Medical informatics: transition from data acquisition to data analysis by means of bioinformatics tools and resources

Mahmood A. Mahdavi

Department of Chemical Engineering,
Ferdowsi University of Mashhad,
Azadi Square, Pardis Campus,
Mashhad, 9177948944, Iran
E-mail: mahdavi@ferdowsi.um.ac.ir

Abstract: Medical informatics has shifted its focus from acquisition and storage of healthcare data by integrating computational, informational, cognitive and organisational sciences to semantic analysis of the data for problem solving and clinical decision-making. In this transition, bioinformatics tools and resources are the most appropriate means to improve the analysis, as major biological databases are now containing clinical data alongside genomics, proteomics and other biological data. This paper briefly reviews bioinformatics tools and resources and then discusses their applications in analysing clinical data for diagnostics.

Keywords: clinical diagnostics; decision-making; microarray; database; homology; bioinformatics.

Reference to this paper should be made as follows: Mahdavi, M.A. (2010) 'Medical informatics: transition from data acquisition to data analysis by means of bioinformatics tools and resources', *Int. J. Data Mining and Bioinformatics*, Vol. 4, No. 2, pp.158–174.

Biographical notes: Mahmood A. Mahdavi received his BSc and MSc in Chemical Engineering from the Isfahan University of Technology, Isfahan, Iran, in 1991 and the Amirkabir University of Technology, Tehran, Iran, in 1994, respectively. He obtained his PhD in Computational Biology and Bioinformatics from the University of Saskatchewan, Saskatoon, Canada, in 2007. He is currently a Faculty in the Department of Chemical Engineering, Ferdowsi University of Mashhad (FUM), Mashhad, Iran. His research interests are computational biology, with emphasis on detection of protein–protein interactions and its application in biomedical research.

1 Introduction

Sequencing technology was a revolutionary discovery in the history of modern biology. Genome projects are completed and publicly available one after another deciphering more complicated organisms from *H. influenza* to *Homo sapiens* (human). The huge accumulation of biological data evolved from the discoveries has made bio-scientists introduce new tools for analysing the data and semantic interpretation of the raw

information (Fell, 2001). Bioinformatics is the discipline emerged from this growing demand. It is a hypothesis-driven science focusing on integrating biological themes together with the help of computer tools and biological databases, and gaining new knowledge from this.

On the other hand, medical research has experienced a shift in direction from *in vivo* to *in silico* investigation, a development building upon bioinformatics and its tools and resources. The primary motivation for this transition is the completion of human genome project in recent years (International Human Genome Sequencing Consortium, IHGSC, 2004). However, the key to success of the integration of bioinformatics and medical research, the so-called medical informatics, is the transformation of biological data including human genomics data into information that is useful for clinical diagnostics and therapies by means of bioinformatics tools. One important output of this transformation is transition from data acquisition to data analysis in the field of medical informatics. A precise definition of medical informatics is presented as follows:

> "The integrative discipline that arises from the synergistic application of computational, informational, cognitive, organizational, and other sciences whose primary focus is the acquisition, storage, and use of information in the health/biomedical domain." (Hersh, 2002)

As stated in this definition, the main focus of medical informatics is acquisition and direct use of collected data in healthcare system. One developing step in the field is analysing the data by means of bioinformatics tools gaining new information about the unidentified diseases. This information is stored in well-organised databases. The analysis, for instance, may result in homology between the unidentified diseased genes and previously known genes in other organisms, or may result in evolutionary or functional correlations for unknown genes. These results assist researchers to identify diseased tissues through *in silico* investigation.

Biological data obtained by means of high-throughput technologies are required to be organised in a way to be understandable by human and analysable by machines. The best way to satisfy these requirements is storing the information in databases. With the availability of database management systems and appropriate programming languages such as MySQL, experimental results are now stored in public databases as open sources. Databases are main resources of bioinformatics (Cattley and Arthur, 2007). They are the organised array of information. One can put things in and is able to get them out again. Databases simplify the information space by specialisation. They are, also, resources for other species-specific databases and tools. Records of information in databases have specific definitions and descriptions. They possess unique keys as identifiers. These identifiers are most often accession numbers for different biological elements including genes, DNA sequences, ORFs, proteins, domains, motifs and so on. Each database has its own set of accession numbers. The accession numbers facilitate searching the database for special information. Update version and links to other databases are amongst other characteristics of biological databases. They present proper documentations and provide submission, update and correction processes.

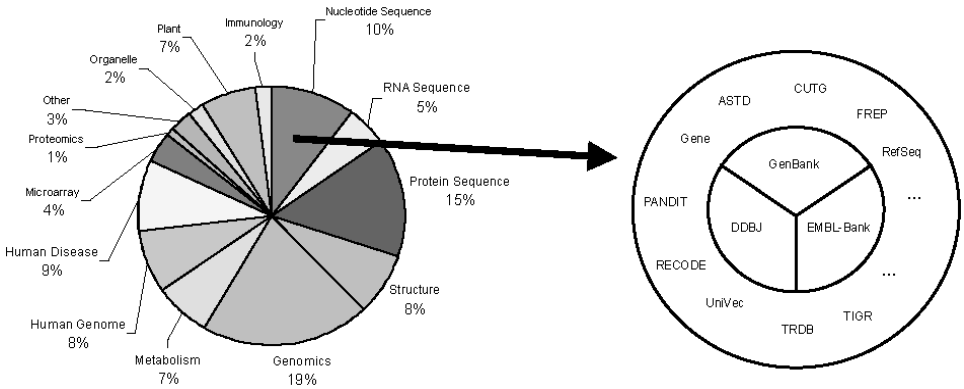
Along with the emergence of databases, appropriate tools have been developed for the manipulation and mining of biological data. These tools enable scientists to search databases for correlations in terms of evolution, homology, function and structure. Computational biology is highly dependent on the bioinformatics tools by which new hypotheses are investigated.

In this paper, major resources and tools of bioinformatics are discussed. The core databases are introduced and the databases evolved from the core databases in recent years are reviewed. Then, more popular bioinformatics tools are reviewed along with their capabilities. Finally, the common applications of bioinformatics tools and resources are briefly discussed and the emerging applications particularly in medical field are described.

2 Databases as resources

Every year, a full list of current molecular biology databases is prepared by NIH and published in *Nucleic Acids Research*. In 2008 update (Galperin, 2008), 1078 databases are reported, 110 more than previous one. Eighty databases updated since last year and 25 obsolete databases have been reported. Biological databases can be divided into different categories in terms of the type of the information they provide. The information ranges from very general and fundamental to very specific or species-specific. For example, National Center for Biotechnology Information (NCBI) contains the DNA sequence of all organisms whose genome is completed. This information is general as well as fundamental. WormBase, or FlyBase are examples of species-specific databases comprising all genomic and functional information related to *C. elegance* worm and *D. melanogaster* fly, respectively. Enzyme is an example of very specific databases that is dedicated to providing information on enzymes and their metabolic activities. There are totally 13 categories of databases in the latest update. Some categories consist of a few core databases and the remaining ones in the category emerge from the core databases. The categories and their subcategories are illustrated in Figure 1. Some of more important categories will be described here.

Figure 1 Distribution of molecular biology databases in terms of their contents (left-hand side pie chart). The figure on the right-hand side shows the subcategories of major databases for nucleotide information as an example



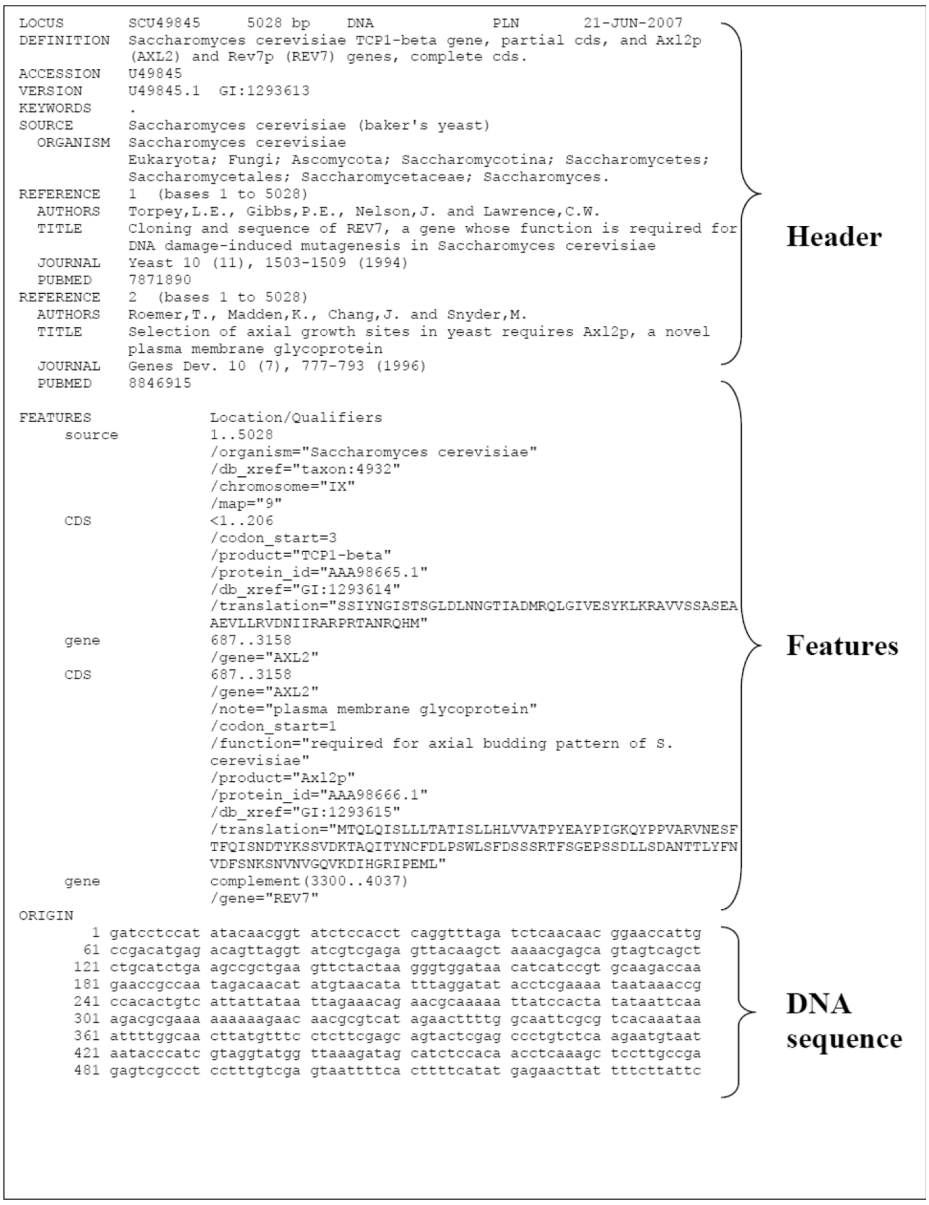
The first and most important category includes nucleotide sequence databases. These databases contain primary data for computational biology as we call them here ‘core databases’. These data account for the backbone of other databases that provide more specific information. Since most biological research often stems from the

nucleotide sequence information, it is globally stored in three major data repositories including GenBank, EMBL-Bank and DDBJ. GenBank is the NIH (US National Institute of Health) genetic sequence database of all publicly available DNA and derived protein sequences with annotations describing the biological information these records contain. This database is accessible through NCBI website. EMBL-Bank is the European Molecular Biology Laboratory DNA database. It is accessible via European Bioinformatics Institute (EBI) website. This database contains DNA sequences of all completed prokaryotic microorganisms and fungi. DDBJ is the DNA Data Bank of Japan operated by National Institute of Genetics (NIG) of Japan. This database also maintains the DNA sequences of many organisms including human genome. The three major repositories contain the raw information required for each biology research and serve as open sources. For the user's convenience, these three databases provide information in identical format (Brunak et al., 2002). Each record of information in each of the three databases is a flat file in text format that consists of three sections as shown in Figure 2: header, features and DNA sequence. Text file is the simplest format readable by machine. The arrangement of the data is in a way that is understandable by human. In the header section, general information about the gene including definition, accession number (common among three databases), organism and citation is presented. In the features section, genetic information such as location, length and amino acid sequence(s) of coding regions are demonstrated. The complete nucleotide sequence of the related gene is placed in the sequence section in the form of 60 residues on each line.

The three core databases are equipped with powerful search engines for the purpose of sequence retrieval and associated data. These search engines not only retrieve the DNA sequence, but also search other associated databases maintained by the websites. The NCBI website provides Entrez for searching DNA sequences in GenBank. It is a global cross database search engine that explores many NCBI's databases when a query term or phrase is entered. The EBI has dedicated Sequence Retrieval System (SRS) to search biological data such as nucleotide sequences. It is used for cross-reference searching of protein sequences, accession numbers, enzymes, organic compounds, etc., as well. NIG of Japan has provided All-round Retrieval of Sequence and Annotation (ARSA) for searching databases including nucleotide sequence databases. This search engine performs cross search for query terms and explores 20 major databases for a specified query.

There are so many other databases in the category built upon the three core repositories. They consider the nucleotide sequences as raw data and after a curation process the data are organised in a more applicable way. During curation process, trained curators try to match separate attempts published in different journals and resolve mismatches in the findings. They finally come up with a non-redundant version of the results and then publish them on the web. These databases also contain information such as coding regions, regulatory sites, gene structures, transcriptional regulators, transcription factors and functional annotations related to each particular sequence. This information is collected from journal publications and assigned to sequences by curators based on the current knowledge. RefSeq, ASTD and Genes are databases as such.

Figure 2 The structure of nucleotide information files in three core nucleotide sequence databases. Each file comprises three sections: Header, Features, DNA sequence. See text for details on each section



Another category of databases are protein sequence databases. This category contains core databases such as UniProt, Protein Information Resources (PIR) and Protein. UniProt knowledgebase consists of Swiss-Prot and TrEMBL. Swiss-Prot is a collection of manually curated annotation information and cross-reference data. The latest version

contains 405,506 entries (Release 56.6, 16 December, 2008). TrEMBL maintains a collection of electronic annotations and includes 6,964,485 entries with the latest version (Release 39.6, 16 December, 2008). Each Uniprot entry contains various types of information in one file such as name and origin of the protein, citation, cross-reference data and the amino acid sequence. Protein Information Resources is another core database in this category that provides variety of information related to proteins. It is the result of integrating more than 90 small and specific databases in one single database. PIR provides structure, family, sequence, expression, variation, function, etc., The NCBI maintains the Protein database containing amino acid sequences either translated from RNA sequences or discovered experimentally. In addition, the Protein contains other protein-related information mentioned earlier.

Other databases in category carry information on protein properties, localisations, motifs and active sites, functional domains and databases dedicated to specific protein families. BLOCKS, eSLDB, InterPro, PRINTS, PDB, Pfam, SMART are examples of such databases. For instance, PDB is a comprehensive database of 3D protein structures and protein functional sites. It contains 55,271 structures (as of December 2008) and collaborates with UK and Japan to ensure that the PDB archive is global and uniform. Most of the structures in the PDB are obtained by experimental techniques such as crystallisation. InterPro is a database of protein families, domains and active sites that is used for identifying new protein sequences by means of searching features of known proteins. It has 11 member databases that share various types of information with the InterPro on proteins. Pfam is a database of protein families currently (release 22 July, 2007) containing 9318 curated families.

The third important category of molecular biology databases is genomics. 19.3% of all databases are rather related to genomic information (see Figure 1). There are some core databases in this category that often contain genomes of various species in three domains of life, prokaryotes, eukaryotes and archaea. EBI Genomes, Entrez Genomes and GIB are three example databases supported by EBI, NCBI and NIG, respectively. Core databases present general genomics data such as ORFs list, comparative genomics and an overview analysis of different genomes. Numerous non-core databases are available on genomes of specific organisms that provide in-depth information. EchoBASE on *E. coli*, BSORF on *B. subtilis*, xanthusBase on *M. xanthus*, CYGD and SGD on *S. cerevisiae*, WormBase on *C. elegans* and FlyBase on *D. melanogaster* are as such. These databases not only carry the nucleotide sequences of the organism's genes, but also contain annotations and genetic information related to each particular gene. Cross-reference data are also included in these databases.

There are more categories of molecular biology databases as bioinformatics resources (see Figure 1). The description of those categories and associated databases can be found elsewhere (Fox et al., 2005).

3 Bioinformatics tools

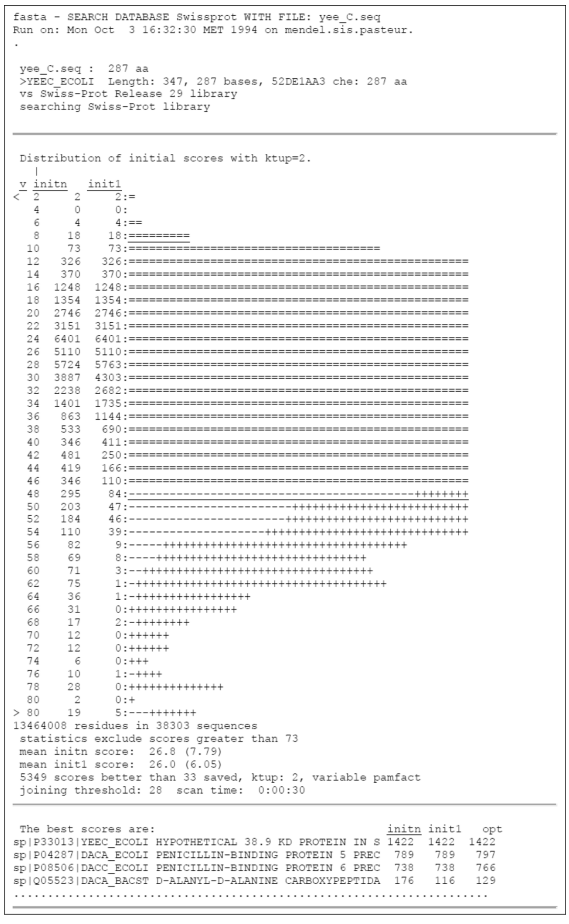
The enormous amount of biological data has required developing appropriate tools by which new knowledge is gained from the data. Bioinformatics is the background of all the growing number of tools. To understand the importance of using bioinformatics tools, one needs to gain an insight into the flow of biological information in the past decade.

In 1999, the bioinformatics information space was consisted of 4,456,822 nucleotide sequences and 706,862 protein sequences. In 2006, the figures were increased to 68,739,698 and 7,861,530 sequences, respectively. In 1995, the first bacterial genome was fully sequenced. Currently, more than 850 genomes are completed, approximately 940 draft assemblies are available and 950 genome projects are in progress (http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html, Revision 31 December, 2008). Thus, without the use of powerful tools, deriving valuable results out of this warehouse of data is impossible.

The most common bioinformatics tool is used to search similarities among sequences. Sequence comparisons lie at the heart of all bioinformatics. Sequence similarity is determined by alignment. Alignment tells us about activity of a new gene, structure and location of a new protein, origin of a protein/organelle/organism. As a result of alignment, similarity between two sequences is quantified. Similarity means that the two sequences share a significant number of bases or residues and thereby the homology of the two sequences is inferred.

FASTA (FAST-All) and BLAST (Basic Local Alignment Search Tool) are the two programmes developed to perform the alignment. These two programmes perform alignments based on heuristic (fast) local alignment method. These methods are based on dynamic programming with a further improvement in considerable reduction in running time through a pre-processing of query and reference sequences. FASTA package (Pearson and Lipman, 1988) is designed for a fast protein or nucleotide comparison. This programme achieves a high level of sensitivity for similarity searching at high speed. Briefly, this programme exploits a four-step algorithm. First, using a look-up table, all dense regions of identities between two sequences are identified. The ktup value determines how many consecutive identities are required for a match to be declared. The lesser the ktup value, the more sensitive the alignment. Often, ktup = 2 is taken for proteins, and ktup = 6 is taken for nucleotides. Once the ktup matches are identified, they are represented as diagonals in a dot plot isolating the dense regions from the background hits. Ktup matches are scored using a substitution matrix (BLOSUM50 for protein sequences and identity matrix for nucleotides). The best 10 local regions are selected from the diagonals. In the second step, the 10 initial regions are re-scored using a scoring matrix (generally PAM250). In this step, shorter than ktup identities are allowed to contribute to score. A subregion with maximum score is identified (init1). In the third step, it is checked to see if several of the initial regions greater than a cut-off value can be joined together. A similarity score that is the sum of the joined regions minus a joining penalty for each gap is calculated (initn). This score is used to rank database sequences. The top ranking database sequences are then further compared using Smith–Waterman algorithm to calculate an optimal score for alignment. The significance of the optimal alignment is evaluated and expressed as *z*-value score, which is defined as the ratio of observed score minus mean of shuffled scores to standard deviation of shuffled scores. FASTA programme has different variations FASTP, FASTN depending on the type of the sequence searches that could be DNA:DNA, Protein:Protein and DNA:Protein. Figure 3 partially shows a typical output of a FASTA searching programme.

Figure 3 A typical output of FASTA program. The header part of the file contains general information on the search including the query protein (an *E. coli* penicillin binding protein in this example), the length of amino acid sequence, and the library database to search in. The second section demonstrates histogram of the distribution of calculated *z*-scores according to the word size or ktup value (ktup = 2 for a protein query sequence). Each line describes one databases sequence matching the query, printed in ascending order of *z*-scores. For each *z*-score bin, the initn and init1 values are specified (see text for description on initn and init1) from the search. Furthermore, when the number of regions reported in columns initn and init1 are equal, it is represented by "=" sign. When the numbers reported on the two columns are different, initn values are shown with a "+" and init1 values are shown with a "–". At the end of this section, some statistics on mean of initn and init1, processing time, etc is reported. In the next section, the best hits are reported. Sequences that have an initn larger than a cut-off value are then used for re-scoring and "opt' scores are calculated. The final section provides the alignment of the best hits and associated information including percent similarity over the region aligned, and length of the region (due to limited space some material deleted from the file)



```
fasta - SEARCH DATABASE Swissprot WITH FILE: yee_C.seq
Run on: Mon Oct  3 16:32:30 MET 1994 on mendel.sis.pasteur.
.

 yee_C.seq :  287 aa
 >YEEC_ECOLI  Length: 347, 287 bases, 52DE1AA3 che: 287 aa
 vs Swiss-Prot Release 29 library
 searching Swiss-Prot library

 Distribution of initial scores with ktup=2.
    |
 v initn   init1
<  2      2     2:=
   4      0     0:
   6      4     4:==
   8     18    18:=========
  10     73    73:====================================
  12    326   326:==================================================
  14    370   370:==================================================
  16   1248  1248:==================================================
  18   1354  1354:==================================================
  20   2746  2746:==================================================
  22   3151  3151:==================================================
  24   6401  6401:==================================================
  26   5110  5110:==================================================
  28   5724  5763:==================================================
  30   3887  4303:==================================================
  32   2238  2682:==================================================
  34   1401  1735:==================================================
  36    863  1144:==================================================
  38    533   690:==================================================
  40    346   411:==================================================
  42    481   250:==================================================
  44    419   166:==================================================
  46    346   110:==================================================
  48    295    84:----------------------------------------++++++++
  50    203    47:-----------------------++++++++++++++++++++++++++
  52    184    46:-----------------------++++++++++++++++++++++++++
  54    110    39:-------------------++++++++++++++++++++++++++++++
  56     82     9:-----+++++++++++++++++++++++++++++++++++
  58     69     8:----++++++++++++++++++++++++++++++
  60     71     3:--++++++++++++++++++++++++++++++++
  62     75     1:-+++++++++++++++++++++++++++++++++
  64     36     1:-+++++++++++++++++
  66     31     0:++++++++++++++++
  68     17     2:-+++++++++
  70     12     0:++++++
  72     12     0:++++++
  74      6     0:+++
  76     10     1:-++++
  78     28     0:+++++++++++++++
  80      2     0:+
> 80     19     5:---+++++++
13464008 residues in 38303 sequences
 statistics exclude scores greater than 73
 mean initn score:  26.8 (7.79)
 mean init1 score:  26.0 (6.05)
 5349 scores better than 33 saved, ktup: 2, variable pamfact
 joining threshold: 28  scan time:  0:00:30

 The best scores are:                                      initn init1   opt
sp|P33013|YEEC_ECOLI HYPOTHETICAL 38.9 KD PROTEIN IN S 1422  1422  1422
sp|P04287|DACA_ECOLI PENICILLIN-BINDING PROTEIN 5 PREC  789   789   797
sp|P08506|DACC_ECOLI PENICILLIN-BINDING PROTEIN 6 PREC  738   738   766
sp|Q05523|DACA_BACST D-ALANYL-D-ALANINE CARBOXYPEPTIDA  176   116   129
.....................................................................
```

BLAST programme is perhaps the most widely used bioinformatics tool ever developed (Altschul et al., 1990). It is an alignment heuristic that determines local alignments between a query and a database. It uses a simplified version of Smith–Waterman algorithm that is, in turn, based on dynamic programming. Similar to FASTA

programme, BLAST consists of two components: a search/scoring algorithm and evaluation of statistical significance of the solution. BLAST starts with a query sequence and a database of sequences along with three given parameters supplied by the user: a word size '*k*' (usually $k = 3$ for amino acid sequence comparison), a similarity threshold *T* and a minimum cut-off score *S*. At the beginning, during a pre-processing operation, low complexity regions of the query sequence are removed and a *k*-letter word list of the query sequence is generated. The words in the list are scored using a scoring matrix such as BLOSUM62 and those words that meet the threshold *T* above the cut-off value *S* are retained in the list. Then, the programme compares the *k*-letter words of the databases sequence with the words in the list in a pairwise fashion and picks the pairs that are highly scored. The *k*-letter pairs are extended from both directions until the score cannot be enlarged. Then, all extensions that have scores equal or greater than *S* are saved. These are highly scoring pairs (HSPs). HSPs with the similarity scores greater than *T* are reported as best hits. Figure 4 shows a typical output of BLAST programme. The statistical significance of HSPs is determined through Expect value (*E*-value). *E*-value is a parameter that describes the number of hits one can expect to see by chance when searching a database of the same size. This means that the lower the *E*-value, or the close it to '0', the more significant the match is. The BLAST family comprises different programmes with different applications. BLASTN compares a nucleotide query sequence against a nucleotide sequence database. BLASTP performs protein/protein comparison. BLASTX compares a DNA query sequence with a protein database. PSI-BLAST is a more recent version of BLAST programme that relies on multiple alignment to save running time with more accuracy.

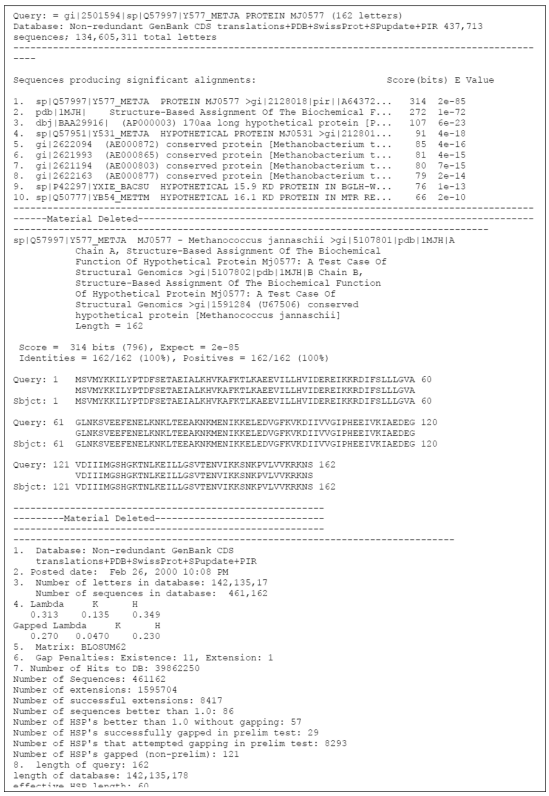
Another bioinformatics tool is microarray technique (Zhou et al., 2005) followed by normalisation and clustering. Microarray is a tool currently used to detect and measure gene expression at the mRNA or protein level, to find mutations, to re-sequence DNA, to locate chromosomal changes, and more. These investigations can be performed without microarrays, however, microarrays promise a high-throughput approach to the tasks. In each cell, only a fraction of genes are turned on and it is the subset that is expressed that confers unique properties of each cell type. The pattern of expression of genes in a cell provides insights into how the cell responds to its environmental changes. DNA microarray is a technology by which scientists can find out which gene is expressed and the level of expression can be measured with the aid of computer. There are many different types of microarrays (called platforms) in use, but all have a high density and number of biomolecules fixed onto a well-defined surface. In general, there are five basic steps of microarrays:

- coupling biomolecules to a platform
- preparing samples for detection
- hybridisation
- scanning
- analysing the data.

Once a platform for the test is determined, RNA sample should be prepared based on specified protocols. There are different sources of RNA including human tissues, model organisms, cell lines, etc. In hybridisation step, both cDNA and RNA sequences, i.e., the single-stranded immobilised DNA target sequence and the fluorescently labelled mobile

RNA probe sequence (usually an oligo sequence), are placed in a solution so that they lock together according to their complements. The cDNA target sequence plays the role of array in which each element represents a gene and RNA is binding to sites on the array corresponding to the genes expressed in each cell. When comparing two different genes, they are tagged with different colours to distinguish them during hybridisation step. Once this step is complete, the microarray slide is placed in a scanner (or reader) that consists of some lasers (special microscope) and a camera. The fluorescent tags are excited by the laser and then a digital image of the array is generated. The data on this image are stored in a computer for further analysis. The most important part of the tool is the special programme that analyses the spots. The intensity of colour on each spot is converted to numbers and the numbers tell us whether the gene of the probe RNA is expressed and the level of expression is quantified. Microarray experiment can be performed with samples taken from different environmental conditions and co-expression of genes under various conditions is sufficiently informative for biologists.

Figure 4 A typical output of BLAST program. The file has four sections. The header section includes general information on the query sequence and the reference database. The second section is a listing of best hits achieved via alignment. The first column is the characteristics of the sequence similar to the query. The next two columns provide the length of the sequence and the E-value of the alignment. The top sequence in the list has the lowest E-value, indicating the most similar sequence in the database to the query sequence. The next section demonstrates the alignment of the query sequence against best hits and the percent of similarity is reported (some alignment deleted due to limited space). The final section is a summary report of search details



```
Query: = gi|2501594|sp|Q57997|Y577_METJA PROTEIN MJ0577 (162 letters)
Database: Non-redundant GenBank CDS translations+PDB+SwissProt+SPupdate+PIR 437,713
sequences; 134,605,311 total letters
-----------------------------------------------------------------------------------------------
----

Sequences producing significant alignments:                          Score(bits) E Value

1.  sp|Q57997|Y577_METJA  PROTEIN MJ0577 >gi|2128018|pir||A64372...   314  2e-85
2.  pdb|1MJH|    Structure-Based Assignment Of The Biochemical F...   272  1e-72
3.  dbj|BAA29916|  (AP000003) 170aa long hypothetical protein [P...   107  6e-23
4.  sp|Q57951|Y531_METJA  HYPOTHETICAL PROTEIN MJ0531 >gi|212801...    91  4e-18
5.  gi|2622094  (AE000872) conserved protein [Methanobacterium t...    85  4e-16
6.  gi|2621993  (AE000865) conserved protein [Methanobacterium t...    81  4e-15
7.  gi|2621194  (AE000803) conserved protein [Methanobacterium t...    80  7e-15
8.  gi|2622163  (AE000877) conserved protein [Methanobacterium t...    79  2e-14
9.  sp|P42297|YXIE_BACSU  HYPOTHETICAL 15.9 KD PROTEIN IN BGLH-W...    76  1e-13
10. sp|Q50777|YB54_METTM  HYPOTHETICAL 16.1 KD PROTEIN IN MTR RE...    66  2e-10
-----------------------------------------------------------------------------------------------
------Material Deleted-------------------------------------------------------------------------
-----------------------------------------------------------------------------

sp|Q57997|Y577_METJA  MJ0577 - Methanococcus jannaschii >gi|5107801|pdb|1MJH|A
          Chain A, Structure-Based Assignment Of The Biochemical
          Function Of Hypothetical Protein Mj0577: A Test Case Of
          Structural Genomics >gi|5107802|pdb|1MJH|B Chain B,
          Structure-Based Assignment Of The Biochemical Function
          Of Hypothetical Protein Mj0577: A Test Case Of
          Structural Genomics >gi|1591284 (U67506) conserved
          hypothetical protein [Methanococcus jannaschii]
          Length = 162

 Score =  314 bits (796), Expect = 2e-85
 Identities = 162/162 (100%), Positives = 162/162 (100%)

Query: 1   MSVMYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIKKRDIFSLLLGVA 60
            MSVMYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIKKRDIFSLLLGVA
Sbjct: 1   MSVMYKKILYPTDFSETAEIALKHVKAFKTLKAEEVILLHVIDEREIKKRDIFSLLLGVA 60

Query: 61  GLNKSVEEFENELKNKLTEEAKNKMENIKKELEDVGFKVKDIIVVGIPHEEIVKIAEDEG 120
            GLNKSVEEFENELKNKLTEEAKNKMENIKKELEDVGFKVKDIIVVGIPHEEIVKIAEDEG
Sbjct: 61  GLNKSVEEFENELKNKLTEEAKNKMENIKKELEDVGFKVKDIIVVGIPHEEIVKIAEDEG 120

Query: 121 VDIIIMGSHGKTNLKEILLGSVTENVIKKSNKPVLVVKRKNS 162
            VDIIIMGSHGKTNLKEILLGSVTENVIKKSNKPVLVVKRKNS
Sbjct: 121 VDIIIMGSHGKTNLKEILLGSVTENVIKKSNKPVLVVKRKNS 162

-----------------------------------------------------
---------Material Deleted----------------------------
-----------------------------------------------------
-----------------------------------------------------------------------------------------

1.  Database: Non-redundant GenBank CDS
    translations+PDB+SwissProt+SPupdate+PIR
2. Posted date:  Feb 26, 2000 10:08 PM
3.  Number of letters in database: 142,135,17
    Number of sequences in database:  461,162
4. Lambda     K      H
   0.313    0.135    0.349
Gapped Lambda      K       H
   0.270    0.0470    0.230
5.  Matrix: BLOSUM62
6.  Gap Penalties: Existence: 11, Extension: 1
7. Number of Hits to DB: 39862250
Number of Sequences: 461162
Number of extensions: 1595704
Number of successful extensions: 8417
Number of sequences better than 1.0: 86
Number of HSP's better than 1.0 without gapping: 57
Number of HSP's successfully gapped in prelim test: 29
Number of HSP's that attempted gapping in prelim test: 8293
Number of HSP's gapped (non-prelim): 121
8.  length of query: 162
length of database: 142,135,178
effective HSP length: 60
```

Owing to variation among the intensity of colours in different slides, the raw expression information is normalised by means of various techniques (Smyth and Speed, 2003). The normalised data are then clustered using available algorithms (D'haeseleer, 2005) and co-expressed genes under same environmental conditions are specified.

4 Bioinformatics applications

4.1 Working with genomes

During past decades, advances in DNA sequencing techniques have enabled researchers to move from sequencing simple prokaryotes such as *Haemophilus influenza* with 1.8 million base pairs to sequencing more complicated higher organisms such as *Homo sapiens* with approximately 3 billion base pairs. This has required biologists to use appropriate bioinformatics tools to analyse and interpret the data. The first post-sequencing analysis is predicting genes and features associated with DNA, RNA and protein sequences. Coding and non-coding regions are identified using gene-finding techniques. Although gene finders are sensitive to species-specific parameters, gene-finding techniques are developed based on two approaches: evidence-based approach and *ab initio* approach (Thomas, 1999). Evidence-based approach relies on the experimental evidence on promoter regions of similar sequences. The corresponding DNA sequences of promoters related to different proteins are determined and stored in a database. The target DNA sequence is compared with the partial sequences in the database using a search tool such as BLAST. A high similarity to a known mRNA or protein is strong evidence that a region of the target DNA is a protein-coding gene. To collect evidence for most or all of the genes, thousands of different cell types must be studied. Furthermore, in complex organisms, all genes are not expressed in a given time and may not be accessible in a single culture. Thus, this approach is somehow difficult to predict genes. Despite these difficulties, extensive transcript and protein sequence databases are generated for human as well as other important model organisms. For example, RefSeq is a database that contains transcript and protein sequences from many different species. However, this database and other databases as such are incomplete and contain significant erroneous data. Because of inherent difficulty of obtaining evidence for many genes, *Ab initio* approach uses genomic DNA sequence alone in a systematic search for certain signs of protein-coding genes. These signs are broadly categorised as either 'signals' or 'content'. Signals are specific sequences that indicate the presence of a gene nearby. For instance, in prokaryotic genomes, promoter sequences are signals that can be easily identified by trained algorithms. Each promoter sequence is basically followed by a coding sequence. Content is related to statistical properties of protein-coding sequence itself. The statistics of stop codons are such that they are easily detectable in sequences of any length. Gene-finding techniques are reviewed in more detail elsewhere (Stormo, 2000).

Another type of analysis performed on genomes by means of bioinformatics tools is comparative analysis (Rubin et al., 2000). This analysis is specifically common when a new genome is completed. Scientists are interested in finding out how similar is the new genome to current genomes. The output of a comparative analysis includes statistics on number of genes, the classification of gene products and an overview of proteins involved in metabolism, regulation and signalling.

4.2 Characterising novel genes and proteins

After gaining the knowledge of sequencing and completing hundreds of genome projects from simple bacteria to complex organisms, the next step is obviously to locate all of the genes and regulatory regions, describe their functions and identify how they differ between groups. When a new gene is identified in a genome, the first attempt to characterise the gene is comparing it with known genes in multiple genomes. BLAST is the most appropriate tool to determine homology between the query gene and previously known genes. When two or more genes (sequences) are found homologous, one can predict that the genes are evolved from a common ancestor (Huynen et al., 2000). This relationship assists in the identification of conserved motifs and domains, structural properties and mutations over time. Transcribed mRNA and expressed protein from the new gene can also be specified and the cellular location of the protein would be predictable. Gene characterisation index has recently been developed for scoring the extent to which a gene is described (Kemmer et al., 2008). Using this bioinformatics tool, research scientists are able to compare levels of understanding of multiple genes.

4.3 Function inference

As the vast majority of genes in model organisms are now identified, the new challenge before life scientists is the determination of functions of the proteins expressed by the genes. Traditionally, intense and directed efforts are applied to decipher the function of novel individual proteins using laboratory techniques. Now with the generation of more and more genomic data, it is required to accelerate the functional characterisation of genes utilising high-throughput technologies for large numbers of genes. Computational approaches play an important role in the extraction of functional information from the raw genomic data. Early function prediction techniques rely on homology-based inference (Gerlt and Babbitt, 2000). When a protein is found homologous of a previously annotated protein, the corresponding annotation is extended to the new protein and the new function is assigned. BLAST is the main bioinformatics tool for searching homologies. Homology-based approaches are able to predict function for merely 30–40% of proteins (Eisenberg et al., 2000). Thus, non-homology-based approaches such as phylogenetic profiles (Pellegrini et al., 1999), protein fusion (Marcotte et al., 1999) and gene neighbourhood (Dandekar et al., 1998) were developed to identify those related proteins, not detectable by homology. Although these methods are not inferring functions upon homology, they use alignment techniques at some points to find conserved sequences.

Microarray experiment findings have also been used to infer function (van Noort et al., 2003). Expression patterns of proteins are suitable indications of proteins relatedness. When two proteins are co-expressed in several environmental occasions, this tells us that the two proteins have more likely functional relationships. Expression profiles derived from a sufficient number of experiments assist in clustering proteins in terms of their expression patterns.

4.4 Drug discovery

The design of new drugs and therapies is the future challenge of bioinformatics. Drugs are more effective with few side effects than conventional medicines. The primary requirement of designing such drugs is understanding life at the molecular level, i.e., understanding genes and gene products and their interactions throughout biological processes. Protein–protein interactions and protein functions are the two interrelated subjects that are developed using bioinformatics tools and resources. There are many computational methods available to predict protein–protein interactions (Yu and Fotouhi, 2006). These methods generate large data sets of interactions on genome scale. They use different sorts of information including primary structure, secondary structure, homology, and evolution and mutation events to predict the interactions.

Examination of protein–protein interaction networks demonstrates that proteins make permanent or transient complexes to accomplish a biological task such as drug delivery (Nooren and Thornton, 2003). Thus, for the design of a new drug, the involved protein complexes and related interactions must be carefully studied. Prediction of robust interactions among all genes in a genome is still to be completed. Statistical figures on interaction networks show that only a small portion of genes in genomes are found with their interactions and much more investigations needed to obtain a rather complete portray of interaction network.

4.5 Clinical diagnostics

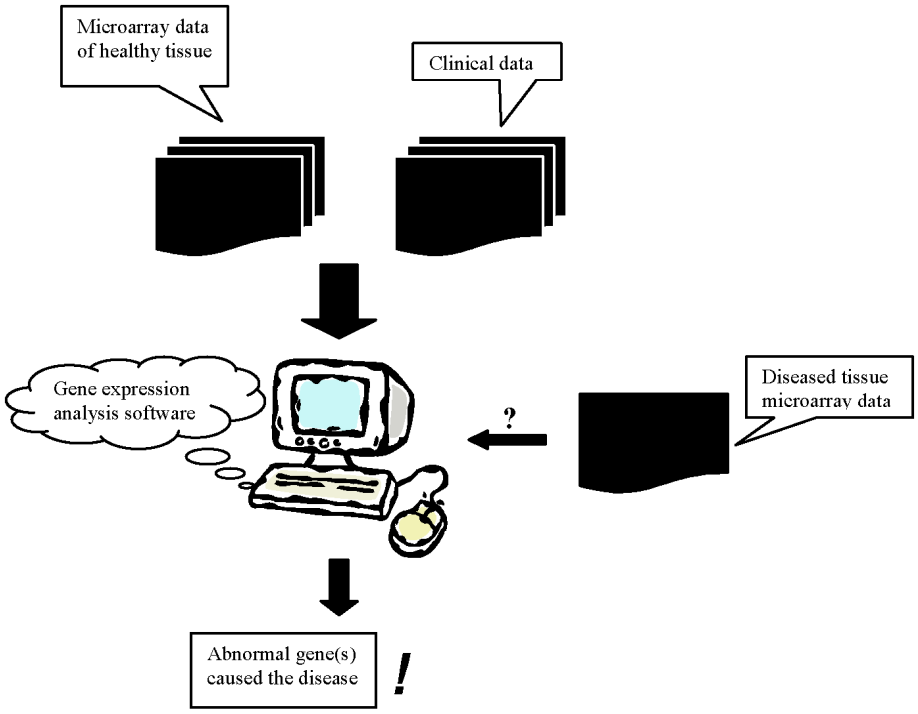
The completion of human genome in less than a decade ago has generated tremendous amount of biological data that is considered as the joining point of bioinformatics and medical research. The availability of this genome, and specifically moving from the acquisition of raw genomic data to finding biological functions and clinical importance of the discovered genes, is more likely eventually leading to qualitative change in the way in which clinical diagnostics is practised (Kohane, 2000). One of the central intellectual assets of this transition is GenBank and its associated genomic and protein databases. Upon the standardisation of bioinformatics and medical informatics resources, clinical records have been merged into GenBank and standardised models have allowed researchers and medical scientists throughout the world to rapidly access the latest information and submit their findings as well. As a result of these standard data models, bioinformatics databases have recently included information on diseased genes and their annotations. It is not surprising that the data model required to capture all this information is extremely complex, as diseased genes may have many different phenotypic disorders. MEDLINE is another database, maintained by National Library of Medicine (NLM), indexing medical literature under a controlled vocabulary called Medical Subject Headings (MeSH) (Darmoni et al., 2001). This system enables interested researchers to mining medical literature to search for particular topics. The database is now providing health information on over 650 diseases and conditions. It is now a major web resource for medical informatics studies.

One of the bioinformatics tools widely used in disease investigation on a molecular level is microarray. Basically, for medical diagnostics, differentially expressed genes, particularly disease-specific genes, are clue to understanding disease causes. The most straightforward microarray experiment is designed for comparing expressions in two different classes of samples from either a normal or an abnormal tissue. The expressions

are then looked for patterns that distinguish them. If the difference is significant (by means of statistical evaluations such as *t*-test), it can be concluded that the abnormal gene may cause the disease (Figure 5). For example, a subtype of cancer was detected by means of expression profiling (Alizadeh et al., 2000). However, the interpretation of expression data is not an easy task because the number of measurement points is much higher than the number of samples and the correlation scheme of the expression levels is unknown. As too many different attributes may contribute to the occurrence of a disease, a systematic approach is required to take into account the effect of different attributes. For example, clinically relevant groups can be determined by a single attribute, e.g., location of the disease, or by a combination of several attributes such as high blood pressure combined with high blood glucose. In these two cases, two distinct research questions may be posed for the same set of expression data. For this reason, a four-step workflow has been proposed for exploratory analysis of microarray data together with clinical data stored in relevant databases (Dugas et al., 2003). These steps consist of: definition of clinically meaningful questions in a masterfile, generation of analysis files, identification and characterisation of genes with differential expressions, and estimation of accuracy of classification. In the first step, a medical expert selects an attribute or attributes upon which patients are assigned to clinically relevant groups. This is the masterfile. Then, research questions are defined based on raw data gathered in the masterfile. The data are divided into a training set and a test set. Next, using various published methods, differentially expressed genes are identified and classified. In this step, up-regulated and down-regulated genes are specified and the patterns of expressions of the samples are searched for correlations. The correlations may be supported by annotations by means of searching tools. In the final step, the accuracy of estimation of classification should be tested using training a support vector machine and applying that to the test set. In another attempt (Hanai et al., 2006), using published microarray data, prognostic prediction software has been developed based on the classification analysis of the data. This analysis benefits from using a support vector machine trained for classification programme.

Bioinformatics and its resources assist in developing or improving Clinical Decision Support System (CDSS). Clinical reasoning and decision-making are phased (Sarbadhikari, 2004). Initially, there is a clinical evaluation (history taking and physical examination), followed by precise laboratory investigations. Then, integration of clinical findings and test results is done. After that, comparative benefits and risks are weighed among the alternative courses of actions, like drug interactions. Finally, the patient's preferences are taken into account, along with ethical and other considerations like cost of therapy, compliance expectations and a therapeutic plan is developed. Right from the first step (history taking) to the final one, bioinformatics resources and tools can be of immense help to the clinician. History taking could be iterative vs. comprehensive. Iterative estimation is often based on experience whereas comprehensive estimation is a more secure way of examination. Utilising bioinformatics resources by means of suitable tools can create a computer-assisted diagnostic system. There are scopes for ambiguities in descriptions, physical examinations, laboratory tests, various drug reactions and specific allergies, patient's non-compliance to the therapy and so on that makes the decision-making towards a particular treatment of a complex process. Appropriate computer algorithms by means of bioinformatics tools and databases may avoid these inconsistencies. The scope and the structure of this knowledge-based support system should be thought and discussed in the medical informatics research community.

Figure 5 Illustration of using microarray data to identify abnormal gene(s) casing a disease. As too many attributes may contribute to the occurrence of a disease, clinical data together with microarray data of a healthy tissue is required to the systematic identification of abnormal genes. Gene expression analysis software assist in the clustering and identification of expression variation (see online version for colours)



5 Conclusion

Production of huge biological information on the scale of genomes has required organising the data in databases in a more useful format and developing appropriate tools to gain knowledge-based discoveries. Bioinformatics is the science of semantic interpretation of the raw biological information by means of the tools. Following the organising resources and developing tools, bioinformatics has emerging application including gene finding, gene characterisation, function inference and drug discovery. With the availability of completed human genome, bioinformatics has gained new application in medical informatics that assists medical scientists to identify diseased genes using bioinformatics capabilities.

Acknowledgement

I thank Dr. Reza Gheshlaghi for his valuable comments on parts of this manuscript.

References

Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Ang, L., Marti, G.E., Moore, T., Hudson, J., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Levy, R., Wilson, W., Grever, M.R., Byrd, J.C., Botstein, D., Brown, P.O. and Staudt, L.M. (2000) 'Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling', *Nature*, Vol. 403, No. 6769, pp.503–512.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) 'Basic local alignment search tool', *J. Mol. Boil.*, Vol. 215, pp.403–410.

Brunak, S., Danchin, A., Hattori, M., Nakamura, H., Shinozaki, K., Matise, T. and Peruss, D. (2002) 'Nucleotide sequence database policies', *Science*, Vol. 298, p.1333.

Cattley, S. and Arthur, J.W. (2007) 'Biomanager: the use of a bioinformatics web application as a teaching tool in undergraduate bioinformatics training', *Briefings In Bioinformatics*, Vol. 8, No. 6, pp.457–465.

D'haeseleer, P. (2005) 'How does gene expression clustering work?', *Nature Biotechnology*, Vol. 23, pp.1499–1501.

Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998) 'Conservation of gene order: a fingerprint of proteins that physically interact', *Trends Biochem. Sci.*, Vol. 9, pp.324–328.

Darmoni, S.J., Thirion, B., Leroy, J.P., Douyere, M. and Piot, J. (2001) 'The use of Dublin core metadata in a structured health resource guide on the internet', *Bull. Med. Libr. Assoc.*, Vol. 89, No.3, pp.297–301.

Dugas, M., Merk, S., Breit, S., Schoch, C., Haferlach, T. and Kaab, S. (2003) 'Bioinformatics for medical diagnostics: assessment of microarray data in the context of clinical databases', *AMIA 2003 Symposium Proceedings*, pp.210–214.

Eisenberg, D., Marcotte, E.M., Xenarios, I. and Yeates, T.O. (2000) 'Protein function in the post genomic era', *Nature*, Vol. 405, pp.823–826.

Fell, D.A. (2001) 'Beyond genomics', *TRENDS in Genetics*, Vol. 17, No. 12, pp.680–682.

Fox, J.A., Butland, S.L., McMillan, S., Campbell, G. and Ouellette, B.F.F. (2005) 'The bioinformatics links directory: a compilation of molecular biology web servers', *Nucleic Acids Research*, *Web Server Issue*, Vol. 33, pp.W3–W24.

Galperin, M.Y. (2008) 'The molecular biology database collection: 2008 update', *Nucleic Acids Research*, Vol. 36 (database issue), pp.D2–D4.

Gerlt, J.A. and Babbitt, P. (2000) 'Can sequence determine function?', *Genome Biology*, Vol. 1, No. 5, pp.1–10.

Hanai, T., Hamada, H. and Okamoto, M. (2006) 'Application of Bioinformatics for DNA microarray data to bioscience, bioengineering and medical fields', *Journal of Bioscience and Bioengineering*, Vol. 101, No. 5, pp.377–384.

Hersh, W. (2002) 'Medical informatics – improving health care through information', *Journal of the American Medical Association*, Vol. 288, pp.1955–1958.

Huynen, M., Snel, B., Lathe, W. and Bork, P. (2000) 'Predicting protein function by genomic context: quantitative evaluation and qualitative inference', *Genome Research*, Vol. 10, pp.1204–1210.

International Human Genome Sequencing Consortium (IHGSC) (2004) 'Finishing the euchromatic sequence of the human genome', *Nature*, Vol. 431, pp.931–945.

Kemmer, D., Podowski, R.M., Yusuf, D., Brumm, J., Cheung, W., Wahlestedt, C., Lenhard, B. and Wasserman, W.W. (2008) 'Gene characterization index: assessing the depth of gene annotation', *PLosOne*, Vol. 3, No. 1, p.e1440.

Kohane, D. (2000) 'Bioinformatics and clinical informatics', *Journal of the American Medical Informatics Association*, Vol. 7, No. 5, pp.512–516.

Marcotte, E.M., Pellegrini, M., Ng, H-L., Rice, D.W., Yeates, T.O. and Eisenberg, D. (1999) 'Detecting protein function and protein-protein interactions from genome sequences', *Science*, Vol. 285, pp.751–753.

Nooren, I.M.A. and Thornton, J.M. (2003) 'Structural characterization and functional significance of transient protein-protein interactions', *J. Mol. Biol.*, Vol. 325, pp.991–1018.

Pearson, W.R. and Lipman, D.J. (1988) 'Improved tools for biological sequence comparison', *Proc. Natl. Acad. Sci., USA*, Vol. 85, pp.2444–2448.

Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (1999) 'Assigning protein functions by comparative genome analysis: protein phylogenetic profiles', *Proc. Natl. Acad. Sci., USA*, Vol. 96, pp.4285–4288.

Rubin, G.M, Yandell, M.D., Wortman, J.R., Miklos, G.L.G., Nelson, C.R., Hariharan, I.K., Fortini, M.E., Li, P.W., Apweiler, R., Fleischmann, W., Cherry, J.M., Henikoff, S., Skupski, M.P., Misra, S., Ashburner, M., Birney, E., Boguski, M.S., Brody, T., Brokstein, P., Celniker, S.E., Chervitz, S.A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R.F., Gelbart, W.M, George, R.A., Goldstein, L.S.B., Gong, F., Guan P., Harris, N.L., Hay, B.A., Hoskins, R.A., Li, J., Li, Z., Hynes, R.O., Jones, S.J.M., Kuehl, P.M., Lemaitre, B., Littleton, J.T., Morrison, D.K., Mungall, C., O'Farrell, P.H., Pickeral, O.K., Shue, C., Vosshall, L.B., Zhang, J., Zhao, Q., Zheng, X.H., Zhong, F., Zhong, W., Gibbs, R., Venter, J.C., Adams, M.D. and Lewis, S. (2000) 'Comparative genomics of the eukaryotes', *Science*, Vol. 287, pp.2204–2215.

Sarbadhikari, S.N. (2004) 'Basic medical science education must include medical informatics', *Indian Journal of Physiol Pharmacol*, Vol. 48, No. 4, pp.395–408.

Smyth, G.K. and Speed, T. (2003) 'Normalization of cDNA microarray data', *Methods*, Vol. 31, pp.265–273.

Stormo, G.D. (2000) 'Gene-finding approaches for eukaryotes', *Genome Research*, Vol. 10, pp.394–397.

Thomas, C.D. (1999) 'Design of gene characterization studies: an overview', *Journal of National Cancer Institute Monographs*, Vol. 26, pp.17–23.

van Noort, V., Snel, B. and Huynen, M.A. (2003) 'Predicting gene functions by conserved co-expression', *TRENDS in Genetics*, Vol. 19, pp.238–242.

Yu, J. and Fotouhi, F. (2006) 'Computational approaches for predicting protein-protein interactions: a survey', *Journal of Medical Systems*, Vol. 30, No. 1, pp.39–44.

Zhou, X.J., Cao, M.C., Huang, H., Wong, A., Nunez-Iglesias, J., Primig, M., Aparicio, O.M., Finch, C.E., Morgan, T.E. and Wong, W.H. (2005) 'Functional annotation and network reconstruction through cross-platform integration of microarray data', *Nature Biotechnology*, Vol. 23, No. 2, pp.238–243.