including the NCBI home page (*http://www.ncbi.nlm.nih.gov*). The best way

Current Protocols in Bioinformatics 1.3.1-1.3.25, June 2011
Published online June 2011 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/0471250953.bi0103s34

Table 1.3.1 Entrez Boolean Search Statements[a,b]

Tag	Significance/usage
[ACCN]	Accession
[AD]	Affiliation
[ALL]	All fields
[AU] or [AUTH]	Author name
	O'Brien J [AUTH] yields all of O'Brien JA, O'Brien JB, etc.
	"O'Brien J" [AUTH] yields only O'Brien J
[ECNO]	Enzyme Commission or Chemical Abstract Service numbers
[FKEY]	Feature key (nucleotide only)
[GENE]	Gene name
[ISS]	Issue of journal
[JOUR]	Journal title, official abbreviation, or ISSN number
	Journal of Biological Chemistry
	J Biol Chem
	0021-9258
[LA]	Language
[MAJR]	MeSH Major Topic
	One of the major topics discussed in the article
[MDAT]	Date of most recent modification to Entrez
	YYYY/MM/DD, YYYY/MM, or YYYY
[MOLWT]	Molecular weight of a protein, in Daltons
[MH]	MeSH Terms
	Controlled vocabulary of biomedical terms (subject)
[ORGN]	Organism
[PS]	Personal name as subject
	Use when name is subject of article, e.g., Varmus H [PS]
[PDAT]	Publication date
	YYYY/MM/DD, YYYY/MM, or YYYY
[PROT]	Protein name (not available in Structure database)
[PT]	Publication type
	Review
	Clinical Trial
	Lectures

continued

Table 1.3.1 Entrez Boolean Search Statements,[a,b] *continued*

Tag	Significance/usage
	`Letter`
	`Technical Publication`
`[SH]`	Subheading
	Used to modify MeSH terms
	`hypertension [MH] AND toxicity [SH]`
`[SUBS]`	Substance name
	Name of chemical discussed in article
`[WORD]`	Text words
	All words and numbers in the title and abstract, MeSH terms, subheadings, chemical substance names, personal name as subject, and MEDLINE secondary sources
`[TITL]`	Title word
	Only words in the definition line (not available in Structure database)
`[UID]`	Unique Identifiers (PMID/MEDLINE numbers)
`[VOL]`	Volume of journal

[a] General syntax is as follows: `search term [tag] [Boolean operator] search term [tag] [Boolean operator]...`

[b] `[Boolean operator]` = `AND`, `OR`, or `NOT`. [tag] represents a tag chosen from the left column of the table above.

to illustrate the integrated nature of the Entrez system and to drive home the power of neighboring (see Commentary) is by considering three biological examples, described in Basic Protocols 1 and 2 and the Alternate Protocol.

Necessary Resources

Software

An up-to-date Web browser, such as Firefox, Internet Explorer, or Safari

Select and search an Entrez database

1. Begin at the Entrez home page (*http://www.ncbi.nlm.nih.gov/Entrez*).
2. In the "Search across databases" text box, enter the following:

```
DCC AND "Vogelstein B"
```

Using Boolean operators such as AND, OR, and NOT is the simplest way to query the Entrez system. Note that all Boolean operators must be capitalized for the query to return the expected results.

In this example, the query will return all available information on the DCC gene in which the term "Vogelstein B" is also found. If the [AU] qualifier were included after the search term "Vogelstein B," the search would instruct Entrez to look only at the author field of entries for the occurrence of "Vogelstein B." A list of this and other available search qualifiers is given in Table 1.3.1. Note that, when using qualifiers, the square parentheses ([ ]) are required. See Table 1.3.1 for examples of how to specify author names and the use of wildcards with author names.

Figure 1.3.1 The Entrez unified results page, showing the number of hits to each of Entrez's component databases fitting the query. Clicking on any of the numbers to the left of the database name takes the user to the results found in that particular database.

3. Click the Go button to submit the query. Running the query in February 2011 returned 14 papers (indicated by the "14" next to the PubMed entry at top of the left-hand column of the list of databases), 18 nucleotide entries (indicated by the "18" next to "Nucleotide" in the list), and 16 protein entries (indicated by the "16" next to "Protein" in the list); see Figure 1.3.1.

   Entrez does a query of all of the available databases, and the number of hits to each database is shown to the left of the name of the database. The user can further narrow down the query by adding additional terms, if interested in a more specific aspect of this gene or if there are simply too many entries returned by the initial query. The reader is encouraged to carefully look at Figure 1.3.1 to see the broad array of biological resources that are available through the Entrez system. The name of each component database is followed by a brief description of its contents. A longer description can be found by clicking on any of the question mark icons to the right of each database entry in the table.

Viewing individual database entries

4. Click on the "14" next to "PubMed."

   This will produce the view seen in Figure 1.3.2. For each of the found papers in PubMed, the user is presented with the title of the paper, the authors on the paper, and the citation.

5. Clicking on any of the hyperlinked titles will take the user to the Abstract view of the selected paper, which presents the name of the paper, the list of authors, their institutional affiliation(s), and the abstract itself, in standard ("Abstract") format. Here, click on the title of the fourth reference in the list, "The DCC gene: structural analysis and mutations in colorectal carcinomas," by Kathy Cho and colleagues (Cho et al., 1994). This will produce the view seen in Figure 1.3.3.

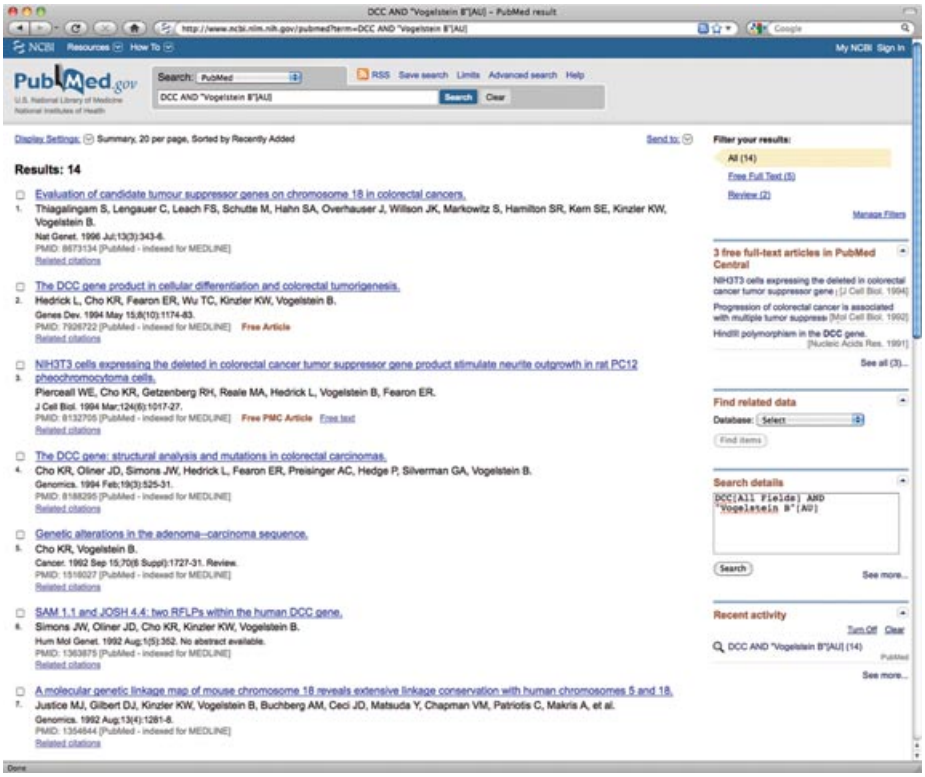

Figure 1.3.2 Results of a text-based Entrez query using Boolean operators against PubMed. The initial query (from Fig. 1.3.1) is shown in the search box near the top of the window. Each entry gives the title of the paper, names of the authors, and the citation information. An individual record can be viewed by clicking on the hyperlinked title of that paper.

6. To expand the display, click on the plus sign below the abstract that is next to the heading "Publication Types, MeSH Terms, Substances, Secondary Source ID, Grant Support." Now, cataloging information, such as the MeSH terms and indexed substances relating to the entry, is displayed below the abstract.

   To see a drop-down menu of additional display options, click on the Display Settings link on the left above the abstract. Choose MEDLINE from this drop-down menu. This selection produces the MEDLINE/MEDLARS layout, with two-letter codes corresponding to the contents of each field going down the left-hand side of the entry (e.g., the author field is denoted by the code AU). Entries in this format can be saved and easily imported into third-party bibliography-management programs, such as EndNote and Reference Manager.

7. Return to the Abstract view by clicking the browser's Back button.

8. To view the full text of the article, click on the Full-Text Article icon located to the right of the journal citation (shown in Fig. 1.3.3). With the proper individual or institutional privileges, the user will be able to view the entire text of a paper, including all figures and tables, as well as download a PDF version of the paper.

Finding related material

9. The right-hand side of the page shown in Figure 1.3.3 provides a partial list of articles that have been deemed similar to the Cho et al. (1994) publication displayed here. To see the complete list of related papers, click on the "See all" link, which can be found at the bottom of the Related Citations section, in the sidebar to the right. Entrez will return a list of 118 references of similar subject matter (again, as of February 2011); the first seven of these are shown in Figure 1.3.4.

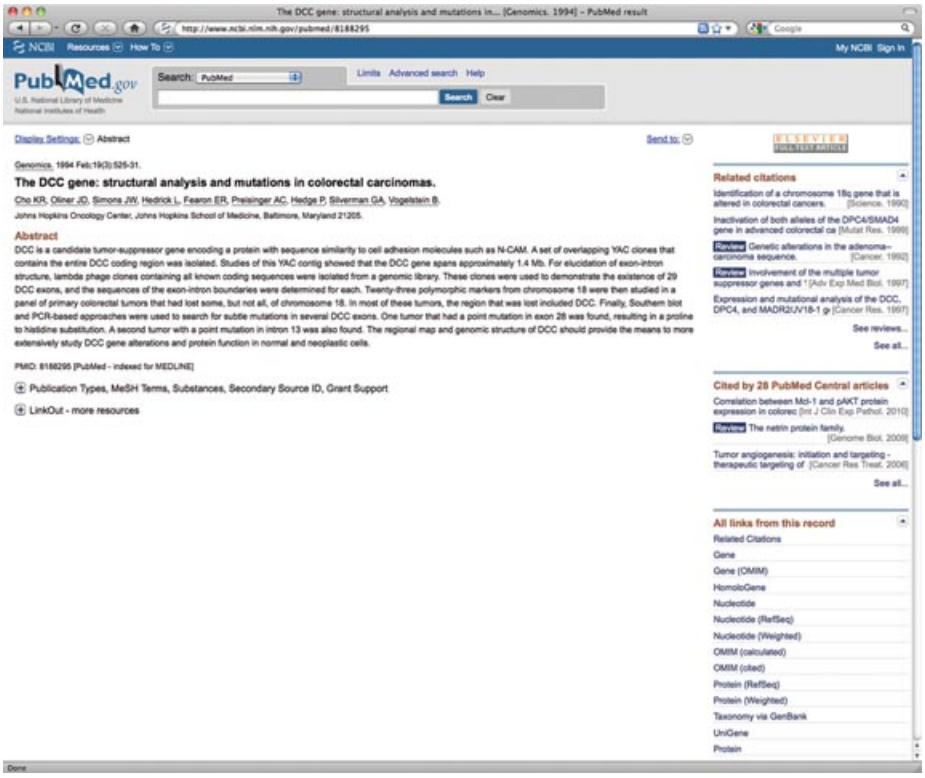

Figure 1.3.3 An example of a PubMed record in Abstract format, as returned through Entrez. This Abstract view is for the fourth reference shown in Figure 1.3.2. The view provides connections to related articles, sequence information, and the actual, full-text journal article. See text for details.

The first paper in the list is the same Cho et al. paper, since, by definition, it is most related to itself (the "parent" entry). The order in which the related papers follow is based on statistical similarity. Thus, the entry closest to the parent is deemed to be the closest in subject matter to the parent. By scanning the titles, the user can easily find related information on other studies, as well as quickly amass a bibliography of relevant references. This can be a useful and time-saving function if the user is writing grants or papers, because abstracts can easily be scanned and papers of real interest identified before the user heads off for the library stacks.

10. Click on the browser's Back button to return to the Abstract view.

11. On the lower right-hand side of the Abstract view page is a section called "All links from this record." The list of items in this section represents hard-link connections to other databases within the Entrez system. Select the Gene link to obtain the page shown in Figure 1.3.5.

    This page is from Entrez Gene, a feature of Entrez that provides links to a wealth of information about the gene in question. The data are gathered from a variety of sources, including NCBI's RefSeq. The screen shows that DCC is the official symbol of a protein-coding gene for a netrin 1 receptor. Summary information on the genomic region, transcripts, and products of the DCC gene is presented graphically, with genomic coordinates provided. Additional content not shown in the figure includes a link that presents DCC in MapViewer, NCBI's genome browser. By scrolling down the Gene page, the user will find, as well, links to functional information, associated phenotypes, information on protein-protein interactions, Gene Ontology assignments, and homologies to selected organisms. Shortcut links to these areas of the page can be found under the Table of Contents section at the top portion of the right-hand sidebar. Further down on the right sidebar are sections titled "Links" and "Links to other resources," which provide extensive lists of links to other related resources provided through NCBI and other sources.

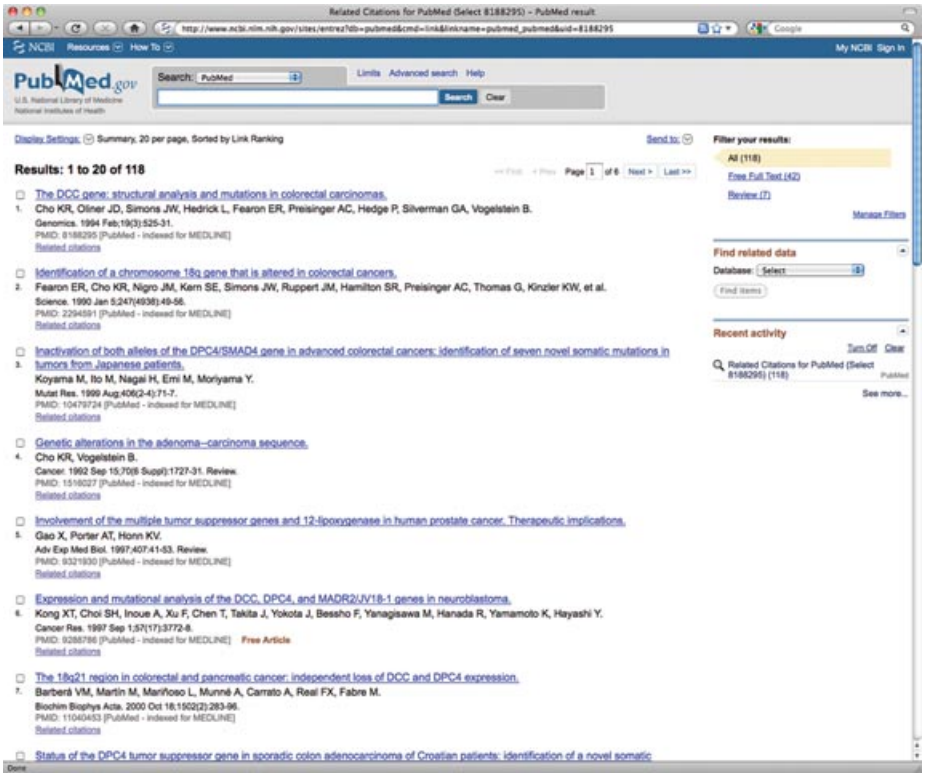

Figure 1.3.4 Related citations for an entry found in PubMed. The original entry from Figure 1.3.3 (Cho et al., 1994) is at the top of the list, indicating that this is the parent entry.

12. To view documented single-nucleotide polymorphisms (SNPs) in this gene, click on the "SNP: GeneView" link located in the Links section of the right sidebar. This will produce the view seen in Figure 1.3.6.

    The information on this page is derived from the Database of Single Nucleotide Polymorphisms (dbSNP; Mullikin and Sherry, 2005). The individual SNPs that occur within the DCC gene are shown in the table at the bottom of the figure; each SNP occupies two lines of the table, with one line showing the "contig reference" (the more common allele) and the other showing the SNP (the less common allele). Details are given in the figure legend. Additional information on dbSNP may be found in UNIT 1.19.

13. Return to the Abstract view, and this time click on Protein (RefSeq) in the Links section of the right-hand sidebar. This retrieves one entry, for the protein sequence having the accession number NP_005206. The RefSeq entry is shown in Figure 1.3.7.

    The NCBI RefSeq project is intended to provide a reference sequence for each molecule in the central dogma (DNA, mRNA, and protein). Additional information on RefSeq can be found in the Commentary at the end of this unit.

14. Return to the Abstract view, and now select OMIM (cited) from the right sidebar links. This retrieves one entry from OMIM, having the accession number *120470, shown in Figure 1.3.8.

    Figure 1.3.8 shows the entry for DCC in the Online Mendelian Inheritance in Man (OMIM) database (McKusick, 1998; Hamosh et al., 2002; UNIT 1.2). OMIM provides concise textual information from the published literature on most human conditions having a genetic basis, as well as pictures illustrating the condition or disorder (where appropriate) and full citation information. Each entry includes information such as the gene symbol, alternate names for the disease, a description of the disease, a clinical synopsis, and references.

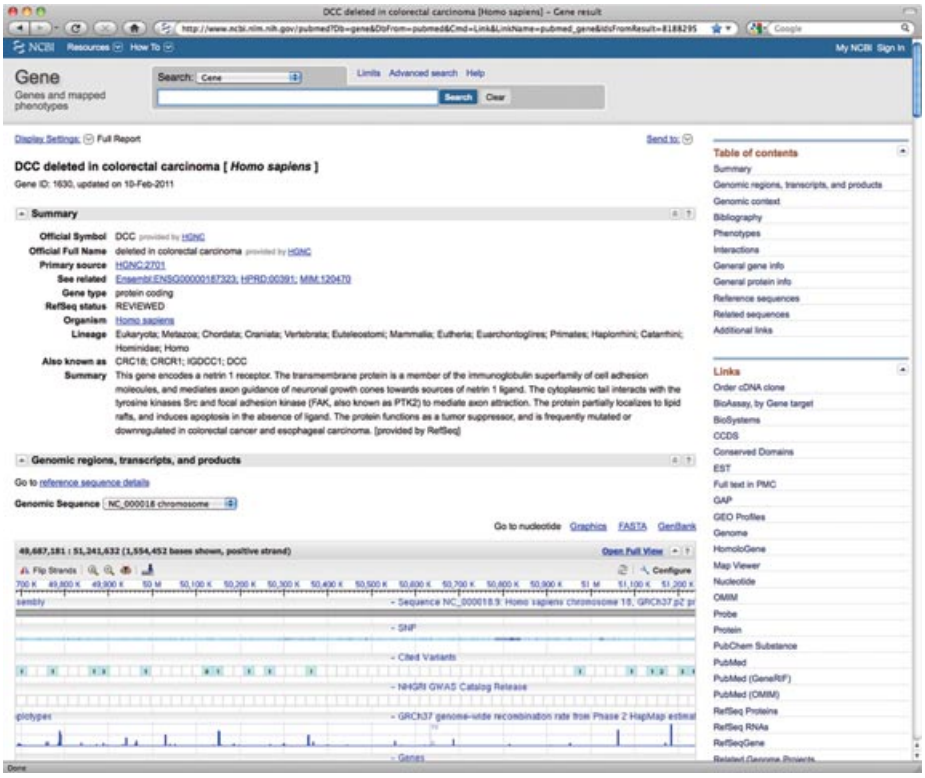

Figure 1.3.5 The Entrez Gene page for the DCC (deleted in colorectal carcinoma) gene. The screen shows that this is a protein-coding gene and provides information on the genomic context of DCC and the encoded protein. An extensive collection of links to other NCBI and external databases is provided along the right-hand side of the window. See text for details.

15. Click on the Allelic Variants link in the Table of Contents in the right sidebar. This will jump the user down to a section on the same Web page labeled Allelic Variants (Fig. 1.3.9).

    A particularly useful feature of OMIM is the list of allelic variants that appears in many OMIM entries. A short description is given, after each allelic variant, of the clinical or biochemical outcome of that particular mutation. There are currently >2500 gene entries containing at least one allelic variant that either causes or is associated with a discrete phenotype in humans. Note that the allelic variants shown in Figure 1.3.9 produce significantly different clinical outcomes (two distinct types of cancer and a brain-related motor disorder).

16. Return to the Abstract view, and this time select GEO Profiles from the links in the right sidebar.

    The Gene Expression Omnibus (GEO) at NCBI serves as a public repository for array-based data, including microarray and chromatin immunoprecipitation (ChIP-chip) data (Barrett et al., 2005). The GEO records shown in Figure 1.3.10 all have accession numbers beginning with the code GDS, indicating that they are "GEO Data Sets," i.e., curated sets of GEO data. The GPL numbers refer to the "GEO Platforms," the actual list of elements on the array (e.g., cDNAs or oligonucleotides). The reader is strongly encouraged to read the online GEO Overview (http://www.ncbi.nlm.nih.gov/geo/info/overview.html), which provides descriptions of the various GEO elements and information on how to browse GEO data.

17. Return to the Abstract view and click on the plus sign below the abstract that is next to the heading, "LinkOut – more resources." This action expands a menu of links to additional resources.

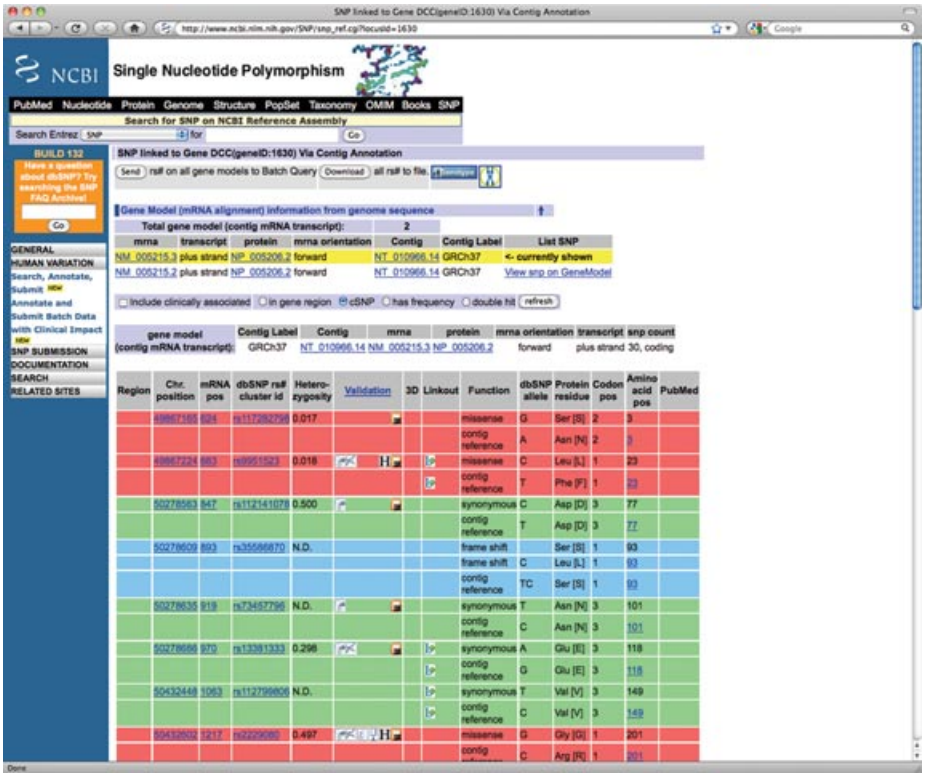

Figure 1.3.6 The dbSNP GeneView page for the DCC gene. The information on individual SNPs is shown in the table towards the bottom of the screen. Each SNP occupies two lines of the table, with one line showing the "contig reference" (the more common allele) and the other showing the SNP (the less common allele). For example, the first two rows in the table show a contig reference A for which there is a documented SNP, changing the A to a G. At the protein level, this changes the amino acid at position 3 of the DCC protein from asparagine to serine. The rows are colored red since this is a "nonsynonymous SNP;" that is, the SNP produces a discrete change at the amino acid level. In contrast, the fifth and sixth rows of the table are shown in green, indicating that this record is for a "synonymous SNP;" the entries describe a SNP where the contig reference (T) and the SNP allele (C) ultimately produce the same amino acid (Asp). For the color version of this figure, go to *http://www.currentprotocols.com/protocol/bi0103*.

LinkOut provides a list of third-party Web sites and resources related to the Entrez query being viewed, such as full text of articles that can be displayed directly through the Web browser, or the ability to order the document through online services. Through LinkOut, users can also obtain information that is particularly useful to patients and clinicians, as will be demonstrated in the next step.

18. From the LinkOut menu, click on the MedlinePlus link for Colorectal Cancer. This generates a page devoted to information for both laymen and health professionals on DCC and disorders related to DCC (Fig. 1.3.11).

    The information available through this page is often appropriate to provide to patients, since the level of writing is geared towards nonprofessionals. There are also interactive tutorials for various procedures related to DCC along the right-hand side of the page.

19. Scroll down the page to find the Clinical Trials section of the MedlinePlus page. Click on the "ClinicalTrials.gov: Colorectal Neoplasms" link, found in this section. This will spawn a new window, taking the user to NIH's central information source for clinical trials, aptly named *ClinicalTrials.gov* (Fig. 1.3.12).

    The listing shown in Figure 1.3.12 gives the first ten of the 796 clinical trials actively recruiting for patients with colorectal neoplasms at the time of this writing. Clicking on

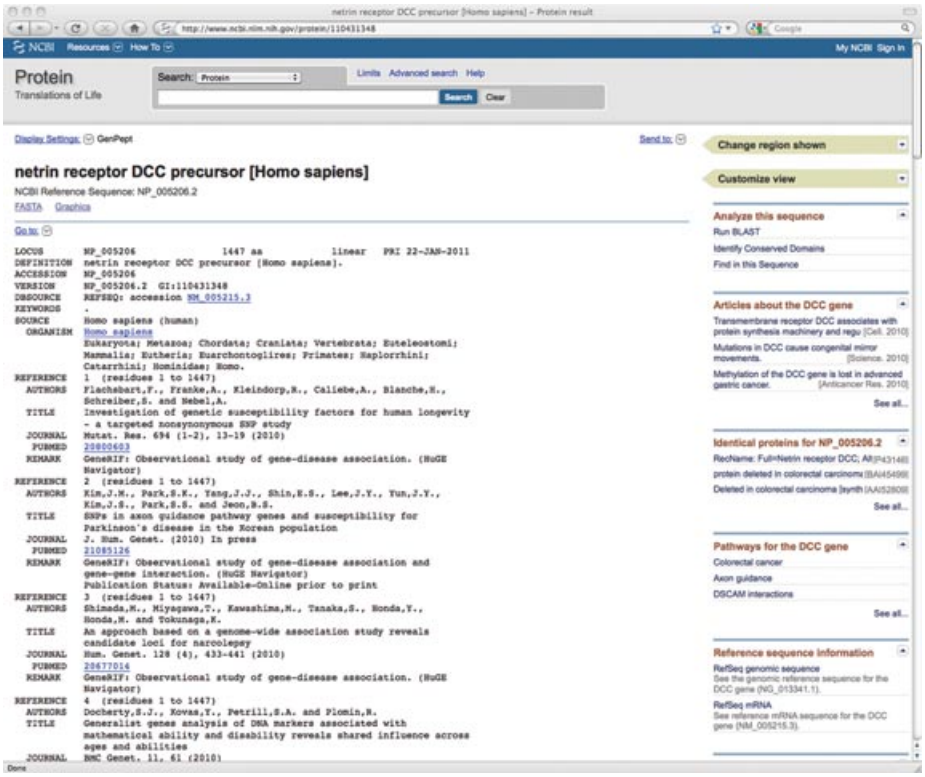

Figure 1.3.7 The RefSeq protein entry corresponding to the original Cho et al. (1994) publication shown in Figure 1.3.2, in GenPept format. See text for details.

the name of the protocol gives information regarding the study, including the principal investigator's name and contact information. Clicking on the Results on the Map tab at the top of the page produces a map of the world, showing how many clinical trials are being conducted in each region or country; this view is useful in identifying trials that are geographically close to a potential study subject's home. While scientists tend to focus on the kind of biological data discussed throughout the rest of this unit, the clinical trials site is, unarguably, the most important of the sites covered here, since it provides a means through which patients suffering from a given genetic or metabolic disorder can receive the latest, cutting-edge treatment, which may make a substantial difference in their quality of life.

SUPPORT PROTOCOL

USING My NCBI TO SAVE SEARCHES AND RESULTS

A storage service called My NCBI is provided to save searches and their corresponding results. The advantage of the My NCBI system is that it can recall the searches that were saved and update them with a click of a mouse, eliminating the need to re-enter the query each time the user wishes to view the most recent results. The system is also capable of e-mailing users updated search results at specified intervals and filtering results based on user-specified criteria. Each user may store up to 100 searches.

Necessary Resources

Software

An up-to-date Web browser, such as Firefox, Internet Explorer, or Safari

Register and log in

1. After executing an Entrez search, such as the one described in Basic Protocol 1, click the Save Search link that appears above the search box at the top of the results page (Fig. 1.3.2).

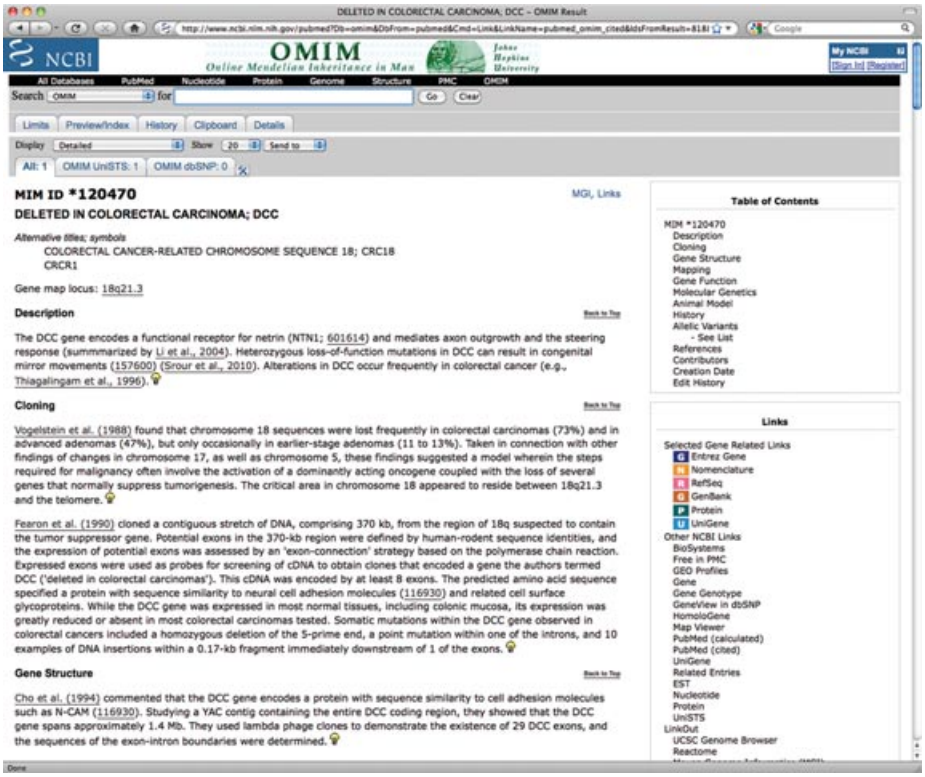

Figure 1.3.8 The OMIM entry for the DCC gene. Each entry includes information such as the gene symbol, alternate names for the disease, a description of the disease, a clinical synopsis, and references.

For purposes of this example, return to the view shown in Figure 1.3.2, which shows the original set of papers found through the initial Entrez search. Clicking the Save Search link will take the user to the My NCBI account page.

2a. *If not registered with My NCBI:* Click the "Register for an account" link, then follow the instructions presented on the screen. Once registered, the user is automatically logged into My NCBI.

2b. *If registered, but not already logged in:* Enter a valid username-password pair and click the Sign In button. The login will remain active for the length of the current session (as long as the browser window is open), unless the "Keep me signed in unless I sign out" box is checked.

Store a My NCBI search

3. The user will be prompted for a name for the search. Upon saving, the window will expand, prompting the user for an e-mail address, the frequency with which to receive updated results, the desired e-mail format, and the number of entries that should be sent.

4. Enter the desired parameters and click Save to complete the process. A confirmatory e-mail will be sent to the e-mail address specified in the Save Search Settings window, asking the user to verify having originated the request and whether to still send the updates via e-mail.

Retrieve and update a My NCBI stored search

5. Click the My NCBI link located at the upper right of most NCBI pages (see Fig. 1.3.2).

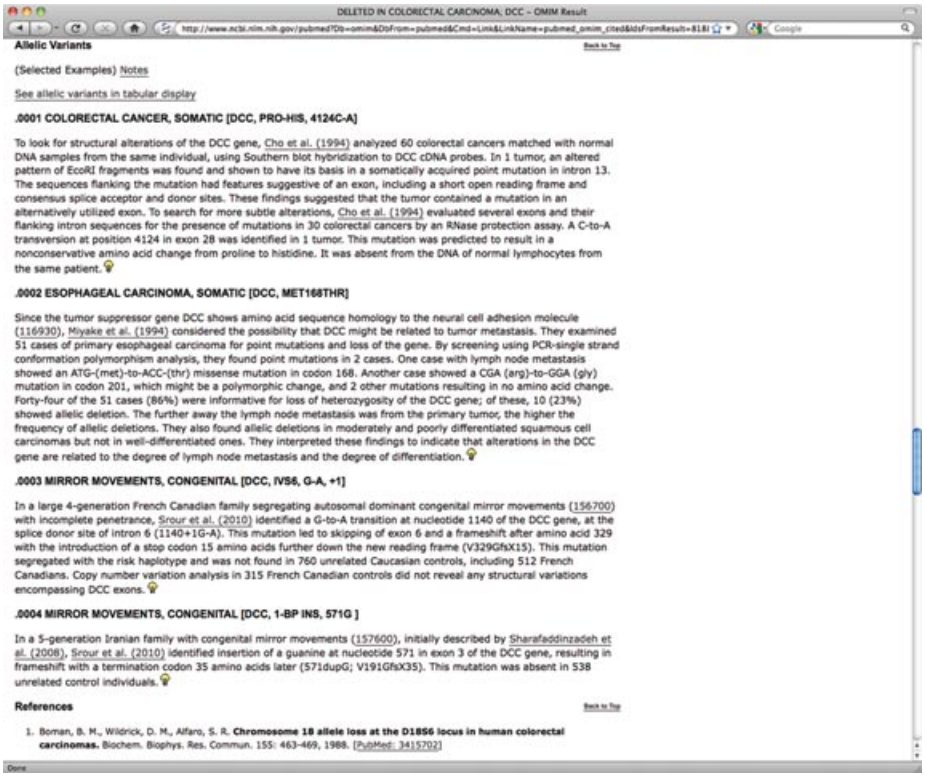

Figure 1.3.9 An example of a list of allelic variants that can be obtained through OMIM. The figure shows the four allelic variants for the DCC gene, two leading to cancers of the digestive tract and two that are associated with a movement disorder. The description under each allelic variant provides information specific to that particular mutation.

The user will be taken to the user's My NCBI page, where clicking on the link to Saved Searches under My Saved Data opens a page showing the user's saved searches. In this case (Fig. 1.3.13), the one search that was stored in step 4 is shown, along with the date it was last updated and the update frequency. The last set of results for this search can be displayed by clicking on the search name, and the update frequency can be changed by clicking on the displayed frequency (here, monthly).

6. Select the search by clicking on the check box to the left of the stored search name, and then click the "Show What's New" button. This will re-run the search and return a count of how many new papers have been added since the last time the query was issued. If no new papers have been found, the message "No new items" will be shown.

7. To view the new papers found, select the number new link. The date and time of the query are now updated to reflect the current date and time. If this link is not selected, the date and time of the query will not be updated.

ALTERNATE PROTOCOL

COMBINING ENTREZ QUERIES

There is another way to perform an Entrez query, involving some built-in features of the system. Consider an example in which one is attempting to find all genes coding for DNA-binding proteins in *Methanothermobacter*. Although this example concentrates on nucleotide sequences, the general strategy works equally well across any of Entrez's component databases.

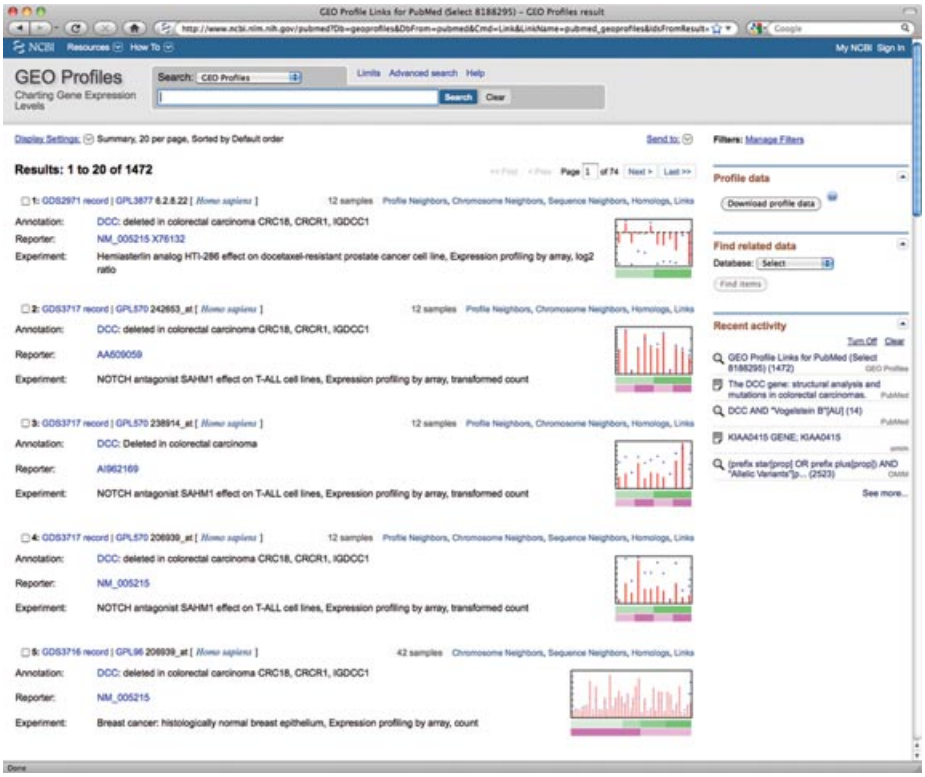

Figure 1.3.10 Gene Expression Omnibus (GEO) DataSets for the DCC gene. For each DataSet, a brief description of the experiment is provided, as well as a schematic of the gene expression profile derived in the study.

Necessary Resources

Software

An up-to-date Web browser, such as Firefox, Internet Explorer, or Safari

Execute multiple queries

1. Go to the NCBI home page (*http://www.ncbi.nlm.nih.gov*). Select Nucleotide from the pull-down menu of the search bar at the top of the page and enter the term `DNA-binding` in the text box. Click Search.

   In February 2011, the query returned 188,948 entries (Fig. 1.3.14).

2. To narrow the query, click on the Limits link, which is located directly above the search box.

   This brings the user to a new page (Fig. 1.3.15) that allows the search to be refined or "limited," as implied by the name of the page.

3. To limit the search by organism, select Organism from the Search Field Tags drop-down menu.

4. Enter `methanothermobacter` in the text box of the search bar at the top of the screen (Fig. 1.3.15), and then click Search.

   In February 2011, the query returned 542 entries (Fig. 1.3.16).

Combine the selected queries

5. Click on the advanced search hyperlink, located above the search box. The Search History section of the Advanced Search page displays the user's most recent queries (Fig. 1.3.17).

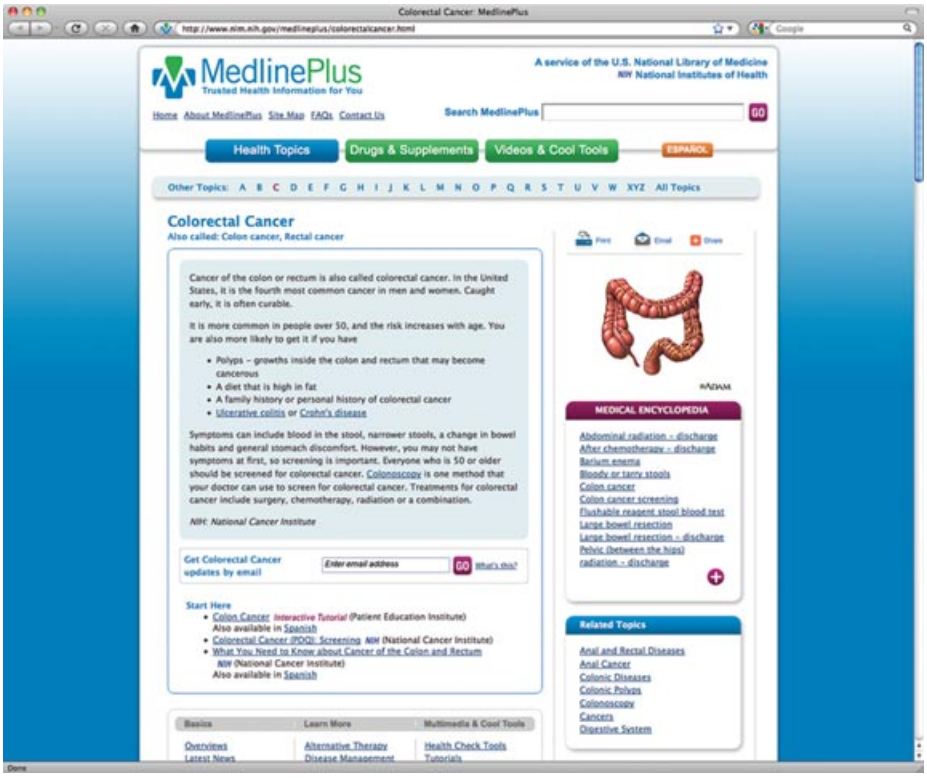

Figure 1.3.11 The MedlinePlus page devoted to information for both laymen and physicians on DCC and disorders related to DCC. The information available through this page is often much more appropriate to provide to patients, since the level of writing is geared towards nonprofessionals. Often, MedlinePlus entries include interactive tutorials for various procedures related to the disease of interest.

The list shows the individual queries, whether those queries were field-limited, the time at which the query was performed, and how many entries that query returned.

6. To combine the two queries into one, use the query numbers (in this particular case, the queries are numbered 17 and 18; readers may see different numbers on their screens). Enter the queries using their number preceded by a pound sign (`#17 AND #18`) in the Search Box at the top of the screen. Click Preview to regenerate a table showing the new, combined query containing six entries. Clicking on the hyperlinked 6 in the Result column takes the user to the six entries common to the two original queries, as shown in Figure 1.3.18.

BASIC PROTOCOL 2

EXAMINING STRUCTURES IN ENTREZ

Structure queries can be accomplished simply by selecting Structure from the Search drop-down menu on the NCBI home page. For the example below, assume that the user is trying to find information regarding the structure of HMG-box B from the rat, whose PDB accession number is 1HMF.

Necessary Resources

Software

An up-to-date Web browser, such as Firefox, Internet Explorer, or Safari

1. Go to the NCBI home page (*http://www.ncbi.nlm.nih.gov*) and select Structure from the Search drop-down menu. Enter `1HMF` in the text box. Click Search.

Figure 1.3.12 The ClinicalTrials.gov page showing actively recruiting clinical trials relating to colorectal neoplasms. Information on each trial, including the principal investigator of the trial and qualification criteria, can be found by clicking on the name of the trial.

2. Click the 1HMF hyperlink on the Structure search results page. The structure summary page displays, and the user will immediately note the format, which is decidedly different from any of the pages displayed so far (Fig. 1.3.19).

   This page shows the description line from the source Molecular Modeling Database (MMDB) document (which is derived from PDB), as well as links to PubMed and to the taxonomy of the source organism. The graphic below the header schematically illustrates the protein as a bar of length 77 (meaning 77 amino acids), below which is a bar showing the position of a defined domain within the protein (here, the HMG box, a DNA-binding domain). A stylized version of the structure of the protein is shown at left.

3. Click on the upper bar corresponding to the full-length protein (marked Sequence A). This displays a table of 33 structure neighbors, as assessed by the Vector Alignment Search Tool (VAST; see Background Information).

4. To glean initial impressions about the shape of the protein, download Cn3D by clicking Download Cn3D. The downloaded application will walk the user through the installation.

   More information on Cn3D is available through the online Cn3D documentation. In addition, the user can save coordinate information to a file and view the data using third-party applications, such as RasMol (UNIT 5.4).

5. Once installed, use the Web browser's Back button to return to the 1HMF structure summary page (Fig. 1.3.19). Click on the Structure View in Cn3D button, directly below the stylized representation of the protein. This will launch the Cn3D viewer once the three-dimensional coordinates of 1HMF have been downloaded from the NCBI server.

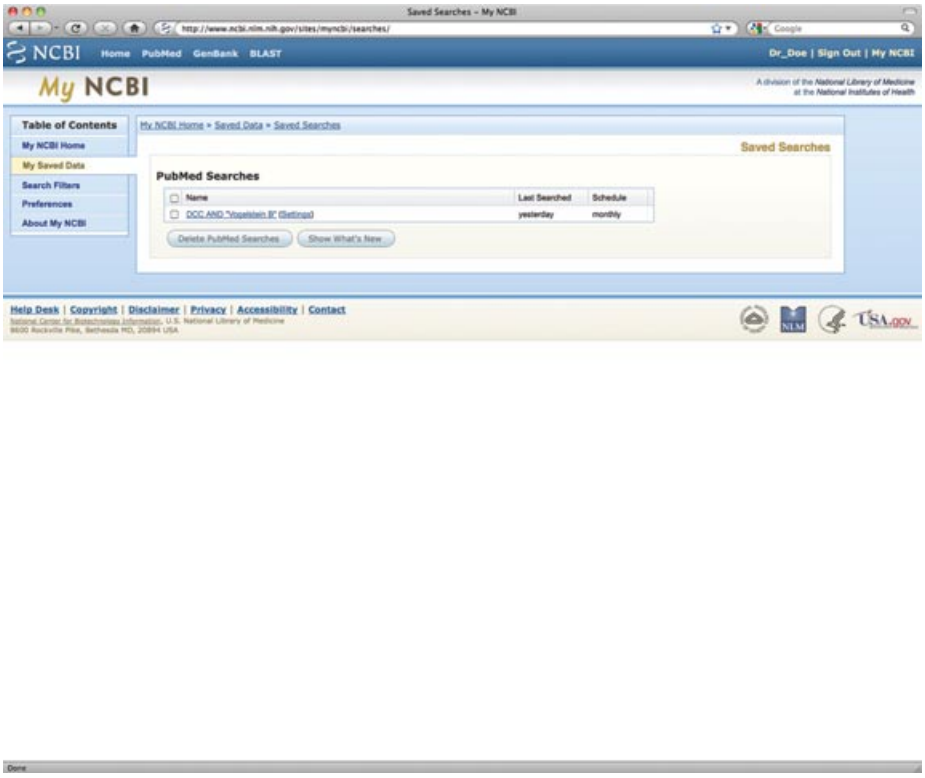

Figure 1.3.13 Searches saved through My NCBI can be recalled, viewed, and updated through the Saved Searches option under the My Saved Data on the user's My NCBI page. See text for details.

If the Cn3D application does not automatically launch, click on the `mmdbsrv.cgi` *file that has been downloaded to the desktop. This will prompt the user to select the Cn3D application, after which Cn3D will launch.*

6. Cn3D will produce two windows, one showing the structure of 1HMF, the other showing the sequence (Fig. 1.3.20A). The user can highlight any part of the sequence shown in the sequence window, and the corresponding part of the structure will appear in yellow. In addition, the user can search the sequence for a specific sequence pattern. Click anywhere in the Sequence/Alignment Viewer window, then select View > Find Pattern. Type `PKRP` into the search box, and then click OK. The portion of the backbone corresponding to these four residues (Pro 7-Pro 10) will be shown in yellow in the structure window.

7. The user can also adjust the display of the structure by selecting options in the Style > Rendering Shortcuts and Style > Coloring Shortcuts submenus. Figure 1.3.20B shows the structure of 1HMF with the Rendering Shortcut set to Spacefill and the Coloring Shortcut set to Charge.

   The view given in Figure 1.3.20B shows the overall C-shape of the protein, which binds to DNA. The blue residues, representing positive charges, indicate the residues that may be responsible for DNA binding through minor-groove interactions; pay particular attention to the two residues that jut into the C-shaped cavity. Negative charges are shown in red, and neutral residues are shown in gray.

8. It is also possible to change the rendering and coloring of defined parts of the molecule.

   a. Reset the Rendering Shortcut to Worms and the Coloring Shortcut to Secondary Structure, restoring the view to that shown in Figure 1.3.20A.

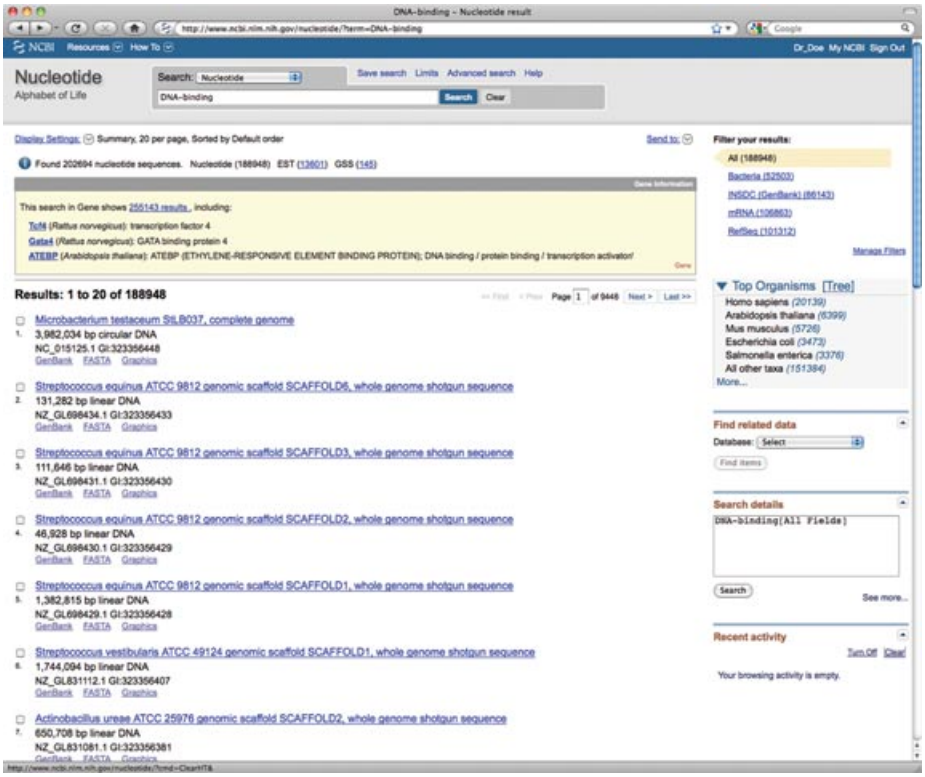

Figure 1.3.14 Formulating a search against the nucleotide portion of Entrez. The initial query is shown in the text box near the top of the window (`DNA-binding`), and the nucleotide entries matching the query are displayed below. See text for details.

b. Highlight the same PKRP residues as before by clicking in the Sequence/Alignment Viewer and selecting the four residues (Pro 7–Pro 10). Click anywhere in the structure window, then select Style > Annotate.

c. In the new User Annotations window, click New. This will produce a new window, titled Edit Annotation.

d. Give the new annotation a name (e.g., PKRP), then click Edit Style. This generates the Style Options window, where settings can be selected for how the PKRP residues should be displayed in the structure window (Fig. 1.3.21, left).

e. Here, change the Protein Backbone rendering to Ball and Stick and the color scheme to Charge. Next, click the box next to Protein Sidechains, change the rendering to Ball and Stick, and the color scheme to Charge. When done, click Done.

f. Click Done in the User Annotations window, then click anywhere in the Sequence/Alignment Viewer to clear the yellow highlighting. The colors will change to correspond to the charges of the individual side chains (Figure 1.3.21, right).

9. Rotate the structure by moving the mouse while holding down the mouse button. To zoom in or out, hold down the Apple key (Mac) or Command key (PC) while dragging the mouse.

COMMENTARY

Background Information

Entrez, to be clear, is not a database itself—rather, it is the interface through which all of its component databases can be accessed and traversed. The Entrez information space includes PubMed records, nucleotide and protein sequence data, three-dimensional structure information, and mapping information.

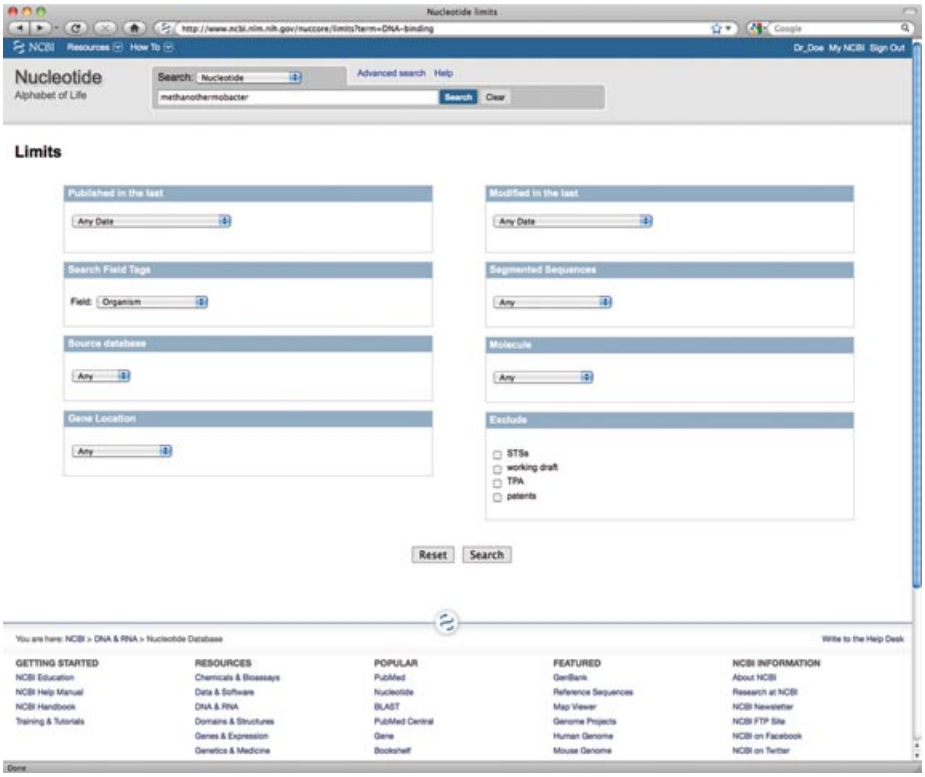

Figure 1.3.15 Using the Limits feature of Entrez to limit a search to a particular organism. See text for details.

The strength of Entrez lies in the fact that all of this information can be accessed by issuing one and only one query. Entrez is able to offer integrated information retrieval through the use of two types of connections between database entries: neighboring and hard links.

Relationships between database entries: Neighboring

The concept of neighboring allows for entries within a given database to be connected to one another. If a user is looking at a particular PubMed entry, the user can ask Entrez to find all of the other papers in PubMed that are similar in subject matter to the original paper (see Basic Protocol 1). Similarly, if a user is looking at a sequence entry, Entrez can return a list of all other sequences that bear similarity to the original sequence. The establishment of neighboring relationships within a database is based on statistical measures of similarity, as follows. While the term "neighboring" is traditionally used to describe these connections, the terminology used on the Entrez Web site will describe neighbors as "related papers," "related sequences," and so forth.

BLAST. Sequence data are compared to one another using the Basic Local Alignment Search Tool, or BLAST (Altschul et al., 1990). This algorithm attempts to find "high-scoring segment pairs" (HSPs), which are pairs of sequences that can be aligned with one another and, when aligned, meet certain scoring and statistical criteria. *UNITS 3.3 & 3.4* discuss the family of BLAST algorithms and their application at length.

VAST. Sets of coordinate data are compared using a vector-based method known as the Vector Alignment Search Tool (VAST; Madej et al., 1995; Gibrat et al., 1996). There are three major steps that take place in the course of a VAST comparison.

First, based on known three-dimensional coordinate data, all of the α-helices and β-sheets that comprise the core of the protein are identified. Straight-line vectors are then calculated based on the position of these secondary-structure elements. VAST keeps track of how one vector is connected to the next (that is, how the C-terminal end of one vector connects to the N-terminal end of the next vector), as well as whether a particular vector represents an α-helix or a β-sheet. Subsequent steps use only these vectors in making comparisons to other proteins. In effect, most of the coordinate data are discarded at this step. The reason for

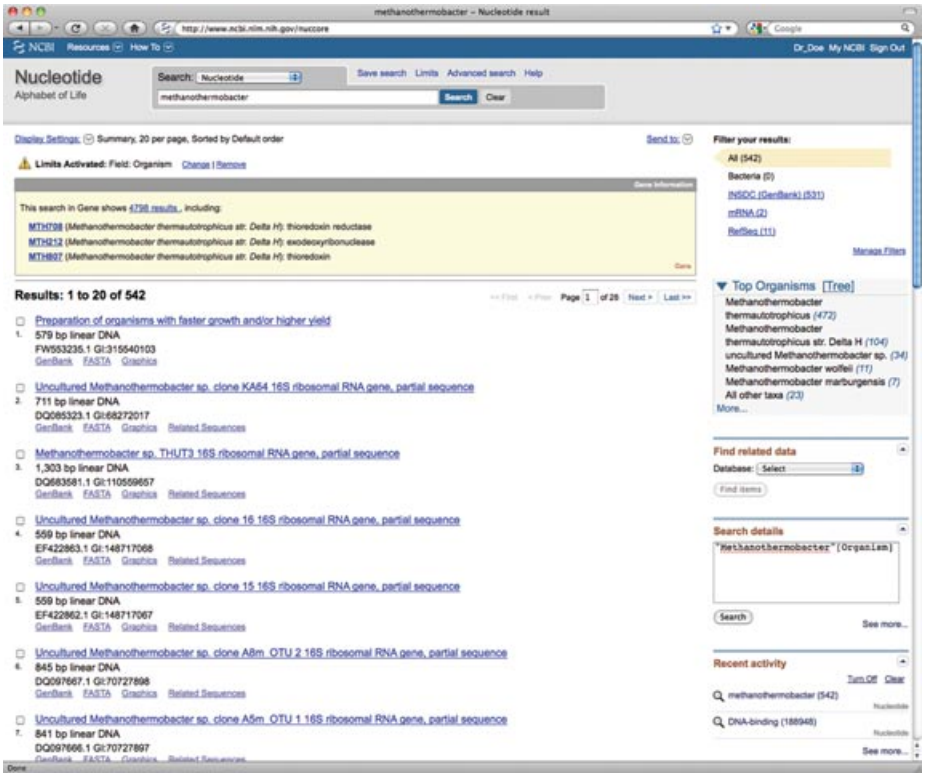

Figure 1.3.16 Results of a limited search against the nucleotide portion of Entrez. The initial query is shown in the text box near the top of the window (`methanothermobacter`), and the nucleotide entries matching the query are displayed below. Note the caution (!) icon next to the words Limits Activated at the top of the results page, indicating that the results displayed have been "limited," here to a particular organism (Fig. 1.3.15). See text for details.

this apparent oversimplification is simply the scale of the problem at hand; with >71,000 structures in PDB that need to be considered, the time that it would take to do an in-depth comparison of each and every structure to all of the other structures in the database would make the calculations both impractical and intractable. The user should keep this simplification in mind when making biological inferences based on the results presented in a VAST table.

Next, the algorithm attempts to optimally align these sets of vectors, looking for pairs of structural elements that are of the same type and relative orientation, with consistent connectivity between the individual elements. The object is to identify highly similar "core substructures," i.e., pairs that represent a statistically significant match above that which would be obtained by comparing randomly chosen proteins to one another.

Finally, a refinement is done using Monte Carlo methods at each residue position in an attempt to optimize the structural alignment.

Through this method, it is possible to find structural (and, presumably, functional) relationships between proteins in cases that may lack overt sequence similarity. The resultant alignment need not be global; matches may be between individual domains of different proteins.

It is important to note here that VAST is not the best method for determining structural similarities. More robust methods, such as homology model building, provide much greater resolving power in determining such relationships, since the raw information within the three-dimensional coordinate file is used to perform more advanced calculations regarding the positions of side chains and the thermodynamic nature of the interactions between side chains. Reducing a structure to a series of vectors necessarily results in a loss of information. However, considering the magnitude of the problem here—again, the number of pairwise comparisons that need to be made—and both the computing power and time needed to employ any of the more advanced methods,

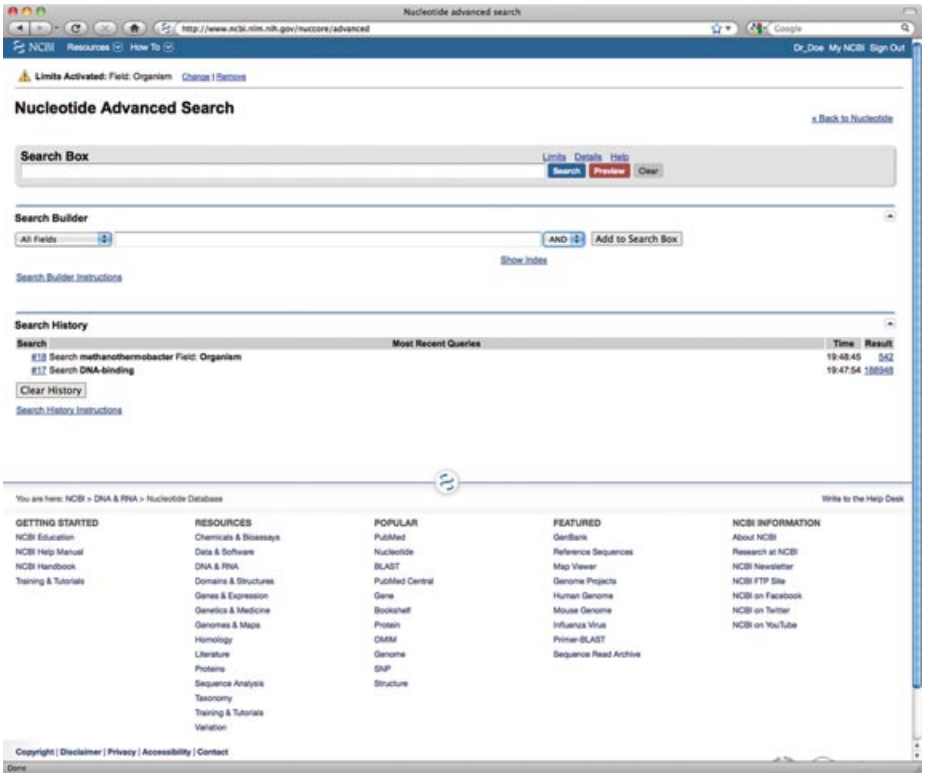

Figure 1.3.17 Combining individual queries using the Advanced Search feature of Entrez. Each search performed in the last 8 hr is saved and given a number in Search History. The searches can be combined using the search numbers and the Boolean operators AND, OR, or NOT. See text for details.

VAST provides a simple and fast first answer to the question of structural similarity. More information on other structure prediction methods based on X-ray or NMR coordinate data can be found in Chapter 5.

Weighted key terms. The problem of comparing sequence data somewhat pales next to that of comparing PubMed entries, free text whose rules of syntax are not necessarily fixed. Given that no two people's writing styles are exactly the same, finding a way to compare seemingly disparate blocks of text poses a substantial problem. Entrez employs a method known as the relevance pairs model of retrieval to make such comparisons, relying on what are known as weighted key terms (Wilbur and Coffee, 1994; Wilbur and Yang, 1996). This concept is best described by example. Consider two manuscripts with the following titles: *BRCA1 as a Genetic Marker for Breast Cancer* and *Genetic Factors in the Familial Transmission of the Breast Cancer BRCA1 Gene*. Both titles contain the terms BRCA1, Breast, and Cancer, and the presence of these common terms may indicate that the manuscripts are similar in their subject matter. The proximity between the words is also taken into account, so that words common to two records that are closer together are scored higher than common words that are further apart. In the current example, the terms Breast and Cancer would score higher based on proximity than either of those words would against BRCA1, since the words are next to each other. Common words found in a title are scored higher than those found in an abstract, since title words are presumed to be "more important" than those found in the body of an abstract. Overall, weighting depends on the frequency of a given word among all the entries in PubMed, with words that occur infrequently in the database as a whole carrying a higher weight.

Hard links

The hard link concept is much easier conceptually than neighboring. Hard links are applied between entries in different databases and exist everywhere there is a logical connection between entries. For instance, if a PubMed

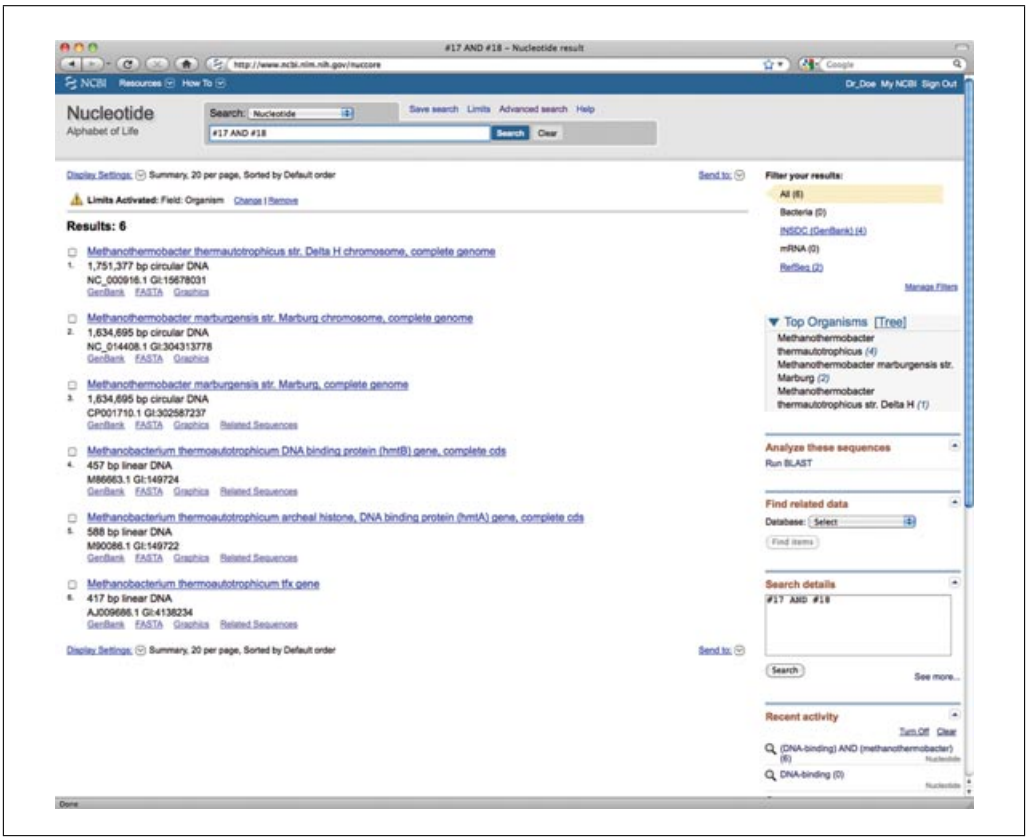

Figure 1.3.18 Entries resulting from the combination of two individual Entrez queries. The query term producing the results is shown in the Search Box near the top of the window (`#17 AND #18`). The numbers correspond to those assigned to the previously performed searches listed in Figure 1.3.17. See text for details.

entry is about the sequencing of a cosmid, a hard link is established between the PubMed entry and the corresponding nucleotide entry. If an open reading frame in that cosmid codes for a known protein, a hard link is established between the nucleotide entry and the protein entry. If, by sheer luck, the protein entry has an experimentally deduced structure, a hard link would be placed between the protein entry and the structural entry. The hard link relationships between databases is illustrated is Figure 1.3.22.

Searches can, in essence, begin anywhere within Entrez—the user has no constraints with respect to where the foray into this information space must begin. However, depending on which database is chosen as the starting point, different fields are available for searching. This stands to reason, inasmuch as the entries in different databases are necessarily organized differently, reflecting the biological nature of the entity that each database is trying to catalog.

RefSeq

While the data derived from systematic sequencing projects and individual investigators' laboratories yields a rich and highly valuable set of sequence-based data, some problems immediately come to the fore. The most important issue in this regard is that a single biological entity may be represented by many different entries in various databases. It also may not be clear whether a given sequence has been experimentally determined or is simply the result of a computational prediction.

To address this issue, NCBI initiated the RefSeq Project, whose major goal has been to provide a reference sequence for each molecule in the central dogma (DNA, mRNA, and protein). Since each biological entity is represented once and only once, RefSeq is, by definition, nonredundant. Nucleotide and protein sequences in RefSeq are explicitly linked to one another. Most importantly, RefSeq entries undergo ongoing curation, assuring that the RefSeq entry represents the

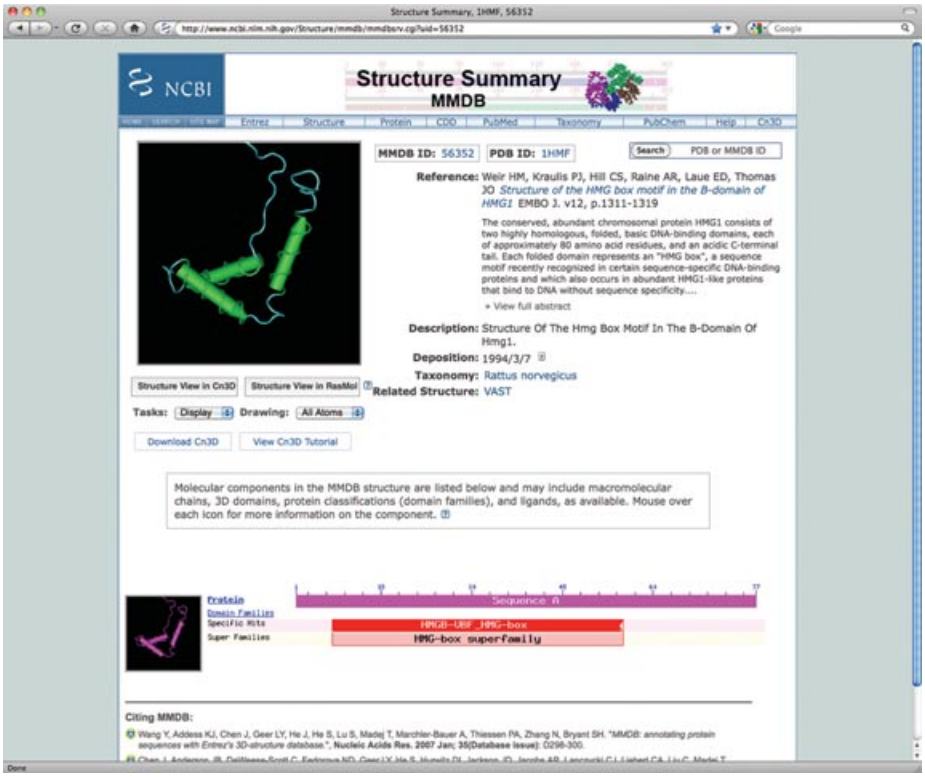

Figure 1.3.19 The structure summary for 1HMF, resulting from a direct query of the structures accessible through the Entrez system. The entry shows header information from the corresponding MMDB entry, links to PubMed, and links to the taxonomy of the source organism. Structure neighbors, as assessed by VAST, can be found by clicking on the long bar (purple on screen) next to the Protein key. The structure itself can be viewed by clicking on the Structure View in Cn3D button, thereby spawning the Cn3D viewer.

most up-to-date state of knowledge regarding a particular DNA, mRNA, or protein sequence.

RefSeq entries are distinguished from other entries in GenBank through the use of a distinct accession number series. RefSeq accession numbers follow a "2 + 6" format: a two-letter code indicating the type of reference sequence, followed by an underscore and a six-digit number. Experimentally determined sequence data are denoted as follows:

`NT_123456` Genomic contigs (DNA)
`NM_123456` mRNAS
`NP_123456` Proteins

Reference sequences derived through genome annotation efforts are denoted as follows:

`XM_123456` Model mRNAs
`XP_123456` Model proteins

It is important that the reader understand the distinction between the "N" numbers and "X" numbers. The former represent actual, experimentally determined sequences, whereas the latter represent computational predictions derived from the raw DNA sequence.

Descriptions of additional types of RefSeq entries, along with more information on the RefSeq Project, can be found on NCBI's RefSeq Web site.

Disclaimer

This unit was written by Drs. Gretchen Gibney and Andreas D. Baxevanis in their private capacity. No official support or endorsement by the National Institutes of Health or the United States Department of Health and Human Services is intended or should be inferred.

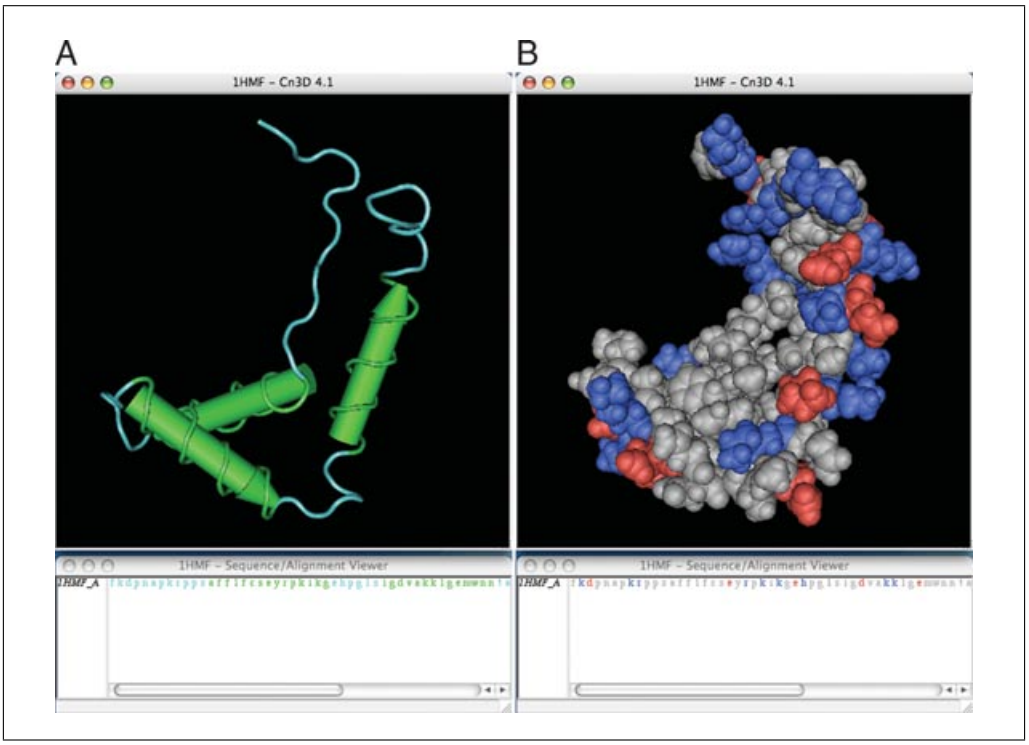

Figure 1.3.20 The structure of 1HMF rendered using Cn3D version 4.1, an interactive molecular viewer. Cn3D can be used as a helper application to any Web browser or as a stand-alone application. In panel A, the backbone of the structure is shown as a worm, with the coloring indicating secondary structural regions; in this case, there are three α-helices, shown in green, with a "crayon" indicating the length and directionality of the helix. Four residues have been highlighted in the sequence window, and those residues are shown in yellow in the structure window. In panel B, the rendering of the structure has been changed, showing the structure in space-filling style, with the coloring being done by charge (red, negative; blue, positive). For both panels, the coloring shown in the structure window is mirrored in the sequence window below. See text for details. For the color version of this figure go to *http://www.currentprotocols.com/protocol/bi0103*.

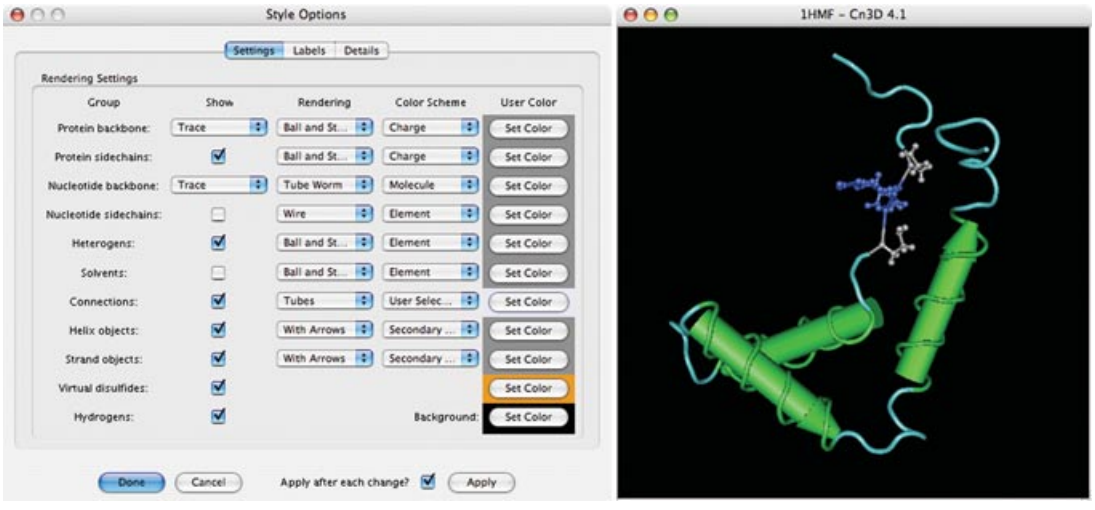

Figure 1.3.21 Changing the rendering and coloring of selected parts of a structure. The Style Options window also allows individual residues to be numbered and the dimensions of side chains and other features to be changed. See text for details. For the color version of this figure go to *http://www.currentprotocols.com/protocol/bi0103*.

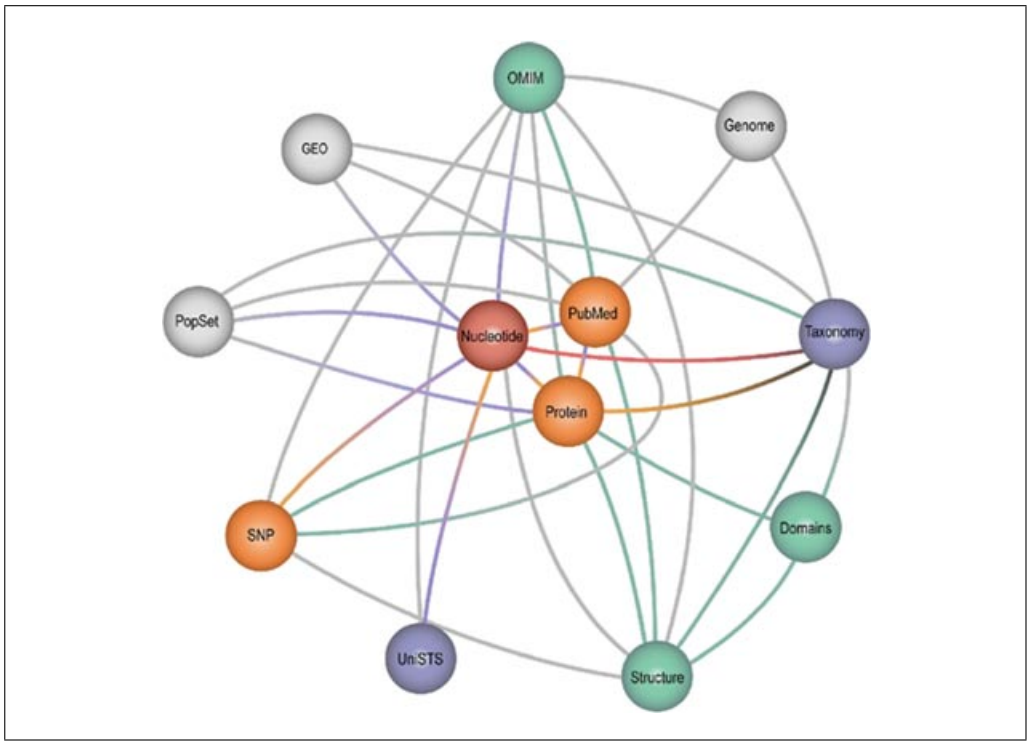

Figure 1.3.22 An overview of the relationships in the Entrez integrated information retrieval system. Each node represents one of the elements that can be accessed through Entrez, and the lines represent how each component database connects to the others. Entrez is under continuous evolution, with new components being added and the interrelationships between the elements changing dynamically. (Figure from The Entrez Search and Retrieval System, The NCBI Handbook; see Internet Resources.) A Flash-based version of this figure can be found at *http://www.ncbi.nlm.nih.gov/Database/datamodel/index.html*.

Literature Cited

Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403-410.

Barrett, T., Suzek, T.O., Troup, D.B., Wilhite, S.E., Ngau, W.C., Ledoux, P., Rudnev, D., Lash, A.E., Fujibuchi, W., and Edgar, R. 2005. NCBI GEO: Mining millions of expression profiles—database and tools. *Nucleic Acids Res.* 33:D562-D566.

Cho, K.R., Oliner, J.D., Simons, J.W., Hedrick, L., Fearon, E.R., Preisinger, A.C., Hedge, P., Silverman, G.A., and Vogelstein, B. 1994. The DCC gene: Structural analysis and mutations in colorectal carcinomas. *Genomics* 19:525-531.

Gibrat, J.-F., Madej, T., and Bryant, S. 1996. Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* 6:377-385.

Hamosh, A., Scott, A.F., Amberger, J., Bocchini, C., Valle, D., and McKusick, V.A. 2002. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 30:52-55.

Madej, T., Gibrat, J.-F., and Bryant, S. 1995. Threading a database of protein cores. *Proteins* 23:356-369.

McKusick, V.A. 1998. Online Mendelian inheritance in man: A catalog of human genes and genetic disorders, 12th Edition. The Johns Hopkins University Press, Baltimore, Maryland.

Mullikin, J.C. and Sherry, S.T. 2005. Sequence polymorphisms. *In* Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, 3rd Edition (A.D. Baxevanis and B.F.F. Ouellette, eds.) pp. 171-193. John Wiley & Sons, Hoboken, New Jersey.

Wilbur, W. and Coffee, L. 1994. The effectiveness of document neighboring in search enhancement. *Inf. Process Manage.* 30:253-266.

Wilbur, W. and Yang, Y. 1996. An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Comput. Biol. Med.* 26:209-222.

Internet Resources

http://www.ncbi.nlm.nih.gov
NCBI Home page.

http://www.ncbi.nlm.nih.gov/Entrez
NCBI Entrez Web page.

http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml
NCBI Cn3D structure viewer.

http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.chapter.ch15
Ostell, J. 2003. The Entrez Search and Retrieval System. The NCBI Handbook, Chapter 15. National Center for Biotechnology Information, Bethesda, Maryland.

http://www.ncbi.nlm.nih.gov/projects/geo/info/overview.html
NCBI GEO Overview.

http://www.ncbi.nlm.nih.gov/RefSeq/
NCBI Reference Sequence (RefSeq) Project.