*Journal of Biochemistry and Molecular Biology*, Vol. 37, No. 1, January 2004, pp. 75-82

JOURNAL OF Biochemistry and Molecular Biology



**Review**

# Bioinformatics in the Post-genome Era

**Ungsik Yu, Sung Hoon Lee, Young Joo Kim and Sangsoo Kim***

National Genome Information Center, Korea Research Institute of Bioscience & Biotechnology, Daejeon 305-333, Korea

*Received 11 December 2003*

**Recent years saw a dramatic increase in genomic and proteomic data in public archives. Now with the complete genome sequences of human and other species in hand, detailed analyses of the genome sequences will undoubtedly improve our understanding of biological systems and at the same time require sophisticated bioinformatic tools. Here we review what computational challenges are ahead and what are the new exciting developments in this exciting field.**

**Keywords:** Bioinformatics, Computational biology, Comparative genomics, Database integration

## Introduction

Bioinformatics is playing a more and more significant role in the study of modern biology. It is now unthinkable to design research projects or experiments without a prior query or consultation of a few bioinformatic databases. In this respect, bioinformatics provides databases and tools to help ordinary biologists. To the other extreme, high throughput data can only be handled properly with some sort of automated analysis pipeline. And there are others developing mathematical or statistical algorithms and computational methodologies to analyze the data and thus uncover biological knowledge underneath the biological data. Here we may need to define bioinformatics and its close relative computational biology. Bioinformatics is the currently popular term for the application of computational and analytical methods to biological problems. Bioinformatics specifically refers to the search for and use of patterns and inherent structure in biological data such as genome sequences, and the development of new methods for database access and queries (NCBI Bioinformatics Definition *URL*). The term computational biology is more frequently used to refer to the physical and mathematical simulation of biological processes. However, as new high-throughput methods for obtaining and storing different types of biological data have developed, the line between these two subdisciplines has blurred.

If the goal of genome research or modern biology is to accumulate knowledge from DNA sequences and proteome information to pathways and networks and all the way to physiology and even to ecological systems, then there are layers of methodologies and expertise to accomplish these goals (Fig. 1). It is the role of bioinformatics to provide the necessary computational and statistical means and data handling capabilities.

In the last 50 years, the pool of publicly available biological data has grown exponentially (GenBank Growth *URL*). Genome projects for many different organisms are producing vast amounts of sequence data. The number of available protein structures may increase even faster than it has been, as initiatives for high-throughput structural biology projects get underway at several research institutions. The availability of new methods has made it possible to assay the rates of transcription and translation of thousands of genes in a single experiment. Plans for a public repository for these data are underway. To approach the study of datasets of this magnitude, advanced computational and data mining methods are required. Here we review methods that were newly introduced for the analysis of genome data and the new trends in this field.

### Completion of the Human Genome Project

After 10 years of international effort to unravel the human genome sequence, an initial draft was announced in 2000, papers described it in 2001 (Int. Human Genome Seq. Consortium, 2001), and the official announcement of completion in the spring 2003 (Collins *et al*., 2003). Except for a small portion of the genome, an extremely high quality sequence has been acquired and deposited in public databases (Golden Path Statistics *URL*). However, access to the genome data and its annotation information has not been an easy task for ordinary biologists, not to mention skilled

*To whom correspondence should be addressed.
Tel: 82-42-879-8500; Fax: 82-42-879-8519
E-mail: sskimb@kribb.re.kr

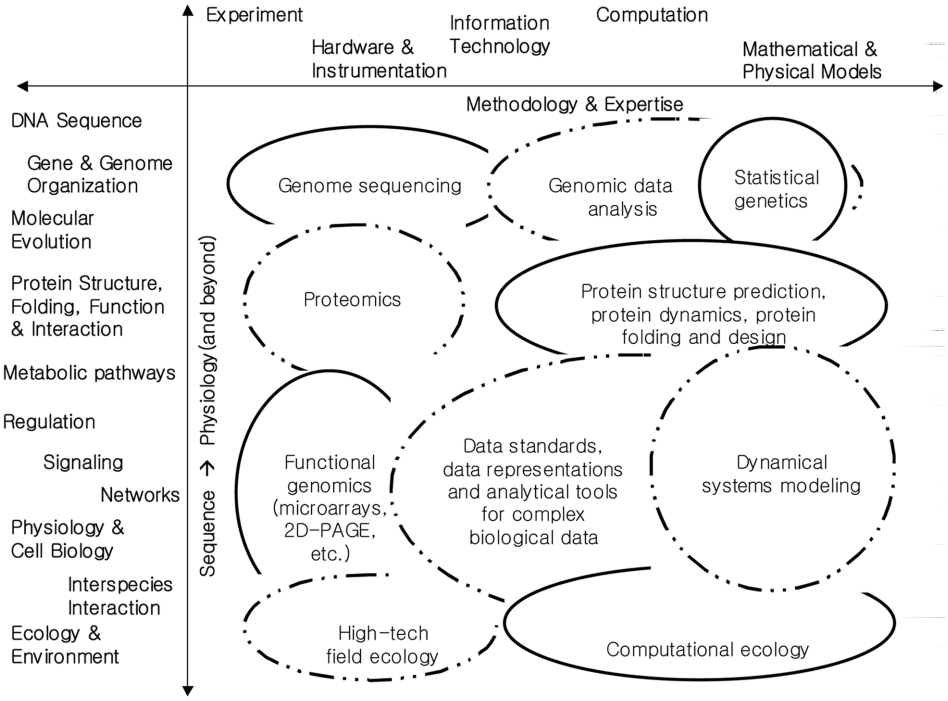


**Fig. 1.** Scope of post-genome research and technology.

bioinformaticians. Following the lead of the team at the University of California, Santa Cruz (Kent *et al*., 2002), NCBI, EBI and others released their own versions of genome browsers. Each has its own strength and design concept, which is very useful in answering questions such as where my gene is in the genome, and what others are nearby, how my gene is spliced compared to the genome, and whether there are other EST sequences supporting it. If one needs the 5' upstream sequence of a gene, it is a matter of a few clicks now with these browsers. A parallel exhibition of the tracks of many features such as CpG islands, mouse homology, repetitive elements and so on helps to understand the patterns in the genome sequence. Genetic markers that were obtained from linkage studies can be easily located on the genome. Genes in the vicinity can then be scrutinized for their role in the phenotypes. This information can be mapped on the mouse synteny region available on NCBI's homology map. The associated mouse mutation and phenotype information can be retrieved from the Jackson Laboratory's database (Jackson Lab. *URL*). The Ensembl browser (Ensembl Genome Browser *URL*), a collaboration of EBI and Sanger Institute, offers unique capabilities such as protein-based queries. For example, for those interested in proteins having certain domains or motifs, Ensembl provides a data mining system called EnsMart. With it, complex queries can be built by combining the restriction on genomic location, novelty, tissue expression pattern, protein domain and motif, SNP-types and so on. These browsers now serve not only human genome but also mouse, rat and other organisms for which complete genome sequences are available. It is certain that these genome browsers will be essential tools for all biologists in the coming years. Nature Genetics published a supplementary issue introducing these browsers with many practical examples (Wolfsberg *et al*., 2002).

## Next Questions to be Answered

The completion of the human genome sequence poses many important research questions that should be answered, as summarized by NHGRI (Collins *et al*., 2003):

- -Comprehensive identification of the structural and functional components encoded in the genome
- -Elucidation of the organization of genetic networks and protein pathways and their contribution to cellular and organismal phenotypes
- -Detailed understanding of the heritable variation in the human genome
- -Understanding of evolutionary variation across species
- -Development of policy options in regard to the widespread use of genome information
- -Identification of the genetic contributions to disease and drug response
- -Identification of gene variants related to good health and resistance to disease
- -Development of genome-based approaches to prediction of disease susceptibility and drug response
- -Early detection of illness and molecular taxonomy of disease states; use of new understanding of genes and pathways to develop powerful new therapeutics

-Investigation of genetic risk information in setting treatment strategy and predicting outcomes
-Development of genome-based tools that improve the health of all

NHGRI also recommended several associated areas in computational biology that need to be studied:

-New approaches are needed to solve problems, such as the identification of different features in a DNA sequence, the analysis of gene expression and regulation, the elucidation of protein structure and protein-protein interactions, the determination of the relationship between genotype and phenotype, and the identification of the patterns of genetic variation in populations and the processes that produced those patterns
-Reusable software modules for the facilitation of interoperability are in great need
-Methods to elucidate the effects of environmental (non-genetic) factors and of gene-environment interactions on health and disease
-New ontologies to describe different data types
-Improved database technologies to facilitate the integration and visualization of different data types, for example, information about pathways, protein structure, gene variation, chemical inhibition and clinical information/ phenotypes
-Improved knowledge management systems and the standardization of data sets to allow the coalescence of knowledge across disciplines.

With these goals in mind, we will review the recent progress made over the years.

## Automatic Annotation of Human and Other Genomes

One of the most interesting developments in the analysis of human genome sequence is provided by the Ensembl project (Ensembl Genome Browser *URL*), which is a joint operation of the European Bioinformatics Institute (EBI) and Sanger Institute. It employs a host of high performance computing servers to automatically annotate the human genome sequence. It predicts genes based on the protein-evidence and as such it provides rich information in relation to protein sequence, structure and function. It is extremely helpful when one wants to locate genes on genome that have certain protein domains. Now the successful operation is extended to other eukaryotic organisms such as mice, rats, fish, and worms.

In the microbial world, The Institute for Genome Research (TIGR) developed an automatic annotation system called CMR (Comprehensive Microbial Resource, TIGR/CMR *URL*). It provides rich annotation information that is regularly updated in order to reflect the new information accumulated in the relevant databases.

## Comparative Genome Analyses and Associated Tools

The availability of many completed genome sequences offers an exciting new opportunity that has never seen before, that is, the comparative genome/proteome analysis of closely and distantly related species. A comparison of genome sequences of the species appropriately apart can be exploited in finding exons and other regulatory elements. Traditionally, eukaryotic gene findings have relied on probabilistic models that have been developed by learning patterns embedded in the known exon sequences. This approach, known to generate many false positives, can be complemented by the method based on the homology with related species. For example, in predicting human genes, the mouse genome sequence has proven extremely useful (Korf *et al*., 2001). Now, it is extended to include other species such as rats and fish, enhancing the performance. A prediction of regulatory elements in the upstream of human genes has been very difficult due to their short length. Now, the cross-species comparison of the promoter predictions can greatly reduce the false positives. M.Q. Zhang recently reported that combining the promoter predictions of humans, mice, and rats that increased the specificity and utilization of even more genomic sequences from other species would be desirable (Zhang, 2003). Eric Davidson also employed this powerful approach to discover cis-regulatory elements of a sea urchin by comparing a wider range of upstream of orthologous genes in a sea star (Davidson, 2003). This information, combined with other experimental data, culminated in constructing a gene regulatory network in embryonic stage of sea urchin (*vide infra*). With the DNA sequencing price becoming cheaper (thanks to the technology development), more sequences will be available in the coming years, which means an exciting era for comparative genome analysis.

Proteins that are conserved in many species may play an essential role in the biology of those species, while the species can be classified based on the profile of conserved proteins. NCBIs COG--Clusters of Orthologous Groups (Tatusov *et al*., 2001) provides precisely that kind of information among microorganisms. Now they have developed a new kid on the block called KOG that is an eukaryotic version of COG. It compares the proteomes of *S. cerevisiae*, *S. pombe*, *C. elegans*, *Drosophila*, *human*, *Arabidopsis*, and *Encephalitozoon cuniculi*. It provides information on 4,852 domains that are conserved in some of these species. It can be a very useful tool in detecting domain structures in novel proteins (see the section on protein motifs).

## Transcriptome Analysis and Gene Discovery

In contrast to the approaches of gene prediction from genome sequences that can be purely hypothetical, the cDNA collection

provides experimental evidence of transcription. Besides, the prediction methods tend to be biased to protein-coding genes, while the cDNA approach does not discriminate non-coding genes that have been drawing attention recently. It has its own weakness in relying on the expression level and thus difficulty in detecting low copy genes. Nevertheless, the complementary use of these two approaches improves the stability and specificity of the gene prediction. More importantly, the contribution of the cDNA-based approach to the understanding of genomics is the experimental collection of alternatively spliced forms and variations in the start and stop sites. This information points to the fact that transcriptome is extremely dynamic and diverse and plays a central role in understanding the gene regulation, function, and expression, bridging genomics and proteomics. Recently mouse (The FANTOM Consortium and RIKEN Genome Exploration Research Group Phase I & II Team, 2002) and human (Gojobori *et al*., 2003) full-length cDNA clones were collected and annotated in terms of functions, expression and structures. Both were organized by two different Japanese groups and offer experimental validation to the genome-based gene predictions. According to Ensembl, the number of genes predicted from human or mouse genomes is around 28,000 each. On the other hand, each consortium collected close to 60,000 cDNAs that can be grouped into 23-30k clusters or transcriptional units. In other words, there are at least two-fold redundancies in the cDNA collections that are a combination of alternative spicing forms and start/stop site variants. Invariably, close to 10,000 cDNAs among the non-redundant clusters lack any coding potential, and are often spliced into exons. While it includes obvious examples of antisense genes, the majority of them cannot be properly characterized at this time. Now with the experimental evidence of various variants, bioinformatic analyses are being pursued to correlate these with observed the phenotypes such as pathological conditions of the samples.

## Gene Expression Analysis and Mining Tools

DNA microarray technologies now become a popular tool for surveying gene expression profiles. The associated statistical analysis methods have also been well established. It is now a common practice to classify samples based on the similarity in expression profiles and also identify clusters of genes characterizing those sample groups. This approach is particularly useful in studying clinical samples such as tumor tissues in various stages. However, cancer is a complicated process and the gene expression profile alone cannot pinpoint oncogenes or discern the tumorigenesis mechanism. Now various downstream analyses may ramify the dysregulated gene lists. One of the most widely used methods is to classify those dysregulated genes according to their functions in order to recognize which functional category is affected the most. Kyoto Encyclopedia of Genes and Genomes (KEGG, Kanehisa *et al*., 2002) accepts a list of gene symbols and returns the metabolic pathways these genes are involved with. As they mark the genes that are indicated by the user in different colors, the recognition of the interrelationship among the genes is straightforward. Another popular database in this category is Gene Ontology (GO, The Gene Ontology Consortium, 2001) that is originally intended to provide controlled vocabulary describing biological terms that are related to biological and biochemical functions as well as cellular localization of gene products. As GO associates each biological term with gene products, it is natural to use this database as the reference for cataloging gene functions. Onto-Express provides this functionality over the web (Khatri *et al*., 2002). Gene MicroArray Pathway Profiler (GenMAPP, Dahlquist *et al*., 2002) combines the best of all and provides a powerful statistical analysis and visual presentation. For a given input of genes and their associated expression level, GenMAPP searches a number of pathways and GO terms and scores the commonly occurring terms in terms of significance over random chance. The resulting pathways are presented graphically as a clickable image. Besides functional classification and pathway mapping, there are other methods to make sense out of a list of dysregulated genes. As the transcriptome analysis profiles the downstream effects of stimulation, these genes are under common transcriptional regulation. EXPANDER (developed by Ron Shamirs group at Weizmann Institute, Israel) is particularly helpful in discovering sets of genes sharing common promoter elements in their upstream regions (Sharan *et al*., 2003). A rather recent approach to the downstream analysis of genes lists out of microarray studies is to use protein-protein interaction information under the proposition that proteins interacting together may be involved in the same biological process and co-regulated. If this approach proves successful, it may revolutionize the whole analysis as it represents the marriage of two high throughput technologies.

## Protein Motif, Domain Finding and Classification

Short blocks or motifs that are found in many proteins may play a significant role in terms of structure and function of the proteins. There are various approaches to detect such motifs, ranging from manual curation of multiple sequence alignments to machine learning approaches such as Hidden Markov Models. The latter is extremely useful in detecting even very large domains in protein sequences. Pfam offers two kinds of domain databases: the first is the curated part containing over 7,000 protein families and domains; the other part is automatically generated from PRODOM that do not overlap with the first (Bateman *et al*., 2002). With these databases, one can scan a protein of interest to detect whether it contains one or more known domains, and if found, what other proteins share the same domains. This is the first typical step in annotating a novel protein, which follows cloning and sequencing the gene. As there are several databases available,

many motifs are redundant among these databases. EBI merged the motifs and domains that are included in these databases, producing a federated database called InterPro along with its associated search program InterProScan (*URL*). NCBI offers a more evolutionarily savvy approach called CDD (Conserved Domain Databases, *URL*) that includes both the Pfam and COG-style profiles that were mentioned previously (see the section on comparative genome analysis). Profiles representing the probability of substituting an amino acid at a particular position with other amino acids were compiled and served as the database against which the query sequence was searched. NCBI implemented this service with their popular BLAST program and called it RPS-BLAST (Reverse Position-Specific BLAST). It is very fast, much like the other protein-based BLAST with various subsets of databases to choose and outputs the result in the familiar BLAST look. If there is no hit to the known families or domains, it would be worth to try PSI-BLAST (Position-Specific Iterated BLAST), which calculates profiles based on initial hits and uses these profiles to search the database again to detect distantly related homologs. This process is iterated by user-specified number of times and the converged results are returned. However, the process is not guaranteed to converge and the optimal setting of the input parameters is not generally known, requiring careful examination of the output.

## Proteome Expression and Localization

Profiling the proteome expression is approached in two different ways: one is a *de novo* discovery using 2D gel electrophoresis or LC/MS, followed by a database search with peptide fragment masses; the other is a chip-based method that is restricted to the probes spotted on the slide but offers parallel high throughput measurement of the protein quantity using the antibody-antigen reaction. Increasing amount and complexity of data generated from typical proteome profiling studies, especially from high throughput analyses, demands a reliable laboratory information management system (LIMS) throughout the course of proteome experiments and analyses. LIMS is often required in every aspect of genome research, but is most critical for a successful proteomics operation due to the facile nature of the process. Coping with complex digestion patterns and post-translational modifications requires a substantial amount of computing power in searching fragment databases. Economic linux clusters are typically employed in this business; for example, 256 Intel Xeon CPUs are used by John Yates group (Yates group *URL*). It would be valuable if the expression levels of all of the available proteins were measured across all tissues and organs. Mathias Uhlens group at the Royal Institute of Technology, Sweden, is building such databases (Uhlen, 2003). It presents enormous storage problems as follows: the size of a typical tissue array image is tens of megabytes, which is easily expanded to tens of gigabyte per antibody across tissue samples, and finally reaches terabytes of information when annotation is included. This causes complications in delivering this important information to the users.

The Nobel laureate in 2002, Sydney Brenner, emphasized the importance of protein localization in the study of protein function (CHI Conference, 2003). According to him, the genome is an inventory of function, but not function, and we need to translate data into knowledge. The 20,000 or so genes that are expressed in a single cell are too complex a problem to solve, and examining orders of magnitude more protein-protein interactions does not help. Instead, Brenner advocates studying functional assemblages of proteins, such as spliceosome, the molecular machine that splices RNA messages. He says that only about 2,000 such machines exist in the cell, a more manageable number. Further confusion vanishes if we think in terms of topographical regions of the cell, such as organelles, the plasma membrane, and the nucleus. Brenner says that about 10 such regions are easier to grasp, although neurons may have greater complexity. In order to realize this approach, we are in urgent need of novel technologies for isolating the facile complexes and tracking proteins down the cell compartments.

## Protein Structure Threading and Prediction

An amino acid sequence of a protein dictates its three-dimensional structure, which in turn defines its biochemical function. It is thus of paramount importance to determine the three-dimensional structure of proteins. A recent issue of Chem. & Eng. News summarized this field very well and its excerpt is given here (Borman, 2003). There are structural genomics project worldwide to determine experimental structures of a host of proteins, for example all of the proteins in a genome or all the representative proteins in the folding space. The computational approaches of a structure prediction that compliments the experimental approaches can be categorized into several classes, depending on the homology to the proteins for which the three-dimensional structures are known (Fig. 2). Comparative homology modeling performs very well for those with greater than a 25% sequence homology to known structures. Currently Modeller, (created by Professor of computational biology, Andrej Sali, and coworkers at the University of California, San Francisco) is the most commonly used comparative modeling software and has had a large impact on the field, while Dunbrack and coworkers developed SCWRL, a side-chain conformation prediction program. For those without detectable homology to known structures, it may still be possible to find a suitable structure using the second technique, fold recognition. For perhaps half or three-quarters of unknown (structurally uncharacterized) proteins, there will be a suitable structure in the database one can use as a basis or template for extrapolating a 3-D model, even in the absence of a sequence match. Rather than finding a structure that's suitable for the sequence, fold recognition methods try to find whether

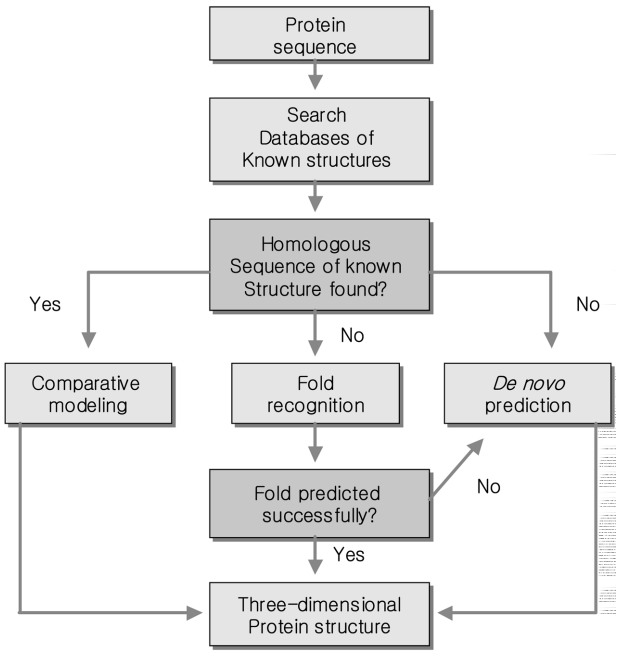


**Fig. 2.** Protein structure prediction techniques are of three main types: comparative modeling, fold recognition, and *de novo* prediction.

the sequence fits on any existing structures. In other words, one threads the sequence onto existing structures to see if these structures bury the sequence's hydrophobic residues well, among a number of other structural considerations. According to CASP5 (The 5th Critical Assessment of Structure Prediction Methods, 2003), the biggest development in recent years was in this area. Various research groups have developed computer servers that accept incoming sequence data and generate models in an automated manner. A new concept was also introduced, that is, a metaserver (a server that sends out sequences to multiple other servers) gets a number of models from them, and then uses them to make consensus models. The result is a significant improvement of results in the fold-recognition category. If there's no known sequence or known structure to match the sequence, the third option is to build a 3-D structure of a protein from scratch, called *de novo* or *ab initio* modeling. This is the only viable option for proteins with completely new folds; structures that have not been previously observed in nature. The success rate in this category varies greatly and depends also upon the size of the protein in question. Recently, Baker *et al*. at Univ. of Washington, Seattle, developed a novel method called Rosetta. Essentially, it makes new proteins by assembling little bits from known protein structures. This hybrid method performs far better than any other *de novo* methods.

## Gene Regulation Network and Systems Biology

Eric Davidson at the California Institute of Technology was able to construct the gene regulatory network (GRN) for endomesoderm specification in sea urchin embryos (Davidson, 2003). This network, which contains ~50 genes (mainly regulatory genes). is a powerful tool for both explanation and prediction, extending from the phenomenology of development to cis-regulatory functions at the nodes of the GRN. The GRN is formulated on the basis of prior knowledge of the developmental process, detailed observations on spatial and temporal patterns of gene expression, and a large-scale perturbation analysis. The GRN is represented in computational models that permit predictions of cis-regulatory inputs at the nodes of the GRN, and that display the gene regulatory transactions that are active or inactive in given spatial domains of the embryo at given times. Among the computational methodologies that were developed to support the GRN analysis are a new application of interspecies sequence comparisons, as mentioned previously (see Comparative Genome Analysis section). It has been proven experimentally to be remarkably useful for the rapid identification of cis-regulatory elements.

According to Leroy Hood at the Institute for Systems Biology, Seattle, USA, systems biology is to determine, analyze, and integrate interrelationships of all the elements in the biological system in response to genetic or environmental perturbations (Hood, 2003). The objective is to model the system graphically and ultimately mathematically so as to understand the systems or emergent properties. Hood's group is developing powerful new tools for gathering extensive genomic and proteomic data sets. They are also developing computational tools for dealing with these data. They have applied these tools to the analyses of several simple systems: galactose utilization in yeast, and the protein and gene regulatory networks for phototrophy and UV repair in halobacterium. It seems that systems approaches, driven by biology, offer powerful insights into the complexities of biological systems.

## Database Integration and Grid Computing

Data integration is a prerequisite for improving the efficacy of extraction and analyses of biological information, particularly for knowledge discovery and research planning. Biologists often ask, "There are so many databases out there. How do I find and access the database for analysis and extraction of information relevant to my research idea?" The goals of data integration are the formulation of querying, reporting, and complex analysis (multidimensional analysis and data mining). However, the fuzziness and complexity of biological data represent major challenges in molecular biology database integration, and require high-level expert interpretations to suffice the requirements of the biological R&D. The traditional approach to this problem has been data warehousing, where all the relevant databases are mirrored, stored locally in a unified format, and mined through a

uniform interface. SRS (Sequence Retrieval System *URL*), developed by EBI, is a good example. However, the overhead costs of this kind of approach are too heavy, in terms of data maintenance. Recently, Lincoln Stein at Cold Spring Harbor Lab. advocated a concept of federated databases, which is based on web services (Stein, 2002). Many specialized databases register their access protocols to a central repository, called the reference server. The client application then refers to the reference server to locate the servers that serve the relevant information and the associated access protocols. Based on these protocols, queries are formulated and sent to the server, and the client analyzes the returned results, all automatically. This new concept can take advantage of distributed and collaborative maintenance of ever growing databases. Currently, DAS (Distributed Annotation System, BioDAS *URL*), a prototype of web services in bioinformatics, is in service and several projects are ongoing to develop client and server applications.

Sequencing and assembling the genome requires tera-flop clusters; proteomics could require 10 to 100 times as much computing power. Optimal target identification with design of intervention may require peta-scale computing. Further, simulation and related optimization efforts place much more emphasis on scalable, massively parallel computing than do informatics tasks. Informatics and simulation pose different architectural requirements on computing: informatics involves almost no floating point arithmetic, needs fairly minimal communication capabilities, and tends to be input/output bound. Simulation and optimization are highly dependent on floating point operations, involve less input/output, and require an effective communications fabric among processors. In the new biology, it is likely that compute farms such as Beowulf clusters, scalable clusters, and massively parallel supercomputers will play important roles.

## Concluding Remarks

Computational methods have become intrinsic to modern biological research. Their importance can only increase as large-scale methods for data generation become more prominent, as the amount and complexity of the data increase, and as the questions being addressed become more sophisticated. All future biological research will integrate computational and experimental components. New computational capabilities will enable the generation of hypotheses and stimulate the development of experimental approaches to test them. The resulting experimental data will, in turn, be used to generate more refined models that will improve the overall understanding and increase opportunities for biotechnological applications.

**Acknowledgments** The work was supported by a grant from the KRIBB Research Initiative Program and in part by the Bio-Infrastructure Program from the Korea Ministry of Science & Technology. S. Kim thanks Dr. Seyon Won for his helpful discussions.

## URLs

-BioDAS (http://www.biodas.org)
-CDD (http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml)
-Ensembl Genome Browser (http://www.ensembl.org)
-GenBank Growth (http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html)
-Golden Path Statistics (http://genome.ucsc.edu/goldenPath/stats.html)
-InterPro (http://www.ebi.ac.uk/interpro/)
-Jackson Lab. Mouse Genome (http://informatics.jax.org)
-NCBI Bioinformatics Definition (http://www.ncbi.nlm.nih.gov/About/primer/bioinformatics.html)
-Sequence Retrieval System (http://www.lionbioscience.com)
-TIGR/CMR (http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl)
-Yates group (http://fields.scripps.edu/)

## References

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M. and Sonnhammer, E. L. (2002) The Pfam Protein Families Database. *Nucleic Acids Res.* **30**, 276-280.
Pfam (http://www.sanger.ac.uk/Software/Pfam/)

Borman, S. (2003) Divining protein architecture. *Chemical & Engineering News* **81**, 26-30.
(http://pubs.acs.org/isubscribe/journals/cen/81/i31/html/8131sci1.html)

CHI Conference (2003) Molecular Medicine Marketplace in Santa Clara, CA, USA.
(http://www.bio-itworld.com/news/050903_report2511.html)

Collins, F. S., Green, E. D., Guttmacher, A. E. and Guyer, M. S. (2003) A vision for the future of genomics research. *Nature* **422**, 835-847.

Dahlquist, K. D., Salomonis, N., Vranizan, K., Lawlor, S. C. and Conklin, B. R. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genetics* **31**, 19-20.
Gene MicroArray Pathway Profiler (http://www.genmapp.org/)

Davidson, E. (2003) Abstract from Transcriptome 2003 in Tokyo, Japan.

Gojobori, T. and the H-Invitational team (2003) Abstract from Transcriptome 2003 in Tokyo, Japan.

Hood, L. (2003) Abstract from Transcriptome 2003 in Tokyo, Japan.

International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921.

Kanehisa, M., Goto, S., Kawashima, S. and Nakaya, A. (2002)

The KEGG databases at GenomeNet. *Nucleic Acids Res.* **30**, 42-46.
Kyoto Encyclopedia of Genes and Genomes (http://www.genome.ad.jp/kegg)

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M. and Haussler, D. (2002) The Human Genome Browser at UCSC. *Genome Res.* **12**, 996-1006.
UCSC Genome Browser (http://genome.ucsc.edu)

Khatri, P., Draghici, S., Ostermeier, G. C. and Krawetz, S. A. (2002) Profiling gene expression using onto-express. *Genomics* **79**, 266-270.
Onto-Express (http://vortex.cs.wayne.edu/Projects.html)

Korf, I., Flicek, P., Duan, D. and Brent, M. R. (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics* **17**, S140-148.

Sharan, R., Maron-Katz, A. and Shamir, R. (2003) CLICK and EXPANDER: A System for Clustering and Visualizing Gene Expression Data. *Bioinformatics* **19**, 1787-1799.

Stein, L. (2002) O'Reilly's Bioinformatics Technology Conference. (http://www.oreillynet.com/pub/a/network/2002/01/29/bioday2.html)

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V., Tatusova, T. A., Shankavaram, U. T., Rao, B. S., Kiryutin, B., Galperin, M. Y., Fedorova, N. D. and Koonin, E. V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* **29**, 22-28.
Clusters of Orthologous Groups (http://www.ncbi.nih.gov/COG/)

The FANTOM Consortium and the RIKEN Genome Exploration Research Group Phase I & II Team (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**, 563-573.

The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.* **11**, 1425-1433.
Gene Ontology Consortium (http://www.geneontology.org)

The 5th Critical Assessment of Structure Prediction Methods (2003)
(http://predictioncenter.llnl.gov/casp5/Casp5.html)

Uhlen, M. (2003) Abstract from Transcriptome 2003 in Tokyo, Japan.

Wolfsberg, T. G., Wetterstrand, K. A., Guyer, M. S., Collins, F. S. and Baxevanis, A. D. (2002) A user's guide to the human genome. *Nat. Genet.* **32**, Suppl. 1-79.

Zhang, M. Q. (2003) Abstract from Transcriptome 2003 in Tokyo, Japan.