
INFORMATION RETRIEVAL FROM BIOLOGICAL DATABASES

Andreas D. Baxevanis

*Genome Technology Branch
National Human Genome Research Institute
National Institutes of Health
Bethesda, Maryland*

As discussed earlier in this book, GenBank was created in response to the explosion in sequence information resulting from a panoply of scientific efforts such as the Human Genome Project. To review, GenBank is an annotated collection of all publicly available DNA and protein sequences and is maintained by the National Center for Biotechnology Information (NCBI). As of this writing, GenBank contains 7 million sequence records covering almost 9 billion nucleotide bases. Sequences find their way into GenBank in several ways, most often by direct submission by individual investigators through tools such as Sequin or through “direct deposit” by large genome sequencing centers.

GenBank, or any other biological database for that matter, serves little purpose unless the database can be easily searched and entries retrieved in a usable, meaningful format. Otherwise, sequencing efforts have no useful end, since the biological community as a whole cannot make use of the information hidden within these millions of bases and amino acids. Much effort has gone into making such data accessible to the average user, and the programs and interfaces resulting from these efforts are the focus of this chapter. The discussion centers on querying the NCBI databases because these more “general” repositories are far and away the ones most often accessed by biologists, but attention is also given to a number of smaller, specialized databases that provide information not necessarily found in GenBank.

INTEGRATED INFORMATION RETRIEVAL: THE ENTREZ SYSTEM

The most widely used interface for the retrieval of information from biological databases is the NCBI Entrez system. Entrez capitalizes on the fact that there are preexisting, 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 information about a particular biological entity could be found without having to sequentially visit and query disparate databases.

Entrez, to be clear, is not a database itself—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. 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 connection between database entries: *neighboring* and *hard links*.

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 other papers in PubMed that are similar in subject matter to the original paper. 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.

BLAST. Sequence data are compared with 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. Chapter 8 discusses the family of BLAST algorithms and their application at length.

VAST. Sets of coordinate data are compared using a vector-based method known as VAST, for Vector Alignment Search Tool (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 is discarded at this step. The reason for this apparent oversimplification is simply due to the scale of the problem at hand; with over 11,000 structures in PDB, the time that it would take to do an in-depth comparison of each and every structure with 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,” pairs that represent a statistically significant match above that which would be obtained by comparing randomly chosen proteins with 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, 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 11.

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
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.

Regardless of the method by which the neighboring relationships are established, the ability to actually code and maintain these relationships is rooted in the format underlying all of the constituent databases. This format, called Abstract Syntax Notation (ASN.1), provides a format in which all similar fields (e.g., those for a bibliographic citation) are all structured identically, regardless of whether the entry is in a protein database, nucleotide database, and so forth. This NCBI data model is discussed in depth in Chapter 2.

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 entry talks 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 in Figure 7.1.

As suggested by the figure, 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 used as the jumping-off point, different fields are available for searching. This stands to reason, inasmuch as the entries in databases of different types are necessarily organized differently, reflecting the biological nature of the entity they are trying to catalog.

Implementations

Regardless of platform, Entrez searches can be performed using one of two interfaces. The first is a client-server implementation known as Network Entrez. This is the fastest of the Entrez programs in that it makes a direct connection to an NCBI “dispatcher.” The graphical user interface features a series of windows, and each time a new piece of information is requested, a new window appears on the user’s screen. Because the client software resides on the user’s machine, it is up to the user to obtain, install, and maintain the software, downloading periodic updates as new features are introduced. The installation process itself is fairly trivial. Network Entrez also comes bundled with interactive, graphical viewers for both genome sequences and three-dimensional structures (Cn3D, cf. Chapter 5).

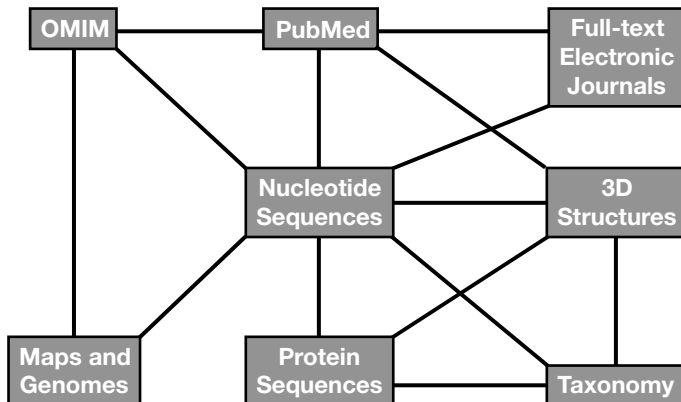


Figure 7.1. Overview of the relationships in the Entrez integrated information retrieval system. Each square represents one of the elements that can be accessed through Entrez, and the lines represent how each element connects to the other elements. Entrez is under continuous evolution, with both new components being added and the interrelationships between the elements changing dynamically.

The second and more widely used implementation is through the World Wide Web. This option makes use of available Web browsers, such as Internet Explorer or Netscape, to deliver search results to the desktop. The use of a Web browser relieves the user of having to make sure that the most current version of Entrez is installed—as long as the browser is of relatively recent vintage, results will always be displayed via the latest Entrez release. The Web naturally lends itself to an application such as this, since all the neighboring and hard link relationships described above can easily be expressed as hypertext, allowing the user to navigate by clicking on selected words in an entry.

The advantage of the Web implementation over the Network version is that the Web allows for the ability to link to external data sources, such as full-text versions of papers maintained by a given journal or press or specialized databases that are not part of Entrez proper. The speed advantage that is gained by the network version causes its limitation in this respect; the direct connection to the NCBI dispatcher means that the user, once connected to NCBI, cannot travel anywhere else. The other main difference between the two methods lies simply in the presentation: the Network version uses a series of windows in presenting search results, whereas the Web version is formatted as sequential pages, following the standard Web paradigm. The final decision is one of personal preference, for both methods will produce the same results *within* the Entrez search space. However, given that Web Entrez can link to external data sources, the remainder of this discussion will focus on the Web implementation.

The Entrez Discovery Pathway: Examples

The best way to illustrate the integrated nature of the Entrez system and to drive home the power of neighboring is by considering two biological examples, using the Web version of Entrez as the interface.

The simplest way to query Entrez is through the use of individual search terms, coupled together by Boolean operators such as AND, OR, or NOT. Consider the case in which one wants to retrieve all papers that discuss aspirin in the context of treating or preventing atherosclerosis. Beginning at the Entrez home page, one would select PubMed from the Search pull-down menu to indicate that the search is to take place in the bibliographic portion of the Entrez search space. Within the text box to the right, one would simply type *atherosclerosis* [MH] AND *aspirin* [NM]. The [MH] qualifying the first term indicates to Entrez that this is a *MeSH term*; MeSH stands for *medical subject heading* and is the qualifier that should be used when searching by subject. The [NM] qualifying the second term indicates that this is the name of a substance or chemical. In this case, the query returned 197 papers (Fig. 7.2; the query is echoed at the top of the new Web page). The user can further narrow down the query by adding additional terms, if the user is interested in a more specific aspect of the pharmacology or if there are quite simply too many papers to read. A list of all available qualifiers is given in Table 7.1.

At this point, to actually look at one of the papers resulting from the search, the user can click on a hyperlinked author's name. By doing so, the user is taken to the Abstract view for the selected paper. Figure 7.3 shows the Abstract view for the first paper in the hit list, by Cayatte et al. The Abstract view presents the name of the paper, the list of authors, their institutional affiliation, and the abstract itself, in standard format. A number of alternative formats are available for displaying this information, and these various formats can be selected using the pull-down menu next to the Display button. Switching to Citation format would produce a very similar-looking entry, the difference being that cataloguing information such as MeSH terms and indexed substances relating to the entry are shown below the abstract. MEDLINE format 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.

At the top of the entry are a number of links that are worth mentioning. First, on the right-hand side is a hyperlink called Related Articles. This is one of the entry points from which the user can take advantage of the neighboring and hard link relationships described earlier. If the user clicks on Related Articles, Entrez will indicate that there are 101 neighbors associated with the original Cayatte reference—that is, 101 references of similar subject matter—and the first six of these papers are shown in Figure 7.4. The first reference in the list is the same Cayatte paper because, by definition, it is most related to itself (the “parent”). The order in which the neighbored entries follows is from most statistically similar downward. 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 that look at the pharmacology of aspirin in atherosclerosis as well as quickly amass a bibliography of relevant references. This is a particularly useful and time-saving functionality when one is writing grants or papers because abstracts can be scanned and papers of real interest identified before one heads off for the library stacks.

The next link in the series is labeled Books, and clicking on that link will take the user to a heavily hyperlinked version of the original citation. The highlighted words in this view correspond to keywords that can take the user to full-text books that are available through NCBI. The first of these books to be made available is

NCBI
 PubMed
 Search

Show:

1: Cayatte AJ, Du Y, Oliver-Krasinski J, Laville G, Verbeuren TJ, Cohen RA. **The thromboxane receptor antagonist S18886 but not aspirin inhibits atherogenesis in apo E-deficient mice: evidence that eicosanoids other than thromboxane contribute to atherosclerosis.** *Atheroscler Thromb Vasc Biol.* 2000 Jul;20(7):1724-8. PMID: 10894809; UI: 200353344 [Related Articles](#)

2: Eincke D. **[ASS or clopidogrel? In high vascular risk it is best to give both].** *MMW Fortschr Med.* 2000 Mar;30:142(13):11. German. No abstract available. PMID: 10783615; UI: 20246045 [Related Articles](#)

3: Walvoort HC. **[Creative mathematics with clopidogrel; exaggeration of the preventive effect by the pharmaceutical company].** *Ned Tijdschr Geneesk.* 2000 Apr 6;144(15):725. Dutch. No abstract available. PMID: 10778723; UI: 20241191 [Related Articles](#)

4: Hachulla E, Piette AM, Hatron PY, Elebry O. **[Aspirin and antiphospholipid syndrome].** *Rev Med Interne.* 2000 Mar;21 Suppl 1:63s-68s. Review. French. PMID: 10763209; UI: 20226419 [Related Articles](#)

5: Tanasesou S, Levesque H, Thuillez C. **[Pharmacology of aspirin].** *Rev Med Interne.* 2000 Mar;21 Suppl 1:168s-26s. Review. French. PMID: 10763201; UI: 20226411 [Related Articles](#)

6: Randi ML, Rossi C, Fabris F, Girolami A. **Essential thrombocythemia in young adults: major thrombotic complications and complications during pregnancy--a follow-up study in 68 patients.** *Clin Appl Thromb Hemost.* 2000 Jan;6(1):31-5. PMID: 10726046; UI: 20190245 [Related Articles](#)

7: Exner M, Herrmann M, Hofbauer R, Kapiotis S, Speiser W, Held J, Seelos C, Gmeiner BM. **The salicylate metabolite gentisic acid, but not the parent drug, inhibits glucose autooxidation-mediated atherogenic modification of low density lipoprotein.** [Related Articles](#)

[About Entrez](#)
[Entrez PubMed Overview](#)
[Help / FAQ](#)
[New/Noteworthy](#)
[PubMed Services](#)
[Journal Browser](#)
[MeSH Browser](#)
[Single Citation](#)
[Batch Citation](#)
[Batch Citation Matcher](#)
[Clinical Queries](#)
[Old PubMed](#)
[Related Resources](#)
[Order Documents](#)
[Grateful Med](#)
[Consumer Health](#)
[Clinical Alerts](#)
[ClinicalTrials.gov](#)
[Privacy Policy](#)

Figure 7.2. A text-based Entrez query using Boolean operators against PubMed. The initial query is shown in the search box near the top of the window. Each entry gives the names of the authors, the title of the paper, and the citation information. The actual record can be retrieved by clicking on the author list.

TABLE 7.1. Entrez Boolean Search Statements

General syntax:

search term [tag] boolean operator search term [tag] . . .

where [tag] =

[AD]	Affiliation
[ALL]	All fields
[AU]	Author name <i>O'Brien J [AU] yields all of O'Brien JA, O'Brien JB, etc. 'O'Brien J' [AU] yields only O'Brien J</i>
[RN]	Enzyme Commission or Chemical Abstract Service numbers
[EDAT]	Entrez date YYYY/MM/DD, YYYY/MM, or YYYY
[IP]	Issue of journal
[TA]	Journal title, official abbreviation, or ISSN number Journal of Biological Chemistry J Biol Chem 0021-9258
[LA]	Language
[MAJR]	MeSH major topic <i>One of the major topics discussed in the article</i>
[MH]	MeSH terms <i>Controlled vocabulary of biomedical terms (subject)</i>
[PS]	Personal name as subject <i>Use when name is subject of article, e.g., Varmus H [PS]</i>
[DP]	Publication date YYYY/MM/DD, YYYY/MM, or YYYY
[PT]	Publication type Review Clinical Trial Lectures Letter Technical Publication
[SH]	Subheading <i>Used to modify MeSH Terms</i> hypertension [MH] AND toxicity [SH]
[NM]	Substance name <i>Name of chemical discussed in article</i>
[TW]	Text words <i>All words and numbers in the title and abstract, MeSH terms, subheadings, chemical substance names, personal name as subject, and MEDLINE secondary sources</i>
[UID]	Unique identifiers (PMID/MEDLINE numbers)
[VI]	Volume of journal

and **boolean operator** = AND, OR, or NOT

The screenshot displays the NCBI PubMed website interface. At the top, there are navigation links for 'National Library of Medicine' and 'PubMed'. Below this, a search bar contains the query 'AT1B' and the search results are displayed. The first result is a paper by Cayatte AJ, Du Y, Oliver-Krasinski J, Lavielle G, Verbeuren TJ, Cohen RA, published in *Arterioscler Thromb Vasc Biol* in 2000. The abstract text is visible, discussing the role of thromboxane in atherosclerosis and the effects of aspirin. The interface includes various toolbars for navigation, such as 'Limits', 'Preview/Index', 'History', and 'Clipboard'. There are also buttons for 'Display', 'Abstract', 'Save', 'Text', 'Order', and 'Add to Clipboard'.

Search Results:

1: *Arterioscler Thromb Vasc Biol* 2000 Jul;20(7):1724-8

AT1B

The thromboxane receptor antagonist S18886 but not aspirin inhibits atherogenesis in apo E-deficient mice: evidence that eicosanoids other than thromboxane contribute to atherosclerosis.

Cayatte AJ, Du Y, Oliver-Krasinski J, Lavielle G, Verbeuren TJ, Cohen RA

Vascular Biology Unit, Whitaker Cardiovascular Institute, Boston University School of Medicine, MA 02118, USA.



Atherosclerosis involves a complex array of factors, including leukocyte adhesion and platelet vasoactive factors. Aspirin, which is used to prevent secondary complications of atherosclerosis, inhibits platelet production of thromboxane (Tx) A(2). The actions of TxA(2) as well as of other arachidonic acid products, such as prostaglandin (PG) H(2), PGE(2), and PGF(2alpha), hydroxyicosatetraenoic acids, and isoprostanes, can be effectively antagonized by blocking thromboxane (TP) receptors. The purpose of this study was to determine the role of platelet-derived TxA(2) in atherosclerotic lesion development by comparing the effects of aspirin and the TP receptor antagonist S18886. The effect of 11 weeks of treatment with aspirin (30 mg kg(-1) d(-1)) or S18886 (5 mg kg(-1) d(-1)) on aortic root atherosclerotic lesions, serum levels of intercellular adhesion molecule-1 (ICAM-1), and the TxA(2) metabolite TxB(2) was determined in apolipoprotein E-deficient mice at 21 weeks of age. Both treatments did not affect body or heart weight or serum cholesterol levels. Aspirin, to a greater extent than S18886, significantly decreased serum TxB(2) levels, indicating the greater efficacy of aspirin in preventing platelet synthesis of TxA(2). S18886, but not aspirin, significantly decreased aortic root lesions as well as serum ICAM-1 levels. S18886 also prevented the increased expression of ICAM-1 in cultured human endothelial cells stimulated by the TP receptor agonist U46619. These results indicate that inhibition of platelet TxA(2) synthesis with aspirin has no significant effect on atherogenesis or adhesion molecule levels. The effects of S18886 suggest that blockade of TP receptors inhibits atherosclerosis by a mechanism independent of platelet-derived TxA(2), perhaps by preventing the expression of adhesion molecules whose expression is stimulated by eicosanoids other than TxA(2).

Comments:

- o Comment in: *Arterioscler Thromb Vasc Biol* 2000 Jul;20(7):1695-8

PMID: 10894809, UI: 20355344

Figure 7.3. An example of a PubMed record in Abstract format as returned through Entrez. This Abstract view for the first reference shown in Figure 7.2. This view provides links to Related Articles, Books, LinkOut, and the actual, full-text journal paper. See text for details.


 National Library of Medicine 

Search for

Nucleotide Protein Structure PopSet

Limits: Preview/Index History Clipboard

Display Summary Save Text Order Add to Clipboard

Show: 20 Items 1-20 of 101 Page 1 of 6 Select page: 1 2 3 4 5 6

1: Cayatte AJ, Du Y, Oliver-Krasinski J, Lavielle G, Verbeuren TJ, Cohen RA. **The thromboxane receptor antagonist S18886 but not aspirin inhibits atherogenesis in apo E-deficient mice: evidence that eicosanoids other than thromboxane contribute to atherosclerosis.** *Arterioscler Thromb Vasc Biol.* 2000 Jul;20(7):1724-8. PMID: 10894809; UI: 20355344 Related Articles

2: Nakashima Y, Raines EW, Plump AS, Breslow JL, Ross R. **Upregulation of VCAM-1 and ICAM-1 at atherosclerosis-prone sites on the endothelium in the ApoE-deficient mouse.** *Arterioscler Thromb Vasc Biol.* 1998 May;18(5):842-51. PMID: 9596845; UI: 98258792 Related Articles

3: Collins RG, Vejji R, Guevara NV, Hicks MJ, Chan L, Beaudet AL. **P-Selectin or intercellular adhesion molecule (ICAM)-1 deficiency substantially protects against atherosclerosis in apolipoprotein E-deficient mice.** *J Exp Med.* 2000 Jan 3;191(1):189-94. PMID: 10620617; UI: 20088904 Related Articles

4: Kyrle PA, Mimar E, Brenner B, Eichler HG, Heistinger M, Marosi L, Lechner K. **Thromboxane A2 and prostacyclin generation in the microvasculature of patients with atherosclerosis--effect of low-dose aspirin.** *Thromb Haemost.* 1989 Jun 30;61(3):374-7. PMID: 2506252; UI: 90020147 Related Articles

5: Brown N, May JA, Wilcox RG, Allan LM, Wilson AM, Kiff PS, Hepburn SJ. **Comparison of antiplatelet activity of microencapsulated aspirin 162.5 Mg (Caspac XL), with enteric coated aspirin 75 mg and 150 mg in patients with atherosclerosis.** *Br J Clin Pharmacol.* 1999 Jul;48(1):57-62. PMID: 10563561; UI: 99315237 Related Articles

6: Dongs ZM, Wagner DD. **Leukocyte-endothelium adhesion molecules in atherosclerosis.** *J Lab Clin Med.* 1996 Nov;132(5):369-75. Review. PMID: 9823930; UI: 99039556 Related Articles

7: Daugherty A, Pure E, Delfat-Bunteiger D, Chen S, Lefterovich J, Roselaar SE, Rader DJ. Related Articles

About Entrez

Entrez PubMed Overview Help FAQ New/Noteworthy

PubMed Services Journal Browser MeSH Browser Single Citation Matcher Batch Citation Matcher Clinical Queries Old PubMed

Related Resources Order Documents Grateful Med Consumer Health Clinical Alerts ClinicalTrials.gov

Privacy Policy


Figure 7.4. Neighbors to an entry found in PubMed. The original entry (Cayette et al., 2000) is at the top of the list, indicating that this is the parent entry. See text for details.

Molecular Biology of the Cell (Alberts et al., 1994). Following the Cayette example, if the user clicks on *atherosclerosis* at this point, it will take them to the relevant part of the textbook, a section devoted to how cells import cholesterol by receptor-mediated endocytosis (Fig. 7.5). From this page, the user can navigate through this particular unit, gathering more general information on transport from the plasma membrane via endosomes and vesicular traffic in the secretory and endocytic pathways.

The final link in the series in the upper right is LinkOut. This feature provides a list of third-party Web sites and resources relating to the Entrez entry being viewed, such as full-text of articles that can be displayed directly through the Web browser, or the capability of ordering the document through services such as Loansome Doc. A “cubby” service for LinkOut enables users to customize which links are displayed in a LinkOut view. Another way of getting to the full text of an article is by following a direct link to the publisher’s Web site. In the Abstract view for the Cayette example (Fig. 7.3), a button directly under the citation is marked ATVB, for *Arteriosclerosis, Thrombosis, and Vascular Biology*, the journal in which the paper is published. With the proper individual or institutional subscription privileges, one would be able to immediately see the full text of the paper, including all figures and tables.

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 methanobacteria. In this case, the search would begin by setting the Search pull-down menu to Nucleotide and then typing the term DNA-binding into the text box. This search returns 23,797 entries in which the search term appears (Fig. 7.6). At this point, to narrow down the search, the user would click on the Limits hyperlink, directly below the text box. This brings the user to a new page that allows the search to be limited, as implied by the name of the hyperlink. Here, the search will be limited by organism, so the Limited To pull-down is changed to Organism, and the word *methanobacterium* is typed into the search box (Fig. 7.7). Clicking Go will now return all of the entries in which *Methanobacterium* is the organism (303). The results from the first search can also be combined with those from the second by clicking on the History hyperlink below the text box, resulting in a list of recent queries (Fig. 7.8). 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 individual query returned. To combine two separate queries into one, the user simply combines the queries by number; in this case, because the queries are numbered #8 and #9, the syntax would be #8 AND #9. Clicking Preview regenerates a table, showing the new, combined query as #10, containing three entries. Alternatively, clicking Go shows the user the three entries, in the now-familiar nucleotide summary format (Fig. 7.9).

As before, there are a series of hyperlinks to the upper right of each entry; three are shown for the first entry, which is for the *M. thermoautotrophicum* tfx gene. The PubMed link takes the user back to the bibliographic entry or entries corresponding to this GenBank entry. Clicking on Protein brings into play one of the hard link relationships, showing the GenPept entry that corresponds to the tfx gene’s conceptual translation (Fig. 7.10). Notice that, within the entry itself, the scientific name of the organism is represented as hypertext; clicking on that link takes the user to the NCBI Taxonomy database, where information on that particular organism’s lineage is available. One of the useful views at this level is the Graphics view; this view

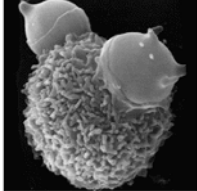


MOLECULAR BIOLOGY OF THE CELL
FOUNDED 1981 • HOME PAGE: www.ncbi.nlm.nih.gov



Vesicular Traffic in the Secretory and Endocytic Pathways

Transport from the Plasma Membrane via Endosomes: Endocytosis¹⁹



Cells Import Cholesterol by Receptor-mediated Endocytosis²³

Many animal cells take up cholesterol through receptor-mediated endocytosis and in this way acquire most of the cholesterol they require to make new membrane. If the uptake is blocked, cholesterol accumulates in the blood and can contribute to the formation in blood vessel walls of atherosclerotic plaques - the deposits of lipid and fibrous tissue that cause strokes and heart attacks by blocking blood flow. In fact, it was through a study of humans with a strong genetic predisposition for atherosclerosis that the mechanism of receptor-mediated endocytosis was first clearly revealed.

Most cholesterol is transported in the blood bound to protein in the form of particles known as low-density lipoproteins, or LDL (Figure 13-29). When a cell needs cholesterol for membrane synthesis, it makes transmembrane receptor proteins for LDL and inserts them into its plasma membrane. Once in the plasma membrane, the LDL receptors diffuse until they associate with clathrin-coated pits that are in the process of forming (Figure 13-30A). Since coated pits constantly pinch off to form coated vesicles, any LDL particles bound to LDL receptors in the coated pits are rapidly internalized in coated vesicles. After shedding their clathrin coats, these vesicles deliver their contents to early endosomes, which are located near the cell periphery. Once in the endosomal compartment, the

Outline
Introduction
Specialized Phagocytic Cells Can Ingest Large Particles
Pinocytic Vesicles Form from Coated Pits in the Plasma Membrane
Clathrin-coated Pits Can Serve as a Concentrating Device for Internalizing Specific Extracellular Macromolecules
Cells Import Cholesterol by Receptor-mediated Endocytosis
Endocytosed Materials Often End Up in Lysosomes

Figure 7.5. Text related to the original Cayette et al. (2000) entry from *Molecular Biology of the Cell* (Alberts et al., 1994). See text for details.

NCBI Search for

Published Nucleotide Protein Structure PopSet

Limits Preview/Index History Clipboard

Display Summary

Show: Items 1-20 of 23797 Page 1 of 1190 Select page: 1 2 3 4 5 6 7 8 9 10 >>

- 1:** AL391222
Arabidopsis thaliana DNA chromosome 5, BAC clone T5K6 (ESSA project)
gi|9795153|emb|AL391222.1|ATT5K6[9795153] Protein
- 2:** AC026238
Arabidopsis thaliana chromosome 1 BAC F25116 genomic sequence, complete sequence
gi|7462023|gb|AC026238.2|AC026238[7462023] Protein, Related Sequences
- 3:** BE551595
7a42d06.x1 NCI CGAP GC6 Homo sapiens cDNA clone IMAGE:3221387 3' similar to TR:Q15553 Q15553 TELOMERIC DNA BINDING PROTEIN ; mRNA sequence
gi|9793207|gb|BE551595.1|BE551595[9793207] PubMed, Protein
- 4:** AB006700
Arabidopsis thaliana genomic DNA, chromosome 5, P1 clone:MHE15
gi|9758404|dbj|AB006700.2|AB006700[9758404] Protein
- 5:** AL391174
S.pombe chromosome II cosmid c317
gi|9757426|emb|AL391174.1|SPBC317[9757426] Protein
- 6:** AL390114
Leishmania major chromosome 12/24 clone Chr.12/24 strain Friedlin, *** SEQUENCING IN PROGRESS *** , in ordered pieces
gi|9756289|emb|AL390114.2|LMFLCHR12[9756289] Protein
- 7:** AL391149
Arabidopsis thaliana DNA chromosome 5, BAC clone T9L3 (ESSA project)
gi|9755738|emb|AL391149.1|ATT9L3[9755738] Protein
- 8:** AL391147
Arabidopsis thaliana DNA chromosome 5, BAC clone F5E19 (ESSA project)
gi|9755718|emb|AL391147.1|ATF5E19[9755718] Protein

About Entrez
Entrez Nucleotide Help | FAQ
Retrieve large data sets
Check sequence revision history
How to create WWW links to Entrez
Related resources BLAST
Reference sequence project
LocusLink
Submit to GenBank

Figure 7.6. Formulating a search against the nucleotide portion of Entrez. The initial query is shown in the text box towards the top of the window, and the nucleotide entries matching the query are shown below. See text for details.

NCBI Nucleotide

Search [Nucleotide] for Methanobacterium

Limits Preview/Index History Clipboard

Use All Fields pull-down menu to specify a field

- Boolean operators AND, OR, NOT must be in upper case
- if search fields tags are used enclose in square brackets, e.g., rubella [t]
- Search limits may exclude ESTs and other data subsets

Limited to:

Organism

exclude ESTs exclude STSs exclude GSS

exclude working draft exclude patents exclude all of the above

Molecule

Only from

Modification Date From To

Gene Location

Modification Date

Use the format YYYY/MM/DD; month and day are optional.

Segmented Sequences

Revised: June 7, 2000.

Disclaimer | Write to the Help Desk
NCBI | NLM | NIH

About Entrez
Entrez Nucleotide
Help | FAQ
Retrieve large data sets
Check sequence revision history
How to create WWW links to Entrez
Related resources
BLAST
Reference sequence project
LocusLink
Submit to GenBank


Figure 7.7. Using the Limits feature of Entrez to limit a search to a particular organism. See text for details.

The screenshot shows the NCBI Entrez Nucleotide search page. At the top, there are navigation tabs for Nucleotide, Protein, Genome, Structure, and PopSet. The search bar contains the query "#8 AND #9". Below the search bar, there are links for "Limits", "Preview/Index", "History", and "Clipboard". A "Clear" button is also present. The search results section includes a warning: "Search History will be lost after one hour of inactivity" and a tip: "To combine searches use # before search number, e.g., #2 AND #6". Below this, the "Search" results are displayed as a table of "Most Recent Queries".

Search	Most Recent Queries	Time	Result
#9	Search Methanobacterium Field: Organism	15:50:30	303
#8	Search DNA-binding	15:45:42	23797

At the bottom of the page, there are links for "About Entrez", "Entrez Nucleotide Help | FAQ", "Retrieve large data sets", "Check sequence revision history", and "How to create WWW links to Entrez".

Figure 7.8. Combining individual queries using the History feature of Entrez. See text for details.


[About Entrez](#)
[Entrez Nucleotide Help | FAQ](#)
[Retrieve large data sets](#)
[Check sequence revision history](#)
[How to create WWW links to Entrez](#)

[Search Nucleotide](#) for #8 AND #9

[Limits](#)
[Preview/Index](#)
[History](#)
[Clipboard](#)

[Nucleotide](#)
[Protein](#)
[Genome](#)
[Structure](#)
[PopSet](#)

Limits

Field: Organism

Display:

Show:

 Items 1-3 of 3

1: A:J009686
 Methanobacterium thermoautotrophicum tfx gene
 gi|4138234|emb|A:J009686.1|MTH9686[4138234]

2: M90086
 Methanobacterium thermoautotrophicum archeal histone, DNA binding protein (hmtA) gene, complete cds

3: M86663
 Methanobacterium thermoautotrophicum DNA binding protein (hmtB) gene, complete cds
 gi|149724|gb|M86663.1|MBFHMTB[149724]

PubMed, Protein, Related Sequences

Protein, Related Sequences

Protein, Related Sequences

Revised: June 7, 2000.

[Disclaimer | Write to the Help Desk](#)
[NCBI | NLM | NIH](#)

[Related resources](#)
 BLAST
 Reference sequence project
 LocusLink
 Submit to GenBank

Figure 7.9. Entries resulting from the combination of two individual Entrez queries. The command producing this Entrez is shown in the text box at the top of the figure, and information on the individual queries that were combined is given in Figure 7.8.

Tfx protein [Methanobacteri...

□ 1 : GI = "4138235" [GenPept]

Related Articles, Protein, Nucleotide

```

LOCUS      CAA08778             138 aa                BCT                22-JAN-1999
DEFINITION Tfx protein [Methanobacterium thermoautotrophicum].
ACCESSION  CAA08778
PID        G4138235
VERSION   CAA08778.1  GI:4138235
DBSOURCE  emb1 locus MTH9686, accession AJ009686.1
KEYWORDS
SOURCE    Methanobacterium thermoautotrophicum.
ORGANISM  Archaea; Euryarchaeota; Methanobacteriales; Methanobacteriaceae;
          Methanobacterium.
REFERENCE  1 (residues 1 to 138)
AUTHORS   Hochheimer,A.
TITLE     Direct Submission
JOURNAL   Submitted (21-JUL-1998) Hochheimer A., Abt. Biochemie, Karl von
          Max-Planck-Institut fuer terrestrische Mikrobiologie, Karl von
          Frisch Strasse, 35043 Marburg, GERMANY
REFERENCE  2 (residues 1 to 138)
AUTHORS   Hochheimer,A., Hedderich,R. and Thauer,R.K.
TITLE     The DNA binding protein Tfx from Methanobacterium
          thermoautotrophicum: structure, DNA binding properties and
          transcriptional regulation
JOURNAL   Mol. Microbiol. 31 (2), 641-650 (1999)
MEDLINE   99157570
FEATURES  Location/Qualifiers
           source                1..138
           /organism="Methanobacterium thermoautotrophicum"
           /strain="Marburg"
           /db_xref="taxon:2166"
           Protein              1..138
           /function="DNA-binding protein"
           mat_peptide          2..138
           /product="Tfx protein"
           CDS                  1..138
           /product="Tfx protein"
           /gene="tfx"
           /coded_by="AJ009686.1:1..417"
           /transl_table=11
ORIGIN
1  mskkfilter qktvlemrer cwsqkkiare lktrtrqmysa ierkamenie ksrintldfyk
61  flksprilc rrgdtldeli klllesnke gihvihsit laflirekas hrivhrvYkS
121  dfeigvtrdg eilvdlns
//
    
```

Figure 7.10. The protein neighbor for the *M. thermoautotrophicum* tfx gene. Clicking on the Protein hyperlink next to the first entry in Figure 7.9 leads the user to this GenPept entry. See text for details.

attempts to show graphically all of the features described within the entry's feature table, providing a very useful overview, particularly when the feature table is very long. The Related Sequences link shows all sequences similar to that of the tfx gene at the nucleotide level, in essence showing the results of a precomputed BLAST search.

The last part of Entrez to be discussed deals with structures. Structure queries can be done directly by specifying *Structure* in the Search pull-down menu. For example, suppose that one wishes to find out information about the structure of HMG-box B from rat, whose accession number is 1HMF. Typing 1HMF into the query box leads the user to the structure summary page for 1HMF, which has a decidedly different format than any of the pages seen so far (Fig. 7.11). This page shows details from the header of the source MMDB document (which is derived from PDB), links to PubMed and to the taxonomy of the source organism, and links to both sequence and structure neighbors. The Sequence Neighbors links show neighbors to 1HMF on the basis of sequence—that is, by BLAST search—thus, although this is a *structure* entry, it is important to realize that sequence neighbors have nothing to do with the structural information, at least not directly. To get information about related structures, one of the Structure Neighbor links can be followed, producing a table of neighbors as assessed by VAST. For a user interested in gleaning initial impressions about the shape of a protein, the Cn3D plug-in, invoked by clicking on *View/Save Structure*, provides a powerful interface, giving far more information than anyone could deduce from simply examining a string of letters (the sequence of the protein). The protein may be rotated along its axes by means of the scroll bars on the bottom, top, and right-hand side of the window or may be freely rotated by clicking and holding down the mouse key while the cursor is within the structure window and then dragging. Users are able to zoom in on particular parts of the structure or change the coloration of the figure, to determine specific structural features about the protein. In Figure 7.12, for instance, *Spacefilling* and *Hydrophobicity* were chosen as the *Render* and *Color* options, respectively. More information on Cn3D is presented in Chapter 5 as well as in the online Cn3D documentation. In addition, users can save coordinate information to a file and view the data using third-party applications such as Kinemage (Richardson and Richardson, 1992) and RasMol (Sayle and Milner-White, 1995).

Finally, at any point along the way in using Entrez, if there are partial or complete search results that the user wishes to retain while moving onto a new query, the *Add to Clipboard* button can be pushed. This stores the results of the current query, which the user can return to by clicking the *Clipboard* hyperlink directly under the text box. The clipboard holds a maximum of 500 items, and the items are held in memory for 1 h.

LOCUSLINK

The Entrez system revolves necessarily around the individual entries making up the various component databases that are part of the Entrez search space. Another way to think about this search space is to organize it around discrete genetic loci. NCBI LocusLink does just this, providing a single query interface to various types of information regarding a given genetic locus, such as phenotypes, map locations, and

MMDB Id: 1217 PDB Id: 1HMF

Protein Chains: (single chain)
MEDLINE: PubMed
Taxonomy: Rattus norvegicus

PDB Authors: H.M.Weir, P.J.Kraulis, C.S.Hill, A.R.C.Raine, E.D.Laue & J.O.Thomas

PDB Deposition: 7-Mar-94

PDB Class: Dna-Binding

PDB Compound: High Mobility Group Protein Fragment-B (Hmgb) (Dna-Binding Hmg-Box Domain B Of Rat Hmg1) (Nmr, 30 Structures)

Sequence Neighbors: (single chain)

Structure Neighbors: (single chain)

[View / Save Structure](#)

NEW *Cn3D 3.0 Released!*

Options:

- Launch Viewer
- See File
- Save File

Viewer:

- Cn3D (asn.1)
- Cn3D v1.0 (asn.1)
- Mage
- RasMol (PDB)

Complexity:

- Cn3D Subset
- Virtual Bond Model
- All Atom Model
- RasMol (PDB)

Figure 7.11. 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 to the taxonomy of the source organism, and links to sequence and structure neighbors.

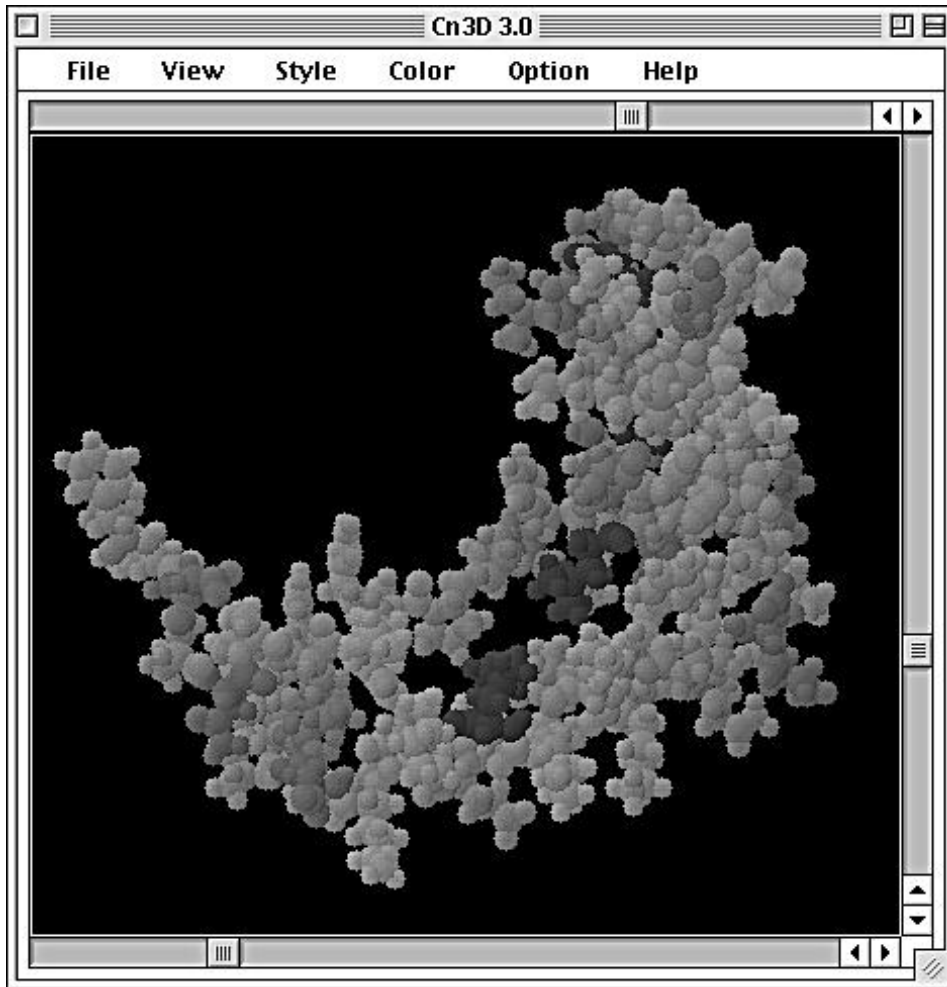


Figure 7.12. The structure of 1HMF rendered using Cn3D version 3.0, an interactive molecular viewer that acts as a plug-in to Web Entrez. Cn3D is also bundled with and can be used with Network Entrez. Details are given in the text.

homologies to other genes. The LocusLink search space currently includes information from humans, mice, rats, fruit flies, and zebrafish.

With the use of the gene for the high-mobility group protein HMG1 as an example, the LocusLink query begins by the user simply typing the name of the gene into the query box appearing at the top of the LocusLink home page. Alternatively, the user could select the gene of interest from an alphabetical list. The query on HMG1 returns three LocusLink entries, from human, mouse, and rat (Fig. 7.13). In this view, the user is given the Locus ID in the first column; the Locus ID is intended to be a unique, stable identifier that is associated with this gene locus. Clicking on the Locus ID for the human (3146) produces the LocusLink Report view, as shown in Figure 7.14. The Report view begins with links to external sources of information, shown as colored buttons at the top of the page. In this particular report, the links

The screenshot shows the NCBI LocusLink search interface. At the top, there is a search bar with 'LocusLink' selected in the search type dropdown, 'HMG1' in the query field, and 'All' in the organism dropdown. The display format is set to 'Brief'. Below the search bar, there is a navigation menu with 'LocusLink Home' and 'Help' links. A horizontal menu shows the alphabet 'A-Z', with 'H' highlighted. Below this, it says '3 loci found'. A table lists the results:

LocusID	Org	Symbol	Description	Position	Links
3146	<i>Hs</i>	HMG1	high-mobility group (nonhistone chromosomal) protein 1	13q12	P O R G H U V
15289	<i>Mm</i>	Hmg1	high mobility group protein 1	5 83.0 cM	P R G H U
25459	<i>Rn</i>	Hmg1	High mobility group 1		P R G H U

Figure 7.13. Results of a LocusLink query, using HMG1 as the search term. The report shows three entries corresponding to HMG1 in human (*Hs*), mouse (*Mm*), and rat (*Rn*). A brief description is given for each found locus, as well as its chromosomal location. A series of blocks is found to the right of each entry, providing a jumping-off point to numerous other sources of data; these links are described in the text.

would lead the user to PubMed (Pub), UniGene (UG, cf. Chapter 12), the dbSNP variation database (VAR, cf. Chapter 12), HomoloGene (HOMOL, see below), and the Genome Database (GDB). These offsite links will change from entry to entry, depending on what information is available for the gene of interest. A complete list of offsite data sources is given in the LocusLink online documentation.

Continuing down the Report view, in the section marked Locus Information, the user is presented with the official gene symbol, along with any alternate symbols that may have traditionally been used in the literature or in sequence records. This section would also include information on the name of the gene product, any aliases for the name of the gene product, the Enzyme Commission number, the name of any diseases that result from variants at this gene locus, and links to OMIM and UniGene. Only those pieces of information that are known or are applicable to this particular gene locus are shown.

In the section labeled Map Information, the report shows what chromosome this locus is on, the cytogenetic and genetic map positions, when known, and any STS markers that are found within the mRNA corresponding to this locus. There is a hyperlink that can take the user to the Entrez Map Viewer, showing the position of this locus and the relationship of this locus to surrounding loci (Fig. 7.15). The Map Viewer shows the chromosomal ideogram to the left, with the highlighted region marked by a thick bar to the right of the ideogram. The user can zoom in or out by clicking on the icon above the ideogram. In the main window, the user is presented

The screenshot displays the NCBI LocusLink interface for the human gene HMG1. At the top, there is a search bar with 'LocusLink' entered and a dropdown menu set to 'All'. Below the search bar are navigation tabs for PubMed, Entrez, BLAST, OMIM, Taxonomy, and Structure. The main content area is divided into several sections:

- Navigation and Search:** Includes a search bar with 'LocusLink' and 'Display: Brief', and a 'Query:' field with 'Go' and 'Clear' buttons.
- Gene Symbol and Name:** Shows 'Homo sapiens Official Gene Symbol and Name (HGNC)' and 'HMG1: high-mobility group (nonhistone chromosomal) protein 1'.
- Locus Information:** Lists LocusID: 3146, Type: gene with protein product, function known or inferred, Alternate Symbols: HMG3, Product: high-mobility group (nonhistone chromosomal) protein 1, UniGene: Hs.189509, and OMIM: 163905.
- Map Information:** Includes 'Entrez Map Viewer: View Genomic Context', Chromosome: 13, and Cytogenetic: 13q12 HUGO.
- RefSeq:** Shows 'Homo sapiens HMG1 Reference Sequence (RefSeq)' with Status: PROVISIONAL, Nucleotide: NM_002128, and Protein: NP_002119 high-mobility group (nonhistone chromosomal) protein 1.
- GenBank Sequences:** A table listing sequences:

Nucleotide	Type	Protein
U51677	g	AAB08987
X12597	m	CAA31110
- Additional Web Resources:** Includes a link to 'GeneCard for HMG1'.

The left sidebar contains various navigation links such as 'LocusLink Home', 'Collaborators', 'Download', 'FAQ', 'Help', 'Statistics', 'NCBI Genome Guides', 'RefSeq', 'Related Resources', and 'Nomenclature'.

Figure 7.14. The LocusLink report view for human HMG1. The report is divided into six sections, providing gene symbol, locus, map, RefSeq, and GenBank information, as well as links to external data sources. See text for details.

with both the cytogenetic and sequence map. In this particular view, 20 genes are shown, with the original locus of interest highlighted. As with most graphical views of this type, the majority of the elements in this view are clickable, taking the user to more information about that particular part of either the cytogenetic or sequence map.

The next section deals with RefSeq information. RefSeq is short for the NCBI Reference Sequence Project, which is an effort aimed at providing a stable, reference sequence for all of the molecules corresponding to the central dogma of biology. The intention is that these reference sequences will provide a framework for making

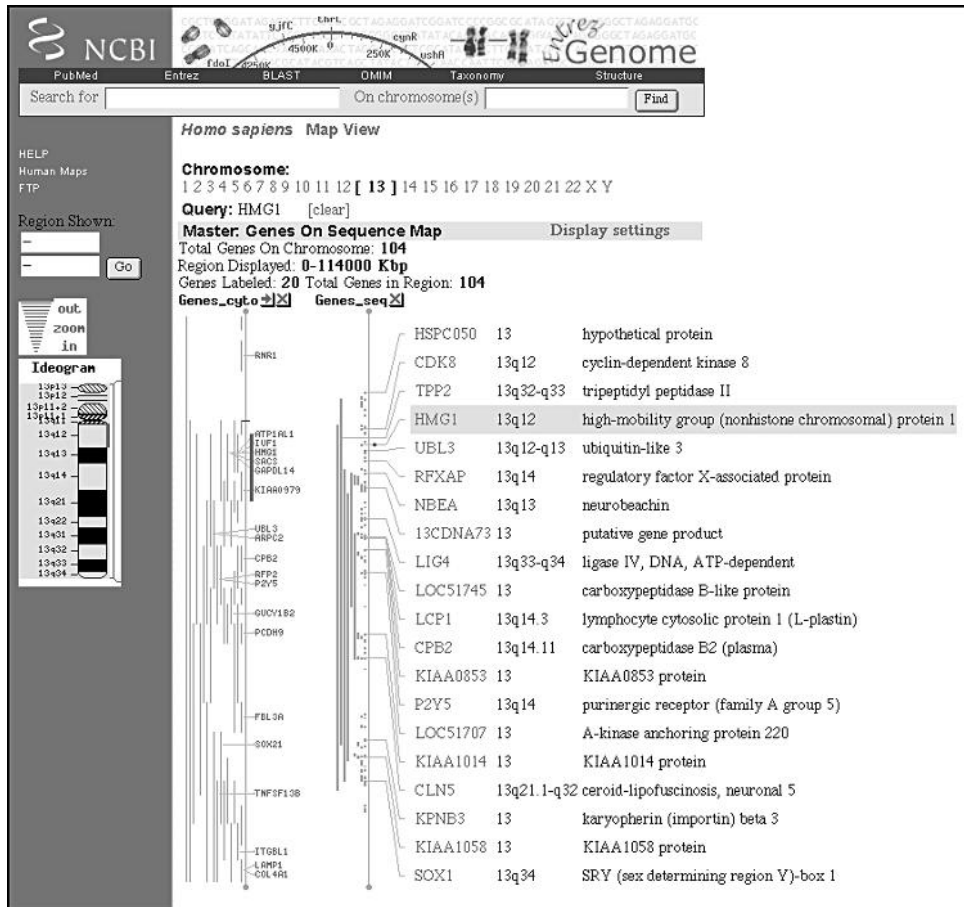


Figure 7.15. The Entrez map view for human HMG1. The chromosomal position is indicated by the ideogram at the left of the window. The main window contains a depiction of both the cytogenetic and sequence map, with the HMG1 gene highlighted. Interestingly, the gene shown at the very bottom of this view (SOX1), like HMG1, is also a member of the high mobility group family of proteins (Baxevanis and Landsman, 1995).

functional annotations, as well as for information regarding mutations, gene expression, and polymorphisms. The sequences listed in this section represent the sequences that were used to build the corresponding RefSeq record. Notice that RefSeq nucleotide records are preceded by NM and that protein records are preceded by NP. A blue button appears next to protein information if there is also structural information available about the protein. The final portion of the LocusLink report shows the GenBank accession numbers that correspond to this locus. The middle column indicates the molecule type for these GenBank entries, with m standing for mRNA, g for genomic DNA, e for an EST, and u for undetermined. In this particular case, there is also a link to the GeneCard for HMG1; clicking on that hyperlink takes the user to the GeneCards database at the Weizmann Institute, providing a concise summary of the function of this gene.

Another way to proceed through the information is to return to the query result screen shown in Figure 7.13 and use the linked alphabet blocks shown to the right of each of the entries. In turn, these links are

- P, for PubMed bibliographic entries corresponding to the locus;
- O, for the Online Mendelian Inheritance in Man summary on this locus;
- R, for the RefSeq entries corresponding to the locus;
- G, for the individual GenBank entries related to the locus; these entries will correspond to the RefSeq entry, as shown in the LocusLink Report view;
- H, for HomoloGene. HomoloGene, which is discussed in greater detail in Chapter 12, allows the user to find both orthologs and homologs for genes represented in LocusLink and UniGene, the assignments being made based on sequence similarity;
- U, whether this locus is part of a UniGene cluster; and
- V, for variation data on this locus contained within dbSNP.

When following either the PubMed or GenBank links, the user is, in essence, returned to the Entrez search space, enabling the user to take advantage of Entrez's navigational features once again.

SEQUENCE DATABASES BEYOND NCBI

Although it may appear from this discussion that NCBI is the center of the sequence universe, many specialized sequence databases throughout the world serve specific groups in the scientific community. Often, these databases provide additional information such as phenotypes, experimental conditions, strain crosses, and map features. The data are of great importance to these subsets of the scientific community, inasmuch as they can influence rational experimental design, but such types of data do not always fit neatly within the confines of the NCBI data model. Development of specialized databases necessarily ensued, but they are intended to be used as an adjunct to GenBank, not in place of it. It is impossible to discuss all of these kinds of databases here, but, to emphasize the sheer number of such databases that exist, *Nucleic Acids Research* devotes its first issue every year to papers describing these databases (cf. Baxevanis, 2001).

An example of a specialized organismal database is the *Saccharomyces* Genome Database (SGD), housed at the Stanford Human Genome Center. The database provides a very simple search interface that allows text-based searches by gene name, gene information, clone, protein information, sequence name, author name, or full text. For example, using Gene Name as the search topic and *hho1* as the name of the gene to be searched for produces a SacchDB information window showing all known information on locus HHO1 (Fig. 7.16). This window provides jumping-off points to other databases, such as GenBank/GenPept, MIPS, and the Yeast Protein Database (YPD). Following the link to Sacch3D for this entry provides information on structural homologs of the HHO1 protein product found in PDB, links to secondary and tertiary structure prediction sites, and precomputed BLAST reports against a number of query databases. Returning to the Locus window and clicking on the map in the upper right-hand corner, the user finds a graphical view of the

[Help](#)

HHO1/YPL127C

[Search SGD](#) | [Gene/Seq Resources](#) | [Help](#) | [Gene Registry](#) | [Maps](#)
[BLAST](#) | [FASTA](#) | [PatMatch](#) | [Sacch3D](#) | [Primers](#) | [SGD Home](#)

HHO1 BASIC INFORMATION

Standard Name	HHO1
Systematic Name	YPL127C
Description	Histone H1
Gene Product	histone H1
Phenotype	Null mutant is viable; other phenotype: Increased basal expression of a CYC1-lacZ reporter gene; nuclear localization of a Hho1-GFP fusion protein
Position	ChrXVI: coordinates 309603 to 308827 old format Sequence details
External Links	MIPS YPD Entrez Protein Entrez Neighbors PIR-DE PIR-JP PIR-US
Primary SGDID	S0006048

HHO1 RESOURCES

Click on map for expanded view

- **Literature**
[Gene Info](#) | [View](#)
- **Retrieve Sequences**
[DNA \(w/ introns\)](#) | [Retrieve](#)
- **Sequence Analysis Tools**
[BLASTP](#) | [Analyze](#)
- **Maps and Displays**
[Chr. Features Map](#) | [View](#)
- **Comparison Resources**
[Worm Homologs](#) | [View](#)
- **Functional Analysis**
[Sheaton Cell Cycle](#) | [View](#)

ADDITIONAL INFORMATION

Global Gene Hunter	Function Junction
Researchers	Protein Info & Composition

SGD™, pages Copyright © 1997-2000 The Board of Trustees of Leland Stanford Junior University. Documents from this server are provided "AS-IS" without any warranty, expressed or implied.

Figure 7.16. A SacchDB Locus view resulting from an SGD query using hho1 as the gene name. The information returned is extensive; it includes the name of the gene product, phenotypic data, and the position of HHO1 within the yeast genome. There are numerous hyperlinks providing access to graphical views and to external database entries related to the current query.

area surrounding the locus in question. Available views include physical maps, genetic maps, and chromosomal physical maps, among others. The chromosomal features map view for HHO1 is shown in Figure 7.17. Note the thick bar at the top of the figure, which gives the position of the current view with respect to the centromere. Clicking on that bar allows the user to move along the chromosome, and clicking on individual gene or ORF name (or, as the authors cite in the figure legend, “any little colorful bar”) gives more detailed information about that particular region.

Another example of an organism-specific database is FlyBase, whose goal is to maintain comprehensive information on the genetics and molecular biology of *Drosophila*. The information found in FlyBase includes an extensive *Drosophila* bibliography, addresses of over 6,000 researchers involved in *Drosophila* projects, a compilation of information on over 51,500 alleles of more than 13,200 genes, information about the expression and properties of over 4,800 transcripts and 2,500 proteins, and descriptions of over 16,700 chromosomal aberrations. Also included is relevant mapping information, functional information on gene products, lists of stock centers and genomic clones, and information from allied databases. Searches on any of these “fields” can be done through a simple search mechanism. For example, searching by gene symbol using `capu` as the search term brings up a record for a gene named cappuccino, which is required for the proper polarity of the developing *Drosophila* oocyte (Emmons et al., 1995). Calling up the cytogenetic map view generates a map showing the gene and cytologic location of cappuccino and other genes in that immediate area, and users can click on any of the gene bars to bring up detailed information on that particular gene (Fig. 7.18). The view can be changed by selecting

Features around YPL127C on chromosome XVI

Spanning a region 10 kb left and 10 kb right
(coordinates 298827 to 319603)

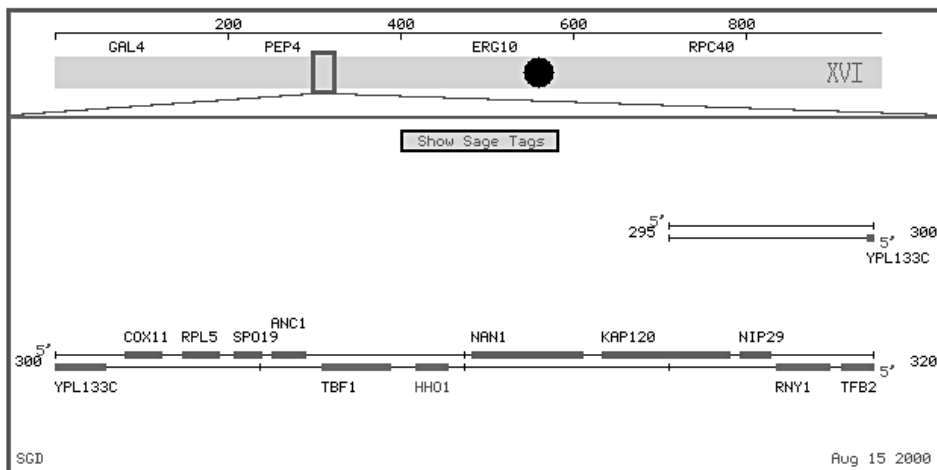


Figure 7.17. A chromosomal features map resulting from the query used to generate the Locus view shown in Figure 7.16. Chromosome XVI is shown at the top of the figure, with the exploded region highlighted by a box. Most items are clickable, returning detailed information about that particular gene, ORF, or construct.

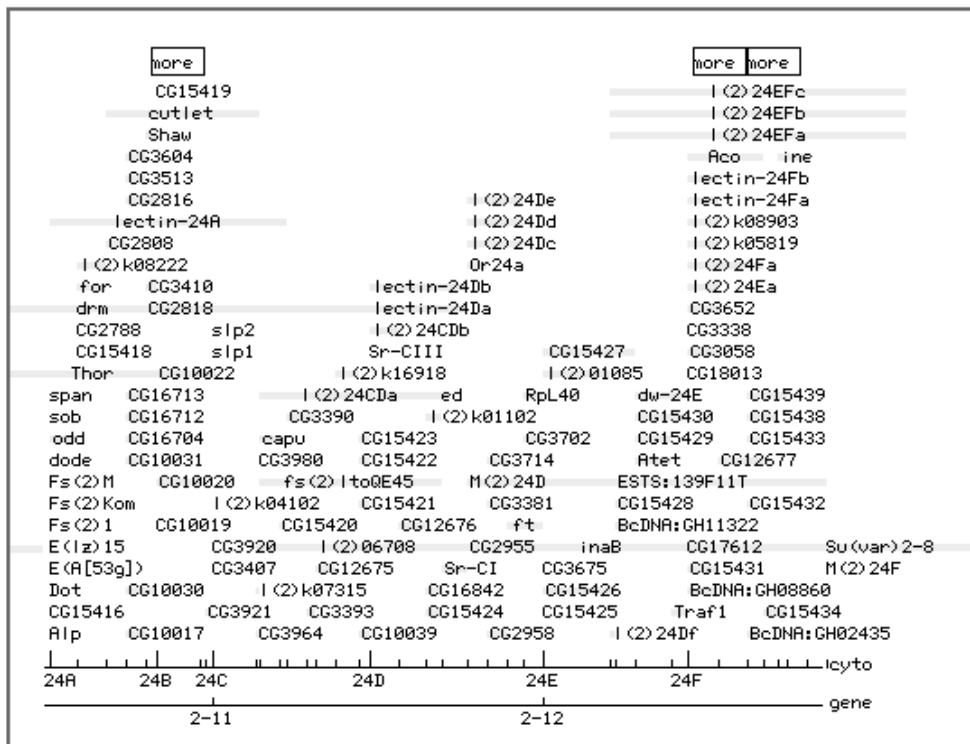


Figure 7.18. Genes view resulting from querying FlyBase for the *capucino* gene (*capu* in the figure, between positions 24C and 24D on the cytologic map). The graphical view can be changed by clicking on any of the Class buttons that appear below the figure, as described in the text. Information on any of the genes shown can be obtained by clicking on the appropriate bar.

one of the Class buttons at the bottom of the window, so that a graphical view of cosmids, deficiencies, duplications, inversions, transpositions, translocations, or other aberrations can be examined instead.

MEDICAL DATABASES

Although the focus of this chapter (and the book in general) is on sequences, databases cataloging and organizing sequence information are not the only kinds of databases useful to the biologist. An example of such a non-sequence-based information resource that is tremendously useful in genomics is Online Mendelian Inheritance in Man (OMIM), the electronic version of the catalog of human genes and genetic disorders founded by Victor McKusick at The Johns Hopkins University (McKusick, 1998; Hamosh et al., 2000). 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. Because the online version of OMIM is housed at NCBI, links to Entrez are provided from all references cited within each OMIM entry.

ALLELIC VARIANTS (selected examples)	
.0001 MCKUSICK-KAUFMAN SYNDROME [MKKS, HIS84TYR]	In an Old Order Amish patient with McKusick-Kaufman syndrome (236700), Stone et al. (2000) identified a C-to-T transition at nucleotide 1137 of the MKKS gene, resulting in a histidine-to-tyrosine substitution at codon 84. This mutation was found in homozygosity; on the same allele there was a second substitution (see 604896.0002). This mutation was found in 1 of 100 Amish control chromosomes, which suggests a carrier frequency of approximately 2%, similar to the estimated carrier frequency calculated using the incidence of this disorder among the Amish. Neither of the Amish substitutions were found in an additional 100 non-Amish control chromosomes. This mutation was predicted to interfere with ATP hydrolysis, which in other chaperonins leads to substantially reduced function.
.0002 MCKUSICK-KAUFMAN SYNDROME [MKKS, ALA242SER]	In an Old Order Amish patient with McKusick-Kaufman syndrome (236700), Stone et al. (2000) identified an alanine-to-serine substitution at codon 242 of the MKKS gene in homozygosity. This mutation is present on the same allele that carries the H84Y mutation (604896.0001). This compound allele was found in all affected individuals in homozygosity among the Old Order Amish. Three individuals homozygous for the affected chromosome had a normal phenotype, consistent with the incomplete penetrance of the MKKS phenotype.
.0003 MCKUSICK-KAUFMAN SYNDROME [MKKS, TYR37CYS]	In a sporadic non-Amish case of McKusick-Kaufman syndrome (236700), Stone et al. (2000) identified an A-to-G transition at nucleotide 997 of the MKKS gene, resulting in a tyrosine-to-cysteine substitution at codon 37. This mutation was not identified in over 200 chromosomes from a non-Amish control group. The patient was a compound heterozygote for a frameshift mutation (604896.0004).
.0004 MCKUSICK-KAUFMAN SYNDROME [MKKS, 2-BP DEL, 2111GG]	In a sporadic non-Amish case of McKusick-Kaufman syndrome (263700), Stone et al. (2000) identified a 2-bp deletion at nucleotide 2111 and 2112 of the MKKS gene, resulting in a frameshift leading to premature termination. This mutation was maternally inherited.

Figure 7.19. An example of a list of allelic variants that can be obtained through OMIM. The figure shows the list of allelic variants for McKusick-Kaufman syndrome.

OMIM has a defined numbering system in which each entry is assigned a unique number, similar to an accession number, but certain positions within that number indicate information about the genetic disorder itself. For example, the first digit represents the mode of inheritance of the disorder: 1 stands for autosomal dominant, 2 for autosomal recessive, 3 for X-linked locus or phenotype, 4 for Y-linked locus or phenotype, 5 for mitochondrial, and 6 for autosomal locus or phenotype. (The distinction between 1 or 2 and 6 is that entries catalogued before May 1994 were assigned either a 1 or 2, whereas entries after that date were assigned a 6 regardless of whether the mode of inheritance was dominant or recessive.) An asterisk preceding a number indicates that the phenotype caused by the gene at this locus is *not* influenced by genes at other loci; however, the disorder itself may be caused by mutations at multiple loci. Disorders for which no mode of inheritance has been determined do not carry asterisks. Finally, a pound sign (#) indicates that the phenotype is caused by two or more genetic mutations.

OMIM searches are very easy to perform. The search engine performs a simple query based on one or more words typed into a search window. A list of documents containing the query words is returned, and users can select one or more disorders from this list to look at the full text of the OMIM entry. The entries include information such as the gene symbol, alternate names for the disease, a description of the disease (including clinical, biochemical, and cytogenetic features), details on the mode of inheritance (including mapping information), a clinical synopsis, and references. A particularly useful feature is lists of allelic variants; a short description is given after each allelic variant of the clinical or biochemical outcome of that particular mutation. There are currently over 1,000 gene entries containing at least one allelic variant that either causes or is associated with a discrete phenotype in humans. Figure 7.19 shows an example of an allelic variant list, in this case for mutations observed in patients with McKusick-Kaufman syndrome (MKKS).

INTERNET RESOURCES FOR TOPICS PRESENTED IN CHAPTER 7

BLAST	http://www.ncbi.nlm.nih.gov/BLAST/
Cn3D	http://www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml
EndNote	http://www.niles.com/
Entrez	http://www.ncbi.nlm.nih.gov/Entrez/
FlyBase	http://flybase.bio.indiana.edu
GDB	http://www.gdb.org/
GeneCards	http://bioinfo.weizmann.ac.il/cards/
HomoloGene	http://www.ncbi.nlm.nih.gov/HomoloGene/
Kinemage	http://www.umass.edu/microbio/rasmol/mage.htm
LocusLink	http://www.ncbi.nlm.nih.gov/LocusLink/
MIPS	http://www.mips.biochem.mpg.de/
MMDB	http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml
OMIM	http://www.ncbi.nlm.nih.gov/Omim
PDB	http://www.rcsb.org/pdb/
RasMol	http://www.umass.edu/microbio/rasmol/
Reference Manager	http://www.risinc.com/
Sacch3D	http://www-genome.stanford.edu/Sacch3D/

SGD	http://genome-www.stanford.edu/Saccharomyces/
VAST	http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml
YPD	http://www.proteome.com/databases/index.html

PROBLEM SET

1. You have been watching the evening news and have just heard an interesting story regarding recent developments on the genetics of colorectal cancer. You would like to get some more information on this research, but the news story was short on details. The only hard information you have is that the principal investigator was Bert Vogelstein at the Johns Hopkins School of Medicine.
 - a. How many of the papers that Dr. Vogelstein has written on the subject of colorectal neoplasms are available through PubMed?
 - b. A paper by Hedrick and colleagues describes the role of the DCC gene product in cellular differentiation and colorectal tumorigenesis. Based on this study, what is the chromosomal location of the DCC gene?
 - c. DCC codes for a cell-surface-localized protein involved in tumor suppression. From what cell line and tissue type was the human tumor suppressor protein (*not* the precursor) isolated?
 - d. In the DCC human tumor suppressor protein *precursor*, what range of amino acids comprise the signal sequence?
2. Online Mendelian Inheritance in Man (OMIM) indicates that the development of colorectal carcinomas involves a dominantly acting oncogene coupled with the loss of several genes (such as DCC) that normally suppress tumorigenesis.
 - a. An allelic variant of DCC also involved in esophageal carcinoma has been cataloged in OMIM. What was the mutation at the amino acid level, and what biological effect did it have in patients?
 - b. Based on the MIM gene map, how many other genes have been mapped to the exact cytogenetic map location as DCC by *PCR of somatic cell hybrid DNA*?
 - c. The OMIM entry for DCC is coupled to the Mouse Genome Database at The Jackson Laboratory, showing that the corresponding mouse gene is located on mouse chromosome 18. What is the resultant phenotype of a *null* mutation of *Dcc* in the mouse?
3. A very active area of commercial research involves the identification and development of new sweeteners for use by the food industry. Whereas traditional sweeteners such as table sugar (sucrose) are carbohydrates, most current research is instead focusing on proteins which have an intrinsically sweet taste. Because these “sweet-tasting proteins” are much sweeter than their carbohydrate counterparts, they are, in essence, calorie free, since so little is used to achieve a sweet taste in food. The most successful example of such a protein is aspartame; however, aspartame is synthetic and does not occur in nature. Alternate, natural protein sources are being investigated, including a sweet tasting protein called monellin.
 - a. According to Ogata and colleagues, how much sweeter than ordinary sugar is monellin on both a molar and weight bases?

- b. Based on the SWISS-PROT entry for monellin chain B from serendipity berry, how many α -helices and β -strands does this protein possess?
- c. What residue (amino acid *and* position), when blocked, abolishes monellin's sweet taste?
- d. Three-dimensional structures are available for monellin. What *other* structure is most closely related to monellin structure 1MOL, as assessed by VAST P-value? Does this structure have the highest *sequence* similarity to 1MOL as well?
- e. The monellin structure is based on a single-chain fusion product. How do the stability and renaturation properties of the fusion product differ from that of the native protein?

REFERENCES

- Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., and Watson, J. D. (1994). *Molecular Biology of the Cell*, Garland Publishing, New York.
- Altschul, S., Gish, W., Miller, W., Myers, E., and Lipman, D. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- Baxevanis, A. D., and Landsman, D. (1995). The HMG-1 box protein family: classification and functional relationships. *Nucleic Acids Res.* 23, 1604–1613.
- Baxevanis, A. D. (2001). The Molecular Biology Database Collection: An Updated Compilation of Biological Database Resources. *Nucleic Acids Res.* 29, 1–7.
- Emmons, S., Phan, H., Calley, J., Chen, W., James, B., and Manseau, L. (1995). Cappuccino, a *Drosophila* maternal effect gene required for polarity of the egg and embryo, is related to the vertebrate limb deformity locus. *Genes Dev.* 9, 2484–2494.
- 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., Valle, D., and McKusick, V. A. (2000). Online Mendelian Inheritance in Man (OMIM). *Human Mutation* 15, 57–61.
- Madej, T., Gibrat, J.-F., and Bryant, S. (1995). Threading a database of protein cores. *Proteins.* 23, 356–369.
- McKusick, V. A. (1998). *Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders* (The Johns Hopkins University Press, Baltimore).
- Richardson, D., and Richardson, J. (1992). The kinemage: A tool for scientific communication. *Protein Sci.* 1, 3–9.
- Sayle, R., and Milner-White, E. (1995). RasMol: biomolecular graphics for all. *Trends Biochem. Sci.* 20, 374.
- Wilbur, W., and Coffee, L. (1994). The effectiveness of document neighboring in search enhancement. *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.

SEQUENCE ALIGNMENT AND DATABASE SEARCHING

Gregory D. Schuler

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

INTRODUCTION

There is a long tradition in biology of comparative analysis leading to discovery. For instance, Darwin's comparison of morphological features of the Galapagos finches and other species ultimately led him to postulate the theory of natural selection. In essence, we are performing the same type of analysis today when we compare the sequences of genes and proteins but in much greater detail. In this activity, the similarities and differences—at the level of individual bases or amino acids—are analyzed, with the aim of inferring structural, functional, and evolutionary relationships among the sequences under study. The most common comparative method is *sequence alignment*, which provides an explicit mapping between the residues of two or more sequences. In this chapter, only *pairwise alignments*, in which only two sequences are compared, will be discussed; the process of constructing *multiple alignments*, which involves more than two sequences, is discussed in Chapter 9. The number of sequences available for comparison has grown explosively since the 1970s, when development of rapid DNA sequencing methodology sparked the “big bang” of sequence information expansion. Comparison of one sequence to the entire database of known sequences is an important discovery technique that should be at the disposal of all molecular biologists. Over the past 30 years, improvements in the speed and sophistication of sequence alignment algorithms, not to mention perfor-

mance of computers, have more than kept pace with the growth in the size of the sequence databases. Today, with the complete genomes and large cDNA sequence collections available for many organisms, we are in the era of “comparative genomics,” in which the full gene complement of two organisms can be compared with one another.

THE EVOLUTIONARY BASIS OF SEQUENCE ALIGNMENT

One goal of sequence alignment is to enable the researcher to determine whether two sequences display sufficient similarity such that an inference of homology is justified. Although these two terms are often interchanged in popular usage, let us distinguish them to avoid confusion in the current discussion. *Similarity* is an observable quantity that might be expressed as, say, percent identity or some other suitable measure. *Homology*, on the other hand, refers to a conclusion drawn from these data that two genes share a common evolutionary history. Genes either are or are not homologous—there are no degrees for homology as there are for similarity. For example, Figure 8.1 shows an alignment between the homologous trypsin proteins from *Mus musculus* (house mouse) and *Astracus astracus* (broad-fingered crayfish), from which it can be calculated that these two sequences have 41% identity.

Bearing in mind the goal of inferring evolutionary relationships, it is fitting that most alignment methods try, at least to some extent, to model the molecular mechanisms by which sequences evolve. Although it is presumed that homologous sequences have diverged from a common ancestral sequence through iterative molecular changes, it is actually known what the ancestral sequence was (barring the possibility that DNA could be recovered from a fossil); all that can be observed are

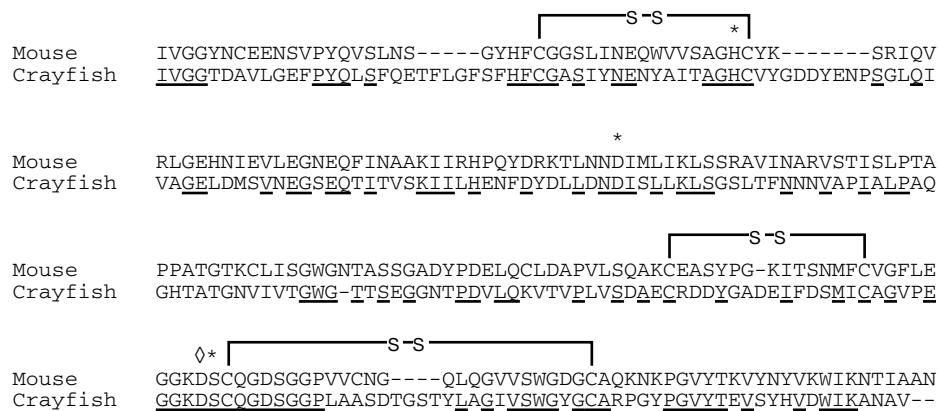


Figure 8.1. Conserved positions are often of functional importance. Alignment of trypsin proteins of mouse (SWISS-PROT P07146) and crayfish (SWISS-PROT P00765). Identical residues are underlined. Indicated above the alignments are three disulfide bonds (–S–S–), with participating cysteine residues conserved, amino acids side chains involved in the charge relay system (asterisk), and active site residue governing substrate specificity (diamond).

the raw sequences from extant organisms. The changes that occur during divergence from the common ancestor can be categorized as substitutions, insertions, and deletions. In the ideal case, in which a sequence alignment genuinely reflects the evolutionary history of two genes or proteins, residues that have been aligned but are not identical would represent substitutions. Regions where the residues of one sequence correspond to nothing in the other would be interpreted as either an insertion into one sequence or a deletion from the other. These *gaps* are usually represented in the alignment as consecutive dashes (or other punctuation character) aligned with letters. For example, the alignment in Figure 8.1 contains five gaps.

In a residue-by-residue alignment, it is often apparent that certain regions of a protein, or perhaps specific amino acids, are more highly conserved than others. This information may be suggestive as to which residues are most crucial for maintaining a protein's structure or function. In the trypsin alignment of Figure 8.1, the active site residues that determine substrate specificity and provide the "charge relay system" of serine proteases correspond to conserved positions, as do the cysteines residues that form several disulfide bonds important for maintaining the enzyme's structure. On the other hand, there may be other positions that do not play a significant functional role yet happen to be identical for historical reasons. Particular caution should be taken when the sequences are taken from very closely related species because similarities may be more reflective of history than of function. For example, regions of high sequence similarity between mouse and rat homologs may simply be those that have not had sufficient time to diverge. Nevertheless, sequence alignments provide a useful way to gain new insights by leveraging existing knowledge, such as deducing structural and functional properties of a novel protein from comparisons to those that have been well studied. It must be emphasized, however, that these inferences should always be tested experimentally and *not* assumed to be correct based on computational analysis alone.

By observing a surprisingly high degree of sequence similarity between two genes or proteins, we might infer that they share a common evolutionary history, and from this it might be anticipated that they should also have similar biological functions. But again, this should be treated as hypothetical until tested experimentally. Zeta-crystallin, for instance, is a component of the transparent lens matrix of the vertebrate eye. However, on the basis of extended sequence similarity, it can be inferred that its homolog in *E. coli* is the metabolic enzyme quinone oxidoreductase (Fig. 8.2). Despite the common ancestry, the function has changed during evolution (Gonzalez et al., 1994). This is analogous to a railroad car that has been converted into a diner: inspection of the exterior structure reveals the structure's history, but relying exclusively on this information may lead to an erroneous conclusion about its current function. When a gene adapts to a new niche, it might also be anticipated that the pattern of conserved positions would change. For example, active site residues should be conserved so long as the protein plays a role in catalysis but could drift once the protein takes on a different function.

The earliest sequence alignment methods were applicable to a simple type of relationship in which the sequences show easily detectable similarity along their entire lengths. An alignment that essentially spans the full extents of the input sequences is called a *global alignment*. The trypsin and quinone oxidoreductase/zeta-crystallin alignments discussed above are both examples of global alignments. Proteins consisting of a single globular domain can often be aligned using a global strategy as can any homologous sequences that have not diverged substantially.

```

Human-ZCr      MATGQKLMRAVRVFEFGGPEVLKLRSDIAVPIPKDHQVLIKVHACGVNPNVETYIRSGTYS
Ecoli-QOR      -----MATRIEFHKHGGPEVLQA-VEFTPADPAENEIQVENKAIGINFIDTYIRSGLYP
                . . . . . * * * * * . . . . . * * * * *
Human-ZCr      RKPLLPTPGSDVAGVIEAVGDNASAFKKGDRVFTSSTISGGYAEYALAADHTVYKLPK
Ecoli-QOR      -PPSLPSGLGTEAAGIVSKVSGVKHIKAGDRVVYAQSALGAYSSVHNI IADKAAILPAA
                * * * * * * * * * * * * * * * * * * * * * *
Human-ZCr      LDFKQGAAGIPYFTAYRALIHSACVKAGESVLVHGASGGVGLAACQIARAYGLKILGTA
Ecoli-QOR      ISFEQAAASFLKGLTVYYLLRKYEIKPDEQFLFHAAAGGVGLIACQWAKALGAKLIGTV
                . * * * * . * * * * . * * * * * * * * * * * * * * *
Human-ZCr      GTEEGQKIVLQNGAHEVFNHREVNYIDKIKKYVGEKGDIDIIEMLANVNLKDLSSLSHG
Ecoli-QOR      GTAQKAQSALKAGAWQVINYREEDLVERLKEITGGKKVRVVYDSVGRDWTWERSLDCLQRR
                ** . . . * . * * * * * * * * * * * * * * * * * *
Human-ZCr      GRVIVVG-SRGTIEINPRDTMAKES---SIIGVTLFSSTKEEFQYYAALQAGMEIGWL
Ecoli-QOR      GLMVSFGNSSGAVTGVNLTGILNKGSLYVTRPSLQGYITTREELTEASNELFSLIASGVI
                * . . * * * * * . . . . . * * * * * * * * * * *
Human-ZCr      KPVIGSQ--YPLEKVAEAHENIIHGSGATGKMILL
Ecoli-QOR      KVDVAEQQKYPLKDAQRAHE-ILESRA TQSSLLIP
                * . * * * * * * * * * * * * *

```

Figure 8.2. Optimal global sequence alignment. Alignment of the amino acid sequences of human zeta-crystallin (SWISS-PROT Q08257) and *E. coli* quinone oxidoreductase (SWISS-PROT P28304). It is an optimal global alignment produced by the CLUSTAL W program (Higgins et al., 1996). Identical residues are marked by asterisks below the alignment, and dots indicate conserved residues.

THE MODULAR NATURE OF PROTEINS

Many proteins do not display global patterns of similarity but instead appear to be mosaics of modular domains (Baron et al., 1991; Doolittle and Bork, 1993; Patthy, 1991). One example of this is illustrated in Figure 8.3, which shows the modular structure of two proteins involved in blood clotting: coagulation factor XII (F12) and tissue-type plasminogen activator (PLAT). Besides the catalytic domain, which provides the serine protease activity, these proteins have different numbers of other structural modules: two types of fibronectin repeats, a domain with similarity to epidermal growth factor, and a module that is called a “kringle” domain. These modules can be repeated or appear in different orders. Patterns of modularity often arises by in-frame exchange of whole exons (Patthy, 1991). Global alignment methods do not take this phenomenon into account, which is understandable considering that they were developed before the exon/intron structure of genes had been discov-

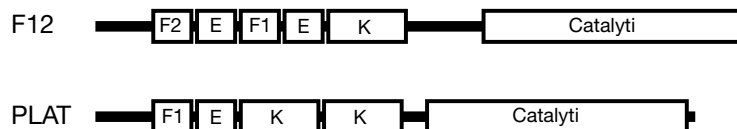


Figure 8.3. Modular structure of two proteins involved in blood clotting. Schematic representation of the modular structure of human tissue plasminogen activator and coagulation factor XII. A module labeled C is shared by several proteins involved in blood clotting. F1 and F2 are frequently repeated units that were first seen in fibronectin. E is a module resembling epidermal growth factor. A module known as a “kringle domain” is denoted K.

ered. In most cases, it is advisable to instead use a sequence comparison method that can produce a *local alignment*. Such an alignment consists of paired subsequences that may be surrounded by residues that are completely unrelated. Consequently, users should bear in mind that some local similarities could be missed if a global alignment strategy is applied inappropriately. Another obvious case in which local alignments are desired is the alignment of the nucleotide sequence of a spliced mRNA to its genomic sequence, where each exon would be a distinct local alignment.

Dot-matrix representations have enjoyed a widespread popularity, in part because of their ability to reveal complex relationships involving multiple regions of local similarity (Fitch, 1969; Gibbs and McIntyre, 1970). An example of this approach is shown in Figure 8.4, in which the F12 and PLAT protein sequences have been compared using dotter (Sonnhammer and Durban, 1996). The basic idea is to use the sequences as the coordinates of a two-dimensional graph and then plot points of correspondence within its interior. Each dot usually indicates that, within some small window, the sequence similarity is above some cutoff (or a range of cutoffs with the use of dotter, each plotted using a different shade of gray). When two sequences are

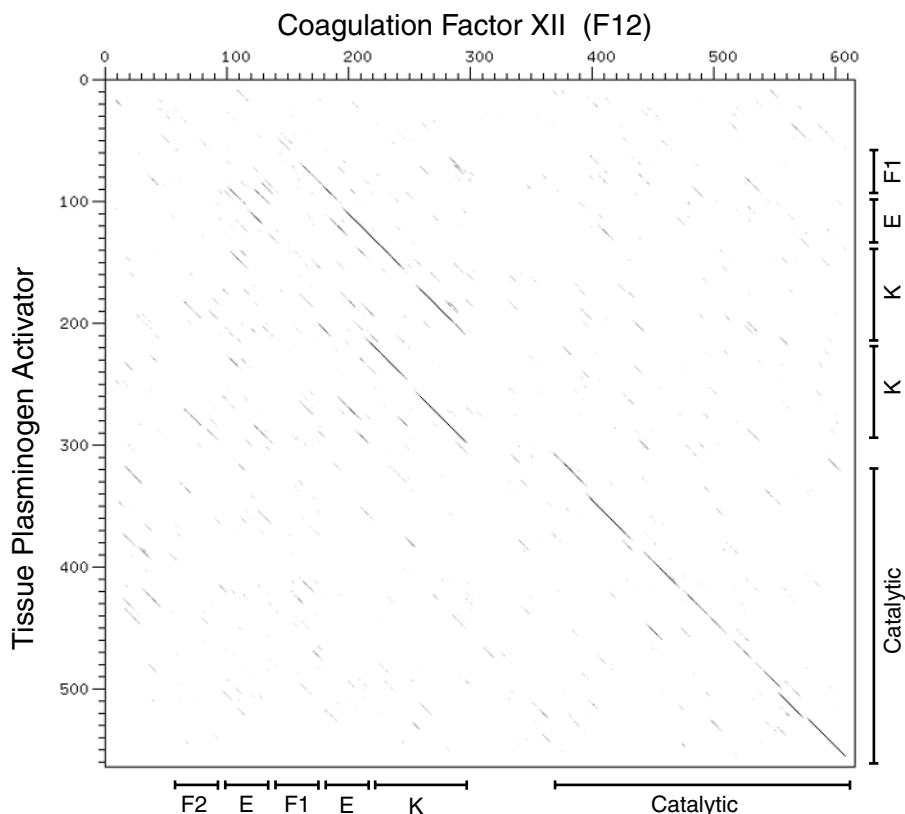


Figure 8.4. Dot matrix sequence comparison. Dot matrix comparison of the amino acid sequences of human coagulation factor XII (F12; SWISS-PROT P00748) and tissue plasminogen activator (PLAT; SWISS-PROT P00750). The figure was generated using the dotter program (Sonnhammer and Durban, 1996).

consistently matching over an extended region, the dots will merge to form a diagonal line segment. It is instructive to compare the positions of the diagonals in dot-matrix of Figure 8.4 with the known modular structure of the two proteins. In particular, note the way in which repeated domains appear: starting with the kringle domain in the PLAT and scanning horizontally, two diagonal segments may be seen, corresponding to the two kringle domains present in the F12 sequence. Although more sophisticated methods for finding local similarities are now available (discussed below), dot-matrix representations have remained popular as illustrative tools.

In a dot-matrix representation, certain patterns of dots may appear to sketch out a “path,” but it is up to the viewer to deduce the alignment from this information. Another graphical representation known as a path graph provides an explicit representation of an alignment. Figure 8.5 illustrates the relationship between the dot-matrix, path graph, and alignment representations for the EGF similarity domain present in both the tissue-type plasminogen activator (PLAT) and the urokinase-type plasminogen activator (PLAU) proteins. To understand a path graph, imagine a two-dimensional lattice in which the vertices represent points between the sequence residues (as opposed to the residues themselves, as in the case of the dot-matrix). An edge that connects two vertices along a diagonal corresponds to the pairing of one

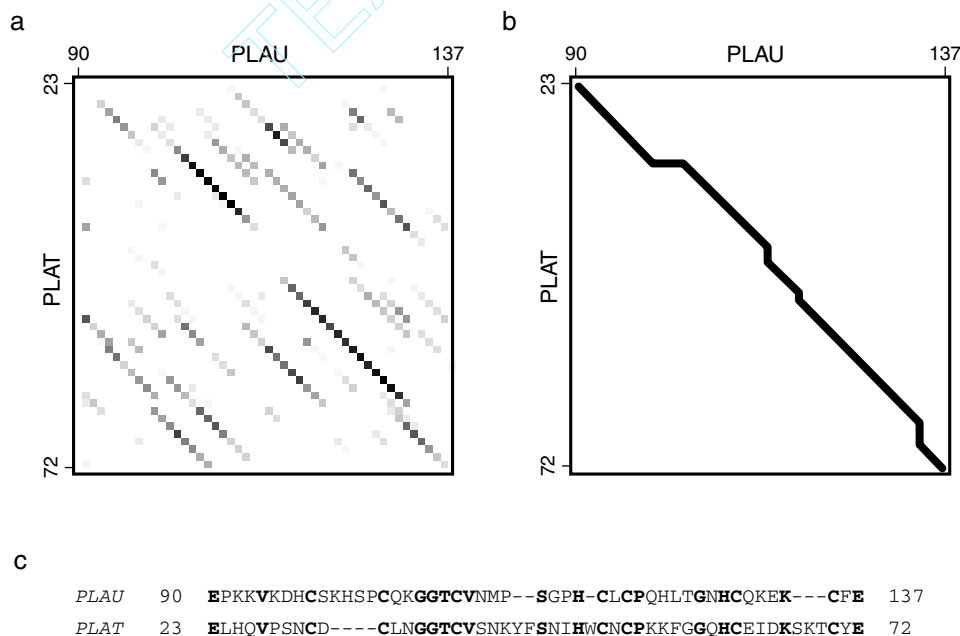


Figure 8.5. Dot-matrix, path graph, and alignment. All three views represent the alignment of the EGF similarity domains in the human urokinase plasminogen activator (PLAU; SWISS-PROT P00749) and tissue plasminogen activator (PLAT; SWISS-PROT P00750) proteins. (a) The entire proteins were compared with dotter and an enlargement of the small region corresponding to the EGF domain is shown here. (b) The path graph representation of the alignment found by BLASTP. (c) The BLASTpgp alignment represented in the familiar text form.

residue from each sequence. Horizontal and vertical edges pair a residue from one sequence with nothing in the other; in other words, these edges constitute a gap in the alignment. The entire graph corresponds to the *search space*, which must be examined for potential alignments. Each possible path through this space corresponds to exactly one alignment.

OPTIMAL ALIGNMENT METHODS

For any but the most trivial problems, the total number of distinct alignments is extraordinarily large, so it is usually of interest to identify the “best” one among them (or the several best ones). This is where the concept of representing an alignment as a path pays off. Many problems in computer science can be reduced to the task of finding the optimal path through a graph (for instance, the problem of finding the most efficient way to route a telephone call from New York to San Francisco), and efficient algorithms have been developed for this purpose. One requirement is a means of assigning a quality score to each possible path (alignment). Normally, this is accomplished by summing the incremental contributions of each step along its route. More sophisticated scoring schemes are discussed below, but for now let us assume that some positive incremental scores will be used for aligning identical residues, with negative scores used for substitutions and gaps. According to this definition of alignment quality, finding the path whose total score is maximal will give us the best sequence alignment.

What is today known as the Needleman-Wunsch algorithm is an application of a best-path strategy called *dynamic programming* to the problem of finding optimal sequence alignments (Needleman and Wunsch, 1970). The basic idea behind dynamic programming comes from the observation that any partial subpath that ends at a point along the true optimal path must itself be the optimal path leading up to that point. Thus, the optimal path can be found by incremental extension of optimal subpaths. In the basic Needleman-Wunsch formulation, the optimal alignment must extend from beginning to end in both sequences, that is, from the top-left corner in the search space to bottom-right (as it is typically drawn). In other words, it seeks global alignments. A simple modification to the basic strategy allows the optimal local alignment to be found (Smith and Waterman, 1981). The path for this alignment does not need to reach the edges of the search graph but may begin and end internally. Such an alignment would be locally optimal if its score cannot be improved either by increasing or decreasing the extent of the alignment. The Smith-Waterman algorithm relies on a property of the scoring system in which the cumulative score for a path will decrease in regions of poorly matching sequences (the scoring systems described below satisfy this criterion). When the score drops to zero, extension of path is terminated and a new one can begin. There can be many individual paths bounded by regions of poorly matching sequence; the one with the highest score is reported as the optimal local alignment.

It is important to bear in mind that optimal methods always report the best alignment that can be achieved, even if it has no biological meaning. On the other hand, when searching for local alignments there may be several significant alignments, so it is a mistake to look only at the optimal one. Refinements to the Smith-Waterman algorithm were proposed for detecting the k best nonintersecting local alignments (Altschul and Erickson, 1986; Sellers, 1984; Waterman and Eggert, 1987).

These ideas were later extended in the development of the SIM algorithm (Huang et al., 1990). A program called lalign (distributed with the FASTA package) provides a useful implementation of SIM (Pearson, 1996). Looking for suboptimal alignments is especially important when comparing multimodule proteins. This is illustrated in Figure 8.6, in which the lalign program was used to find the three best local alignments of the human coagulation factor IX and factor XII proteins. The second and

```

Comparison of:
(A) f9-human.aa >F9 gi|119772|sp|P00740|FA9_HUMAN COAGULATION FA - 461 aa
(B) f12-hum.aa >F12 gi|119763|sp|P00748|FA12_HUMAN COAGULATION - 615 aa
using protein matrix

① 35.4% identity in 254 aa overlap; score: 358

      220      230      240      250      260      270
F9   QSFNDFTRVVGEDAKPGQFPWQVVLNGKVDAFCGGSIVNEKWIIVTAAHCVE---TGVKI
     .....: : : : : : : : : : : : : : : : : : : : : : : : : : :
F12  KSLSSMTRVVGGLVALRGAHPYIAALY-WGHSFCAGSLIAPCWVLTAAHCLQDRPAPEDL
     370      380      390      400      410      420

      280      290      300      310      320      330
F9   TVVAGEHNIEETEHETEQKRNVIRIIPHHNYNAAINKYNHDIALLELDEPL----VLNSY
     ::: : . . . . . : : : : : : : : : : : : : : : : : : : : :
F12  TVVLQQERRNHSCEPCQTLAVRSYRLHEAFSPV--SYQHDLALLRLQEDADGSCALLSPY
     430      440      450      460      470      480

      340      350      360      370      380
F9   VTPICIAADKEYTNIFLKFSGYVSGWGRVFKGRS-ALVLQYLRVPLVDRATCLRSTKF-
     : : : : : : : : : : : : : : : : : : : : : : : : : :
F12  VQPVCLPSGAARPSETTLCQ--VAGWGHQFEGAEYASFLQEAQVPFSLERCSAPDVHG
     490      500      510      520      530

      390      400      410      420      430      440
F9   -TIYNNMFCAGFHEGGRDSCQGDSSGGPHVTEVEGTS---FLTGIISWGEECAMKGYGIY
     . : : : : : : : : : : : : : : : : : : : : : : : : : :
F12  SSILPGMLCAGFLEGGTDACQGDSSGGLVCEQAERRLTLQGIISWGS CGDRNKPGVY
     540      550      560      570      580      590

      450
F9   TKVSRVYVNWIKEKT
     : : : : : : : : : :
F12  TDVAYYLAWIREHT
     600      610

-----

② 34.7% identity in 49 aa overlap; score: 120

      100      110      120      130      140
F9   VDGDCQCESNPCLNGGSKDDINSYECWCPFGFEGKNCLELDVTCNIKNGR
     .....: : : : : : : : : : : : : : : : : : : : :
F12  LASQACRTNPCLHGGRCLVEVEGHRLCHCPVGYTGPFCDVDTKASCYDGR
     180      190      200      210      220

-----

③ 33.3% identity in 36 aa overlap; score: 87

      100      110      120
F9   DQCESN-PCLNGGSKDDINSYECWCPFGFEGKNCE
     .....: : : : : : : : : :
F12  DHCSKHSPCQKGGTCVNMPSPGPHCLCPQHLTGNGHCQ
     100      110      120      130

-----

```

Figure 8.6. Optimal and suboptimal local alignments. The three best alignments found when using lalign to align the sequences of human coagulation factor IX (F9; SWISS-PROT 900740) and coagulation factor XII (F12; SWISS-PROT P00748).

third alignments represent functional modules that would have been missed by a standard Smith-Waterman search, which would have reported only the first (optimal) alignment.

SUBSTITUTION SCORES AND GAP PENALTIES

The scoring system described above made use of a simple match/mismatch scheme, but, when comparing proteins, we can increase sensitivity to weak alignments through the use of a *substitution matrix*. It is well known that certain amino acids can substitute easily for one another in related proteins, presumably because of their similar physicochemical properties. Examples of these “conservative substitutions” include isoleucine for valine (both small and hydrophobic) and serine for threonine (both polar). When calculating alignment scores, identical amino acids should be given greater value than substitutions, but conservative substitutions should also be greater than nonconservative changes. In other words, a range of values is desired. Furthermore, different sets of values may be desired for comparing very similar sequences (e.g., a mouse gene and its rat homolog) as opposed to highly divergent sequences (e.g., mouse and yeast genes). These considerations can be dealt with in a flexible manner through the use of a substitution matrix, in which the score for any pair of amino acids can be easily looked up.

The first substitution matrices to gain widespread usage were those based on the point accepted mutation (PAM) model of evolution (Dayhoff et al., 1978). One PAM is a unit of evolutionary divergence in which 1% of the amino acids have been changed. This does not imply that after 100 PAMs every amino acid will be different; some positions may change several times, perhaps even reverting to the original amino acid, whereas others may not change at all. If there were no selection for fitness, the frequencies of each possible substitution would be primarily influenced by the overall frequencies of the different amino acids (called the *background frequencies*). However, in related proteins, the observed substitution frequencies (called the *target frequencies*) are biased toward those that do not seriously disrupt the protein’s function. In other words, these are point mutations that have been “accepted” during evolution. Dayhoff and coworkers were the first to explicitly use a *log-odds* approach, in which the substitution scores in the matrix are proportional to the natural log of the ratio of target frequencies to background frequencies. To estimate the target frequencies, pairs of very closely related sequences (which could be aligned unambiguously without the aid of a substitution matrix) were used to collect mutation frequencies corresponding to 1 PAM, and these data were then extrapolated to a distance of 250 PAMs. The resulting PAM250 matrix is shown in Figure 8.7. Although PAM250 was the only matrix published by Dayhoff et al. (1978), the underlying mutation data can be extrapolated to other PAM distances to produce a family of matrices. When aligning sequences that are highly divergent, best results are obtained at higher PAM values, such as PAM200 or PAM250. Matrices constructed from lower PAM values can be used if the sequences have a greater degree of similarity (Altschul, 1991).

The BLOSUM substitution matrices have been constructed in a similar fashion, but make use of a different strategy for estimating the target frequencies (Henikoff and Henikoff, 1992). The underlying data are derived from the BLOCKS database (Henikoff and Henikoff, 1991), which contains local multiple alignments (“blocks”)

STATISTICAL SIGNIFICANCE OF ALIGNMENTS

For any given alignment, one can calculate a score representing the quality of the alignment, but an important question is whether or not this score is high enough to provide evidence of homology. In addressing this question, it is helpful to have some notion of how high of a score can be expected due purely to chance alone. Unfortunately, there is no mathematical theory to describe the expected distribution of scores for global alignments. One of the few methods available for assessing their significance is to compare the observed alignment score with those of many alignments made from random sequences of the same length and composition as those under study (Altschul and Erickson, 1985; Fitch, 1983).

However, for local alignments, the situation is much better. A statistical model advanced by Karlin and Altschul provides a mathematical theory to describe the expected distribution of random local alignment scores (Dembo et al., 1984; Karlin and Altschul, 1990). The form of the probability density function is known as the *extreme value distribution*. This is worth noting because application of the more familiar normal distribution can result in greatly exaggerated claims of significance. The extreme value distribution is characterized by two parameters, K and λ , which should be tailored for the particular set of alignment scoring rules and residue background frequencies at hand. Although analytical calculation of these parameters can currently be done only for alignments that lack gaps, methods have been developed to estimate appropriate values of K and λ for gapped alignments (Altschul and Gish, 1996; Waterman and Vingron, 1994). By relating an observed alignment score S to the expected distribution, it is possible to calculate statistical significance in the form of an *E value*. The simple interpretation of an *E value* is the number of alignments with scores at least equal to S that would be expected by chance alone. The significance of an alignment also depends on the size of the search space that was used; larger databases produce more chance alignments. The search space has typically been calculated as the product of the sequence lengths, but, for correct statistics, the lengths must be reduced by the expected length of a local alignment to avoid an “edge effect” (Altschul and Gish, 1996). This is due to the fact that an alignment that begins near the edge of the search space will run out of sequence before it can achieve a significant score.

DATABASE SIMILARITY SEARCHING

The discussion so far has focused on the alignment of specific pairs of sequences, but, for a newly determined sequence, one would generally have no way of knowing the appropriate sequence (or sequences) to use in such a comparison. Database similarity searching allows us to determine which of the hundreds of thousands of sequences present in the database are potentially related to a particular sequence of interest. This process sometimes leads to unexpected discoveries. The first “eureka moment” with this strategy came when the viral oncogene *v-sis* was found to be a modified form of the normal cellular gene that encodes platelet-derived growth factor (Doolittle et al., 1983; Waterfield et al., 1983). At the time of this discovery, sequence databases were small enough that such a finding might have been considered surprising. Today, however, it would be much more surprising to perform a database search and *not* get a hit. Large numbers of partial sequences representing novel

impact on the effectiveness of a search. Furthermore, there are various interfaces to these facilities such as console-style commands, Web-based forms, and E-mail. Figure 8.10 shows an example of performing a database search using the BLAST Web interface. One advantage of this approach is that, for any interesting alignment observed, complete annotation and literature citations can be obtained simply by following hypertext links to the original sequences entries and related on-line literature.

Current sequence databases are immense and have continued to increase at an exponential rate, making straightforward application of dynamic programming methods impractical for database searching. One solution is to use massively parallel computers and other specialized hardware, but, for the purposes of this discussion, we will consider only what can be done using general-purpose computers. With optimal methods being impractical, it is necessary to resort to heuristic methods, which make use of approximations to significantly speed up sequence comparisons, but with a small risk that true alignments can be missed. One heuristic method is based on the strategy of breaking a sequence up into short runs of consecutive letters called *words*. Word-based methods were introduced in the early 1980s and are used by virtually all popular search programs in use today (Wilbur and Lipman, 1983). The basic idea is that an alignment representing a true sequence relationship will contain at least one word that is common to both sequences. These *word hits* can be identified extremely rapidly by preindexing all words from the query and then consulting the index as the database is scanned.

FASTA

The first widely-used program for database similarity searching was FASTA (Lipman and Pearson, 1985; Pearson and Lipman, 1988; Pearson, 2000). To achieve a high degree of sensitivity, this program performs optimized searches for local alignments using a substitution matrix. However, as noted above, it would take a substantial amount of time to apply this strategy exhaustively. To improve speed, the program uses the observed pattern of word hits to identify potential matches before attempting the more time-consuming optimized search. The trade-off between speed and sensitivity is controlled by the *ktup* parameter, which specifies the size of a word. Increasing the value of *ktup* decreases the number of background word hits (i.e., those that do not mark the position of an optimal alignment). This, in turn, decreases the amount of optimized searching required and improves overall search speed. The default *ktup* value for comparing proteins is 2, but, for finding very distant relationships, it is recommended that it be reduced to 1.

The FASTA program does not investigate every word hit encountered, but instead looks initially for segments containing several nearby hits. By using a heuristic method, these segments are assigned scores and the score of the best segment found appears in the output as the *init1* score (Figure 8.9a). Several segments may then be combined and a new *initn* score is calculated from the ensemble. Most potential matches are then further evaluated by performing a search for a gapped local alignment that is constrained to a diagonal band centered around the best initial segment. The score of this optimized alignment is shown in the output as the *opt* score. For those alignments finally reported (a user-specified number from the top of the hit list), a full Smith-Waterman alignment search (without the constraining band) is performed. An example, of such an alignment is shown in Figure 8.9b. It should be noted that only the single optimal alignment is produced for each database sequence.

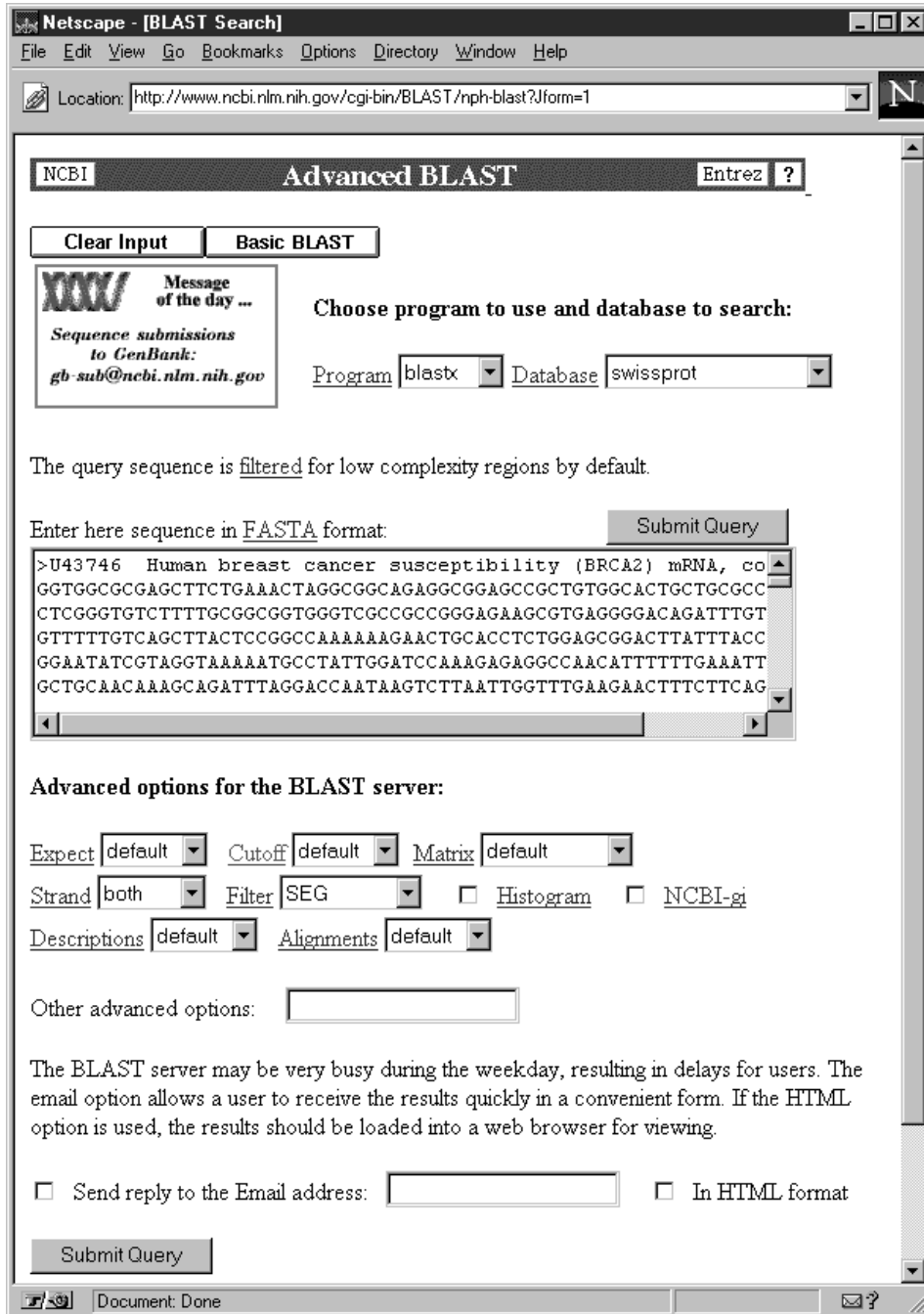


Figure 8.10. Database similarity search on the World Wide Web. The figure illustrates the use of the NCBI BLAST Web front end. The query sequence should be pasted from the clipboard into the large text field (where the sequence of U43746 is shown in this figure). Other essential elements of the search are the name of the search program and the database, both of which may be selected from drop-down lists. Additional optional parameters may be set if desired. In addition to this “Advanced BLAST” form, there is also a “Basic BLAST” form in which the advanced options are hidden. In either case, simply press the Submit Query button to begin the search.

As pointed out above, meaningful alignments can be missed by this approach if the proteins contain multiple modules. Consequently, it is recommended that matching sequences be further analyzed with the lalign program.

Beginning with version 2.0, FASTA provides an estimate of the statistical significance of each alignment found. The program assumes an extreme value distribution for random scores but with the use of a rewritten form of the probability density function in which the expected score is a linear function of the natural log of the length of the database sequence. Simple linear regression can then be used to calculate a normalized Z-score for each alignment. Finally, an expectation E is calculated, which gives the expected number of random alignments with Z-scores greater than or equal to the value observed.

BLAST

The BLAST programs introduced a number of refinements to database searching that improved overall search speed and put database searching on a firm statistical foundation (Altschul et al., 1990). One innovation introduced in BLAST is the idea of *neighborhood words*. Instead of requiring words to match exactly, a word hit is achieved if the word taken from the subject sequence has a score of at least T when a comparison is made using a substitution matrix to the word from the query. This strategy allows the word size (W) to be kept high (for speed) without sacrificing sensitivity. Thus, T becomes the critical parameter determining speed and sensitivity and W is rarely varied. If the value of T is increased, the number of background word hits will go down and the program will run faster. Reducing T allows more distant relationships to be found.

The occurrence of a word hit is followed by an attempt to find a locally optimal alignment whose score is at least equal to a score cutoff S . This is accomplished by iteratively extending the alignment both to the left and to the right, with accumulation of incremental scores for matches, mismatches, and the introduction of gaps. In practice, it is more convenient to specify an expectation cutoff E , which the program internally converts to an appropriate value of S (which would depend on the search context). In regions where matching residues are scarce, the cumulative score will begin to drop. As the mismatch and gap penalties mount, it becomes less likely that the score will rebound and ultimately reach S . This observation provides the basis for an additional heuristic whereby the extension of a hit is terminated when the reduction in score (relative to the maximum value encountered) exceeds the score dropoff threshold X . Using smaller values of X improves performance by reducing the time spent on unpromising hit extensions, at the expense of occasionally missing some true alignments.

There are several variants of BLAST, each distinguished by the type of sequence (DNA or protein) of the query and database sequences (see Table 8.1). The BLASTP program compares a protein query to a protein database. The corresponding program for nucleotide sequences is BLASTN. If the sequence types differ, the DNA sequence can be translated by the program (in all six reading frames) and compared to the protein sequence. BLASTX compares a DNA query sequence to the protein database, which is useful for analyzing new sequence data and ESTs. For a protein query against a nucleotide database, use the TBLASTN program. This is useful for finding unannotated coding regions in database sequences. A final variant is used only in

TABLE 8.1. BLAST Programs

Program	Query	Database	Comments
BLASTP	Protein	Protein	Uses substitution matrix for finding distant relationships; SEG filtering available
BLASTN	Nucleotide	Nucleotide	Tuned for very high-scoring matches, not distant relationships
BLASTX	Nucleotide (translated)	Protein	Useful for analysis of new DNA sequences and ESTs
TBLASTN	Protein	Nucleotide (translated)	Useful for finding unannotated coding regions in database sequences
TBLASTX	Nucleotide (translated)	Nucleotide (translated)	May be useful for EST analysis, but computationally intensive

specialized situations but is mentioned here for the sake of completeness: TBLASTX takes DNA query and database sequences, translates them both, and compares them as protein sequences. This program is mainly useful for comparisons of ESTs, where it is suspected that sequences may have coding potential even though the exact coding region has not been determined.

All of these programs make use of sequence databases located on server machines, which obviates the need for any local database maintenance. Some protein and nucleotide sequences databases currently available from the NCBI for BLAST searching are listed in Tables 8.2 and 8.3. For routine searches, the *nr* database provides comprehensive collections of both amino acid and nucleotide sequence data, with redundancy reduced by merging sequences that are completely identical. To examine all sequences submitted or updated within the last 30 days, a database called *month* is provided. Both *nr* and *month* are updated on a daily basis. Several other databases listed in Tables 8.2 and 8.3 are useful in more specialized situations, such as comparing against the complete genomes of model organisms (*ecoli* or *yeast*), searching specific classes of sequences (*est* or *sts*), or testing for the presence of contaminating or otherwise problematic sequences (*vector*, *alu*, or *mito*).

TABLE 8.2. Protein Sequence Databases for use with BLAST

Database	Description
<i>nr</i>	Non-redundant merge of SWISS-PROT, PIR, PRF, and proteins derived from GenBank coding sequences and PDB atomic coordinates
<i>month</i>	Subset of <i>nr</i> which is new or modified within the last 30 days
<i>swissprot</i>	The SWISS-PROT database
<i>pdb</i>	Amino acid sequences parsed from atomic coordinates of three-dimensional structures
<i>ecoli</i>	Complete set of proteins encoded by the <i>E. coli</i> genome
<i>yeast</i>	Complete set of proteins encoded by the <i>S. cerevisiae</i> genome
<i>drosoph</i>	Complete set of proteins encoded by the <i>D. melanogaster</i> genome

TABLE 8.3. Nucleotide Sequence Databases for use with BLAST

Database	Description
<i>nr</i>	Nonredundant GenBank, excluding the EST, STS, and GSS divisions
<i>month</i>	Subset of <i>nr</i> , which is new or modified within the last 30 days
<i>est</i>	GenBank EST division (expressed sequence tags)
<i>sts</i>	GenBank STS division (sequence tagged sites)
<i>htgs</i>	GenBank HTG division (high-throughput genomic sequences)
<i>gss</i>	GenBank GSS division (genome survey sequences)
<i>ecoli</i>	Complete genomic sequence of <i>E. coli</i>
<i>yeast</i>	Complete genomic sequence of <i>S. cerevisiae</i>
<i>drosoph</i>	Complete genomic sequence of <i>D. melanogaster</i>
<i>mito</i>	Complete genomic sequences of vertebrate mitochondria
<i>alu</i>	Collection of primate Alu repeat sequences
<i>vector</i>	Collection of popular cloning vectors

An example of a BLAST search will serve to introduce various elements of a search output. For the example in Figure 8.11, the amino acid sequence of one of the Alzheimer's disease susceptibility proteins (conceptual translation of GenBank L43964) was used as the query in a TBLASTX search of the *est* database. One goal of such a search would be to identify cDNA clones for potential homologs in model organisms, thereby opening the door for experimental studies that would not be practical in humans (the clones corresponding to EST sequences are readily available). Each of the EST sequences in the database is translated in all reading frames before they are compared against the Alzheimer's protein sequence. Figure 8.11a shows the hit list produced by this search. The first two columns give the identifiers and descriptions for each sequence having a significant match. Although the definitions are truncated in this overview, the figures shows that sequences from both mouse and *Drosophila* are represented. The next column gives the reading frame that produced the best alignment (although there may be hits to translations from other frames as well). The next three columns provide the score of the best alignment, the sum *P*-value, and the number of HSPs that were used in the *P*-value calculation. The alignment involving one of the *Drosophila* ESTs (marked by the arrow) is shown in Figure 8.11b. There are actually two alignments involved, and scores are provided for each. In each case, the conceptual translation of the EST is shown aligned with the query sequence. Identical amino acids are echoed to the text line in between the sequences, and plus (+) symbols are used to indicate nonidentical residues that have positive substitution scores (i.e., conservative substitutions). It is noteworthy that the two alignments arise from different reading frames and are adjacent to one another, as can be seen from the sequence coordinates. This pattern is indicative of a reading frame error in the EST sequence. When analyzing sequence single-pass data, it is extremely useful to have tools that are relatively error tolerant.

DATABASE SEARCHING ARTIFACTS

A query sequence that contains repetitive elements is likely to produce many false and confounding database matches. One clue that this may be a problem is the

a

Sequences producing High-scoring Segment Pairs:	Reading Frame	High Score	Smallest Sum Probability P(N)	N
gb AA056325 AA056325 zf53a03.s1 Soares retina N2b4HR H...	+3	724	3.4e-102	2
gb T03796 T03796 IB913 Infant brain, Bento Soares ...	+3	567	2.6e-78	2
gb AA260597 AA260597 mx76g09.r1 Soares mouse NML Mus m...	+2	239	4.9e-53	4
gb H86456 H86456 yt01b06.s1 Homo sapiens cDNA clon...	+2	323	4.3e-52	4
gb N24576 N24576 yx72a04.s1 Homo sapiens cDNA clon...	+1	365	5.5e-47	2
gb AA265273 AA265273 mx91c12.r1 Soares mouse NML Mus m...	+2	239	6.4e-41	2
gb AA237206 AA237206 mx18e01.r1 Soares mouse NML Mus m...	+3	159	1.5e-40	3
gb R14600 R14600 yf34b10.r1 Homo sapiens cDNA clon...	+1	278	1.5e-40	2
gb AA200706 AA200706 mu03f12.r1 Soares mouse 3NbMS Mus...	+1	343	1.9e-40	1
gb AA045064 AA045064 zk77f12.s1 Soares pregnant uterus...	-3	269	2.3e-37	2
gb AA087434 AA087434 mm28a04.r1 Stratagene mouse skin ...	+3	322	3.6e-37	1
gb R05907 R05907 ye93h02.r1 Homo sapiens cDNA clon...	+3	252	7.7e-37	2
gb AA268820 AA268820 vb01c10.r1 Soares mouse NML Mus m...	+2	234	7.7e-35	2
gb AA162310 AA162310 mn44a07.r1 Beddington mouse embry...	+1	134	8.3e-34	3
gb N27820 N27820 yx54h10.r1 Homo sapiens cDNA clon...	+3	154	7.8e-29	2
gb AA234907 AA234907 zs38f03.r1 Soares NhHMPu S1 Homo ...	+2	155	1.8e-28	2
gb AA231081 AA231081 mw11d11.r1 Soares mouse 3NME12 5 ...	+3	134	8.8e-23	2
gb H91652 H91652 ys80c04.s1 Homo sapiens cDNA clon...	-3	215	3.7e-22	1
gb H50532 H50532 yo30h08.s1 Homo sapiens cDNA clon...	-2	211	1.2e-21	1
gb AA150236 AA150236 zl03c01.r1 Soares pregnant uterus...	+1	159	5.0e-21	2
gb AA144382 AA144382 mr15d12.r1 Soares mouse 3NbMS Mus...	+3	159	7.6e-21	2
gb AA390557 AA390557 LD09473.5prime LD Drosophila Embr...	+3	130	1.6e-20	2
gb AA210480 AA210480 mo86b03.r1 Beddington mouse embry...	+2	128	2.0e-20	3
gb H19012 H19012 ym44b02.r1 Homo sapiens cDNA clon...	+2	134	5.9e-20	2
gb AA283084 AA283084 zt14g09.s1 Soares NbHTGBC Homo sa...	-3	175	2.3e-19	2
gb H25759 H25759 yl14d01.s1 Homo sapiens cDNA clon...	-2	185	5.0e-18	1
gb H33787 H33787 EST110123 Rattus sp. cDNA 5' end ...	+1	137	6.7e-17	2
gb AA201988 AA201988 LD05058.5prime LD Drosophila Embr...	+3	175	5.5e-15	1
gb AA263526 AA263526 LD06652.5prime LD Drosophila Embr...	+1	167	7.0e-14	1
gb R46340 R46340 yj52c04.s1 Homo sapiens cDNA clon...	-1	151	5.6e-13	1
gb AA246675 AA246675 LD05588.5prime LD Drosophila Embr...	+2	117	2.8e-10	2
gb AA282899 AA282899 zt14g09.r1 Soares NbHTGBC Homo sa...	+3	118	6.1e-07	1
gb AA247705 AA247705 csh0941.seq.F Human fetal heart, ...	+3	56	0.0039	2

b

```

gb|AA390557|AA390557 LD09473.5prime LD Drosophila Embryo Drosophila
melanogaster cDNA clone LD09473 5'
Length = 659

Score = 130 (60.4 bits), Expect = 1.6e-20, Sum P(2) = 1.6e-20
Identities = 25/60 (41%), Positives = 40/60 (66%), Frame = +3

Query: 105 TIKSVRFYTEKNGQLIYTTFTEDTPSVGQRLNLSVLNLTLMISVIVVMTIFLVVLYKYRC 164
+I S+ FY + L+YT F E +P + +++ ++LI++SV+VVM T L+VLYK RC
Sbjct: 480 SINSISFYNSTDVVLLLYTPFHEQSPSPVKFWSALGSSLILMSVVVMTFLLIVLYKKRC 659

Score = 117 (54.3 bits), Expect = 1.6e-20, Sum P(2) = 1.6e-20
Identities = 23/30 (76%), Positives = 27/30 (90%), Frame = +1

Query: 75 LEEELTLKYGAKHVIMLVFPVPTLCMIVVVA 104
+EEE LKYGA+HVI LFPVPV+LCM+VVVA
Sbjct: 391 MEEEQGLKYGAQHVIKLVFPVSLCMLVVVA 480
    
```

Figure 8.11. Output of a TBLASTN search. The protein product of the Alzheimer's disease susceptibility gene (GenBank L43964) was used as the query in a TBLASTN search against the est database. The goal was to identify cDNA clones from other organisms that may represent homologs of the human gene. (a) Portion of the hit list showing the 25 best hits. Each sequence is identified by GenBank accession number and a portion of the definition line. The reading frame and score of the best HSP are shown, together with the sum probability of a chance occurrence. The value in the last column gives the number of HSPs that were used in the sum probability calculation. At least 10 sequences from mouse and one from *Drosophila* may be seen on the hit list. (b) Match to the conceptual translation of the *Drosophila* EST sequence (GenBank AA390557). Two HSPs were found, each in a different reading frame. Identical residues are echoed to the central line, and plus (+) symbols indicate pairs of nonidentical amino acids with positive substitution scores.

finding of significant matches to repeat “warning sequences” that have been included in both GenBank and SWISS-PROT. These entries are consensus sequences (or translations thereof) for different subfamilies of human Alu repeats. However, with the large amount of human genomic sequence now present in the database, it is common to have many hits to individual repeats with scores greater than those for any consensus repeat. Consequently, hits to Alu-warning entries are less striking than when the database was smaller. Other indications of likely artifacts would be finding hits to many proteins that seem to have no functional relationship to one another or hits to genomic sequences from many different chromosomes. These patterns might also be seen if both query and database are contaminated with foreign sequences from the same source, for instance, cloning vectors.

Although it is always good practice to critically evaluate database search results and be suspicious of artifacts when the data don’t make sense, a more proactive approach involves *masking* problematic sequences in the query before doing the search. The problem of repetitive elements is ably handled by the popular program RepeatMasker, which identifies, classifies, and masks several types of repetitive elements and simple repeats (A. F. A. Smit and P. Green, unpublished). A masking strategy, which we will call “hard masking,” is to replace subsequences with an ambiguity character (“N” for nucleotide sequences or “X” for proteins). Alternatively, a “soft-masking” approach, in which the residues are instead converted to lowercase letters, may be used with certain search programs. Because ambiguous residues are treated as mismatches (even when aligned to themselves), hard-masking effectively prohibits the identified repeats from making a positive contribution to the alignment score. Although hard masking is excellent for avoiding false hits, the fact that even the true alignments may be altered can present problems, particularly when alignment scores and lengths are used to classify alignments. The solution to this dilemma is to use soft masking. Recent versions of the BLAST programs have an option that ignores regions of the query sequence that are lowercase when constructing the word dictionary. However, an alignment that is initiated in unique sequence may be extended through a repeat and would have the same alignment score as it would with unmasked sequence. With RepeatMasker, the `-xsmall` command-line option may be used for soft masking.

Both proteins and nucleic acids contain regions of biased composition, which can lead to confusing database search results. These *low-complexity regions* (LCRs) range from the obvious homopolymeric runs and short-period repeats to the more subtle cases where one or a few different residues may be overrepresented. Alignment of LCR-containing sequences is problematic because they do not fit the model of residue-by-residue sequence conservation. In some cases, the functionally relevant attributes may be only the periodicity or composition and not any specific sequence. Furthermore, methods for assessing the statistical significance of alignments are based on certain notions of randomness which LCRs do not obey. Consequently, many false positives may be observed in the output of a database search with an LCR-containing query sequence because the significance of matches can be overestimated (Altschul et al., 1994).

A program called seg has been developed to partition a protein sequence into segments of low and high compositional complexity (Wootton and Federhen, 1996; Wootton and Federhen, 1993). Using this program, it has been shown that more than half of the proteins in the database contain at least one LCR (Wootton, 1994; Wootton and Federhen, 1993). The evolutionary, functional, and structural properties of LCRs

are not well understood. Perhaps LCRs arise by such mechanisms as polymerase slippage, biased nucleotide substitution, or unequal crossing-over. In proteins, LCRs are likely to exist structurally as nonglobular regions. Regions that have been defined physicochemically as nonglobular are usually identified correctly using seg (Wootton, 1994). In DNA, there are many classes of satellite and microsatellite sequences that consist of many copies of a simple repeat unit.

The protein product of the human homolog of the *Drosophila* achaete-scute gene provides a good example of an LCR-containing protein. When analyzed with seg, two regions of low compositional complexity were identified. Figure 8.12a shows



Figure 8.12. Identifying low-complexity regions with SEG. Analysis of the human achaete-scute protein (SWISS-PROT P50553) using seg reveals two regions of low compositional complexity. (a) Program output in the default “tree” format shows the low-complexity sequences in lower-case letters on the left and high-complexity in upper-case on the right. (b) Using the -x command-line switch, the seg program will generate a version of the sequence in which the low-complexity sequences have been masked. (c) For convenience, the BLAST programs can be instructed to perform the masking automatically. When a masked query sequence is used in a database search, some of the alignments may contain masked segments, as shown in this BLASTP output.

the default “tree” output, in which the low-complexity sequences are shown in lowercase letters on the left and the high-complexity sequences in uppercase on the right. The first region is a 61-residue segment containing homopolymeric tracts of glutamine and alanine. The second is a 14-residue segment with a bias toward arginine. Without filtering, many database sequences with biased regions involving these amino acids would be reported. Using a command-line option, `seg` can generate the masked version of the sequence for use as a search query (Figure 8.12b). Alternatively, filtering can be performed automatically by the BLAST programs through the use of optional parameters. Note that, in some implementations of BLAST, such as the Web version, filtering may be enabled by default.

POSITION-SPECIFIC SCORING MATRICES

In a standard substitution matrix, such as BLOSUM62, the substitution of one amino acid with another is associated with a single score—an obvious simplification given that the same amino acid may have different conservation patterns in one context than another in accordance with differing roles in biological function. Database searches can be tailored to find specific proteins families or domains through the use of substitution scores that reflect the substitution frequencies of each individual amino acid position in a domain. There is a large literature on the construction and application of these *position-specific scoring matrices* (PSSMs), which may also be called hidden Markov models (HMMs), motifs, or profiles (Bucher et al., 1996; Gribskov et al., 1987; Schneider et al., 1986; Staden, 1988; Tatusov et al., 1994). In its simplest form, a PSSM consists of a set of 20 substitution scores at each position along the motif—one for each of the amino acids. Amino acids that are commonly found at a particular position receive higher scores, whereas lower scores correspond to amino acids unlikely to appear at that position. It is also possible to assign scores to insertions and deletions in a position-specific manner.

A commonly used software package, HMMER (Eddy et al., 1995), contains a set of related programs for constructing and using PSSMs. Given a multiple alignment of several related proteins (e.g., one made using CLUSTAL W), the `hmmbuild` program may be used to calculate the position-specific scores and save it to a file (HMM file format). Using the `hmmsearch` program, the HMM file may be used as a query against a sequence database. Conversely, `hmmpfam` is used to compare a single query sequence against a database of PSSMs (HMMs). A comprehensive database of protein domains, Pfam (Bateman et al., 2000), is often used for this purpose.

The power of PSSMs in database searches can be further enhanced by iterative approaches in which the highest scoring matches in one search are incorporated into a PSSM used in successive searches. Position-Specific Iterated BLAST (PSI-BLAST) provides an automated facility for constructing, refining, and searching PSSMs within the context of a single program. Starting with a query sequence provided by the user, the process begins with a standard BLASTP search of a sequence database. Highly significant alignments found in this search are then used to construct a PSSM on-the-fly. Comparisons of the PSSM against the sequence database are performed using a variation of the word-based BLAST algorithm used for standard sequence comparisons. The process continues until no new matches are found or a specified limit on number of iterations is reached.

To demonstrate the improved sensitivity of the PSI-BLAST approach, the sequence of histidine triad (HIT) protein was used as a database search query. Simi-

Sequences producing significant alignments:				High	E
				Score	Value
Pass 1:					
sp	P49789	FHIT_HUMAN	FRAGILE HISTIDINE TRIAD PROTEIN	290	7e-79
sp	P49776	APH1_SCHPO	BIS(5'-NUCLEOSYL)-TETRAPHOSPHATASE (ASYMME...	117	8e-27
sp	P49775	YD15_YEAST	HYPOTHETICAL 24.8 KD HIT-LIKE PROTEIN	88.0	6e-18
sp	Q11066	YHIT_MYCTU	HYPOTHETICAL 20.0 KD HIT-LIKE PROTEIN	52.7	3e-07
sp	Q04344	HIT_YEAST	HIT1 PROTEIN (ORF U)	45.3	4e-05
Pass 2:					
sp	P47378	YHIT_MYCGE	HYPOTHETICAL 15.6 KD HIT-LIKE PROTEIN	70.5	1e-12
sp	P32083	YHIT_MYCHR	HYPOTHETICAL 13.1 KD HIT-LIKE PROTEIN IN P...	59.0	3e-09
sp	P26724	YHIT_AZOBR	HYPOTHETICAL 13.2 KD HIT-LIKE PROTEIN IN H...	57.6	9e-09
sp	P32084	YHIT_SYNP7	HYPOTHETICAL 12.4 KD HIT-LIKE PROTEIN IN P...	55.7	3e-08
sp	P53795	YHIT_CAEEL	HYPOTHETICAL HIT-LIKE PROTEIN F21C3.3	54.3	9e-08
sp	P42856	ZB14_MAIZE	14 KD ZINC-BINDING PROTEIN (PROTEIN KINASE...	52.8	2e-07
sp	P42855	ZB14_BRAJU	14 KD ZINC-BINDING PROTEIN (PROTEIN KINASE...	50.2	1e-06
sp	P49774	YHIT_MYCLE	HYPOTHETICAL 17.0 KD PROTEIN HIT-LIKE PROT...	49.5	2e-06
sp	P49773	IPK1_HUMAN	PROTEIN KINASE C INHIBITOR 1 (PKCI-1)	49.1	3e-06
sp	P16436	IPK1_BOVIN	PROTEIN KINASE C INHIBITOR 1 (PKCI-1) (17 ...	48.7	4e-06
sp	P44956	YCFH_HAEIN	HYPOTHETICAL HIT-LIKE PROTEIN HI0961	47.3	1e-05
sp	P43424	GAL7_RAT	GALACTOSE-1-PHOSPHATE URIDYLTRANSFERASE	41.0	8e-04
Pass 3:					
sp	Q03249	GAL7_MOUSE	GALACTOSE-1-PHOSPHATE URIDYLTRANSFERASE	87.2	1e-17
sp	P07902	GAL7_HUMAN	GALACTOSE-1-PHOSPHATE URIDYLTRANSFERASE	79.8	2e-15
sp	P31764	GAL7_HAEIN	GALACTOSE-1-PHOSPHATE URIDYLTRANSFERASE	64.7	6e-11
sp	P09148	GAL7_ECOLI	GALACTOSE-1-PHOSPHATE URIDYLTRANSFERASE	62.5	3e-10
sp	P22714	GAL7_SALTY	GALACTOSE-1-PHOSPHATE URIDYLTRANSFERASE	58.1	6e-09
sp	P09580	GAL7_KLULA	GALACTOSE-1-PHOSPHATE URIDYLTRANSFERASE	48.5	4e-06
sp	P08431	GAL7_YEAST	GALACTOSE-1-PHOSPHATE URIDYLTRANSFERASE	40.8	0.001
Pass 4:					
sp	P40908	GAL7_CRYNE	GALACTOSE-1-PHOSPHATE URIDYLTRANSFERASE	71.0	8e-13
sp	P13212	GAL7_STRLI	GALACTOSE-1-PHOSPHATE URIDYLTRANSFERASE	57.0	1e-08

Figure 8.13. Increased sensitivity using PSI-BLAST. The human histidine triad (HIT) protein (SWISS-PROT P49789) was used as the query in a BLASTP search with the PSI-BLAST functionality enabled. Definition lines, scores, and *E* values are shown for all statistically significant matches newly identified in each iteration.

larity between HIT and galactose-1-phosphate uridylyltransferase (GalT) has recently been described based on superimposition of their three-dimensional structures (Holm and Sander, 1997). However, sequence similarity between these two proteins is extremely weak. With a standard (single-pass) BLASTP search, no significant hits to GalT sequences are observed. However, with a multipass search, new relationships are discovered at each iteration, as shown in Figure 8.13. The rat GalT protein is found in the second iteration and, after information from this alignment is incorporated into the profile, several additional homologs from other organisms are also identified.

SPLICED ALIGNMENTS

The identification of genes within long stretches of DNA sequence is a central problem for automatic annotation of complete genomes. For the very compact genomes of viruses and bacteria, this work amounts to little more than the enumeration of open reading frames. However, gene identification in eukaryotic genomes is signif-

icantly more challenging because of the larger amount of intergenic sequence and the fact that protein-coding regions may be interrupted by introns. The two fundamental strategies seek to identify genes using either the intrinsic signals in the DNA sequence (see Chapter 10) or alignments to mRNA and protein sequences.

At first glance, the problem of aligning mRNA and genomic sequences seems trivial—using a local alignment strategy each exon would come out as a separate locally optimal alignment separated by large ‘deletions’ in the mRNA sequence corresponding to the introns that have been spliced out. If the goal is merely to obtain a crude sense of how a gene is organized, a simple alignment is sufficient. However, for the purpose of genome annotation, it is important that all exons be found with precise endpoints or a correct protein translation cannot be obtained. The *sim4* program is designed to address genome annotation needs by performing mRNA/genomic alignments rapidly and accurately (Florea et al., 1998). It begins with a BLAST-like search for finding the obvious exons—those with very high alignment scores—and follows this with search at lower stringency to identify any missed (usually short) exons. Splice donor and acceptor signals in the genomic sequences are used to adjust the exon boundaries (see Fig. 8.14). To avoid problems caused by tandemly repeated genes, an additional constraint is imposed to require that the order of the exons found in the genome match that implied by the mRNA. Other programs available for performing mRNA/genomic alignments are *est_genome* (Mott, 1997) and the *est2gen* program from the *Wise* package (Birney et al., 1996).

It should be noted that an mRNA sequence may align perfectly well to a pseudogene and that such an alignment may be difficult to distinguish from a functional gene. Certain features may be indicative of retropseudogenes, that is those resulting from integration of a reverse-transcribed mRNA. For example, a poly(A) tract found in the genomic sequence at the 3' end of the gene is indicative of a pseudogene produced through an mRNA intermediate. Such an mRNA alignment will also lack the large gaps corresponding to introns, although this alone cannot be used to conclude that it is a retropseudogene because there many authentic genes that consist of a single exon. However, once all mRNA alignments have been generated for a genome, one strategy for retropseudogene identification involves looking for pairs of highly similar alignments in which one lacks and the other contains introns. Other types of pseudogenes may arise by gene duplication events followed by inactivation of one of the copies. Such cases can be very difficult to diagnose. Programs that perform mRNA/genomic alignments usually do not use any knowledge of reading frame; therefore, determining that a potential gene contains a frameshift will require subsequent analysis of the protein translations. In the case of genes predicted by alignment with ESTs only, the protein-coding sequence is not known so any interruptions of the reading frame will not be apparent. The possibility that an apparent frame shift may actually be a sequencing error in the genomic sequence should also be considered, particularly in the analysis of a working draft sequence. In many cases, determining whether an mRNA-predicted gene is a functional gene or a pseudogene must await experimental validation.

CONCLUSIONS

Sequence alignment and database searching are performed tens of thousands of times per day by scientists around the world and represent critical techniques that all mo-

seq1 = mrna (>gi|7661723|ref|NM_015372.1), 1247 bp
seq2 = genomic (>gi|1941922|emb|Z82248.1|HSN44A4), 40662 bp

(complement)
1-118 (15628-15745) 100% ->
119-318 (22863-23062) 100% ->
319-1247 (26529-27457) 100%

0 . : . : . : . : . : . :
1 CCCCAGGCGTGGGAAGATGGAACCAGAACAATTCGAACGAGCAGAGCAAA
15628 CCCCAGGCGTGGGAAGATGGAACCAGAACAATTCGAACGAGCAGAGCAAA
50 . : . : . : . : . : . :
51 ACAGATCGGAATTGCAGACTTCAGGTCGTGGCAGAGAAAACCAGCTGAGA
15678 ACAGATCGGAATTGCAGACTTCAGGTCGTGGCAGAGAAAACCAGCTGAGA
100 . : . : . : . : . : . :
101 CAGGGCGCCACTTACTAG CTCTGAAAGTCTAGGATATTTTG
15728 CAGGGCGCCACTTACTAGGTG...CAGCTCTGAAAGTCTAGGATATTTTG
150 . : . : . : . : . : . :
142 CCACTGGAAGACCAGCAGACAATGTCATGACAACCTCAAGAGGATACAACA
22886 CCACTGGAAGACCAGCAGACAATGTCATGACAACCTCAAGAGGATACAACA
200 . : . : . : . : . : . :
192 GGGCTGCATCAAAAGACAAGTCTTTGGACCATGTCAAGACCTGGAGCGAA
22936 GGGCTGCATCAAAAGACAAGTCTTTGGACCATGTCAAGACCTGGAGCGAA
250 . : . : . : . : . : . :
242 GAAGGTAATGAACTCCTACTTTCATAGCAGGCTGTGGGCCAGCAGTTTGCT
22986 GAAGGTAATGAACTCCTACTTTCATAGCAGGCTGTGGGCCAGCAGTTTGCT
300 . : . : . : . : . : . :
292 ACTACGCTGTCTCTTGGTTAAGGCAAG GTTTCAGTATCAAC
23036 ACTACGCTGTCTCTTGGTTAAGGCAAGGTC...CAGGTTTCAGTATCAAC
350 . : . : . : . : . : . :
333 CTGACTTCTTTTGAAGGATCCCTTGGCCTCACGCTGGAGTGGGCACCTG
26543 CTGACTTCTTTTGAAGGATCCCTTGGCCTCACGCTGGAGTGGGCACCTG
400 . : . : . : . : . : . :
383 CCCTAGCCCACAGAGCTGGATTTCTCCCTTTCTCAATCACACAGGGAGC
26593 CCCTAGCCCACAGAGCTGGATTTCTCCCTTTCTCAATCACACAGGGAGC

ouput truncated for brevity

Figure 8.14. Spliced alignment. The sim4 program was used to align a novel human mRNA (RefSeq NM_015372) to the genomic sequence of a cosmid from chromosome 22 (EMBL Z82248). Three exons were identified on the complementary strand (the third one has been truncated for brevity). The ">>>" symbols indicate splice sites found at the exon/intron boundaries.

lecular biologists should be familiar with. It can be expected that these methods will continue to evolve to meet the challenges of an ever-increasing database size. This chapter has described some of the fundamental concepts involved, but it is useful to consult the documentation of the various programs for more detailed information. Researchers should have a basic understanding of how the programs work so that parameters can be intelligently selected. In addition, they should be aware of potential artifacts and know how to avoid them. Above all, it is important to apply the same powers of observation and critical evaluation that are used with any experimental method.

INTERNET RESOURCES FOR TOPICS PRESENTED IN CHAPTER 8

BLAST	http://ncbi.nlm.nih.gov/BLAST/
CLUSTAL W	ftp://ftp.ebi.ac.uk/pub/software/
dotter	ftp://ftp.sanger.ac.uk/pub/dotter/
FASTA, lalign	ftp://ftp.virginia.edu/pub/fasta/
hmmer	http://hmmer.wustl.edu/
RepeatMasker	http://ftp.genome.washington.edu/RM/RepeatMasker.html
seg	ftp://ncbi.nlm.nih.gov/pub/seg/
sim4	http://globin.cse.psu.edu
Wise package	http://www.sanger.ac.uk/Software/Wise2/

PROBLEM SET

1. What is the difference between a global and a local alignment strategy?
2. Calculate the score of the DNA sequence alignment shown below using the following scoring rules: +1 for a match, -2 for a mismatch, -3 for opening a gap, and -1 for each position in the gap.

```
GACTACGATCCGTATACGCACA--GGTTCAGAC
||||||| ||||||||| |||||||
GACTACGAGCCGTATACGCACACAGGTTTCAGAC
```

3. If a match from a database search is reported to have a E-value of 0.0, should it be considered highly insignificant or highly significant?

REFERENCES

- Altschul, S. F. (1991). Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* 219, 555–565.
- Altschul, S. F., Boguski, M. S., Gish, W., and Wootton, J. C. (1994). Issues in searching molecular sequence databases. *Nature Genet.* 6, 119–29.
- Altschul, S. F., and Erickson, B. W. (1986). Locally optimal subalignments using nonlinear similarity functions. *Bull. Math. Biol.* 48, 633–660.

- Altschul, S. F., and Erickson, B. W. (1985). Significance of nucleotide sequence alignments: A method for random sequence permutation that preserves dinucleotide and codon usage. *Mol. Biol. Evol.* 2, 526–538.
- Altschul, S. F., and Gish, W. (1996). Local alignment statistics. *Methods Enzymol.* 266, 460–480.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- Baron, M., Norman, D. G., and Campbell, I. D. (1991). Protein modules. *Trends Biochem. Sci.* 16, 13–17.
- Bateman, A., Birney, E., Durbin, R., Eddy, S. R., Howe, K. L., and Sonnhammer, E. L. (2000). The Pfam protein families database. *Nucleic Acids Res.* 28, 263–266.
- Birney, E., Thompson, J. D., and Gibson, T. J. (1996). PairWise and SearchWise: finding the optimal alignment in a simultaneous comparison of a protein profile against all DNA translation frames. *Nucleic Acids Res.* 24, 2730–2739.
- Bucher, P., Karplus, K., Moeri, N., and Hofmann, K. (1996). A flexible motif search technique based on generalized profiles. *Comput. Chem.* 20, 3–23.
- Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. (1978). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*, M. O. Dayhoff, ed. (Washington: National Biomedical Research Foundation), p. 345–352.
- Dembo, A., Karlin, S., and Zeitouni, O. (1984). Limit distribution of maximal non-aligned two-sequence segmental score. *Ann. Prob.* 22, 2022–2039.
- Doolittle, R. F., Hunkapiller, M. W., Hood, L. E., Devare, S. G., Robbins, K. C., Aaronson, S. A., and Antoniades, H. N. (1983). Simian sarcoma virus onc gene, v-sis, is derived from the gene (or genes) encoding a platelet-derived growth factor. *Science* 221, 275–277.
- Doolittle, R. J., and Bork, P. (1993). Evolutionarily mobile modules in proteins. *Sci. Am.* 269, 50–56.
- Eddy, S. R., Mitchison, G., and Durbin, R. (1995). Maximum discrimination hidden Markov models of sequence consensus. *J. Comput. Biol.* 2, 9–23.
- Fitch, W. M. (1969). Locating gaps in amino acid sequences to optimize the homology between two proteins. *Biochem. Genet.* 3, 99–108.
- Fitch, W. M. (1983). Random sequences. *J. Mol. Biol.* 163, 171–176.
- Florea, L., Hartzell, G., Zhang, Z., Rubin, G. M., and Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. *Genome Res.* 8, 967–974.
- Gibbs, A. J., and McIntyre, G. A. (1970). The diagram: a method for comparing sequences. Its use with amino acid and nucleotide sequences. *Eur. J. Biochem.* 16, 1–11.
- Gonzalez, P., Hernandez-Calzadilla, C., Rao, P. V., Rodriguez, I. R., Zigler, J. S., Jr., and Borrás, T. (1994). Comparative analysis of the zeta-crystallin/quinone reductase gene in guinea pig and mouse. *Mol. Biol. Evol.* 11, 305–315.
- Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* 84, 4355–4358.
- Henikoff, S., and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* 89, 10915–10919.
- Henikoff, S., and Henikoff, J. G. (1991). Automated assembly of protein blocks for database searching. *Nucleic Acids Res.* 19, 6565–6572.
- Higgins, D. G., Thompson, J. D., and Gibson, T. J. (1996). Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.* 266, 383–402.
- Holm, L., and Sander, C. (1997). Enzyme HIT. *Trends Biochem. Sci.* 22, 16–117.
- Huang, X., Hardison, R. C., and Miller, W. (1990). A space-efficient algorithm for local similarities. *Comput. Applic. Biosci.* 6, 373–381.

- Karlin, S., and Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA* 87, 2264–2268.
- Lipman, D. J., and Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science* 227, 1435–1441.
- Mott, R. (1997). EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput. Appl. Biosci.* 13, 477–478.
- Needleman, S. B., and Wunsch, C. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48, 443–453.
- Pathy, L. (1991). Modular exchange principles in proteins. *Curr. Opin. Struct. Biol.* 1, 351–361.
- Pearson, W. (2000). Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.* 132, 185–219.
- Pearson, W. R. (1996). Effective protein sequence comparison. *Methods Enzymol.* 266, 227–258.
- Pearson, W. R., and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85, 2444–2448.
- Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* 188, 415–31.
- Sellers, P. H. (1984). Pattern recognition in genetic sequences by mismatch density. *Bull. Math. Biol.* 46, 510–514.
- Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* 147, 195–197.
- Sonnhammer, E. L. L., and Durban, R. (1996). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 167, GC1–GC10.
- Staden, R. (1988). Methods to define and locate patterns of motifs in sequences. *Comput. Appl. Biosci.* 4, 53–60.
- Tatusov, R. L., Altschul, S. F., and Koonin, E. V. (1994). Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc. Natl. Acad. Sci. USA* 91, 12091–12095.
- Waterfield, M. D., Scrace, G. T., Whittle, N., Stroobant, P., Johnsson, A., Wasteson, A., Westermark, B., Heldin, C. H., Huang, J. S., and Deuel, T. F. (1983). Platelet-derived growth factor is structurally related to the putative transforming protein p28sis of simian sarcoma virus. *Nature* 304, 35–39.
- Waterman, M. S., and Eggert, M. (1987). A new algorithm for best subsequence alignments with applications to tRNA-rRNA comparisons. *J. Mol. Biol.* 197, 723–728.
- Waterman, M. S., and Vingron, M. (1994). Rapid and accurate estimates of statistical significance for sequence data base searches. *Proc. Natl. Acad. Sci. USA* 91, 4625–4628.
- Wilbur, W. J., and Lipman, D. J. (1983). Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA* 80, 726–730.
- Wootton, J. C. (1994). Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput. Chem.* 18, 269–285.
- Wootton, J. C., and Federhen, S. (1996). Analysis of compositionally biased regions in sequence databases. *Methods Enzymol.* 266, 554–571.
- Wootton, J. C., and Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* 17, 149–163.