Journal of Computational Intelligence in Bioinformatics
ISSN: 0973-385X Volume 4 Number 1 (2011) pp. 7–17

http://www.ripublication.com/jcib.htm

Bioinformatics New Era: Introduction and Overview

Nabeel Ahmad[1], Akhilesh Bind[2] and Sanjiv Kumar Maheshwari[1]*

[1]Plant Molecular Biology Lab,
Department of Biotechnology,
College of Engineering and Technology,
IFTM Campus, Delhi Road, Moradabad (UP) India 244001,
[2]AIIDU Allahabad, India
**Corresponding author Email: sanjiv08@gmail.com*

Abstract

A flood of data means that many of the challenges in biology are now in computing. Bioinformatics, the application of computational techniques to analyze the information associated with bio-molecules on a large-scale, has now firmly established itself as a discipline in molecular biology, and encompasses a wide range of subject areas from structural biology, genomics to gene expression studies. In this review we provide an introduction and overview of the current state of the field. We discuss the main principles that underpin bioinformatics analyses, look at the types of biological information and databases that are commonly used, and finally examine some of the studies that are being conducted, particularly with reference to transcription regulatory systems.

Introduction

Biological data are being produced at a phenomenal rate [1]. For example as of January 2010, the GenBank repository of nucleic acid sequences contain 9,252,000 entries [2] and the Swiss-Prot database of protein sequences contained 99,163 [3]. On an average these databases are doubling in a year time [2]. After the completion of H. influenza genome [4], complete genome sequence of over 50 organisms have been released (ranging from 450 to 100,000 genes). In addition to this, the data from many other projects that study the gene expression, protein structures and their interactions are also being added. We can imagine the amount of information in terms of quantity and variety is being added. As a result of this surge in data, computers have become inevitable constituents of biological research. This approach is required to handle such a large databases. Bioinformatics is often defined as the application of computational

techniques to understand and to organize the information associated with biological macromolecules, also known as biomolecucles. This unexpected union between two subjects is largely attributed to the fact that life itself is an information technology: organisms physiology is largely determined by its genes, which is the most basic can be viewed as the digital information. At the same time there have been major advances in the technologies that supply the initial data. The development in the processing and maintenance of the data is the another area which is supporting the development of Bio-informatics. This incredible processing power has been matched by developments in computer technology; the most important areas of improvements have been in CPU, disk storage and internet for rapid computation and better data storage and handling of easy access and exchange of the data.

Objectives of Bioinformatics

The objectives of Bioinformatics are three folds. First at its simplest bioinformatics organizes data in a way that allows the researcher to access existing information and to submit the new entries as they are generated, e. g. Protein data bank for 3D macromolecular structure [6, 7]. Data curation is another essential task; the information stored in these databases is essentially useless until it is analyzed. Thus the purpose of Bioinformatics extends much farther then creation of data to develop tools and resources that aid in the analysis of the data. For example after sequencing a particular protein, it is of interest to compare with previously characterized sequences. This needs more than just a simple text based search and programs such as FASTA [8] and PSI-FASTA [9] must consider what comprises a biologically significant match. Development of such resources dictates expertise in computational theory as well as a through understanding of Biology. The third aim is to use the tools to analyze the data and interpret the results in a biologically meaningful manner. In Bioinformatics, we can now conduct global analyses of all the data available to find out the common principles that apply across many systems and highlight novel features.

This review intends to provide the introduction to Bioinformatics. We are focusing on the First and third aims of the bioinformatics just described above. The types of data are available for analysis by bioinformatics and the range of topics that we can cover fall within the field (Table-1). We take a broad view and to include subjects that may not normally be listed.

We are presenting an overview of the sources of information: First raw DNA sequences, protein sequences, macromolecular structures, genome sequences and other whole genome data. Raw DNA sequences are strings of the four base letters comprising of genes. Normally each gene is composed of 1000 to 1500 bases long. The GenBank repository of nucleic acid sequences currently holds a total of 9.5 billion bases in 8.8 million entries (all databases). At the next level are protein sequences comprising of 20 amino acids- letters. At present there are about 300,000 known protein sequences and the list is ever expanding, typically bacterial protein consists of 300 amino acids. Macromolecular structural data represents a more complex form of information. There are currently 13,000 entries in the Protein Data

Bank (PDB). A typical PDB file for a medium sized protein contains the xyz coordinates of approximately 2,000 atoms.

Table 1: Sources of data used in bioinformatics, the quantity of each type of data that is currently available, and bioinformatics subject areas that utilise this data.

Data source	Data size	Bioinformatics Topics
Raw DNA sequence	8.2 million sequences (9.5 billion bases)	Separating coding and non-coding regions Identification of introns and exons Gene product prediction Forensic analysis
Protein sequence	300,000 sequences (~300 amino acids each)	Sequence comparison algorithms Multiple sequence alignments algorithms Identification of conserved sequence motifs
Macromolecular structure	13,000 structures (~1,000 atomic coordinates each)	Secondary, tertiary structure prediction 3D structural alignment algorithms Protein geometry measurements Surface and volume shape calculations Intermolecular interactions Molecular simulations (force-field calculations). molecular movements, docking predictions)
Genomes	40 complete genomes (1.6 million – 3 billion bases each)	Characterisation of repeats Structural assignments to genes Phylogenetic analysis Genomic-scale censuses (characterisation of protein content, metabolic pathways)
Gene expression	largest: ~20 time point measurements for ~6,000 genes	Linkage analysis relating specific genes to diseases Correlating expression patterns Mapping expression data to sequence, structural and biochemical data
Other data		
Literature	11 million citations	Digital libraries for automated bibliographical searches
Metabolic pathways		Knowledge databases of data from literature Pathway simulations

As with the raw DNA sequences, genomes consist of strings of base- letters, ranging from 1.6 million bases in *Haemophilus influenzae* to 3 billion in humans. An important aspect of complete genomes is the distinction between coding regions and non- coding regions –'junk' repetitive sequences making up the bulk of base sequences especially in eukaryotes. We can now measure expression levels of almost every gene in a given cell on a whole-genome level although public availability of such data is still limited. Expression level measurements are made under different environmental conditions; different stages of the cell cycle and different cell types in

multi cellular organisms. Currently the largest dataset for yeast has made approximately 20 time-point measurements for 6,000 genes [10]. Other genomic-scale data include biochemical information on metabolic pathways, regulatory networks, protein-protein interaction data from two-hybrid experiments, and systematic knockouts of individual genes to test the viability of an organism.

What is apparent from this list is the diversity in the size and complexity of different datasets. There are invariably more sequence-based data than structural data because of the relative ease with which they can be produced. This is partly related to the greater complexity and information-content of individual structures compared to individual sequences. While more biological information can be derived from a single structure than a protein sequence, the lack of depth in the latter is remedied by analyzing larger quantities of data

Redundancy and multiplicity of data

A concept that underpins most research methods in bioinformatics is that much of this data can be grouped together based on biologically meaningful similarities. For example, sequence segments are often repeated at different positions of genomic DNA [11]. Genes can be clustered into those with particular functions (e. g, enzymatic actions) or according to the metabolic pathway to which they belong [12], although here, single genes may actually possess several functions [13]. Going further, distinct proteins frequently have comparable sequences – organisms often have multiple copies of a particular gene through duplication while different species have equivalent or similar proteins that were inherited when they diverged from each other in evolution. At a structural level, we predict there to be a finite number of different tertiary structures – estimates range between 1,000 and 10,000 folds [14, 15] – and proteins adopt equivalent structures even when they differ greatly in sequence [16]. As a result, although the number of structures in the PDB has increased exponentially, the rate of discovery of novel folds has actually decreased.

There are common terms to describe the relationship between pairs of proteins or the genes from which they are derived: analogous proteins have related folds, but unrelated sequences, while homologous proteins are both sequentially and structurally similar.

The two categories can sometimes be difficult to distinguish especially if the relationship between the two proteins is remote [17, 18]. Among homologues, it is useful to distinguish between orthologues, proteins in different species that have evolved from common ancestral gene and prologues, proteins that are related by gene duplication with in a genome [19]. Normally orthologues retain the same function while prologues evolve distinct but related functions [20].

An important concept that arises from these observations is that of a finite part lists for different organisms [21, 22]: an inventory of proteins contained within an organism arranged according to different properties such as gene sequence, protein fold or function. Taking protein fold as an example, we mentioned that with a few exceptions, the tertiary structures of proteins adopt one of a limited repertoire of folds. As the number of different fold families is considerably smaller than the number of

gene families, categorizing the proteins by fold provides a substantial simplification of the content of a genome. Similar simplification can be provided by other attributes such as protein function. As such we except this notion of finite parts lists become increasingly common in the future for genome analysis.

Clearly, ancestral aspect of managing this large volume of data lies in developing methods for assessing similarities between different biomoleucles and identifying those that are related. We are providing the list of primary sources data providers and also introducing the some secondary databases that systematically group the data (Table-2). These classifications ease comparisons between genomes and their products, allowing the identification of common themes between those that are related and highlighting features that are unique to some.

Table 2: List of different URLs which are important for Bioinformatics work.

Database	URL
Protein sequence (primary)	
SWISS-PROT	www.expasy.ch/sprot/sprot-top.html
PIR-International	www.mips.biochem.mpg.de/proj/protseqdb
Protein sequence (composite)	
OWL	www.bioinf.man.ac.uk/dbbrowser/OWL
NRDB	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein
Protein sequence (secondary)	www.expasy.ch/prosite
PROSITE	www.bioinf.man.ac.uk/dbbrowser/PRINTS/PRINTS.html
PRINTS	www.sanger.ac.uk/Pfam/
Pfam	
Macromolecular structures	www.rcsb.org/pdb
Protein Data Bank (PDB)	ndbserver.rutgers.edu/
Nucleic Acids Database (NDB)	www.ncifcrf.gov/CRYS/HIVdb/NEW_DATABASE
HIV Protease Database	www.ebi.ac.uk:8081/home.html
ReLiBase	www.biochem.ucl.ac.uk/bsm/pdbsum
PDBsum	www.biochem.ucl.ac.uk/bsm/cath
CATH	scop.mrc-lmb.cam.ac.uk/scop
SCOP	www.embl-ebi.ac.uk/dali/fssp
FSSP	
	www.ncbi.nlm.nih.gov/Genbank
Nucleotide sequences	www.ebi.ac.uk/embl
GenBank	www.ddbj.nig.ac.jp
EMBL	
DDBJ	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome
Genome sequences	bioinfo.mbb.yale.edu/genome
Entrez genomes	www.ncbi.nlm.nih.gov/COG
GeneCensus	
COGs	
	www.ebi.ac.uk/interpro
Integrated databases	www.expasy.ch/srs5
InterPro	www.ncbi.nlm.nih.gov/Entrez
Sequence retrieval system (SRS)	
Entrez	

Protein sequence databases

Protein sequence databases are categorized as primary, composite or secondary. Primary databases contain over 300,000 protein sequences and function as a repository for the raw data. Some more common repositories, such as SWISS-PROT [3] and Protein Information Resources (PIR) - International [23], annotate the sequences as well as describe the protein's functions, its domain structure and post-translational modifications. Composite databases such as OWL [24] and the NRDB [25] compile and filter sequence data from different primary databases to produce combined non-redundant sets that are more complete than the individual databases and also include protein sequence data from the translated coding regions in DNA sequence databases (see below). Secondary databases contain information derived from protein sequences and help the user determine whether a new sequence belongs to a known protein family. One of the most popular is PROSITE [26], a database of short sequence patterns and profiles that characterize biologically significant site sin proteins. PRINTS [27] expand, on this concept and provides a compendium of protein fingerprints – groups of conserved motifs that characterize a protein family. Motifs are usually separated along a protein sequence, but may be contiguous in 3D space when the protein is folded. By using multiple motifs, finger prints can encode protein folds and functionalities more flexibly than PROSITE. Finally, Pfam [28] contains a large collection of multiple sequence alignments and profile Hidden Markov Models covering many common protein domains. Pfam-A, comprises of accurate manually compiled alignments while Pfam-B is an automated clustering of the whole SWISS-PORT database. These different databases are known as secondary databases. The different secondary structures have recently been incorporated into single resources named InterPro [29].

Structural databases:

Next we look at the data bases of macro molecular structures. The protein data bank [6, 7], provides a primary archive of all 3D structures for macro molecules such as proteins, DNA and RNA and various complexes. Most of the 13,000 structures are solved by x-ray crystallography and NMR, but some theoretical models have also been included. As the information provided in the individual PDB entries can be difficult to extract, PDBsum [30] provides a separate web page for every structure in the PDB displaying detailed structural analyses, schematic diagrams and data on interactions between different molecules in a given enter. Three major databases classify proteins by structure in order to identify the structural and evolutionary relationships: CATH [31], SCOP [32] and FSSP [33] databases. All comprises of hierarchical structural taxonomy where groups of proteins increase in similarity at lower levels of the classification tree. Some other databases which includes the Nucleic acids Database, NDB [34] for structures related to nucleic acids, the HIV protease database [35] for HIV-1, HIV-2 and SIV protease structures and their complexes and ReLiBase [36] for receptor ligand binding complexes.

Nucleotide and Genome sequences

The biggest excitement currently lies with the availability of complete genome sequences for different organisms. The GenBAnk [2], EMBL [37] and DDBJ [38] database contain DNA sequences for individual genes that encode proteins and RNA products. Much like composite protein sequence database, the Enterz nucleotide database [39] compiles sequences data from these primary databases.

Sequencing of whole genome is normally involves International collaborations, hence individual genomes are published at different sites. The Enterz genome database [40] brings together all complete and partial genomes in a single location and currently represents over 1000 organisms.

In addition to providing the raw nucleotide sequence, information is presented at several levels of detail including: a list of completed genomes, all chromosomes in an organism, detailed view of single chromosomes making coding and non-coding regions and single genes. At each level there are graphical presentations, pre computed analysis and links to other sections of Enterz. For example, annotations for single genes include the translated protein sequence, sequence alignments with similar genes in other genomes and summaries of the experimentally characterized or predicted function. GeneCensus [41] also provides an entry point for genome analysis with an interactive whole geneme comparison from an evolutionary perspective. The database allows building of phylogenetic trees based on different criteria such as ribosomal RNA or protein fold occurrence. The site also enables multiple genome comparisons, analysis of single genomes and retrieval of information individual genes. The COGs databases [20] classifies protein encoded in 21 completed genomes n the basis of similarity.

Gene Expression Data

The most recent source of genomic scale data has been from expression experiments, which quantify the expression levels of individual genes. These experiments measure the amount of mRNA or protein products that are produced by the cell. For the former there are three technologies: the cDNA microarray [42-44], AffymatrixGenChip [45] and SAGE methods [46]. The first method measures the relative levels of mRNA abundance between different samples, while the last two measures the absolute levels. Most of the effort in gene expression analysis has concentrated on the yeast and human genome and there is no central repository for this data. For humans, the main application has been to understand expression in tumor and cancer cells. The molecular mechanisms of breast tumors [52], Lymphoma and Leukemia, molecular profiling projects provide data for microarray experiments on human cell lines.

The technologies for measuring the protein abundance are currently limited to 2D gel electrophoresis followed by mass spectroscopy [54]. As gels can only resolve about 1000 proteins [55], only the most abundant can be visualized. At present, data from these experiments are available from literature [56-57].

Data integration

The most economic activity in the bioinformatics often results from integrating multiple sources of data [58]. For instance, the 3D coordinates of a protein are more useful if combined with data about the protein's functions, occurrence in different genomes and interactions with other molecules. In this way, individual piece of information are put in context with respect to other data. Unfortunately, it is not always straightforward to access and cross-reference these sources of information because of the differences in nomenclature and file formats.

At a basic level, this problem is frequently addressed by providing external links to other databases, for example in PDBsum, web-pages for individual structures direct the users towards corresponding entries in the PDB, NDB, CATH, SCOP and SWISS-PROT. At a more advanced level, there have been efforts to integrate access across several data sources. One is the sequence Retrieval System (SRS) which allows flat file databases to be indexed to each other; this allows the users to retrieve, link and access entries from nucleic acid, protein sequence, protein motif, protein structure.

Conclusion

With the explosion of the data in the biological sciences, computational methods have become indispensible to the research related to biotechnology. Bioinformatics was initially developed for analysis of biological sequences. However now it encompasses a wide range of subjects such as biology, genomics and gene expression studies. In this review, we have attempted to provide introduction and overview of the current status of the bioinformatics. In particular we have discussed the types of biological information and databases that are commonly used. Two principal approached underpin all studies in the bioinformatics. First is that of comparing and grouping the data according to biologically meaningful similarities and second, that of analyzing one type of data to infer and understand the observations for another type of data.

Acknowledgement

The authors wish to record the acknowledgement for the Dr. R. M. Dubey, Managing Director and Dr. Anupam Srivastav, Director, College of Engineering and Technology, Moradabad, for providing all the support and encouragement to the authors to complete this task.

References

[1] Reichhardt T, It's sink or swim as a tidal wave of data approaches. *Nature 399* (1999) 517-520.

[2] Benson D A, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA & Wheeler DL. GenBank. *Nucleic Acids Res* 28 (2000) 15-18.

[3] Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 28(2000)45-48.

[4] Fleischmann R D, Adams M D, White O, Clayton R A, Kirkness E F, Kerlavage A R, *et al*. Whole-genome random sequencing and assembly of *Haemophilus influenza* Rd. *Science* 269 (1995) 496-512.

[5] Drowning in data. The Economist 26 June 1999.

[6] Bernstein F C, Koetzle T F, Williams G J, Meyer E F, Jr., Brice M D, Rodgers J R, et al. The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur J Biochem* 80 (1977) 319-324.

[7] Berman H M, Westbrook J, Feng Z, Gilliland G, Bhat T N, Weissig H, et al. The Protein Data Bank. *Nucleic Acids Res* 28(2000) 235-242.

[8] Pearson W R & Lipman D J. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* 85(1988) 2444-2448.

[9] Altschul S F, Madden T L, Schaffer A A, Zhang J, Zhang Z, Miller W, *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25 (1997) 3389-3402.

[10] Holstege F C J E, Wyrick J J, Lee T I, Hengartner C J, Green M R, Golub T R, Lander E S, Young R A. Dissecting the regulatory circuitry of a eukaryotic genome. *Cell* 95(1998) 717-728.

[11] Pedersendagger A G, Jensendagger L J, Brunak S, Staerfeldt H H, Ussery D W. A DNA structural atlas for *Escherichia coli*. *J Mol Biol* 299 (2000) 907-930.

[12] Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28 (2000) 27-30.

[13] Jeffery C J. Moonlighting proteins. *TIBS* 24 (1999) 8-11.

[14] Chothia C. Proteins. One thousand families for the molecular biologist [news]. *Nature* 357(1992) 543-544.

[15] Orengo C A, Jones D T, Thornton J M. Protein superfamilies and domain superfolds. *Nature* 372 (1994) 631-634.

[16] Lesk A M, Chothia C. How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. *J Mol Biol* 136 (1980;) 225-270.

[17] Russell R B, Saqi M A, Sayle R A, Bates P A, Sternberg M J. Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *J Mol Biol* 269 (1997) 423-439.

[18] Russell R B, Saqi M A, Bates P A, Sayle R A, Sternberg M J. Recognition of analogous and homologous protein folds—assessment of prediction success and associated alignment accuracy using empirical substitution matrices. *Protein Eng* 11(1998) 1-9.

[19] Fitch W M. Distinguishing homologous from analogous proteins. *Syst Zool* 19 (1970) 99-110.

[20] Tatusov R L, Koonin E V, Lipman D J. A genomic perspective on protein families. *Science* 278 (1997) 631-637.

[21] Gerstein M, Hegyi H. Comparing genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol Rev* 22 (1998) 277-304.

[22] Skolnick J, Fetrow J S. From genes to protein structure and function: novel applications of computational approaches in the genomic era. *TIBTECH* 18 (2000) 34-39.

[23] McGarvey P B, Huang H, Barker W C, Orcutt B C, Garavelli J S, Srinivasarao G Y, *et al.* PIR: a new resource for bioinformatics. *Bioinformatics* 16 (2000) 290-291.

[24] Bleasby A J, Akrigg D, Attwood T K. OWL—a non-redundant composite protein sequence database. *Nucleic Acids Res* 22(1994) 3574-3577.

[25] Bleasby A J, Wootton J C. Construction of validated, non-redundant composite protein sequence databases. *Protein Eng* 3 (1990) 153-159.

[26] Hofmann K, Bucher P, Falquet L, Bairoch A. The PROSITE database, its status in 1999. *Nucleic Acids Res* 27 (1999) 215-219.

[27] Attwood T K, Croning M D, Flower D R, Lewis A P, Mabey J E, Scordis P, *et al.* PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res* 28 (2000) 225-227.

[28] Bateman A, Birney E, Durbin R, Eddy S R, Howe K L, Sonnhammer E L. The Pfam protein families database. *Nucleic Acids Res* 28(2000) 263-266.

[29] Attwood T K, Flower D R, Lewis A P, Mabey J E, Morgan S R, Scordis P, *et al.* PRINTS prepares for the new millennium. *Nucleic Acids Res* 27 (1999) 220-225.

[30] Laskowski R A, Hutchinson E G, Michie A D, Wallace A C, Jones M L, Thornton J M. PDBsum: a Web-based database of summaries and analyses of all PDB structures. *TIBS* 22(1997) 488-490.

[31] Pearl F M, Lee D, Bray J E, Sillitoe I, Todd AE, Harrison AP, *et al.* Assigning genomic sequences to CATH. *Nucleic Acids Res* 28(2000) 277-282.

[32] Lo Conte L, Ailey B, Hubbard T J, Brenner S E, Murzin A G, Chothia C. SCOP: a structural classification of proteins database. *Nucleic Acids Res* 28(2000) 257-259.

[33] Holm L, Sander C. Touring protein fold space with Dali/FSSP. *Nucleic Acids Res* 26 (1998) 316-319.

[34] Berman H M, Olson W K, Beveridge D L, Westbrook J, Gelbin A, Demeny T, *et al.* The Nucleic Acid Database. A comprehensive relational database of three dimensional structures of nucleic acids. *Biophys J* 63(1992) 751-759.

[35] Vondrasek J, Wlodawer A. Database of HIV proteinase structures. *TIBS* 22 (1997) 183.

[36] Hendlich M. Databases for protein-ligand complexes. *Acta Cryst D* 54 (1998) 1178- 1182.

[37] Baker W, van den Broek A, Camon E, Hingamp P, Sterk P, Stoesser G, *et al.* The EMBL nucleotide sequence database. *Nucleic Acids Res* 28 (2000) 19-23.

[38] Okayama T, Tamura T, Gojobori T, Tateno Y, Ikeo K, Miyazaki S, *et al.* Formal design and implementation of an improved DDBJ DNA database with a new schema and object-oriented library. *Bioinformatics* 14 (1998) 472-478.

[39] Schuler G D, Epstein J A, Ohkawa H, Kans J A. Entrez: molecular biology database and retrieval system. *Methods Enzymol* 266 (1996) 141-162.

[40] Tatusova T A, Karsch-Mizrachi I, Ostell J A. Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics* 15 (1999) 536-543.
[41] Lin J, Gerstein M. Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Res* 10 (2000) 808-818.
[42] Eisen M B, Brown P O. DNA arrays for analysis of gene expression. *Methods Enzymol* 303 (1999) 179-205.
[43] Cheung V G, Morley M, Aguilar F, Massimi A, Kucherlapati R, Childs G. Making and reading microarrays. *Nat Genet* 21(1999) 15-19.
[44] Duggan D J, Bittner M, Chen Y, Meltzer P, Trent JM. Expression profiling using cDNA microarrays. *Nat Genet* 21(1999) 10-14.
[45] Lipshutz R J F S, Gingeras TR , Lockhart D J. High density synthetic oligonucleotide arrays. *Nat Gen* 21(1999) 20-24.
[46] Velculescu V E Z L, Zhou, W Traverso, G St Croix, B Vogelstein B, Kinzler K W. Serial Analysis of Gene Expression Detailed Protocol. 1999.
[47] Roth F P H J, Estep P W, Church G M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat Biotechnol* 16(1998) 939-45.
[48] Jelinsky S A, Samson L D. Global response of *Saccharomyces cerevisiae* to an alkylating agent. *Proc Natl Acad Sci U S A* 96(199;) 1486-1491.
[49] Cho R J, Campbell M J, Winzeler E A, Steinmetz L, Conway A, Wodicka L, *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* 2 (1998) 65-73.
[50] DeRisi J L, Iyer V R, Brown P O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278(1997) 680-686.
[51] Winzeler E A, Shoemaker D D, Astromoff A, Liang H, Anderson K, Andre B, *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285 (1999) 901-906.
[52] Perou C M, Sorlie T, Eisen M B, van de Rijn M, Jeffrey S S, Rees C A, *et al.* Molecular portraits of human breast tumours. *Nature* 406 (2000) 747-752.
[53] Golub T R, Slonim D K, Tamayo P, Huard C, Gaasenbeek M, Mesirov J P, *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286 (1999) 531-537.
[54] Celis J E, Gromov P, 2D protein electrophoresis: can it be perfected? *Curr Opin Biotechnol* 10 (1999) 16-21.
[55] Pandey A, Mann M. Proteomics to study genes and genomes. *Nature* 405 (2000) 837-846.
[56] Futcher B, Latter G I, Monardo P, McLaughlin C S, Garrels J I. A sampling of the yeast proteome. *Mol Cell Biol* 19 (1999) 7357-7368.
[57] Gygi S P, Rist B, Gerber S A, Turecek F, Gelb M H, Aebersold R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 17(1999) 994-999.
[58] Gerstein M. Integrative database analysis in structural genomics. *Nature Struct Biol* 7 (2000) 960-963.