Current bioinformatics tools in genomic biomedical research (Review)

ANDREAS TEUFEL, MARKUS KRUPP, ARNDT WEINMANN and PETER R. GALLE

Department of Medicine I, Johannes Gutenberg University, Langenbeckstr. 1, D-55101 Mainz, Germany

Received December 2, 2005; Accepted February 8, 2006

Abstract. On the advent of a completely assembled human genome, modern biology and molecular medicine stepped into an era of increasingly rich sequence database information and high-throughput genomic analysis. However, as sequence entries in the major genomic databases currently rise exponentially, the gap between available, deposited sequence data and analysis by means of conventional molecular biology is rapidly widening, making new approaches of high-throughput genomic analysis necessary. At present, the only effective way to keep abreast of the dramatic increase in sequence and related information is to apply biocomputational approaches. Thus, over recent years, the field of bioinformatics has rapidly developed into an essential aid for genomic data analysis and powerful bioinformatics tools have been developed, many of them publicly available through the World Wide Web. In this review, we summarize and describe the basic bioinformatics tools for genomic research such as: genomic databases, genome browsers, tools for sequence alignment, single nucleotide polymorphism (SNP) databases, tools for ab initio gene prediction, expression databases, and algorithms for promoter prediction.

Contents

Correspondence to: Dr Andreas Teufel, Department of Medicine I, Johannes Gutenberg University, Building 301, Langenbeckstr. 1, D-55101 Mainz, Germany
E-mail: teufel@uni-mainz.de

Key words: bioinformatics, computational biology, transcription regulation, genomics, oncogenomics, Human Genome Project

1. Introduction

Officially initiated in 1990, it was not until the end of the last decade that the Human Genome Project began to effectively deliver sequence data to the research community. However, since then, and complimented by the commercial operations of Celera Genomics (1), these two major sequencing initiatives have generated a vast amount of genomic sequence data. The latest release of GenBank (Build 35) contained 47 million sequences with a current exponential increase of novel submissions (2). This enormous increase in available sequence information leads to a rapidly widening gap between the amount of raw sequence data and their analyses by means of molecular biology or other genetic approaches.

Thus, over recent years, an increasing need for high-throughput analysis methods has led to the development of sophisticated bioinformatics approaches. Such computational tools are currently the only way to rapidly and cost-effectively screen and analyze large quantities of sequence and gene expression information in order to close the gap between the generation of genomics data and their analysis by conventional biological approaches.

The field of bioinformatics is currently developing rapidly and multifaceted tools and approaches are being established for genomic biology and medicine applications. Many of these approaches have become excellent aids to answer detailed genome-related questions. In this review, we focus on widely-used bioinformatics tools that are readily accessible over the World Wide Web. Importantly, in relation to general use by biomedical investigators, these tools do not require extensive computational knowledge.

2. Genomic databases

Besides the large sequencing facilities, multiple individual laboratories across the world contributed sequence information to the public Human Genome Project. The generated sequence data are stored in large genomic repositories, of which the most commonly used are the database of the European Molecular Biology Laboratory (EMBL)/European Bioinformatics Institute (EBI) (3), the National Center for Biotechnology Information (NCBI, GenBank) database (4) and the DNA Database of Japan (DDBJ) (5). These three main repositories work in close collaboration, exchanging their sequence information on a daily basis. Furthermore, these organisations have agreed upon a common terminology, making sequence information

and files highly compatible between the individual databases. At present, each individual database contains between 46 (EBI) and 48 (DDJB) million sequence entries, consisting of 45-79 billion nucleotides. Given the primary focus of the Human Genome Project, i.e. the accurate sequencing of the human genome, and the general research bias towards health-related questions, most of the sequences currently deposited are human. Thus, approximately 30 million entries within GenBank are of human origin. However, more recently, an increasing number of sequences of other species, especially the biological model organisms of the mouse, rat, fish, fly, frog and worm have been deposited in these genomic databases. With the completion of the human genome, the main sequencing effort has now shifted towards these model organisms, the analysis of expressed sequence tags (ESTs) for gene expression/characterization studies and also towards an examination of human cancer genomes in an attempt to identify disease-linked mutations.

3. Genome browsers

Making these millions of sequences available to the entire biomedical community, through easily accessible and user-friendly systems has become essential. Hence, in recent years several web-accessible tools, so-called genome browsers, have been developed in order to provide such easy access. Currently, the most commonly used browsers are the EntrezGene browser (6), the UCSC genome browser (7) and the EBI/Ensembl browser (3). The focus of these genome browsers differs slightly; EntrezGene focuses more on individual sequences whereas the UCSC and EBI/Ensembl genome browsers have advantages when browsing large genomic contigs (of essentially up to chromosomal length) and also for comparative genomics when comparing data from different species. However, all three genome browsers provide essential genomic data, such as genomic sequence, exon structure, mRNA sequence, and EST or SNP data, through simple web-based text search interfaces. In addition, these databases may be installed locally, an interesting option for high-throughput analyses, which can be considerably more time-consuming if carried out via the World Wide Web.

4. Sequence alignment

Sequence comparison and alignment programs are essential bioinformatics tools. To date, the most widely utilized algorithms are the basic local alignment search tool (BLAST) (8) and its derivative, Gapped BLAST (9), as well as the FASTA (10,11) and ClustalW (12,13) algorithms. In contrast to earlier sequence alignment tools, such as the Smith *et al* (14) or Needleman-Wunsch (15) algorithms, BLAST uses a heuristic approach by first finding short contiguous matches. Each match is subsequently extended in order to yield higher scoring alignments, resulting in an optimal final alignment. This heuristic approach represents a balance between speed and sensitivity, which may be varied by changing a threshold value, T. Besides its advantage in speed, the algorithm and indeed the complete family of BLAST programs are easily accessible through the NCBI website (http://www.ncbi.nlm.nih.gov/BLAST). The sequence of interest may simply be pasted into a web form for further analysis. The database to be searched may be chosen; with nr (non-redundant), htgs (high-throughput genome sequence) and EST (expressed sequence tags) databases of several organisms being the most commonly used options for genomic research. In addition, the size of the short initial contigs to be searched ('word size') may be chosen to be 7, 11, or 15, changing the stringency of the search.

Looking at the BLAST output, the quality of the alignment may be estimated by the alignment score and the e-value. The BLAST alignment score is a measure of the extent of the local ungapped alignments and is partly dependent on the underlying scoring matrix; the higher the score, the better the alignment. The e-value measures the statistical significance threshold for reporting sequence matches against the individual genome database; for example, a default threshold value of 1E-5 means that, in 1E-5, matches would be expected to occur by chance (16).

In addition to scores and e-values of the individual alignment, BLAST also returns the individual alignments along with percentage rates of identity and similarity of nucleotides or amino acids. Finally, these individual alignments are then linked to other NCBI genomic databases, such as EntrezGene, Geo Profiles, or UniGene (4), providing easy access to genomic data for the matched sequence of interest.

In a similar way, FASTA (10,11) examines only identities that occur in a run of an adjustable number of consecutive matches. During further steps in the comparison of sequences, the regions with the most matches and smallest distance between the matches are further evaluated, allowing replacements and, thus, leading to a completed optimum alignment of sequences. Applying different scoring matrices for insertions, deletions and mismatches of the alignment, a measure of similarity is provided. Depending on the applied matrix, the stringency may be varied due to individual needs. FASTA programs may be installed locally through the Virginia Bioinformatics web server (fasta.bioch.virginia.edu) or can be conveniently accessed through a web interface at the server of the EBI (www.ebi.ac.uk).

In order to demonstrate or visualize conserved structural features of a number of related sequences, a multiple alignment tool is necessary. The most commonly used program for multiple sequence alignment is currently the ClustalW (12) algorithm. ClustalW was implemented as a combination of a phylogenetic and an heuristic approach. In an initial step, a phylogenetic tree is generated utilizing neighbour-joining methods. Thereafter, beginning with the two least distant sequences, all neighbouring sequences are subsequently aligned by a heuristic algorithm, finally leading to a complete alignment of all sequences. ClustalW is publicly available through the EBI or GeneBee (http://www.genebee.msu.su) web servers.

5. Mutation repositories and single nucleotide polymorphism (SNP) databases

Human genome and EST sequencing programs have generated large amounts of overlapping sequence data for both coding and non-coding regions. Such multiple sequence reads, complemented by specific programs of SNP discovery, have resulted in the submission of over 10,000,000 human SNPs

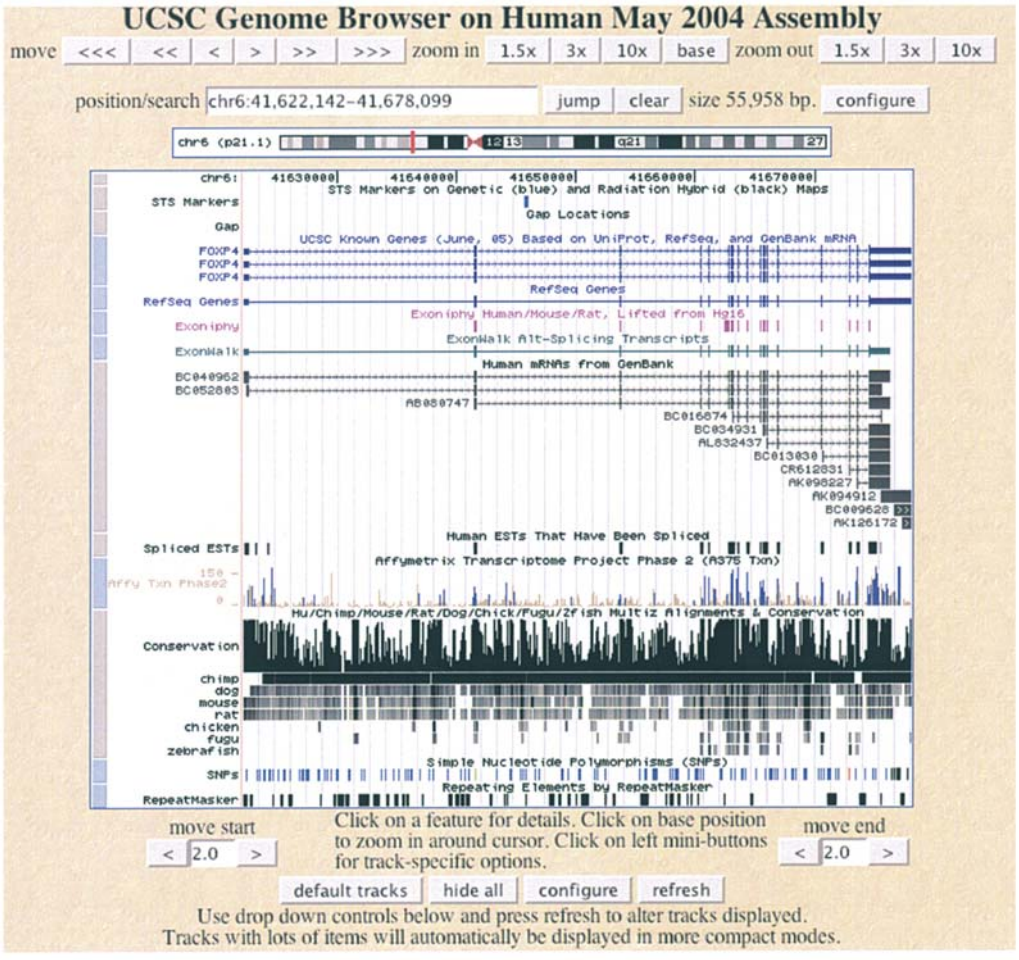

Figure 1. Screenshot of the UCSC genome browser displaying the human FOXP4 genomic region (37).

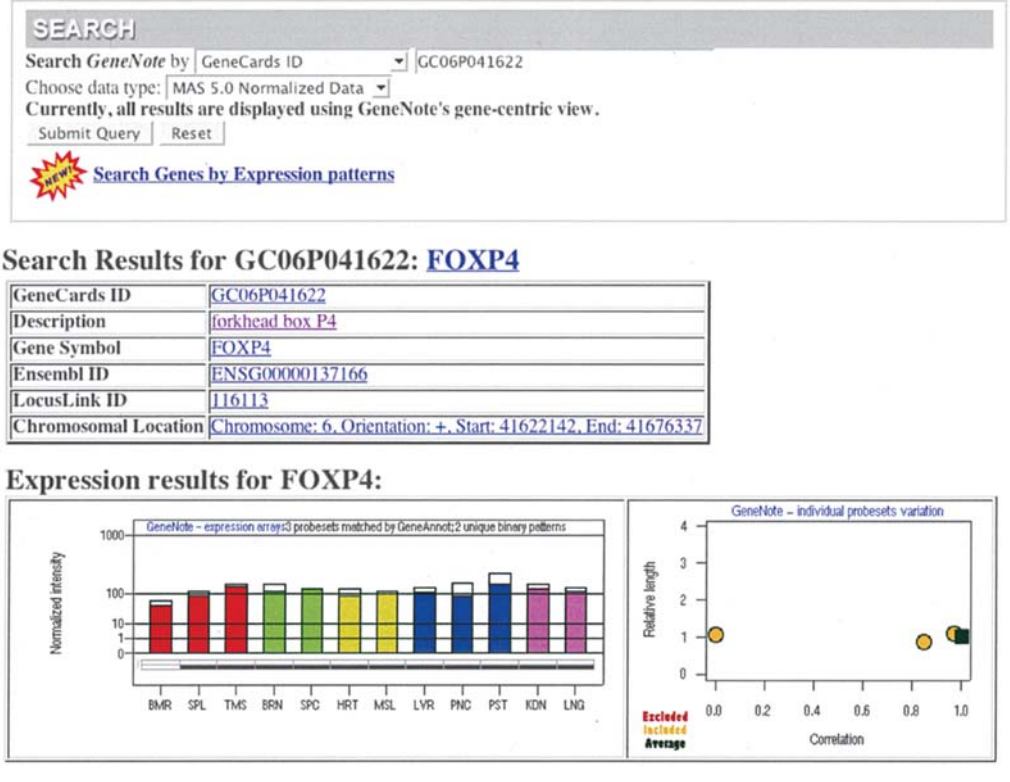

Figure 2. Screenshot of the GeneNote expression profile for human FOXP4 (37).

to the NCBI dbSNP database (17). Although they may be very useful in relation to genetic linkage studies, most of these genetic variations are likely to be functionally silent as they will be localized in intron sequences or else in coding sequence positions that do not lead to amino acid changes in the translated protein. However, a considerable number of these SNPs, especially those in amino acid coding positions or the regulatory regions of genes will have a biological and/or medical impact. For example, hemochromatosis or sickle cell anaemia are prominent examples of diseases caused by single nucleotide changes. Furthermore, the efficacy of several drugs has been demonstrated to be significantly modulated by single nucleotide changes. Thus, the identification of SNPs and investigations of their role in disease development and predisposition as well as their pharmacological impact is of high relevance.

One of the most comprehensive and commonly searched SNP repositories is the dbSNP database, which contains ~10 million human SNPs . This magnitude of identified polymorphisms obviously provides a rich source of largely unstudied genetic variations, all of which are potentially informative with respect to disease development or drug interaction. However, the scale of such variation may also represent something of a problem as the investigation of a large number of SNPs per gene may be difficult and time consuming. For example, searching for SNPs in the human TP53 suppressor gene results in 1436 entries in dbSNP, a number far to big for conventional manual analysis aimed at an evaluation of their possible clinical relevance. Increasingly therefore, rational large-scale population-based technology-driven solutions to such issues will be required. However, excellent links within dbSNP to other NCBI databases, especially the sequence databases, makes the transfer from the identification of an interesting variant to a more detailed positional examination and molecular verification convenient.

A further source of validated and reported mutations is the Online Mendelian Inheritance in Man (OMIM) database (18). Entering a gene of interest, the OMIM database report page, as well as directly detailing SNPs and mutations in some cases, allows the user to quickly link to the relevant dbSNP entry for the gene of interest via a drop-down menu. Gene-specific reports provided by OMIM are of excellent quality and provide direct links to key Pubmed citations. However, perhaps one slight disadvantage of OMIM is that hyperlinks to the corresponding sequence information of other SNP databases are not provided.

Furthermore, mutation databases such as the Human Gene Mutation Database (HGMD) (19), the Japanese Single Nucleotide Polymorphism database (JSNP) (20) and HGVBase (21) may be helpful for individual questions but do not display major advantages over NCBI dbSNP.

6. Ab initio gene prediction

Although sequence alignment is very useful, in the context of the localization of a human gene or EST to a particular region of the genome (for example through the use of the BLAT search program in the UCSC genome browser) (7), such approaches contribute little to the task of gene discovery from non-annotated regions of genomic sequence, where no EST matches are present. Such regions may often be identified from genetic linkage or similar studies. Consequently, they are often large, containing several tens or hundreds of kilobases of sequence, the bulk of which will usually be non-coding. It is therefore essential for further analysis to identify the coding regions of any novel genes within such genomic regions.

The latest generation of gene prediction programs such as GENESCAN (22) or AUGUSTUS (23) provide an acceptable level of accuracy coupled with ready usability, and are therefore of significant use in the analysis of uncharacterized genomic sequences. These programs essentially rely on a statistical model, the Hidden Markov Model (24), for predicting either an intronic or exonic state of the sequence in a given region. Thus, the algorithm does not rely directly on sequence homology to known genes. However, the statistical models that underpin the operation of the software have been trained on a set of training data relating to known genes and characteristic structure. Therefore, parts of the genome that are of significantly different structure compared to these training data may be only poorly annotated. However, exon level specificity was reported to be as high as 80% for such approaches and, thus, for most genomic sequences, these algorithms provide reasonable accuracy for an initial scan of exons within a given region.

GENESCAN and AUGUSTUS may be accessed via the World Wide Web. Both take large genomic fragments via a web interface and return the predicted exon-intron structure as well as predicted promoters for both strands as a graphical output and a table detailing the exact start and end nucleotide positions of any exon, intron, and promoter within the given sequence.

7. Expression profiling

The determination of the expression level of a gene in one tissue over another may be critically important in the context of evaluating its importance in normal function or disease processes and in highlighting whether a particular gene product might have potential as a future drug target. As the expression level of a gene relies on many complex factors, including specific transcription factor levels, epigenetic modification, genetic variation and somatic or constitutional mutation, there are no efficient algorithms for ab initio gene expression profiling and expression data are therefore derived from observational studies. Whilst it has been possible to readily and accurately examine the expression of small numbers of genes for many years, the rapid increase in the amount of genomics data, with 25-30,000 human genes now described, has necessitated the development of efficient ways to investigate genome-wide expression, such as high-throughput EST sequencing, serial analysis of gene expression (SAGE), and gene expression microarray analysis. Data so derived are now generally stored in large, web-accessible database repositories.

SAGE tags are short sequence tags derived from the 3' end of mRNA molecules. The advantage of this approach lies in the small size of the tags. After an intermediate cloning step, 10-20 of these tags are sequenced with one single sequencing

reaction, significantly saving time and money. These tags in many cases allow an allocation to individual genes through database sequence alignment and thus allow high-throughput expression analysis by analysing the number of sequenced tags per gene and tissue. Large SAGE and EST sequencing efforts are currently underway and data from multiple experiments may be accessed through open-source web interfaces. Two of the most commonly used are the Stanford SOURCE web tool (25) and the NCI CGAP (4) sites. The advantage of the SOURCE site lies in its convenient web interface allowing the user to search SAGE and EST expression profiles by gene names or accession numbers. The resulting predicted expression profile can also be viewed as a graphical output, sorting the individual tissues by the level of a gene's prediction/normalized number of sequenced ESTs per tissue. The NCI CGAP suite is more sequence oriented, giving the user the opportunity of accessing the individual EST sequences and additional background information, e.g. the RNA library and vectors used in the EST profiling experiment that the tag was derived from. However, although these EST-based expression profiling tools may complement future expression profiling strategies, they suffer from a number of practical drawbacks and so have been largely superseded for global analyses by gene expression micorarrays.

Each gene expression microarray may contain tens of thousands of specific elements, each representing an individual gene or splice variant. Such arrays can simultaneously measure absolute expression for each element in a single sample (or relative expression in paired samples). Throughout the past decade, a vast number of microarray studies have been published. Mostly analyzed from only one or a small number of perspectives, these data contain a vast amount of untapped expression information. Lately, several databases have been established in order to make the corresponding sets of raw data from these experiments openly accessible. Most widely used are the Stanford Microarray Database (26), the EBI ArrayExpress (27,28), and the NCBI Gene express omnibus (4). Thus, these data may now be downloaded and examined in different and especially new ways and perspectives.

In order to make expression data from normal tissues available, the GeneNote (29) group at the Weizman Institute of Science established 12 microarray data sets from normal tissues and made the data freely available via the internet. Users may search these data by entering a gene name and retrieving the normalized levels of expression in graphical and numerical form. Similarly, tissue-specific and disease-specific gene expression microarray information is now associated with many of the annotated gene descriptions in the UCSC genome browser (7). To reach such data, following a sequence-based alignment search, the user need simply click on the gene map in the browser window. Despite the documented problem of variations of expression in multiple microarray experiments, these data must currently be regarded to be more reliable than the SAGE and EST sequence-based estimations of gene expression.

8. Promoter prediction

Essential to the regulation of a gene's expression level is the corresponding promoter region. The prediction of a promoter remains problematic since the sequence may not correspond closely to a standard consensus sequence. Adding to this problem, unless the gene is very well characterised, the position of the promoter may not be obvious, as the transcription initiation point, which represents the extreme 5' end of the mRNA, may not be known with accuracy. Coupled with the possibility of splicing within the 5' leader at any position beyond the initiation point, this means that the promoter may be some considerable distance, perhaps many kilobases, upstream of the actual coding sequence of the gene. Furthermore, promoters of mammalian genes do not share many regions of sequence similarity, making the recognition of promoters even more difficult. To solve this structural problem, most bioinformatic (and molecular biological) searches focus on the 'core region' of the promoter.

Some of the main structural features of interest are the TATA box [a consensus of TATA(T/A)A(T/A)] and the potential binding sites for transcription factors. The TATA box is recognized by the TATA binding protein (TBP), which is a part of the TFIID transcription initiation factor. Thus, the consensus sequence of the TATA box or the position weight matrix is often used to recognize the location of the promoter region. In addition, around the transcription start site, a loosely conserved so called initiator region has been reported, which may be bound by several proteins. These proteins may be capable of initiating transcription even in absence of TBP. Specific transcription factor binding sites are thought to be typically between 5 and 15 bp long and the presence of such sites in the vicinity of a putative coding sequence may signal the presence of a promoter. Thus, the potential promoter sequences may be searched either by consensus sequences or by positional weight matrices. However, significant improvements in promoter prediction have only been made within the last few years. PromoterScan (30) has been viewed as one of the first promoter prediction algorithms with acceptably high specificity. Recently, PromoterInspector (31) and Dragon Promoter Finder (32) made further progress in specificity and sensitivity of promoter prediction algorithms.

PromoterScan (30) identifies promoters using a TATA box positional weight matrix combined with the density of specific transcription factor binding sites. The algorithm has been demonstrated to be of comparatively high specificity but low sensitivity.

9. Future prospects; data integration and gene ontology

With the availability of large quantities of biological information for many individual genes, from multiple bioinformatics sources, efficiently integrating these data into the context of specific analyses has become essential to enhance the speed and quality of genomic research. With the increasing use of high-throughput methods, the field of biology is increasingly faced with the problem of storing, indexing and retrieving vast amounts of data (of often related but different forms) from a range of sources. Furthermore, the nature of the data to be integrated is becoming increasingly diverse; including genomic, mRNA and protein sequences, protein structures and modifications, protein function, biomolecular interactions, gene expression, alternative splicing, epigenetic modification, DNA polymorphisms, taxonomic

data, molecular pathways, genetic networks, bibliographic data, and evolution.

As such data derive from different and often independent fields of biological research, these areas generally have their own terminology and data requirements. Much of the information that the biological researcher is interested in is available in public reference databases and in the millions of articles of scientific research literature, mostly accessible on the World Wide Web. It is estimated that 80% of biological data are in text form, and even the abstracts are written in free text utilizing a complex biological vocabulary, which may vary significantly in different areas of research. Thus, despite their wide availability, these data are not generally machine-readable. Consequently, no promising, significant increase in the efficiency of data integration may be expected from automation of Pubmed data retrieval (33,34).

In order to make these data machine accessible and to integrate different data sources, there is an obvious need for a standard vocabulary and to translate the available data into clear defined standard data sets and terms. At present, several approaches to meet these obvious needs are underway and the most commonly used annotation standards currently rely essentially on ontologies. These ontologies provide conceptualizations of domains of knowledge and facilitate both communication between researchers and the use of domain knowledge by computers for multiple purposes (35).

The functional data annotated for individual genes are currently mostly dependent on the available gene ontology annotations of the Gene Ontology (GO) Consortium (36). The Gene Ontology project uses standardized GO terms, which describe three major aspects of a gene's biological information: its biological processes, molecular function and cellular localization. Although GO annotations may develop into a powerful tool in the future, they are currently often limited and incomplete as they are dependent on database entries by individual investigators. This is particularly problematic, as the translation of biological data into GO terms is highly dependent on a researchers view and scientific background on a given subject. In many cases, there is no perfectly fitting GO term available to describe the biological data, making a 'best fit' annotation necessary. Even more concerning is the very limited quality control of the data entries. In addition, just as the initial GO assignments may initially be time consuming and difficult to make, the problem of keeping results and annotations up to date adds another layer of complexity to the problem.

However, the advantages of such annotations are obvious. Once a gene's GO characteristics are recorded, the ontology is thereby optimized for computational high-throughput analysis, allowing a highly time-saving comparison of a large amount of functional data. Thus, despite any disadvantages, GO annotations may develop into a powerful tool in the future, especially with a rapidly increasing amount of available datasets.

In summary, many sophisticated, extremely valuable, easily-accessible and user-friendly tools have been developed for bioinformatics analysis in genomic biomedical research. However, the integration of these bioinformatics tools and data is the current challenge and, if successfully met, will significantly accelerate genomic biomedical research.

References

1. Private vs public genomics? Nature 403: 117.
2. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J and Wheeler DL: GenBank: update. Nucleic Acids Res 32: D23-D26, 2004.
3. Hubbard T, Andrews D, Caccamo M, Cameron G, Chen Y, Clamp M, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Gilbert J, Hammond M, Herrero J, Hotz H, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Kokocinsci F, London D, Longden I, McVicker G, Melsopp C, Meidl P, Potter S, Proctor G, Rae M, Rios D, Schuster M, Searle S, Severin J, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Trevanion S, Ureta-Vidal A, Vogel J, White S, Woodwark C and Birney E: Ensembl 2005. Nucleic Acids Res 33: D447-D453, 2005.
4. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Church DM, Di Cuccio M, Edgar R, Federhen S, Helmberg W, Kenton DL, Khovayko O, Lipman DJ, Madden TL, Maglott DR, Ostell J, Pontius JU, Pruitt KD, Schuler GD, Schriml LM, Sequeira E, Sherry ST, Sirotkin K, Starchenko G, Suzek TO, Tatusov R, Tatusova TA, Wagner L and Yaschenko E: Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 33: D39-D45, 2005.
5. Tateno Y, Imanishi T, Miyazaki S, Fukami-Kobayashi K, Saitou N, Sugawara H and Gojobori T: DNA Data Bank of Japan (DDBJ) for genome scale research in life science. Nucleic Acids Res 30: 27-30, 2002.
6. Maglott D, Ostell J, Pruitt KD and Tatusova T: Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 33: D54-D58, 2005.
7. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM and Haussler D: The human genome browser at UCSC. Genome Res 12: 996-1006, 2002.
8. Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ: Basic local alignment search tool. J Mol Biol 215: 403-410, 1990.
9. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402, 1997.
10. Pearson WR: Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 132: 185-219, 2000.
11. Pearson WR and Lipman DJ: Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85: 2444-2448, 1988.
12. Thompson JD, Higgins DG and Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680, 1994.
13. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG and Thompson JD: Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 31: 3497-3500, 2003.
14. Smith TF, Waterman MS and Fitch WM: Comparative biosequence metrics. J Mol Evol 18: 38-46, 1981.
15. Needleman SB and Wunsch CD: A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48: 443-453, 1970.
16. Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L and Altschul SF: IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. Bioinformatics 15: 1000-1011, 1999.
17. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM and Sirotkin K: dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29: 308-311, 2001.
18. Hamosh A, Scott AF, Amberger J, Valle D and McKusick VA: Online Mendelian Inheritance in Man (OMIM). Hum Mutat 15: 57-61, 2000.
19. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M and Cooper DN: Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21: 577-581, 2003.
20. Hirakawa M, Tanaka T, Hashimoto Y, Kuroda M, Takagi T and Nakamura Y: JSNP: a database of common gene variations in the Japanese population. Nucleic Acids Res 30: 158-162, 2002.

21. Fredman D, Munns G, Rios D, Sjoholm F, Siegfried M, Lenhard B, Lehvaslaiho H and Brookes AJ: HGVbase: a curated resource describing human DNA variation and phenotype relationships. Nucleic Acids Res 32: D516-D519, 2004.
22. Burge C and Karlin S: Prediction of complete gene structures in human genomic DNA. J Mol Biol 268: 78-94, 1997.
23. Stanke M, Steinkamp R, Waack S and Morgenstern B: AUGUSTUS: a web server for gene finding in eukaryotes. Nucleic Acids Res 32: W309-W312, 2004.
24. Mukherjee S and Mitra S: Hidden Markov Models, grammars and biology: a tutorial. J Bioinform Comput Biol 3: 491-526, 2005.
25. Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, Rees CA, Cherry JM, Botstein D, Brown PO and Alizadeh AA: SOURCE: a unified genomic resource of functional annotations, ontologies and gene expression data. Nucleic Acids Res 31: 219-223, 2003.
26. Sherlock G, Hernandez-Boussard T, Kasarskis A, Binkley G, Matese JC, Dwight SS, Kaloper M, Weng S, Jin H, Ball CA, Eisen MB, Spellman PT, Brown PO, Botstein D and Cherry JM: The Stanford Microarray Database. Nucleic Acids Res 29: 152-155, 2001.
27. Brazma A, Parkinson H, Sarkans U, Shojatalab M, Vilo J, Abeygunawardena N, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A, Rocca-Serra P and Sansone SA: ArrayExpress - a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 31: 68-71, 2003.
28. Rocca-Serra P, Brazma A, Parkinson H, Sarkans U, Shojatalab M, Contrino S, Vilo J, Abeygunawardena N, Mukherjee G, Holloway E, Kapushesky M, Kemmeren P, Lara GG, Oezcimen A and Sansone SA: ArrayExpress: a public database of gene expression data at EBI. C R Biol 326: 1075-1078, 2003.
29. Shmueli O, Horn-Saban S, Chalifa-Caspi V, Shmoish M, Ophir R, Benjamin-Rodrig H, Safran M, Domany E and Lancet D: GeneNote: whole genome expression profiles in normal human tissues. C R Biol 326: 1067-1072, 2003.
30. Prestridge DS: Predicting Pol II promoter sequences using transcription factor binding sites. J Mol Biol 249: 923-932, 1995.
31. Scherf M, Klingenhoff A and Werner T: Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. J Mol Biol 297: 599-606, 2000.
32. Bajic VB, Seah SH, Chong A, Zhang G, Koh JL and Brusic V: Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters. Bioinformatics 18: 198-199, 2002.
33. Bittner M, Meltzer P and Trent J: Data analysis and integration: of steps and arrows. Nat Genet 22: 213-215, 1999.
34. Lacroix Z: Biological data integration: wrapping data and tools. IEEE Trans Inf Technol Biomed 6: 123-128, 2002.
35. Schulze-Kremer S: Ontologies for molecular biology and bioinformatics. In Silico Biol 2: 179-193, 2002.
36. Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C, Richter J, Rubin GM, Blake JA, Bult C, Dolan M, Drabkin H, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Cherry JM, Christie KR, Costanzo MC, Dwight SS, Engel S, Fisk DG, Hirschman JE, Hong EL, Nash RS, Sethuraman A, Theesfeld CL, Botstein D, Dolinski K, Feierbach B, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Lee V, Chisholm R, Gaudet P, Kibbe W, Kishore R, Schwarz EM, Sternberg P, Gwinn M, Hannick L, Wortman J, Berriman M, Wood V, De la Cruz N, Tonellato P, Jaiswal P, Seigfried T and White R: The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32: D258-D261, 2004.
37. Teufel A, Wong EA, Mukhopadhyay M, Malik N and Westphal H: FoxP4, a novel forkhead transcription factor. Biochim Biophys Acta 1627: 147-152, 2003.