



## GeneCards: integrating information about genes, proteins and diseases

The information produced by the Human Genome Project is of outstanding importance to many researchers in biology and medicine. However, the data are scattered over a variety of heterogeneous databases that are specialized for certain aspects. It can often be difficult to find comprehensive information that gives an adequate impression of current biomedical knowledge about particular genes and their products. To overcome this problem, we have developed a new type of database called 'GeneCards' (Ref. 1). This resource integrates information extracted from 'lower level databases' dealing with human genes, the functions of the proteins they encode, and the diseases in which they are involved.

### What can be found at the GeneCards site?

To get an idea about the problems that can be solved with the GeneCards database, it is important to understand what a 'GeneCard' actually is. The GeneCard for the recently discovered breast cancer gene 1 (*BRCA1*; Ref. 2) is a good example because there is already a lot of interesting information about this gene from different fields. The GeneCard provides the following information: (1) the official gene name as approved by the HUGO/GDB nomenclature committee<sup>3</sup>; (2) a list of synonyms; (3) homologous genes in the mouse; (4) the chromosomal locations of the gene and of its homologues; (5) the name of the protein(s) encoded by the gene, including brief descriptions about its cellular function, expression patterns, similarities

with other proteins and involvement in diseases; (6) a list of disorders in which the gene is involved according to genetic evidence; (7) new diagnoses and therapies that can be regarded as being applications of the knowledge about this gene; and (8) links to sites that contain further information [e.g. the Genome Database (GDB), SWISS-PROT, the Human Gene Mutation Database (HGMD), Online Mendelian Inheritance in Man (OMIM), and so on; Ref. 4].

In addition to the extracts directly visible on the GeneCard, links to further information, in most cases to the entry in the source database, are included. If you are interested in knowing more about the disorders listed in section 6, the associated link (Ref. 2: 'More about this disorder') takes you directly to a list of Web sites that covers related articles, reviews, homepages of researchers, conference proceedings and announcements, and medical and scientific news articles. In a similar fashion, more information about the gene and its products can be retrieved by using the Web search link in the section called 'Additional Sources of Information' (Ref. 2: 'Search the web').

An individual GeneCard, which offers fast access to current biomedical information about a particular gene and its product(s), can be accessed by using our powerful, but simple-to-use, search engine. A special feature of this search engine is that it returns 'verbose extracts' of the data that might be relevant, and that it highlights those words that matched the query. Thus, by browsing through these extracts, the user can quickly select the GeneCards that might be most interesting, without losing time by clicking into links that later turn out to be of limited relevance to the question asked (e.g. try a GeneCards search for 'Alzheimer\*', 'p53' or 'apopto\*').

If your search does not produce any result, the database suggests strategies for improving your search that depend on the type of query you have entered, and offers direct links to search engines of other databases. For example, if you type a word that produces no results, it suggests using the wildcard '\*' for truncation, and gives some examples of how this can be done. You can then start a new search directly by clicking on one of these suggestions.

### The underlying software technology

The data presented in the GeneCards knowledge database are extracted periodically in an automatic fashion from the sources that are indicated in the left column of the main table in each GeneCard (Refs 2, 4). We use a package of PERL scripts (PLUK, package for locating useful knowledge) developed by us for this purpose. PLUK examines texts for contextual relationships, then extracts and processes the relevant information and stores it in the GeneCards source files.

### Future prospects

At present, GeneCards contains the more than 6000 genes that have an officially approved gene name and symbol (Ref. 3). We are currently developing the following features, which will soon be integrated into the database (for details see Ref. 4). (1) A user guidance system that utilizes artificial intelligence tools to afford fast and reliable information retrieval even to poorly specified queries. This will be based on a continuous interaction with GeneCards users. (2) A 'second level' GeneCards system that offers a detailed, but easy-to-browse, summary of current knowledge about the involvement of a gene and its products in the pathogenesis of a particular disease. This information will be extracted from Medline articles by an extended PLUK version able to build lists and tables from free text information.

Since the publication of the GeneCards database in early 1997, we have received many stimulating comments from the scientific community. We are optimistic that the GeneCards project, which is based on a combination of standard bioinformatics technology with newly developed knowledge extraction tools, will be useful to the biomedical and genome communities. Hopefully, it will help to organize the rapidly accumulating biological and medical knowledge about genes, and will become an essential tool for presenting the complexity of genomics to the whole scientific community.

**Michael Rebhan\***  
lvrebhan@bioinformatics.  
weizmann.ac.il

**Vered Chalifa-Caspi\***  
lchalifa@bioinformatics.  
weizmann.ac.il

**Jaime Prilusky\***  
lprilus@weizmann.  
weizmann.ac.il

**Doron Lancet\***  
bmlancet@weizmann.weizmann.ac.il

\*Department of Membrane Research  
and Biophysics.

†Department of Biological Services  
(Bioinformatics Unit).

‡Weizmann Institute Genome Project,  
Weizmann Institute of Science,  
76100 Rehovot, Israel.

### References

- 1 <http://bioinfo.weizmann.ac.il/cards>
- 2 <http://bioinfo.weizmann.ac.il/cgi-bin/lvrebhan/carddisp?BRCA1>
- 3 <http://www.gene.ucl.ac.uk/nomenclature/>
- 4 <http://bioinfo.weizmann.ac.il/cards/background.html>