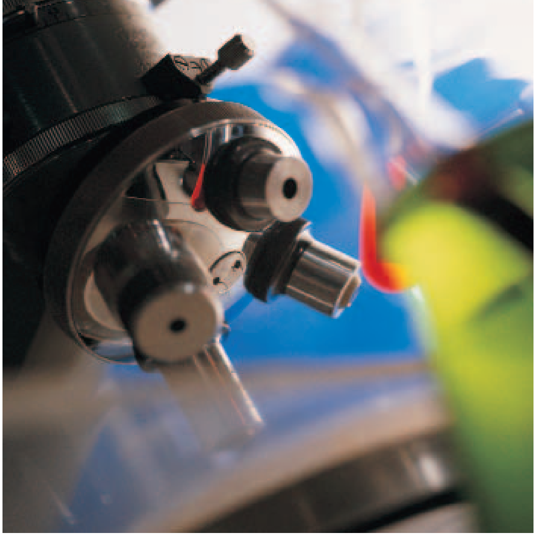
Research in the "postgenome era" examines the genomic data produced by DNA sequencing efforts, seeking a greater understanding of biological life.

See-Kiong Ng and Limsoon Wong



Accomplishments and Challenges in Bioinformatics

Informatics has helped launch molecular biology into the genome era. The use of informatics to organize, manage, and analyze genomic data (the genetic material of an organism) has become an important element of biology and medical research. A new IT discipline—*bioinformatics*—fuses computing, mathematics, and biology to meet the many computational challenges in modern molecular biology and medical research. The two major themes in bioinformatics—data management and knowledge discovery—rely on effectively adopting techniques developed in IT for biological data, with IT scientists playing an essential role.

In the 1990s, the Human Genome Project and other genome sequencing efforts generated large quantities of DNA sequence data. Informatics projects in algorithms, software, and databases were crucial in the automated assembly and analysis of the genomic data. The "Road to Unraveling the Human Genetic Blueprint" sidebar lists key advances in human genome research.

The Internet also played a critical role: the World Wide Web let researchers throughout the world instantaneously share and access biological data captured in online community databases. Information technologies produced the necessary speedup for collaborative research efforts in biology, helping genome researchers complete their projects on time.

We're now in the "postgenome" era. Many genomes have already been completely sequenced, and genome research has migrated from raw data generation to scientific knowledge discovery. Likewise, informatics has shifted from managing and integrating sequence databases to discovering knowledge from such biological data. Informatics' role in biological research has increased and it will certainly become increasingly important in extending our future understanding of biological life.

DATA MANAGEMENT

The many genome mapping and sequencing initiatives of the 1990s resulted in numerous databases. The hot topics then were managing and integrating these databases and comparing and assembling the sequences they contained.

Data integration

No single data source can provide answers to many of biologists' questions; however, information from several sources can help satisfactorily solve some of them. Unfortunately, this has proved difficult in practice. In fact, in 1993 the US Department of Energy published a list of queries it considered unsolvable. What's interesting about these queries was that a conceptually straightforward answer to each of them existed in databases. They were unsolvable because the databases were geographically distributed, ran on different com-

Published by the IEEE Computer Society

puter systems with different capabilities, and had very different formats.

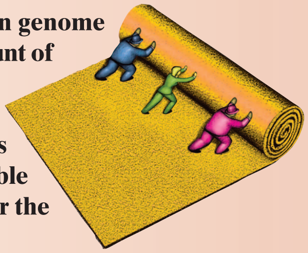
One of the US Department of Energy's "impossible queries" was: "For each gene on a given cytogenetic band, find its nonhuman homologs." Answering this query required two databases: the Genome Database, GDB, (www.gdb.org) for information on which gene was on which cytogenetic band, and the National Center for Biotechnology Information's Entrez database (www.ncbi.nlm.nih.gov/Entrez) for information on which gene was a homolog of which other genes. GDB, a relational database from the company Sybase supporting Structured Query Language (SQL) queries, was located in Baltimore, Maryland. Entrez, which users accessed through an ASN.1 (Abstract Syntax Notation One) interface supporting simple keyword indexing, was in Bethesda, approximately 38 miles south.

Kleisli, a powerful general query system developed at the University of Pennsylvania in the mid-1990s, solved this problem. Kleisli lets users view many data sources as if they reside within a federated nested relational database system. It automatically handles heterogeneity, letting users formulate queries in an SQL-like high-level way independent of

- the data sources' geographic location,
- whether the data source is a sophisticated relational database system or a dumb flat file, and
- the access protocols to the data sources.

Kleisli's query optimizer lets users formulate queries clearly and succinctly without having to worry about whether the queries will run fast. Figure 1 shows Kleisli's solution to the Department of Energy's "impossible query."

Road to Unraveling the Human Genetic Blueprint



The race to mapping the human genome generated an unprecedented amount of data and information, requiring the organizational and analytical power of computers. Computers and biology thus became inseparable partners in the journey to discover the genetic basis of life.

Several key historical events led to the complete sequencing of the human genome:

➤ **1865—Gregor Mendel discovers laws of genetics.**

➤ **1953—James Watson and Francis Crick describe the double-helical structure of DNA.**

➤ **1977—Frederik Sanger, Allan Maxam, and Walter Gilbert pioneer DNA sequencing.**

➤ **1982—US National Institutes of Health establishes GenBank, an international clearinghouse for all publicly available genetic sequence data.**

➤ **1985—Kary Mullis invents polymerase chain reaction (PCR) for DNA amplification.**

➤ **1985—Leroy Hood develops the first automatic DNA sequencing machine.**

➤ **1990—Human Genome Project begins, with the goal of sequencing human and model organism genomes.**

➤ **1999—First human chromosome sequence published.**

➤ **2001—Draft version of human genome sequence published.**

➤ **2003—Human Genome Project ends with the completed version of human genome sequence.**

A detailed graphic timeline is available at http://www.genome.gov/11007569.

Several additional approaches to the biological data integration problem exist today. Ensembl, SRS, and DiscoveryLink are some of the better-known examples.

- EnsEMBL (http://www.ensembl.org) provides easy access to eukaryotic genomic sequence data. It also automatically predicts genes in these data and assembles supporting annotations for its predictions. Not quite an integration technology, it's nonetheless an excellent example of successfully integrating data and tools for the highly demanding purpose of genome browsing.
- SRS (http://srs.ebi.ac.uk) is arguably the most widely used database query and navigation system in the life science community. In terms of querying power, SRS is an information retrieval system and doesn't organize or transform the retrieved results in a way that facilitates setting up an analytical pipeline. However, SRS provides easy-to-use graphical user interface access to various scientific databases. For this reason, SRS is sometimes considered more of a user interface integration tool than a true data integration tool.
- IBM's DiscoveryLink (http://www.ibm.com/discoverylink) goes a step beyond SRS as a general data integration system in that it contains an explicit data model—the relational data model. Consequently, it also offers SQL-like queries for access to biological sources, albeit in a more restrictive manner than Kleisli, which supports the nested relational data model.

Recently, XML has become the de facto standard for data exchange between applications on the Web. XML is a standard for formatting documents rather than a data integration system. However, taken as a whole, the growing suite of tools based on XML can serve as a data integration system. Designed to allow for hierarchical nesting (the ability to enclose one data object within another) and flexible tag definition, XML is a powerful data model and useful data exchange format, especially suitable for the complex and evolving nature of biological data. It's therefore not surprising that the bioinformatics database community has rapidly embraced XML.

Many bioinformatics resource and databases such as the Gene Ontology Consortium (GO, http://www.geneontology.org), Entrez, and the Protein Information Resource (PIR, http://pir.georgetown.edu) now offer access to data using XML. The database community's intense interest in developing query languages for semistructured data has also resulted in several powerful XML query languages such as XQL and XQuery. These new languages let users query across multiple bioinformatics data sources and transform the results into a more suitable form for subsequent biocomputing analysis steps.

Research and development work on XML query optimization and XML data stores is also in progress. We can anticipate robust and stable XML-based general data integrating and warehousing systems in the near future. Consequently, XML and the growing suite of XML-based tools could soon mature into an alternative data integration system in bioinformatics comparable to Kleisli in generality and sophistication.

Figure 1. Kleisli solution.

```
sybase-add (name: “gdb”, ...);
create view locus from locus_cyto_location using gdb;
create view eref from object_genbank_eref using gdb;
select
       accn: g.genbank_ref,
       nonhuman-homologs: H
from
       locus c,
       eref g,
       {g.genbank_ref} r,
       {select u
        from r.na-get-homolog-summary u
        where not(u.title like “%Human%”)
           and not(u.title like “%H.sapien%”)} H
where c.chrom_num = “22” and
       g.object_id = c.locus_id  and not (H = {});
```

This Kleisli query answers the US Department of Energy query “list nonhuman homologs of genes on human chromosome 22.” The first three statements connect to GDB and map two tables in GDB to Kleisli. The next few lines extract from these tables the accession numbers of genes on Chromosome 22, use the Entrez function na-get-homolog-summary to obtain their homologs, and filter the homologs for nonhuman homologs. Underlying this simple SQL-like query, Kleisli automatically handles the heterogeneity and geographical distribution of the two underlying sources, and automatically optimizes, makes concurrent, and coordinates the various query execution threads.

Figure 2. A GenBank data record.

```
{(#uid: 6138971,
       #title: “Homo sapiens adrenergic ...”,
       #accession: “NM_001619”,
       #organism: “Homo sapiens”,
       #taxon: 9606,
       #lineage: [“Eukaryota”, “Metazoa”, ... ],
       #seq: “CTCGGCCTCGGGCGCGGC...”,
       #feature: {
            (#name: “source”,
            #continuous: true,
            #position: [
                   (#accn: “NM_001619”,
                   #start: 0, #end: 3602,
                   #negative: false)],
            #anno: [
                   (#anno_name: “organism”,
                   #descr: “Homo sapiens”), ... ]),
            ...}, ...)}
```

Data warehousing

In addition to querying data sources on the fly, biologists and biotechnology companies must create their own customized data warehouses. Several factors motivate such warehouses:

- Query execution can be more efficient, assuming data reside locally on a powerful database system.
- Query execution can be more reliable, assuming data reside locally on a high-availability database system and a high-availability network.
- Query execution on a local warehouse avoids unintended denial-of-service attacks on the original sources.
- Most importantly, many public sources contain errors. Some of these errors can't be corrected or detected on the fly. Hence, humans—perhaps assisted by computers—must cleanse the data, which are then warehoused to avoid repeating this task.

A biological data warehouse should be efficient to query, easy to update, and should model data naturally. This last requirement is important because biological data, such as the GenBank report in Figure 2, have a complex nesting structure. Warehousing such data in a radically different form tends to complicate their effective use.

Biological data's complex structure makes relational database management systems such as Sybase unsuitable as a warehouse. Such DBMSs force us to fragment our data into many pieces to satisfy the third normal form requirement. Only a skilled expert can perform this normalization process correctly. The final user, however, is rarely the same expert. Thus, a user wanting to ask questions on the data might first have to figure out how the original data was fragmented in the warehouse. The fragmentation can also pose efficiency problems, as a query can cause the DBMS to perform many joins to reassemble the fragments into the original data.

Kleisli can turn a relational DBMS into a nested relational DBMS. It can use flat DBMSs such as Sybase, Oracle, and MySQL as its updateable complex object store. In fact, it can use all of these varieties of DBMSs simultaneously. This capability makes Kleisli a good system for warehousing complex biological data. XML, with its built-in expressive power and flexibility, is also a great contender for biological data warehousing. More recently, some commercial relational DBMSs such as Oracle have begun offering better support for complex objects. Hopefully, they'll soon be able to perform complex biological data warehousing more conveniently and naturally.

The Details: Further Readings

Data integration

➤ L. Wong, "Technologies for Integrating Biological Data," *Briefings in Bioinformatics*, vol. 3, no. 4, 2002, pp. 389–404.

➤ L. Wong, "Kleisli, a Functional Query System," *J. Functional Programming*, vol. 10, no. 1, 2000, pp. 19–56.

➤ S. Davidson and colleagues, "BioKleisli: A Digital Library for Biomedical Researchers," *Int'l J. Digital Libraries*, vol. 1, no. 1, Apr. 1997, pp.36–53.

➤ F. Achard, G. Vaysseix, and E. Barillot, "XML, Bioinformatics and Data Integration," *Bioinformatics,* vol. 17, no. 2, 2001, pp. 115–125.

Biological sequence analysis

➤ F. Zeng, R. Yap, and L. Wong, "Using Feature Generation and Feature Selection for Accurate Prediction of Translation Initiation Sites," *Proc. 13th Int'l Conf. Genome Informatics*, Universal Academy Press, 2002, pp. 192–200.

➤ H. Liu and L. Wong, "Data Mining Tools for Biological Sequences," *J. Bioinformatics and Computational Biology*, vol. 1, no. 1, 2003, pp. 139–168.

Gene expression analysis

➤ J. Li and colleagues, "Simple Rules Underlying Gene Expression Profiles of More Than Six Subtypes of Acute Lymphoblastic Leukemia (ALL) Patients, *Bioinformatics*, vol. 19, 2003, pp. 71–78.

➤ J. Li and L. Wong, "Identifying Good Diagnostic Genes or Genes Groups from Gene Expression Data by Using the Concept of Emerging Patterns," *Bioinformatics*, vol. 18, 2002, pp. 725–734.

Scientific literature mining

➤ S.-K. Ng and M. Wong, "Toward Routine Automatic Pathway Discovery from Online Scientific Text Abstracts," *Genome Informatics*, vol. 10, Dec. 1999, pp. 104–112.

➤ L. Wong, "Pies, A Protein Interaction Extraction System," Proc. Pacific Symp. Biocomputing, *World Scientific*, 2001, pp. 520–531.

KNOWLEDGE DISCOVERY

As we entered the era of postgenome knowledge discovery, scientists began asking many probing questions about the genome data such as, "What does a genome sequence do in a cell?" and, "Does it play an important role in a particular disease?" The genome projects' success depends on the ease with which they can obtain accu-

Figure 3. Recognizing translation initiation sites.

```
299 HSU27655.1 CAT U27655 Homo sapiens
CGTGTGTGCAGCAGCCTGCAGCTGCCCCAAGCCATGGCTGAACACTGACTCCCAGCTGTG   80
CCCAGGGCTTCAAAGACTTCTCAGCTTCGAGCATGGCTTTTGGCTGTCAGGGCAGCTGTA   160
GGAGGCAGATGAGAAGAGGGAGATGGCCTTGGAGGAAGGGAAGGGGCCTGGTGCCGAGGA   240
CCTCTCCTGGCCAGGAGCTTCCTCCAGGACAAGACCTTCCACCCAACA...........
```

What makes the second ATG the translation initiation site?

rate and timely answers to these questions. Informatics therefore plays a more important role in upstream genomic research.

Three case studies illustrate how informatics can help turn a diverse range of biological data into useful information and valuable knowledge. This can include recognizing useful gene structures from biological sequence data, deriving diagnostic knowledge from postgenome experimental data, and extracting scientific information from literature data. In all three examples, researchers used various IT techniques plus some biological knowledge to solve the problems effectively. Indeed, bioinformatics is moving beyond data management into a more involved domain that often demands in-depth biological knowledge; postgenome bioinformaticists are now required to be not just computationally sophisticated but also biologically knowledgeable.

Biological sequence analysis

In addition to having a draft human genome sequence (thanks to the Human Genome Project), we now know many genes' approximate positions. Each gene appears to be a simple-looking linear sequence of four letter types (or *nucleotides*)—As, Cs, Gs, and Ts—along the genome. To understand how a gene works, however, we must discover the gene's underlying structures along the genetic sequence, such as its transcription start site (point at which transcription into nuclear RNA begins), transcription factor binding site, translation initiation site (point at which translation into protein sequence begins), splice points, and poly(A) signals. Many genes' precise structures are still unknown, and determining these features through traditional wet-laboratory experiments is costly and slow. Computational analysis tools that accurately reveal some of these features will therefore be useful, if not necessary.

Informatics lets us solve the TIS recognition problem using computers. Translation is the biological process of synthesizing proteins from mRNAs. The TIS is the region where the process initiates. As Figure 3 shows, although a TIS starts with the three-nucleotide signature "ATG" in cDNAs, not all ATGs in the genetic sequence are translation start sites. Automatically recognizing which of these ATGs is a gene's actual TIS is a challenging machine-learning problem.

In 1997, Pedersen and Nielsen addressed this problem by applying an artificial neural network (ANN) trained on a 203-nucleotide window. They obtained results of 78-percent sensitivity and 87-percent specificity, giving an overall accuracy of 85 percent. In 1999 and 2000, Zien and colleagues worked on the same problem using support vector machines instead. Combining the support vector machine (SVM) with polynomial kernels, they achieved performance similar to Pedersen and Nielsen. When they used SVM with specially engineered locality-improved kernels, they obtained 69.9-percent sensitivity and 94.1-percent specificity, giving an improved overall accuracy of 88.1 percent.

Because the accuracy obtained by these and many other systems is already sufficiently high, much of today's research on the TIS recognition problem aims to better understand TISs' underlying biological mechanisms and characteristics.

Our approach comprises three steps:

- feature generation,
- feature selection, and
- feature integration by a machine-learning algorithm for decision-making.

This approach achieves 80.19-percent sensitivity and 96.48-percent specificity, giving an overall accuracy of 92.45 percent. Furthermore, it yields a few explicit features for understanding TISs, such as:

- The presence of A or G three nucleotides to a target ATG is favorable for translation initiation.
- The presence of an in-frame ATG upstream near a target ATG is unfavorable for translation initiation.
- The presence of an in-frame stop codon (a three-nucleotide signature that signals termination of the translation process) downstream near a target ATG is also unfavorable for translation initiation.

Such understanding of biological patterns acquired by machine-learning algorithms is becoming increasingly important as the bioinformatics endgame elevates into the discovery of new knowledge and providing accurate computation results is no longer sufficient. Bioinformatics users require explainable results and usable decision rules instead of unexplained yes/no output.

Gene expression analysis

Medical records analysis is another postgenome application aimed mainly at diagnosis, prognosis, and treatment planning. Medical records also require understandable outputs from machine-learning algorithms. Here we're looking for patterns that are

- *Valid.* They also occur in new data with high certainty.
- *Novel.* They aren't obvious to experts and provide new insights.
- *Useful*. They enable reliable predictions.
- *Understandable*. They pose no obstacle in their interpretation, particular by clinicians.

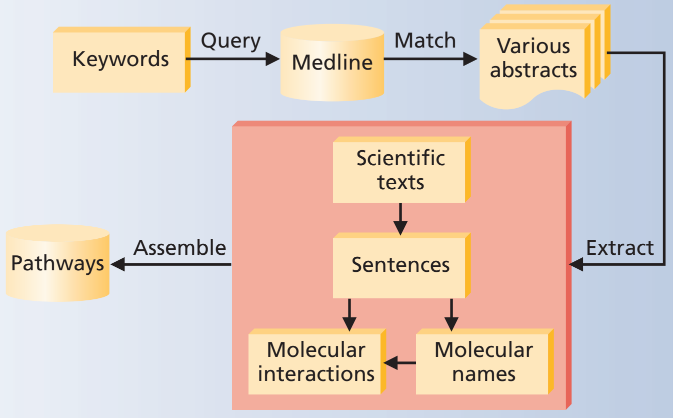
Scientists now use microarrays (miniaturized 2D arrays of DNA or protein samples, typically on a glass slide or microchip, that can be tested with biological probes) to measure the expression level of thousands of genes simultaneously. The gene expression profiles thus obtained might help us understand gene interactions under various experimental conditions and the correlation of gene expressions to disease states, provided we can successfully achieve gene expression analysis. Gene expression data measured by microarrays or other means will likely soon be part of patients' medical records.

Many methods for analyzing medical records exist, such as decision-tree induction, Bayesian networks (a class of probabilistic inference networks) neural networks, and SVMs. Although decision trees are easy to understand, construct, and use, they're usually inaccurate with nonlinear decision boundaries. Bayesian networks, neural networks, and SVMs perform better in nonlinear situations. However, their resultant models are "black boxes" that might not be easy to understand and therefore limited in their use for medical diagnosis.

PCL is a new data-mining method combining high accuracy and high understandability. It focuses on fast techniques for identifying patterns whose frequencies in two classes differ by a large ratio—the *emerging patterns*—and on combining these patterns to make a decision.

The PCL classifier effectively analyzes gene expression data. One successful application was the classification of heterogeneous acute lymphoblastic leukemia (ALL) samples. Accurately classifying an ALL sample into one of six known subtypes is important for prescribing the right treatment for leukemia patients and thus enhancing their prognosis. However, few hospitals have all the expertise necessary to correctly diagnose their leukemia patients. An accurate and automated classifier such as PCL, together with microarray technologies, could lead to more accurate diagnoses.

Figure 4. Mining literature for protein interactions.



We've tested PCL on a data set consisting of gene expression profiles of 327 ALL samples, obtained by hybridization on the Affymetrix U95A GeneChip microarray containing probes for 12,558 genes. The samples contain all the known ALL subtypes. We used 215 samples as training data for constructing the classification model using PCL and 112 samples for blinded testing. PCL made considerably fewer false predictions than other conventional methods. More importantly, the top emerging patterns in the PCL method also serve as high-level rules for understanding the differences between ALL subtypes. Hospitals can also use these rules to suggest treatment plans.

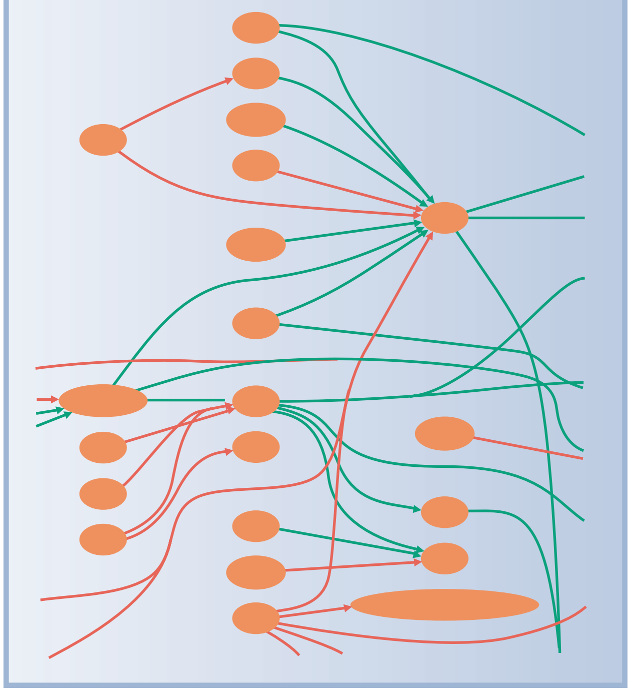
Scientific literature mining

Other than the molecular sequence databases generated by the genome projects, much of the scientific data reported in the literature have not been captured in structured databases for easy automated analysis. For instance, molecular interaction information for genes and proteins is still primarily reported in scientific journals in free-text formats.

Molecular interaction information is important in postgenome research. Biomedical scientists have therefore expended much effort in creating curated online databases of proteins and their interactions, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG, www.kegg.org) and the Cell Signaling Networks Database (CSNDB, geo.nihs.go.jp/csndb). However, such hand-curated databases are laborious and unlikely to scale.

Natural language processing (NLP) of biomedical literature is one alternative to manual text processing. Figure 4 shows a typical workflow for mining the biomedical literature for protein interaction pathways. The system collects numerous abstracts and texts from biological research papers in scientific literature databases such as NCBI's

Figure 5. Pathway extracted by Pies.



Medline, the main online biomedical literature repository. It then applies NLP algorithms to recognize names of proteins and other molecules in the texts.

Sentences containing multiple occurrences of protein names and some action words—such as "inhibit" or "activate"—are extracted. Natural language parsers then analyze the sentences to determine the exact relationships between the proteins mentioned. Lastly, it automatically assembles these relationships into a network for us, so we know exactly which protein is acting directly or indirectly on which other proteins and in what way.

Pies is one of the first systems capable of analyzing and extracting interaction information from English-language biology research papers. Pies is a rule-based system that recognizes names of proteins and molecules and their interactions. Figure 5 shows approximately 20 percent of the system's output given a protein Syk with a pathway of interest. Pies downloaded and examined several hundred scientific abstracts from Medline, recognizing several hundred interactions involving hundreds of proteins and molecules mentioned in the abstracts.

Understandably, the complex nature of linguistics and biology makes biomedical text mining especially difficult. This challenging task has recently attracted increased interest from the bioinformatics and other computational communities (such as computational linguistics). Hopefully, a combined effort by researchers in bioinformatics and other information technologies will fill some of the gaps.

The future of molecular biology and biomedicine will greatly depend on advances in informatics. As we review researchers' many achievements in bioinformatics, we're confident that the marriage between molecular biology and information technology is a happy one. Accomplishments in bioinformatics have advanced molecular biology and information technology. Although many computational challenges lie ahead, more fruitful outcomes of this successful multidisciplinary marriage are likely. ■

See-Kiong Ng *is head of the Decision Systems Laboratory at the Institute for Infocomm Research, Singapore. Contact him at skng@i2r.a-star.edu.sg.*

Limsoon Wong *is deputy executive director, research, of the Institute for Infocomm Research. Contact him at limsoon@i2r.a-star.edu.sg.*