BIOBIBLIOMETRICS: INFORMATION RETRIEVAL AND VISUALIZATION FROM CO-OCCURRENCES OF GENE NAMES IN MEDLINE ABSTRACTS.

B.J. STAPLEY

Kentucky Center for Structural Biology & Kentucky Center for Computational Science, University of Kentucky, 800 Rose St, Lexington, KY, 40536-0298, USA

G. BENOIT

College of Communications & Information Studies, School of Library & Information Science, 502 King Library Building, So.,University of Kentucky, Lexington, KY, 40506-0039, USA

Successful information retrieval from biomedical literature databases is becoming increasingly difficult. We have developed a prototype system for retrieving and visualizing information from literature and genomic databases using gene names. The premise of our work is that, if two genes have a related biological function, the co-occurrence of two gene names (or aliases of those genes) within the biomedical literature is more likely. From a collection of Medline documents, we have extracted the number of co-occurrences of every pair of *Saccharomyces cerevisiae* genes. The query is automatically conflated to include gene aliases as well. In addition, the retrieved document set can be filtered by the user with a MeSH term. From this co-occurrence data we construct a matrix that contains dissimilarity measurements of every pair of genes, based on their joint and individual occurrence statistics. A graph is generated from this matrix, with node and edge inclusion being determined by a user-defined threshold. Nodes of the graph represent genes, while edge lengths are a function of the occurrence of the two genes within the literature. Nodes can be hypertext-linked to sequence databases, while edges are linked to those Medline documents that generated them. The system is a tool for efficiently exploring the biomedical information landscape and may act as a inference network.

1 Introduction

1.1 The Nature of Textual Biomedical Information

Biomedical information is growing explosively and thus effective information retrieval is becoming more difficult. Crudely, biological information can be classified into two types: biomolecular primary, secondary, and tertiary structural data; and natural language text contained in databases of biomedical literature abstracts. The relationship between these two forms of information is important because raw sequence/structure data contains little information without accurate annotation and expert knowledge of the properties of genes involved. Conversely, the stuff of biomedical literature is very often genes and gene products, for which sequence and chemical properties are rarely stated in the literature abstracts.

In recent years, biomolecular sequence information has be accumulating from

systematic and non-systematic sequencing of DNA and protein. Much of the sequence data of biomedical importance is contained in large annotated protein and gene databases such as Swiss-Prot [1] and GenBank [2]. Many tool exists for analysis and retrieval of this kind of information.

With respect to biological natural language information, the foremost tool is Medline [3][a]. Medline is the largest English language biomedical bibliographic database. The mechanism by which information is retrieved from Medline has remained largely unchanged since its inception. User queries are terms matched against documents containing fields derived directly from articles in the medical literature and indirectly from a pre-coordinate, hierarchical thesaurus of medical subject headings (MeSH headings and terms) that describe the general and specific content of the document. Information retrieval on MeSH's can be extremely effective when the content of the document can be accurately described by a set of keywords. Within the medical literature this is often the case; however, in molecular biology this is less true because a controlled vocabulary index may have trouble keeping up with a rapidly evolving field [4]. The sheer size of Medline can be daunting to many scientists involved in biomedical research. For example, a text query for "cell cycle" AND "Saccharomyces" of the Medline database retrieves 4909 documents - a most disheartening number to the neophyte. An individual who had knowledge of even half this number of papers would be considered an expert, and yet might be unfamiliar with a substantial part of literature. It is clear that distillation of the literature is becoming difficult.

The study of the relationship between biological, textual and sequence data has included systematic naming and identification of protein fold families [5][6][7], the linking of sequence database entries to literature database entries [8][9] and the manual and automated annotation of gene sequences in databases [10][11][12]. Linking literature and sequence databases has been attempted by ENTREZ [8]. Data-mining across sequence databases has been attempted by the GeneCards project [13] and DBGET [14]. Sequence documents are frequently manually annotated with literature citations – several tools exist to assist in this task [11][12].

What is the answer to the problem of biomedical informational overload? The solution we propose - in a discrete and limited field - is to simultaneously integrate, synthesize and visualize the information contained within Medline and sequence databases in a way that automatically represents some of the knowledge structure latent within the literature and links it to sequence data. We achieve this by information retrieval based of queries formed from gene naming terms.

1.2 Lexical and Semantic Properties of Gene Terms

Much of the effort of the individual molecular biologist is focused on the

---

elucidation of the properties and function of a single cellular process or phenomenon. At a deeper level, this involves characterization of the separate components that generate that phenomenon. These components are genes, gene products (proteins and RNA) and products enzymatically generated.

Very often the naming term for a characterized gene consists of a three-letter abbreviation or acronym followed by a number. The specificity of the gene referent and the ability to match its presence in structured fields or full-text suggests that gene terms are excellent query components for information retrieval. Theoretically, the specificity of gene abbreviations likely favors high precision rates in the retrieval set. For example, a query such as "membrane" or "protease" is much more likely to retrieve non-relevant documents (high recall, low precision) than one formed from terms like "sic1" or "cdc28." In addition, a complete set of gene terms from a genome represents a controlled vocabulary, without additional manipulation of terms. Hypothetically, many of the problems associated with full-text medical data retrieval are minimized when using gene terms.

The simplicity of the above hypothesis is moderated by the fact that a gene may be represented semantically as a single preferred term, or the same gene represented by a large number of aliases. Gene aliases arise because a gene was discovered by separate researchers, or because what was once thought to be two separate genes is ultimately shown to be a single entity. Additionally, a gene may be referenced by other semantic forms, which makes precise retrieval more difficult. For gene products, the problems are greater because a given entity may be referred to in a variety of constructions. Recently, however, Fukuda et al. described a method for the extraction of protein names using "proper-noun phrase extracting rules." [15]

Efforts toward the systematic naming of genes have been made, but the problem still exists [16][17]. For those wishing to extract information about a particular gene and its role in a cellular phenomenon the haphazard way in which gene names arise poses a problem. A query formed on a single gene term may generate a set of appropriate records, but it will fail to retrieve all relevant ones because of false drops of documents relying on aliases that are not associated with a preferred gene name or by a phrase. Conflation of the query term can help here if knowledge of the aliases is available. A further problem is that gene names are sometimes degenerate – that is, several genes may have the same name or alias. (An example of this is the ORF2 term from *S. cerevisiae*, which is an alias for both the Serine/Threonine kinase cdc7 and pip2, a protein involved in peroxisome proliferation.) In addition, a gene name/alias may be the same as something other than another gene (typically an amino acid). Thus extraction on a single gene term may also result in poor precision. Here conflation of the query will work against accurate extraction.

Despite the above caveats, gene terms are superior to natural-language searching English for information retrieval. In this paper we exploit this property

and along with one further premise, create a tool for information retrieval and visualization. This additional premise is the following: if two genes are related biologically – the nature of this relationship is somewhat vague – then there is an increase in the likelihood of those two gene names occurring in the same document or document abstract. We exploit this idea by generating graphs in which nodes represent genes and edges are generated from the co-occurrences of two genes. We suggest that the graphs thus generated to some extent reflect biological relationships between genes. The graph and its web interface constitute a prototype information retrieval and visualization system for rapid and precise exploration of biomedical literature.

2. System and Methods

We have constructed a prototype system for biomedical information retrieval and visualization using the *Saccharomyces cerevisiae* genome (SGD) [b] [18] and a set of Medline documents published between 1997 and 1998 and containing the MeSH term '*Saccharomyces cerevisiae*'.

2.1 Bibliometric Distance.

Initially, we consider the possible significance of the co-occurrence of two gene terms in a Medline document and the informational value of such occurrences. Two gene names can occur in the same text for the following reasons:
1) Evidence of a physiological relationship between the two genes. We can further subdivide this kind of relationship.
   a) A direct physical interaction between the genes. This may be between the gene products or between the product and the DNA.
   b) There is a abstract functional link between the two elements. It may be that the genes perform similar functions or are involved in the same underlying process (e.g. DNA repair, glycolysis, etc).
2) An evolutionary relationship between the two genes. The two genes have a detectable sequence or structural similarity that implies a common origin.
3) An artifact of the experimental method. Some gene names are linked to particular experimental techniques; as promoters or reporters for genetic assays of gene function and expression (e.g. GAL4, TRP)
4) A negation of the 1. or 2. Somewhat uncommon because scientists do not often report negative results.
5) Genomic proximity. The method by which sequence data is obtained and reported may results in the joint citation of gene names that are close together

---

[b] Available at http://genome-www.stanford.edu/Saccharomyces/

on a chromosome.

Given two genes, how can we assess the relationship between the pair by analysis of the literature? What is required is some function of the distribution of occurrences of the two gene names amongst the documents. The situation is pictured in Figure 1. A certain number of all the documents contain the gene name *i* and another set contain gene *j*. Some of the documents contain both gene *i* and gene *j*. The optimal function for describing the dissimilarity based on the occurrence statistics of genes and will be dealt within a forthcoming paper. For the present, we use the reciprocal of the Dice coefficient between the two genes as a measure of their *BioBibliometric* distance (eq 1) [19]. For a whole genome, we can measure the similarity/dissimilarity of every pair of genes and place them in a symmetric matrix which describes the relationships between them.

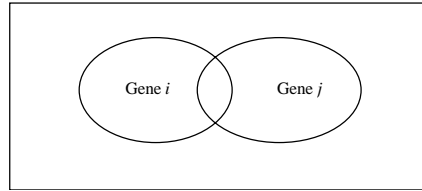$$d_{ij} = \frac{|i| + |j|}{|i \cap j|} \qquad \text{eq.1}$$

Figure 1: Venn diagram of a set of Medline documents showing the intersection of documents containing both gene *i* and gene *j*.

2.2 Extraction of Gene Terms from the Saccharomyces Genome Database

The controlled vocabulary of gene terms from the SGD was generated thus: first, any transfer RNA genes were removed by discarding any gene for which the locus name matches 't.(…)'. (a 't' followed by a single letter, followed by three characters enclosed in paratheses). The set was further reduced by removing any gene terms that are identical to the ORF name. ORF names are formed as follows: 'Y' for yeast, followed by a single letter A to P, for the chromosome, a three digit number followed by W or C (for Watson or Crick strand). Genes which have their ORF name as the sole naming term are unlikely to have a known function and thus probably occur very rarely in the literature. We also do not consider such terms as aliases for genes.

2.3 Extraction and Processing of Medline Abstracts

A set of 2524 Medline documents containing the MeSH term '*Saccharomyces cerevisiae*' and published between 1997 and 1998 consitutes the document collection for the prototype system. From this set, an abbreviated document

collection was formed by discarding documents which contained genes terms for fewer than two genes within their title, MeSH terms, compound registry or abstract fields. A matched gene term is any any case-insensitive occurrence of the gene term preceeded by whitespace and proceeded by 'p' (for protein), and then a non-alphanumeric, or a non-alphanumeric alone. The abbreviated Medline document consists of Medline unique identifier, title, MeSH terms fields and fields for the SGD genes that occurred in the original Medline document. Fast and accurate generation of a gene dissimilarity matrix can be achieved from a retrieved set of these abbreviated medline documents.

3. Implementation

The prototype system is implemented as a Java/Perl application with information retrieval and query formation performed by Perl on the server-side and data visualization by a Java applet on the client-side [c]. A schema for the system is illustrated in Figure 2 and discussed below. A user-defined query consists of a regular expression for retrieval on the MeSH term field of the abbreviated documents (Medline key-word filter). The user also sets the maximum bibliometric distance for the co-occurrence of a pair of genes that will result in edge generation (user-defined display threshold). A minimum number of co-occurrences of pairs genes can also be specified. The user parameters are passed to a perl script which initially retrieves all the abbreviated medline documents that contain the specified MeSH term expression within their MeSH field. From this set, the script evaluates the numbers of occurrences of every gene and co-occurrences of every pair of genes. Pairs of genes that do not fufill the two user-defined criteria are discarded.

The nodes, edges, and edge lengths thus generated are passed to a Java applet which is executed in a HTML document that also contains a table that links genes to the SGD and pairs of genes to those medline documents that generated the edge between the pair (Web-based display). The Java applet initially places the nodes of the graph at random positions within the applet window and then performs a minimization to bring edge lengths as close to their calculated values as possible. The user can drag nodes around the display so as to better see the edges between nodes and also to lift the areas of the graph out of local energy minima. The table within the application mimics the dissimilarity matrix. Row and column labels represent retrieved genes. These labels are linked to the SGD gene entries. The presence of an edge between a pair of genes is displayed as an element of the table which contains the number of co-occurrences of the two genes and is also hypertext-linked to documents on the NLM Medline website that contain the gene

---

[c] Available at http://sophocles.gws.uky.edu/~ben/Interface.html.

names or aliases of both genes.

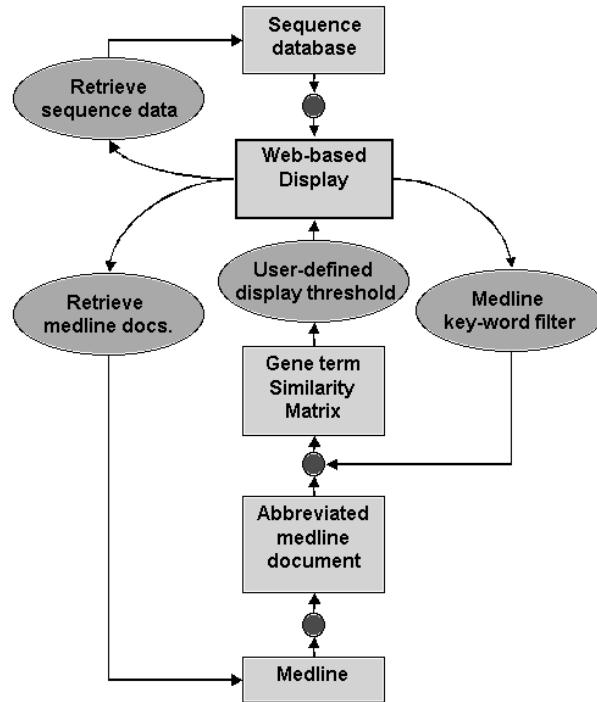Figure 2: A Schema for the BioBibliometric Information Retrieval and Visualization System.

4. Example

We illustrate the utility of our system using a retrieved document collection from the MeSH filter 'DNA repair' and an edge inclusion threshold of less than 20 for the reciprocal of the Dice coefficient between a pair of genes. Additionally, an edge is only created if at least three documents contain a joint occurrence of the gene pair. The display generated by this query is shown in Figure 3. We investigated the veracity of the links generated by our system by studying the retrieved Medline and SGD documents.

The *rad50*, *MRE11* and *xrs2* cluster consists of genes involved in DNA double-stranded break (DSB) repair [20]. These proteins form a heterotrimer and mutation of any one of these genes results in very similar phenotypes. This complex is involved in repair implemented through both homologous
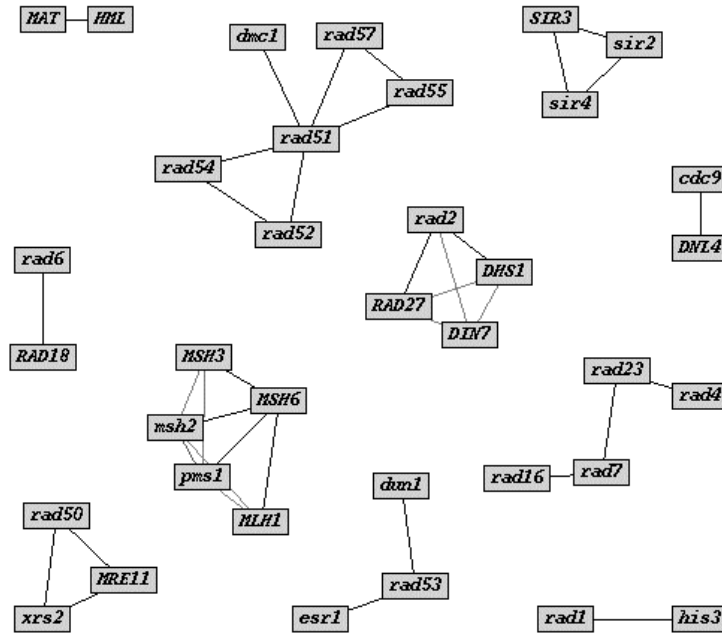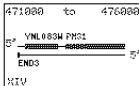
Figure 3: Graph generated for Query "DNA repair ", $d_{ij} < 20$ and $|i \cap j| > 2$.

recombination and DNA end-joining. The former is achieved by a evolutionary conserved complex of *Rads 51*, *52*, *55* and *57* which is called the recombinosome [21] [22]. Our system has successfully linked the members of this complex and in addition, linked the proteins *rad54* and *dmc1* to the *cerevisiae* recA homologue *rad51*. *Rad54* is an ATPase of the SWI2/MOT1 family that physically interacts with *Rad51* and has been implicated in the recombinosome complex. *Dmc1* is a sequence homologue of recA and *rad51* that has a distinct yet overlapping role in meiotic recombination [23].

Rads 4, 7, 16 and 23 represent four members of an eighteen protein complex called the repairosome which is involved in nucleotide excision repair (NER) [24]; the other elements of this assembly have not been retrieved. *Rad7* has been shown by yeast two-hybrid assay to interact with the other members of this cluster. *Rad4* and *rad23* form what is called the nucleotide excision repair factor 2 (NEF2).

The *rad2*, *RAD27*, *DHS1*, and *DIN7* grouping is an example of an evolutionary relationship between genes resulting in co-citation (see 2 under *2.1 Bibliometric Distance*) but they are also all involved in DNA mismatch repair [25]. *Rad27* is a single-stranded DNA endonuclease and 5'-3' exonuclease that is 32% identical to *din7* and 27% identical to *DHS1* with which it is synthetically lethal.

Locus : pms1

```
471000   to    476000
    YNL083W PMS1
5'  -===-====-=====-  5'
    ENDS
XIV
```

[Help]

```
Locus_info: Other_name YNL082W
            Gene_class PMS
            Gene_Info PMS1
            Description required for mismatch repair in mitosis and meiosis, low
                        levels of postmeiotic segregation, and high spore viability,
                        dispensable for homeologous recombination
            Gene_product MutL homolog, similar to Mlh1p, associates with Mlh1p,
                        possibly forming a heterodimer, Pms1p and Msh1p act in
                        concert to bind to a Msh2p-heteroduplex complex containing a
                        G-T mismatch
            Phenotype Null mutant is viable; postmeiotic segregation increased
            Locus_notes 10 pms1 data in 1985 mapping review; new reference
Position_info: Chromosome XIV
            Map XIV Position -66
            ORF_name YNL082W
Sequence_info: ORF_sequence YNL082W [Gene/Sequence Resources|MIPS ORF Info]
```

Figure 4: Retrieved SGD Document for Gene Node pms1.

Entrez medline Query

[Details] [97321279 97347244 97404375 97265419 9821] [Search] [Clear]

Docs Per Page: [20 ⊟]   Entrez Date limit: [No Limit ⊟]

5 citations found

[Display] [Abstract report ⊟]   for the articles selected (default all).

[Order] documents on this page through Loansome Doc

Petes TD, et al.   [See Related Articles]
     Stabilization of microsatellite sequences by variant repeats in the yeast Saccharomyces cerevisiae.
     Genetics. 1997 Jun;146(2):491–8.
     PMID: 9178000; UI: 97321279.

Hunter N, et al.   [See Related Articles]
     Mlh1 is unique among mismatch repair proteins in its ability to promote crossing-over during meiosis.
     Genes Dev. 1997 Jun 15;11(12):1573–82.
     PMID: 9203583; UI: 97347244.

Sugawara N, et al.   [See Related Articles]
     Role of Saccharomyces cerevisiae Msh2 and Msh3 repair proteins in double-strand break–induced
     recombination.
     Proc Natl Acad Sci U S A. 1997 Aug 19;94(17):9214–9.
     PMID: 9256462; UI: 97404375.

Greene CN, et al.   [See Related Articles]
     Frameshift intermediates in homopolymer runs are removed efficiently by yeast mismatch repair proteins.
     Mol Cell Biol. 1997 May;17(5):2844–50.
     PMID: 9111356; UI: 97265419.

Habraken Y, et al.   [See Related Articles]
     ATP-dependent assembly of a ternary complex consisting of a DNA mismatch and the yeast
     MSH2-MSH6 and MLH1-PMS1 protein complexes.
     J Biol Chem. 1998 Apr 17;273(16):9837–41.
     PMID: 9545323; UI: 98212020.

Figure 5: Retrieved Information from the edge between pms1 and msh6.

Rad2 is related evolutionarily to the other members of this cluster but is also an element of the repairosome - a linkage that has not been recognized by the system.

MSH2 is a DNA mismatch binding factor involved in repair of single base mismatches and short insertions/deletions that interacts with the other members of this cluster [26]. Overproduced and purified *Msh2p-Msh6p* complex binds DNA substrates containing a G/T mismatch and insertion/deletion mismatches, consequent ATP-binding by *Msh2p-Msh6p* induces a conformation change that leads to the formation of the ternary structure with *Mlh1p-Pms1p*. We illustrate hypertext node and edge linking using *pms1* and *MSH6*. The annotations of the sequences of these two genes confirm the validity of the links that were automatically generated by the system (Figure 4). The retrieved literature citations that linked these two genes are shown in Figure 5. These articles also include references to other genes in the cluster.

5. Discussion

Our original premise was that gene terms that occur together in the same document with statistically significant frequency may do so because there is functional relationship between the two; this has been borne out. Obviously, the system is imperfect and many of the edges may not reflect *in vivo* relationships between the genes, however; the graph is a starting point for the user to investigate these relationships.

Firstly, the system to some extent extracts knowledge latent within the retrieved information. For example, the proteins involved in the recombinosome or repairosome have been clustered. The extraction of this knowledge was not specified by the user and may not been explicitly stated in the literature. Knowledge is represented graphically by the clustering of genes with related functions and or physical interactions– nucleotide excision repair, DNA end-joining etc. Many elements are missing in the retrieved data but very few of the generated links between genes carry no semantic (high precision). The method by which the initial Medline document set is retrieved is significant to the precision of the generated edges. 'DNA repair' may be an ideal case and other queries produce poorer retrieval.

Secondly, the system generates accurate and automatic links between the literature and sequence databases. This permits the user to efficiently investigate gene-gene and gene-literature relationships. Thus one can rapidly assess the precision and nature of the links between gene nodes. The system's retrieval representation moves away from the traditional ranked list of relevant documents that standard Medline searching generates. The user is now free to browse the retrieved information across multiple databases so as to formulate their information requirements dynamically.

A final perhaps contentious – facet of the tool is as an inference network. The graphical display, to some degree has made apparent to the user connections or

patterns of connections of which the user may have been previously unaware. In addition, gene nodes within the graph are often under stress; that is their edge lengths are drawn away from the bibliometric distance in the similarity matrix by the network of nodes around them. The implication here is that there is an inconsistency in the citation of these gene terms in the literature that maybe worthy of further investigation. Similarly, two genes that never occur together in Medline documents may be brought close in the graph through other genes - possibly implying an uncharacterized physiological relationship. Such a suggestion has a parallel in the ARROWSMITH system [28] which detects relationships between clinical conditions and physiological states through analysis of Medline document titles. ARROWSMITH has be criticized both in its inception and its implementation; however, it has successfully predicted a link between migraine headaches and magnesium deficiency and between Raynaud's disease and dietary fish oil, both of which have be subsequently validated experimentally [29] [30]. If biobibliometrics does not generate plausible hypotheses, it can at least clarify and classify gene relationships so as to assist the user in hypothesis generation. It can also alter the user's perception of the relationships between genes in such a way as to stimulate new experiments and methods.

The theoretical aspect of this graphic approach can be discussed only in brief terms. At the most basic, graphic representations can encapsulate large amounts of information, particularly in the use of space (to suggest lack of relationship) and the use of graphic primitives (lines, boxes, etc., which create explicit graphic associations). One limit to the graphic approach is the volume of primitives on screen and mixed iconic messages. In one example, Mapuccino [31], the net is a combination of graphics, texts, colors, and entity-relationship codes, but this kind of display does not provide the user with information about the contents of the pages, only their link structure. Other examples, such as BRAQUE [32] require considerable subject expertise and experience using the interface to profit from the retrieval set. Other graphic representations by design or by accident impose a sense of value by the placement of icons. Our program requires no training to use and demonstrates graphically the *associations* of documents without implying a scientific value to them by placement.

As an information storage and retrieval model, the prototype offers several benefits. This version of the prototype offers graphic and textual representations of gene co-occurrences. The graphic version provides researchers a more intuitive method of assessing the value of information sources. Additionally, the nodes (which simultaneously represent genes and documents) provide a novel document retrieval effort by automatically displaying document attributes or the document itself with links to supplemental databases, integrated into a single interface. The search behavior of this version is predicated on the presence of genes in the source document set and a binary matching function.

Acknowledgments

References

1. A. Bairoch & R. Apweiler, *Nucleic Acids Res.* **26**, 38 (1998)
2. W.M. Gelbart, *Science* **282**, 659 (1998)
3. D. Hutchinson. *Medline for health professionals: how to search PubMed on the Internet*, (Sacramento, New Wind, 1998).
4. M.L. Pao, *Concepts of Information Retrieval*, (Englewood, Colo., Libraries Unlimited, 1989)
5. A.G. Murzin et al, *J. Mol. Biol.* **247**, 536 (1995)
6. C.A. Orengo et al, *Structure* **5**, 1093 (1997)
7. J.C. Tamames et al, *Bioinformatics* **14**, 542 (1998)
8. G.D. Schuler et al, *Methods in Enzymology* **266**, 141 (1996)
9. D.A. Benson et al, *Nucleic Acids Res.* **26**, 1 (1998)
10. M.A. Andrade & A. Valencia, *ISMB* **5**, 25 (1997)
11. G.C. Overton et al, *Pac. Symp. Biocomput. 1998*, 291
12. E.C. Uberbacher et al, *Pac. Symp. Biocomput. 1998*, 217
13. M.V. Rebhan et al, *Bioinformatics* **14**, 656 (1998)
14. W. Fujibuchi et al, *Pac. Symp. Biocomput. 1998*, 683
15. K.A. Fukuda et al, *Pac. Symp. Biocomput. 1998*, 707
16. J.A. Blake et al, *Nucleic Acids Res.* **27**, 95 (1999)
17. D. Lonsdale & C. Price, *TIBS* **21**, 443 (1996)
18. J.M. Cherry et al, *Nucleic Acids Res.* **26**, 73 (1998)
19. C.J. van Rijsbergen, *Information Retrieval* (London: Butterworths, 1979)
20. C.I. Nugent et al, *Curr. Biol.* **21**, 657 (1998)
21. S.L. Hays et al, *Proc. Natl. Acad. Sci. U S A.* **92**, 6925 (1995)
22. P. Baumann & S.C. West, *TIBS* **23**, 247 (1998)
23. A. Shinohara et al, *Cell* **69**, 457 (1992)
24. Z. Wang et al, *Mol. Cell. Biol.* **17**, 635 (1997)
25. P.A. Mieczkowski et al, *Mol. Gen. Genet.* **253**, 655 (1997)
26. R.D. Kolodner and G.T. Marsischky, *Curr. Opin. Genet. Dev.* **9**, 89 (1999)
27. N.R. Smalheiser & D.R. Swanson, *Comput. Methods Programs Biomed.* **57,** 149 (1998)
28. D.R. Swanson, *Perspect Biol. Med.* **31**, 526 (1988)
29. D.R. Swanson, *Perspect Biol. Med.* **30**, 7 (1986)
30. Y. Maarek et al, *6ᵗʰ WWW Conference*, p. 713-722, Santa Clara, CA (1997)
31. P.G. Marchetti et al, *ACM SIGIR 93*. 358 (1993)