BIOINFORMATICS

Vol. 14 no. 8 1998
Pages 656–664

GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support

Michael Rebhan[1,2,3], Vered Chalifa-Caspi[1,2,3], Jaime Prilusky[2,3] and Doron Lancet[1,3,4]

[1]Department of Molecular Genetics, [2]Department of Biological Services (Bioinformatics Unit) and [3]The Genome Center, Weizmann Institute of Science, 76100 Rehovot, Israel

Received on May 30, 1997; revised on June 26, 1998; accepted on June 29, 1998

Abstract

Motivation: *Modern biology is shifting from the 'one gene one postdoc' approach to genomic analyses that include the simultaneous monitoring of thousands of genes. The importance of efficient access to concise and integrated biomedical information to support data analysis and decision making is therefore increasing rapidly, in both academic and industrial research. However, knowledge discovery in the widely scattered resources relevant for biomedical research is often a cumbersome and non-trivial task, one that requires a significant amount of training and effort.*

Results: *To develop a model for a new type of topic-specific overview resource that provides efficient access to distributed information, we designed a database called 'GeneCards'. It is a freely accessible Web resource that offers one hypertext 'card' for each of the more than 7000 human genes that currently have an approved gene symbol published by the HUGO/GDB nomenclature committee. The presented information aims at giving immediate insight into current knowledge about the respective gene, including a focus on its functions in health and disease. It is compiled by Perl scripts that automatically extract relevant information from several databases, including SWISS-PROT, OMIM, Genatlas and GDB.*
Analyses of the interactions of users with the Web interface of GeneCards triggered development of easy-to-scan displays optimized for human browsing. Also, we developed algorithms that offer 'ready-to-click' query reformulation support, to facilitate information retrieval and exploration. Many of the long-term users turn to GeneCards to quickly access information about the function of very large sets of genes, for example in the realm of large-scale expression studies using 'DNA chip' technology or two-dimensional protein electrophoresis.

Availability: *Freely available at http://bioinformatics.weizmann.ac.il/cards/*
Contact: *cards@bioinformatics.weizmann.ac.il*

Introduction

In recent years, the scientific community has witnessed the establishment of a number of very useful Internet-accessible resources for biological and medical information (Recipon and Makalowski, 1997), a trend that has been stimulated by efforts to store the data produced by the Human Genome Project, and by decreasing costs for data storage and transfer. In particular, the World-Wide Web (WWW) has evolved into an 'electronic library' that already contains a vast amount of useful biomedical information. However, simply making data available does not guarantee that the scientific user will find the requested information in an acceptable amount of time. The sheer number of different resources, their heterogeneity and complexity, often cause frustration among those scientists who wish to rapidly navigate to interesting information, but frequently 'get lost' in a labyrinth of hypertext links.

As long as these problems do not cause severe bottlenecks in scientific research, current techniques for providing biomedical information are largely sufficient to meet demands. However, the rapidly growing fields of functional genomics (Dujon, 1998; Hieter and Boguski, 1997; Lander, 1996) and proteomics (Humphery-Smith *et al.*, 1997), may change not only the way we do research in the laboratory, but also the way we look at biomedical information in general. Comprehending the data that involve complex networks composed of numerous genes requires timely access to well-organized, concise information. In particular, information about the known functions of genes should be presented to scientists in a way that corresponds to their 'conceptual information space'. When navigating a related Web resource, the user should be able to follow a chain of thought without major

[4]To whom correspondence should be addressed

© Oxford University Press

interruptions caused by technical problems such as spelling errors in query keywords. In other words, intelligent and user-oriented resources that offer concise, well-organized, integrated information are urgently needed.

We recently established a novel gene function resource, 'GeneCards' (Rebhan *et al*., 1997; Rebhan and Prilusky, 1997). It contains comprehensive information about human genes, including data about the cellular functions of their products, and their involvement in diseases.

A brief overview of the user interface of GeneCards, with a focus on implications for the two dimensional electrophoretic analysis of proteins, has been published recently (Rebhan and Prilusky, 1997).

System and methods

Platform

Both the scripts that extract data from distributed resources (see below) and the public CGI (http://hoohoo.ncsa.uiuc.edu/cgi/) interface of GeneCards have been written in Perl 5.004_04 (Wall *et al*., 1996), on a Sun Ultra-Enterprise-10000 supercomputer (http://www.sun.com/servers/enterprise/10000/), running SunOS 5.5.1.

Availability

GeneCards can be accessed freely at http://bioinformatics.weizmann.ac.il/cards. Information about the availability of mirror site packages can be obtained at http://bioinformatics.weizmann.ac.il/cards/mirror.html.

Data storage, retrieval and display

The GeneCards data are stored in flat files, which are indexed by two methods: (1) The Glimpse (http://glimpse.cs.arizona.edu/) package, which uses 'agrep', a variant of the UNIX command 'grep', to search large sets of files for patterns. Agrep can also be used to search files for patterns with a given number of misspellings (see spell check algorithm below). (2) The Excite index (http://www.excite.com/navigate/) is used for the concept search option, enabling the user to expand their search by performing a less precise query that often produces a higher recall (http://www.excite.com/navigate/). Both indices are used by a CGI script that enables users to search the collection.

The Web page that presents the information related to one particular gene, referred to as the 'GeneCard' (Figure 1), is created on-the-fly by a CGI script that accesses the corresponding flat file and displays it to the user. To help users identify the meaning of hyperlinks, short descriptions are displayed in the status line of most browsers when the mouse is moved over the respective link (Javascript 'window.status' function; (http://bioinformatics.weizmann.ac.il/comp/javascript_status.html)), whenever space on the screen was considered too limited to display link descriptions beneath the links.

The layout of a GeneCard was designed to facilitate quick human browsing. In principle, tables and bulleted lists were used, in addition to a relatively high degree of formatting of the text itself with different font sizes, bold characters and colors (Figure 1). Short paragraphs, separators in table cells and concise link descriptions are employed to strive for 'scannability'. To allow for easy navigation inside different GeneCards pages, frequently requested links are accessible directly from the bottom of each GeneCard.

To enable the user to assess the relationship between the entered query and the information returned by the Web interface, detailed search results are returned that not only list the ID, the name of an entry, and the locus of the gene, but also those lines in the entry text that matched the query (with those words that matched the query pattern highlighted) (Figure 2).

Algorithms

Extracting navigation-related information from heterogeneous databases

Automatic extraction of information about a particular subject from various resources is difficult without standard nomenclature. Although a general standard nomenclature for biology and medicine is far from being achieved, nomenclature committees for various organisms have been established (http://ash.gene.ucl.ac.uk/nomenclature/FAQ.shtml), among them the HUGO/GDB nomenclature committee (http://ash.gene.ucl.ac.uk/nomenclature/). The latter regularily publishes a list of approved gene symbols and names that are useful to find most of the known human genes for which we have some information about cellular functions and medical implications. Therefore, we chose to use this list as a starting point to develop scripts for the automatic extraction of information from sources that use this nomenclature. We studied different strategies for the extraction and integration of information from heterogeneous, Web-accessible resources in a way that provides immediate insight into the function of a gene. Importantly, we did not attempt to integrate all the information available, but to select the data that would be most helpful for providing an overview of current knowledge.

Using the text processing capabilities of Perl, we were able to develop a package of scripts we termed 'PLUK' (Package for Locating Useful Knowledge) that search different data sources for information about human genes with approved symbols (http://www.gene.ucl.ac.uk/nomenclature/).

Sources for data extraction have been selected according to the following criteria: reliability of information content, usage of approved gene symbols, ease of information extraction, standardization of data format, degree of integration of

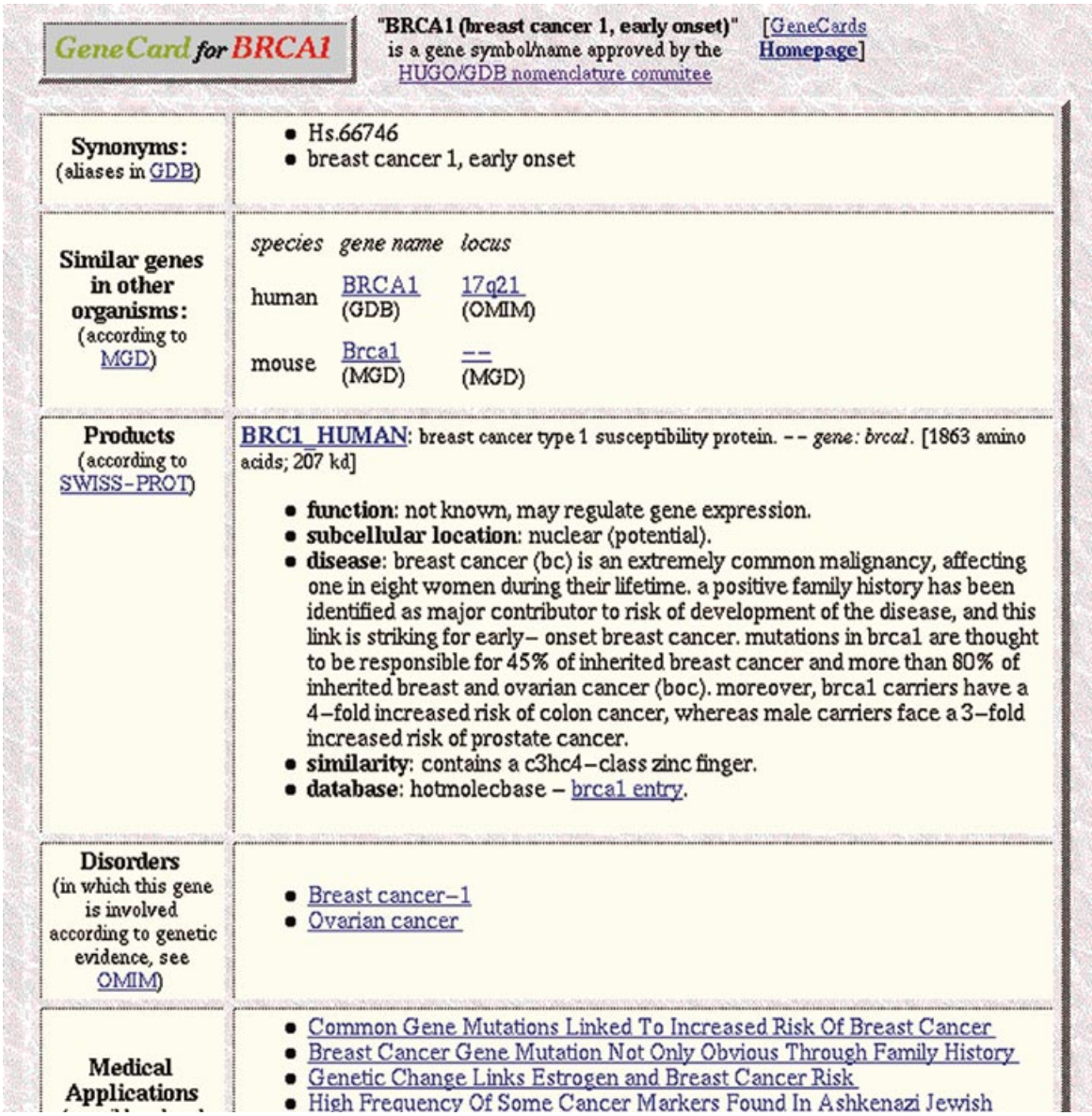

Fig. 1. A screenshot of the topmost part of the GeneCard for BRCA1, a breast cancer related gene. Note that the layout is specifically designed for human browsing (see 'Strategy').

the data and availability. Due to the heterogeneity of the data sources currently selected (GDB, MGD, OMIM, SWISS-PROT, HGMD, Genatlas and Doctor's Guide), special information extraction scripts had to be written for each data source. Data retrieval is accomplished using one of two methods, to retrieve the data related to a particular gene symbol: (1) Text extraction from database text file deposits ('dumps'), which are mirrored at the Weizmann Institute Bioinformatics FTP server (ftp://bioinfo.weizmann.ac.il/pub/databases). (2) If the first option is not available, we use a Web query mechanism which accesses the Web interface of the source database, at its closest mirror site. The latter mechanism employs LWP (Library for WWW access in Perl) (http://www.sn.no/libwww-perl/).

The retrieved data sets are then used to parse information, i.e. those data that were considered to be helpful in providing an overview of current knowledge about the function of the respective gene. This information is subsequently integrated into preliminary GeneCards data files. Then, the data are checked manually by browsing them with a CGI script developed by us that allows one to rapidly scan a considerable portion of the whole set. If we do not find data that are appar-

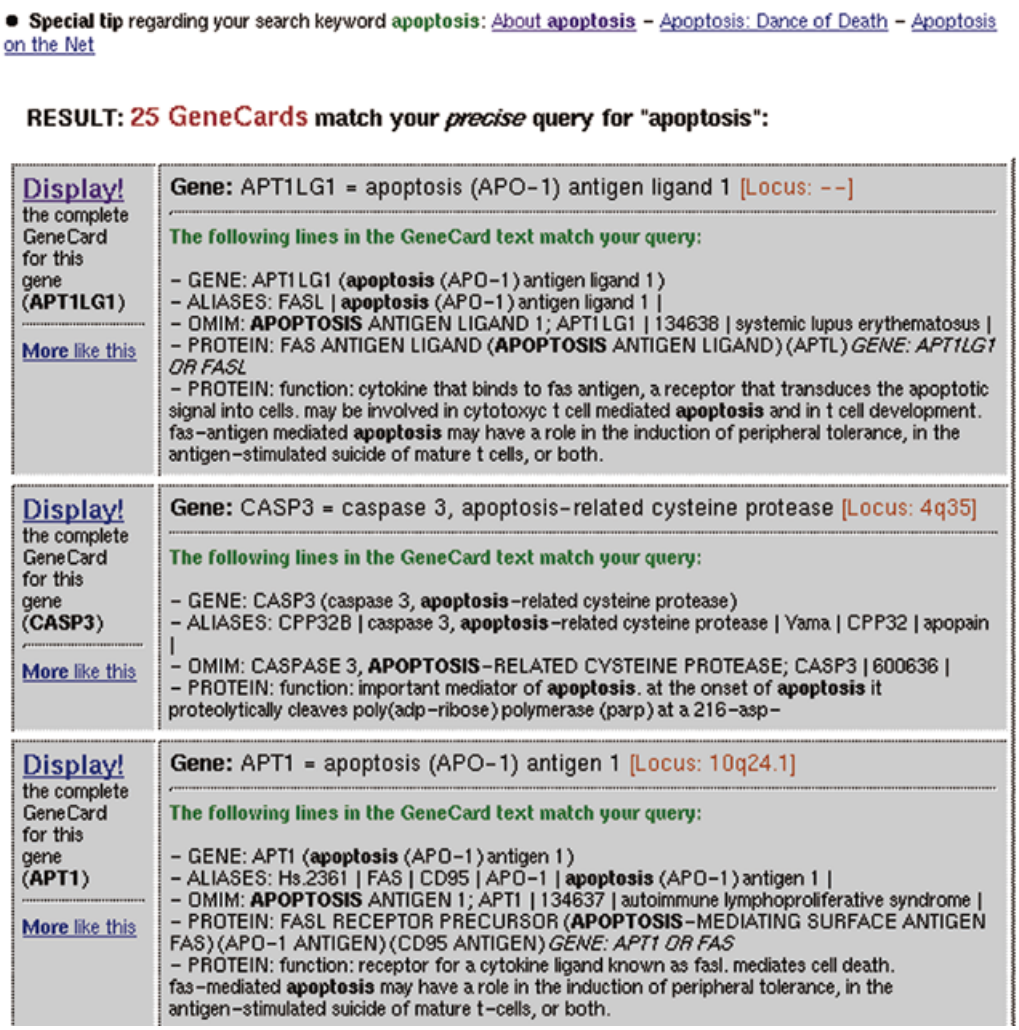

Fig. 2. A screenshot of the topmost part of the search result for the query 'apoptosis'. Note the 'special tips' on the top, the detai led search results, and the 'more like this' link.

ently not related to the respective gene, but which were retrieved because of the use of non-unique gene symbols (false positives), we include those data in the next public release of GeneCards. If false positives have been retrieved by our scripts, we either try to modify the extraction algorithms, or we ask the maintainers of the relevant data source to change their data, depending on the nature of the problem. In most cases, false positives are returned because the data source does not distinguish between approved and non-approved gene symbols. In general, our extraction scripts seem to be able to avoid false positives, while the number of false negatives — missing data from one data source, although this resource contains information about the gene — may be in the range of 0–10%, depending on the data source.

Query reformulation

Boolean Expansion: The algorithm that proved to be most effective in helping users expand their queries (see 'Analysis of user interactions' below) works as follows: after splitting the query into separate keywords, two of them which have the lowest non-zero frequency in the Glimpse index are combined with Boolean 'AND', and offered as a query reformulation option with and without the wildcard '*' (word truncation). For example, if the user enters a zero-hit query like 'alpha smooth muscle actin', the algorithm offers queries for 'smooth AND actin' and '*smooth* AND *actin*', which both match the string 'actin, alpha 2, smooth muscle' in the GeneCard for the gene ACTA2 (Figure 3). In

Fig. 3. A screenshot of the topmost part of the search result for the query 'alpha smooth muscle actin'. Note the suggestions for query reformulation, which include the query 'smooth AND actin'.

Fig. 4. A screenshot of the topmost part of the search result for the query 'Altshimer'. Note the alternative spellings suggested by the interface.

some cases, the wildcard option was helpful as the default option for a GeneCards search is the word match.

Spell Corrector: Whenever the user enters a search, the script that processes the input determines the frequency of the different words that compose the query, searching the Glimpse index. If one of the keywords in the search is not found there, the script performs an 'agrep' search in another file that simply lists all words occurring in the GeneCards data files to find words with *n* misspellings. The algorithm begins with small values for *n*, which are successively increased until a certain number of words is found. The maximal value of *n* depends on the length of the keyword. On the resulting 'query reformulation page' (Figure 4), a list of ready-to-click agrep output keywords is presented.

Relevance feedback: In information retrieval, relevance feedback methods have been used to improve performance for a particular query by modifying the query, based on the user's reaction to the initial retrieved documents (Haines and Croft, 1993). The algorithm behind GeneCards' 'More like this' link, which is offered for every search result (Figure 2), selects keywords from particular parts of the flat file storing the data for the respective gene, based on their frequency in the Glimpse index (only those below a particular threshold are taken). These keywords, combined with Boolean 'OR', are then used to retrieve a list of possibly related GeneCards. Therefore, this algorithm helps to retrieve genes involved in the same cellular process or disease, or factors that interact with the gene of interest.

Special Tips: To provide additional information to the user that is related to particular queries, we compiled a list of more than 100 'search tips' (Figure 2). One or more of these tips are displayed on the search result page if one of the search keywords entered matches a certain pattern. Using the user interaction records, we tried to concentrate on providing solutions for common problems. In many cases, the search tips contain links to outside resources specialized on a particular set of genes, or a specific disease.

Implementation

An overview of the implementation is shown in Figure 5.

Usage examples

A user entering the query 'apoptose' will receive a result page that not only states that the search was unsuccessful, but that also offers a few suggestions for query reformulation (Figure 5). Among them, the keywords 'apoptosis' and 'apoptotic' appear (cf. Figure 4). In this example, a search for 'apoptosis' appears to be the most sensible option. Choosing it, the user receives the following (Figure 2):

(1) A table displaying information about human genes that are involved in this cellular process, with the standard name of the gene (e.g. CASP3 = caspase 3, apoptosis-related cysteine protease), its cytogenetic location (4q53, in this case), and extracts taken from this GeneCards entry that matched the query (e.g. the line: 'PROTEIN: function: important mediator of apoptosis. at the onset of apoptosis it proteolytically cleaves poly(adp-ribose) polymerase (parp) at a 216-asp');

(2) Hyperlinks to the respective GeneCards entries (e.g. the GeneCard for CASP3; cf. Figure 1);

(3) Hyperlinks called 'More like this' which retrieve genes that are involved in similar cellular processes (apoptosis), that have a similar name (e.g. caspase), or that belong to the same gene family (e.g. CASP6, CASP9);

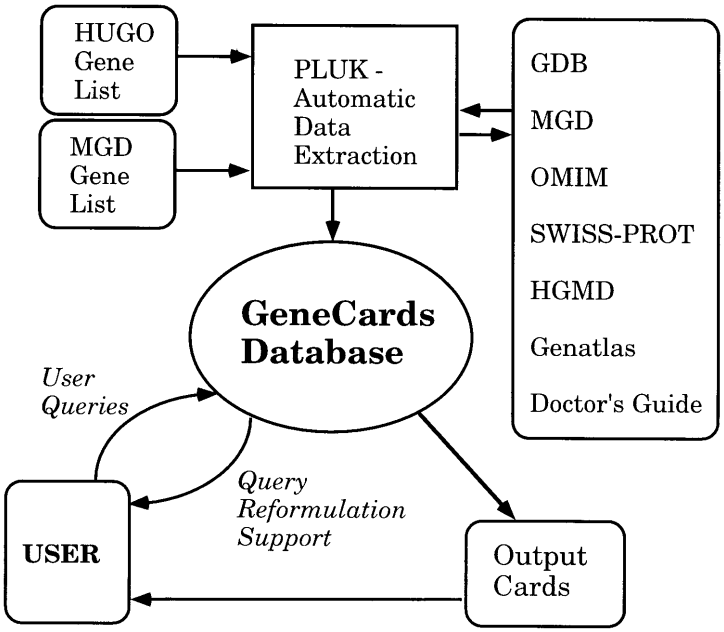

Fig. 5. An overview of the implementation. The HUGO list of approved gene symbols, or another list of genes, is used by PLUK to extract information about each gene from GDB, MGD, OMIM etc.; the results are stored in the GeneCards database, which offers query reformulation support (including a spell corrector specific for the GeneCards data).

(4) In many cases, 'special tips' that provide links to highly selected Web sites related to the query will be returned. In this case, one of the offered links is entitled 'About apoptosis', and retrieves a page on our site that explains the meaning of the word apoptosis in a few words (such pages, called 'GeneCards dictionary' entries, are available for some of the most frequently requested search keywords). In addition, it offers links to information about online overview articles, scientific news and reviews;

(5) A link entitled 'How to find more information' that retrieves a page offering query reformulation support. Here, you will find a page that offers the suggestion to search for 'apopto*', to make a 'concept search', or to directly query outside resources like PubMed, GDB or Excite by clicking on a search button that searches those resources with predefined options.

Selecting one of the hyperlinks entitled 'Display the complete GeneCard for this gene' (see step (2) above), the user accesses one of the more than 7000 'GeneCards' which offer overview information about a particular human gene. Such a Web page is illustrated in Figure 1, which represents the 'GeneCard for BRCA1'. To facilitate human browsing, the information is concise, and highly structured. On the top of the page, the official name and symbol as approved by the HUGO/GDB nomenclature committee (http://www.gene.ucl.ac.uk/nomenclature/) is given. The first row of the table contains the synonyms extracted from GDB, followed by the row with information about similar genes in other organisms (currently only mouse, taken from MGD). Below, information about the name of the protein product(s) from SWISS-PROT, including functions, subcellular localization and role in diseases may be found. Detailed information, as with the data mentioned above, can be accessed via the hyperlinks, in this case by clicking on the link entitled 'BRC1_HUMAN' (the SWISS-PROT ID of this protein). Whenever you move the mouse over a link on a GeneCard, a short description of the link will appear in the status line of browsers that support Javascript.

Analysis of user interactions

Following the initial implementation of GeneCards (Rebhan *et al.*, 1997), we began to thoroughly analyze the interactions between users and the Web interface. Analysis of such interactions during a 26 day period in summer 1997 showed that the total number of entered queries during this period was 2368, an average of about 91 queries per day (this figure rose to more than 300 queries per day in June 1998). Each of these queries is a string of keywords entered by a GeneCards user into one of the search boxes on the interface (the display of retrieved entries is not counted). In 900 of these queries (38%), the database returned zero results. Of those 'zero-hit' searches, a non-redundant set of 213 was selected for further

analysis; queries which did not produce any results after extensive modification of the query by us were not included.

This sample of 213 zero-hit queries (9% of all entered queries), likely to represent typical cases of unsuccessful query formulation, was used to develop algorithms for query reformulation support. The output of those algorithms should then help the user to expand the query.

In 32% of the above mentioned sample of 213 zero-hit searches, the Boolean Expansion algorithm (Figure 3) calculated suggestions for query expansion that would have helped the user to find meaningful results. Spelling errors in at least one of the search keywords were found to be the major problem in 22% of the queries in the evaluated set of 213 queries. Our Spell Corrector (Figure 4) was able to present the apparently correct version of the word among relatively small sets of alternatives in 97% of those cases. In 23% of the query sample, the suggestion of using wildcards in different forms (with or without truncation of the end of the word) was found to be useful. Also, we found that in 18% of the cases it was helpful to suggest splitting the query into the words that compose it, and to suggest the user to search for each of them separately.

In 86% of the queries in our test set, the above mentioned query reformulation algorithms were able to offer the user at least one ready-to-click suggestion that directly led to meaningful results.

Discussion

An overview of the strategy used to build GeneCards is shown in Table 1. The various items mentioned there are addressed below, and in the section 'Algorithms'. Figure 5 shows an overview of the implementation.

Data extraction and integration

From the user's point of view, data integration means an ability to efficiently discover new relationships among data items. To achieve various levels of data integration, several approaches are commonly employed (Jamison *et al*., 1996): (a) complete non-redundant integration into one single format (warehouse, e.g. IGD, ENTREZ and OWL); (b) creation of hyperlinks between related database entries (virtual federation, e.g. SRS); (c) parallel query of multiple databases (e.g. ENQUire). GeneCards may be considered to be halfway between the warehouse and the virtual federation, because more than 'just the link' is offered, although no complete integration is achieved. Such a strategy helps the user to quickly gain an overview of current knowledge related to a query, which is difficult to obtain by browsing virtual federations alone. On the other hand, it avoids the substantial costs associated with the development and maintenance of large data warehouses. The latter advantage may help small academic groups to offer useful, specialized compendia as a service to the research community.

When compared with a parallel query resource like ENQUire, GeneCards has the advantage of a quicker response, because data are stored locally. Also, it is easier to gain an overview of the 'query-related information space' in such a 'cleaned' set of data, especially in the case of queries that return numerous results (see 'Displaying Search Results' below). Specifically, GeneCards offers rapid insight into relationships between protein- and nucleotide-related data.

Usability analysis

A systematic and continuous analysis of user interactions is crucial to find frequent and important problems, and has been advocated by researchers in the field of Human–Computer Interaction (for an introduction, see Shneiderman, 1998; Nielsen, 1995; Buckingham Shum and McKnight, 1997). Many Web information providers rely largely on email feedback sent by users, although such a feedback mechanism probably will not give the designer a comprehensive impression about the problems that different user groups have when accessing the resource. Moreover, it is almost impossible to distinguish frequent and important problems from rare ones. A thorough analysis of logging data (Shneiderman, 1997; Smith *et al*., 1997) can expose 'hidden links', cumbersome sequences of action, unclear commands and explanations, and common query reformulation problems. As a consequence, the developers can concentrate on the most important problems first.

Table 1. Our strategy to address common usage problems with biomedical information on the Internet

1	Topic-specific compendia provide concise information gathered from distributed sources
2	Promotion of standard nomenclature (especially gene symbols and disease names)
3	Hyperlink descriptions facilitate hypertext navigation
4	It matters how you display the data — make them easy to scan and understand
5	Search result pages should contain query-related extracts from the fetched database entries
6	Guiding the user through the retrieval process, e.g. with query reformulation support
7	Provide context with dictionary lookup tools and 'special tips'
8	Thorough analysis of user interactions to reveal common problems

Web navigation

For efficient navigation of hypertext documents, it is important that the user is able to identify the 'meaning' of a hyperlink, so that the fetched page matches the expectations raised by this link (Smith *et al*., 1997). A link title that only contains the acronym of the database and the ID number of the entry, as offered in some database federations, may not be sufficient for users who are not familiar with its meaning. Short descriptions of the linked page should help users select links more accurately. Widespread use of HTML META tags for Web site description (http://bioinfo.weizmann.ac.il/comp/meta_tags.html) would help to automatically associate links with descriptions.

Another problem is that people rarely read Web pages word by word; instead, they scan the page, picking out individual words and sentences (Nielsen, 1997; Morkes and Nielsen, 1997). As a result, Nielsen and coworkers, and others, suggest that Web pages should employ 'scannable' text, using: (a) highlighted keywords (hypertext links serve as one form of highlighting; typeface variations and color are others; see also Shneiderman, 1997); (b) meaningful subheadings; (c) bulleted lists; (d) one idea per paragraph (users will skip over any additional ideas if they are not caught by the first few words in the paragraph); (e) the inverted pyramid style, starting with the conclusion; (f) half the word count (or less) than conventional writing (Nielsen, 1997). These usability concerns have played an important role during the development of GeneCards (Figure 1).

Displaying search results

Our experience with GeneCards has shown that another simple feature is important for the usabililty of a searchable Web resource: the display of detailed search results that contain those sentences from the scanned files which contain the query keywords (Figure 2). This can help users get a quick impression about the query-related information space. Also, it will be easier for the user to find relevant entries, a feature that is especially helpful for queries that return large numbers of results. For example, when searching a database for the keyword 'p53', the search result will contain not only the title of the retrieved pages, but also a few lines of those parts of the text that contain the word 'p53', for example sentences like 'Gene: tumor protein p53.', 'Interacts with p53 to form a complex...', and 'Human homolog of p53-binding protein.'. Because GeneCards contains selected extracts from various data sources, the density of interesting information on such search result pages can be relatively high.

Query reformulation support

Many users have problems to efficiently reformulate their queries, and do not read extensive help instructions about advanced search strategies due to time constraints (Smith *et al*., 1997; Pollock and Hockley, 1997). An intelligent user interface for a search engine should be able to react to common usage problems, like misspelled keywords, and propose options to expand or narrow the search, e.g. by providing reformulated, ready-to-click queries that are related to the original query entered by the user. The GeneCards Web interface features such suggestions for query reformulation, and our user interaction analysis shows that they offer sensible options in the vast majority of cases. However, not all users seem to notice and understand the query reformulation page; many do not select one of the suggestions offered. This effect may be at least partly due to the novelty of the approach.

Future directions

We believe that many of the tools and concepts developed in the realm of the GeneCards project may be successfully applied to other scientific resources. Therefore, we are planning to make some of our more generally applicable tools available. We furthermore will try to integrate more data into the current system, depending on the suggestions we get, and on the availability of new resources. To facilitate the parsing of information stored in GeneCards, we will provide those data in boulder IO format (a simple TAG=VALUE data format designed for sharing data between programs (http://www.genome.wi.mit.edu/genome_software/), XML, and possibly also via an object-oriented broker system based on the CORBA standard. An SQL query interface is also planned.

Conclusions

Because GeneCards offers concise information about the functions of human genes, it has become a valuable tool for many researchers, especially for those working on the analysis of gene networks that contain hundreds or thousands of different genes and proteins (e.g. in the realm of studies employing gene chip technology or two-dimensional gel electrophoresis). Moreover, it may serve as a prototype for new types of compendia that help to cope with the avalanche of biomedical information by automatically extracting essential information from the Internet, and by offering the user a Web interface that makes the difficult process of information exploration as efficient as possible.

Acknowledgements

This work has been supported by a fellowship from the Minerva Stiftung fuer die Forschung (to M.R.), the Israel Academy of Science, the Israeli Ministry of Science, the Henri and Francoise Glasberg Foundation, the Dr. Ernst Nathan Fund for Dermatological Research, La Fondation Raphael et Regina Levy, Mr Marc S. Levine (Houston, Texas), and the Kalman & Ida Wolens Foundation.

References

Buckingham Shum,S. and McKnight,C. (eds) (1997) Web usability (special issue). *Int. J. Human-Computer Studies*, **47**(1). (http://ijhcs.open.ac.uk/)

Dujon,B. (1998) European Functional Analysis Network (EUROFAN) and the functional analysis of the *Saccharomyces cerevisiae* genome. *Electrophoresis*, **19,** 617–624.

Haines,D. and Croft,W.B. (1993) Relevance feedback and inference networks. *Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2–11.

Hieter,P. and Boguski,M. (1997) Functional genomics: it's all how you read it. *Science*, **278,** 601–602.

Humphery-Smith,I., Cordwell,S.J. and Blackstock,W.P. (1997) Proteome research: complementarity and limitations with respect to the RNA and DNA worlds. *Electophoresis*, **18,** 1217–1242. (http://www.proteome.usyd.edu.au/proteome_review.html)

Jamison,D.C., Mills,B. and Schatz,B. (1996) An extensible network query unification system for biological databases. *Comput. Appl. Biosci*., **12,** 145–150.

Lander,E.S. (1996) The new genomics: global views of biology. *Science*, **274,** 536.

Morkes,J. and Nielsen,J. (1997) Concise, scannable, and objective: how to write for the Web. SunSoft paper. (http://www.useit.com/papers/webwriting/writing.html)

Nielsen,J. (1995) *Advances in Human Computer Interaction*, Vol. 5. Ablex, Norwood, NJ. (http://www.useit.com/jakob/ahci5book.html)

Nielsen,J. (1997) *How People Read on the Web*. Alertbox for Web Usability, Oct. 1. (http://www.useit.com/alertbox/)

Pollock,A. and Hockley,A. (1997) What's wrong with internet searching. D-Lib Magazine, March 1997. (http://www.dlib.org/dlib/march97/bt/03pollock.html)

Rebhan,M. and Prilusky,J. (1997) Rapid access to biomedical knowledge with GeneCards and HotMolecBase: Implications for the electrophoretic analysis of large sets of gene products. Electrophoresis. *Biomed. Biocomp*., **18,** 2774–2780.

Rebhan,M., Chalifa-Caspi,V., Prilusky,J. and Lancet,D. (1997) GeneCards: integrating information about genes, proteins and diseases. *Trends Genetics*, **13,** 163.

Recipon,H. and Makalowski,W. (1997) The biologist and the World Wide Web: an overview of the search engines technology, current status and future perspectives. *Curr. Op. Biotechnol*., **8,** 115–118.

Shneiderman,B. (1997) Designing information-abundant web sites: issues and recommendations. In Buckingham Shum,S. and McKnight,C. (eds), Web usability (special issue), *Int. J. Human-Computer Studies*, **47**(1), 5–30. (http://ijhcs.open.ac.uk/)

Shneiderman,B. (1998) *Designing the User Interface*, 3rd edn. Addison-Wesley, Reading. (http://www.aw.com/DTUI)

Smith,P.A., Newman,I.A. and Parks,L.M. (1997) Virtual hierarchies and virtual networks: some lessons from hypermedia usability research applied to the World Wide Web. In Buckingham Shum,S. and McKnight,C. (eds), Web usability (special issue), *Int. J. Human-Computer Studies*, **47,** 67–96. (http://ijhcs.open.ac.uk/)

Wall,L., Christiansen,T. and Schwartz,R.L. (1996) *Programming Perl*, 2nd edn. O'Reilly & Associates, Inc., Sebastopol, CA.