Bioinformatics approach to extract information from genes

G. R. Sridhar*, C. H. Divakar, T. Hanuman***, Allam Appa Rao******

*Endocrine and Diabetes Centre, Visakhapatnam, **Department of Computer Sciences, Visakhapatnam, ***Department of Computer Sciences, Acharya Nagarjuna University, Guntur, ****Department of Computer Sciences and System Engineering, Andhra University College of Engineering, Visakhapatnam, India

With the widespread availability of nucleotide and amino acid sequences, novel methods for extracting biologically and clinically relevant knowledge are feasible. Data is deposited on the Internet on websites such as GeneCards, available at http://www.genecards.org/mirror.shtml. Further information can be obtained from related sites - UniProt (http://www.uniprot.org) and SwissProt (http://www.expasy.org/sprot/). Using FASTA and CLUSTAL_X programs, similarity scores can be calculated to choose items of interest. Further information can be obtained by mining text, either manually or increasingly using text-mining tools such as PathBinderH and GENIA corpus.

KEY WORDS: Bioinformatics, GeneCards

Nucleotide and amino acid sequences of diverse life forms, including plants, bacteria and animals, are now widely accessible.[1] The current focus is shifting to gather, annotate and make associations using the available data. Collaboration among scientists from varying backgrounds is now commonplace.

Here we present a summary of bioinformatics resources available to extract information from nucleotide and amino acid sequence data.

GeneCards

In the 1990s the Internet became an enormous repository of useful biomedical information; however, mere availability of data did not always allow the user to access it efficiently, and the user was often 'lost in a labyrinth of hypertext links.'

Correspondence to **Dr. G. R. Sridhar,** Endocrine and Diabetes Centre, 15-12-16 Krishnagar, Visakhapatnam - 530 002, India.
E-mail: sridharvizag@gmail.com / grsridhar@hotmail.com

The GeneCards, available at http://www.genecards.org/mirror.shtml, was developed to present information about known function of genes 'in a way that corresponds to their conceptual information space.' It contained information about human genes, including their cellular function and involvement in diseases.[2]

Data are stored in flat files, indexed by the Glimpse package (http://glimpse.cs.arizona.edu). Information related to a gene is created 'on-the-fly by a CGI script,' which allowed quick browsing. To facilitate easy grasp, bulleted lists and tables were generated.[2]

The site is freely available for educational and research purposes as a system of 'human genes, proteins and diseases' to integrate, search and display gene-centered human genetic information.[3] It is rapidly becoming a major resource for biological data mining and integration.

How to use GeneCards (http://www.genecards.org/background.shtml - accessed on 21st May 2006)

"The information in GeneCards is organized in a gene-centric fashion, where each gene has a virtual card displaying the information associated with that gene. Each card can be accessed from the main page search box through several mechanisms:

- Selecting the 'Symbol only' button and typing the gene symbol
- Selecting the 'Symbol/alias' button and typing a symbol or alias for the gene
- Selecting the 'GC id' button and typing the requested GCid
- Selecting the (default) 'keywords' button to do a full free text search

*In the first three cases, use of a wild card (*) will result in retrieval of a list of symbols, and use of space in the

query term will result in a keywords search instead of the selected search type.

*The GeneCards search is case insensitive."

Further information about proteins of interest can be obtained from databases such as UniProt or SwissProt.

UniProt and SwissProt

UniProt has been described as the 'Universal Protein knowledgebase.' The aim is to 'provide a comprehensive, fully classified, richly and accurately annotated protein sequence knowledgebase, with extensive cross-references and query interfaces.'[4] Accessible online at http://www.uniprot.org, it consists of three components: 'The UniProt Archive (UniParc) stores complete body of publicly available protein sequence data; the UniProt Knowledgebase (UniProt) provides the central database of protein sequences with accurate, consistent and rich sequence and functional annotation and the UniProt NREF databases (UniRef) provide nonredundant data collections based on the UniProt knowledgebase in order to obtain complete coverage of sequence space at several resolutions.'[4] The aim is to ultimately allow 'researchers to integrate the enormous amount of data from the Human Genome Project and from structural and functional genomics and proteomics.'

The database continues to grow and have new features. The new recent evolution of the database has been reviewed in 2006.[5]

The SwissProt protein knowledgebase (http://www.expasy.org/sprot/ and http://www.ebi.ac.uk/swissprot/) 'connects amino acid sequences with the current knowledge in the Life Sciences.'[6] It is noted for 'its high-quality annotation, the usage of standardized nomenclature, direct links to specialized databases and minimal redundancy.' 'The extensive integration of SwissProt with specialized databases enables users to navigate through the current knowledge in the Life Sciences, providing an insight into the universe of proteins.'[6]

FASTA for Sequence Comparison

Once the amino acid or nucleotide sequences of interest are selected from the databases, computer programs such as FASTA[7] can be used to compare the sequences. The program evaluates similarity scores and identifies structures based on sequence similarity. In essence, it locates groups of identities between sequences during the first step of comparison, followed by rescoring of regions by a score matrix that allows shorter identities to contribute to the similarity score. It then checks the single best scoring initial region and then calculates an optimal alignment of initial regions to rank the sequences. Finally, the highest scoring sequences are aligned by a modified optimization method.[7]

The sequences are then stored as a word file and results interpreted. To facilitate and hasten the analysis, interactive programs have been developed, viz., Visual BLAST and Visual FASTA.[8] They are available at (http://www.lmcp.jussieu.fr/).

Multiple-sequence alignments can then be performed using the CLUSTAL_X windows interface.[9] It is a Windows-based program for multiple-sequence and profile alignments and analysis. CLUSTAL_X Windows interface is particularly useful when the compared sequences are not very homologous.

Annotation and Synthesis of Biologically Relevant Knowledge

Choosing sequences based on bioinformatics tools, one can then check published information to synthesize functional or pathological expression and pathways of the genes. Increasingly, methods are now being developed to automate this process.[10] Broadly, the aims are to identify mentions of relevant biological entities (genes, proteins, etc.) in running text and to automatically annotate the protein. PathBinderH[11] and GENIA corpus[12] are being developed using Natural Language Processing methods to mine text data.

References

1. Sridhar GR. Impact of human genome project on medical practice. J Assoc Physicians India 2001;49:905-8.
2. Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. GeneCards: A novel functional genomics compendium with automated data mining and query reformulation support. Bioinformatics 1998;14:656-64.
3. Safran M, Solomon I, Shmueli O, Lapidot M, Shen-Orr S, Adato A, *et al*. GeneCards 2002: Towards a complete, object-oriented, human gene compendium. Bioinformatics 2002;18:1542-3.
4. Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, *et al*. UniProt: The Universal Protein knowledgebase. Nucl Acids Res 2004;32:D115-9.
5. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, *et al*. The Universal Protein Resource (UniProt): An expanding universe of protein information. Nucleic Acids Res 2006;34:D187-91.
6. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A,

Gasteiger E, *et al* The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 2003;31:365-70.
7. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 1988;85:2444-8.
8. Durand P, Canard L, Mornon JP. Visual BLAST and visual FASTA: graphic workbenches for interactive analysis of full BLAST and FASTA outputs under MICROSOFT WINDOWS 95/NT. Comput Appl Biosci 1997;13:407-13.
9. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 1997;25:4876-82.
10. Mukherjea S, Sahay S. Discovering biomedical relations utilizing the world-wide web. Pac Symp Biocomput 2006;11:164-75.
11. Ding J, Viswanathan K, Berleant D, *et al*. PathBinderH: A tool for sentence-focused, plant taxonomy-sensitive access to the biological literature. Software Artifact Research and Development Laboratory, Technical Report SARD11-19-04. [Last accessed on 2006 Jun 06]. Available from: http://class.ee.iastate.edu/berleant/s/paperPathBinderHreport.pdf.
12. Kim JD, Ohta T, Tateisi Y, Tsujii J. GENIA corpus--semantically annotated corpus for bio-textmining. Bioinformatics 2003;19:i180-2.

Source of Support: Nil, **Conflict of Interest:** None declared.