PREFACE

Expert Commentary - Expressed sequence tag (EST) databases are a well established and continuously growing source to study gene expression, alternative splicing, genome sequences, gene-associated polymorphisms and sequence homologies through bioinformatic approaches. Here, the authors examine recent efforts to identify and characterize cancer genes and tumor markers using ESTs. Limitations of EST mining strategies and directions for future research are also summarized.

Short Commentary A - The field of protein bioinformatics analyzes the sequence and structure of protein; it plays a critical role in the discovery of small therapeutic agents as well as protein drugs. Here the authors present our recent progress in this field, including the new concepts of concavity druggability and antibody druggability, which are expected to raise the drug discovery success ratio.

Short Commentary B - Linkage disequilibrium (LD), the nonrandom association of alleles from different loci, can provide valuable information about the structure of human genome haplotypes. Because haplotype-based methods offer a powerful approach for disease gene mapping, this information may facilitate studies of the association between genomic variation and human traits. Single nucleotide polymorphism (SNP) alleles at multiple loci produce an LD pattern resulting from gene–gene interactions that can provide a foundation for developing statistics to detect other such interactions. Although several studies have used LD to address the role of gene interactions in various phenotypes and complex diseases, the current lack of formal statistics and the potential importance of data resulting from this research have motivated us to develop LD-based statistics. The authors chose to examine skin pigmentation because it is a complex trait, and SNP alleles at multiple loci may play a role in determining normal variation in skin pigmentation. The main purpose of this chapter is to outline the development of LD-based statistics for detecting interactions among SNP alleles at multiple loci that contribute to variation in human skin pigmentation. To accomplish this, the authors developed a general theory to study LD patterns in gene-interaction trait models. They then developed a definition of gene interaction and a measure of interactions among SNP alleles at multiple loci contributing to the trait in the framework of LD analysis.

Short Commentary C - In this paper, the authors use the penalty approach in order to study two constrained minimization problems on complete metric spaces. A penalty function is said to have the generalized exact penalty property if there is a penalty coefficient for which approximate solutions of the unconstrained penalized problem are close enough to

approximate solutions of the corresponding constrained problem. In this paper they establish sufficient conditions for the generalized exact penalty property.

Short Commentary D - With the advent of computational research, several applications in science can be seen. The application in medicine is also documented. Computational medicine study can help answer a complicated query in medicine within a short period. How to create a computational medicine study is a common question from the beginner. In this article, the author will describe the steps for creating computational medicine research. Briefly, a simple process as that for simple in vivo and in vitro research can be used. The setting up of a conceptional framework based on literature review is the first necessary step. Next, selection of the proper database and tool for manipulation is needed. Simulating based on the designed framework can help one reach the answer. These steps must be thoroughly followed to complete computational medicine research.

Short Commentary E - *Background*: Human cancer is often classified according to the anatomic site at which it occurs, and researchers are often taught these cancer types are actually a spectrum of disease. A review in 2000 (Hanahan and Weinberg; Cell 2000 100:57-70) reported that all cancers share six characteristics: (1) self-sufficiency in growth signaling, (2) the ability to ignore external anti-growth signals, (3) the ability to avoid apoptosis, (4) sustained angiogenesis, (5) the capacity for limitless reproduction and (6) the ability to invade tissue and spread to other anatomic sites. Our goal was to identify related cancer types using different observational strategies. *Methods*: The authors employed one method that used text-mining of online information about genes and disease. A second method used medical records of patients in British Columbia who were diagnosed with multiple cancer types between 1970 and 2004. A third method correlated Canadian provincial cancer rates for various cancer types. *Results*: Several pairs of related cancer types were identified using each method, although no pair was identified by all three strategies. The pairs of cancer types lung/bladder and lung/kidney were both identified by the text-mining and correlation studies. Esophageal cancer and melanoma were identified as related cancer types by both the analysis of patients with multiple primary cancers and the correlation study. Discussion: If cancer types are related, patients with one cancer might increase surveillance for other related cancer types, and drugs that are effective for treating one cancer might be successfully adapted for the related cancer types.

Chapter 1 - High-throughput laboratory measurement technologies, including microarrays for genomics and proteomics, are now typically used in biomedical studies ranging from animal models to human clinical trials. These methods, such as gene expression microarrays, aim to capture the global expression patterns of thousands of genes or proteins simultaneously. A common feature is the high-dimensionality of the resulting data. Post-study analytical challenges involve methods to extract the meaningful information from the millions of data points. However, there is a need to develop systematic approaches to planning such studies. In this work, the authors provide a synthesis of the available and current trends in sample size and power analysis for genomics studies. Their emphasis will be on clarifying assumptions of the available methods as well as their applicability in practice, including the assumption of independent gene expression. The authors also emphasize emerging sample size design methods that focus on the false discovery rate (FDR) over the traditional family-wise error rate as a criterion.

Chapter 2 - Completion of the human genome sequencing has provided us with opportunities to understand the molecular complexity of the human body. The transcriptional

regulatory circuits of gene expressions are one of the most promising matters to be resolved by exploring the human genome sequence. The authors have been interested in human cell fate regulated by the transcription factor E2F. To accelerate the investigation, the authors need to develop a strategy that can efficiently identify E2F target genes. Basically, their approach is to combine computational and experimental analysis. Annotated data of gene expression profiles deposited in the public database and knowledge accumulated in the published literature are a treasure-house of E2F candidate target genes. Next, a promoter region based on the information of the transcriptional start site can be used for motif searching of E2F regulatory elements. Finally, a set of predicted E2F regulatory elements are tested by molecular biological and biochemical assays. In this chapter, the author gives a basic introduction of our recent strategy for computational and experimental analysis for the prediction of transcription factor E2F regulatory elements in the human gene promoter. In addition, recent progress in unrevealing E2F functions achieved by genome wide approaches is discussed.

Chapter 3 – In this chapter, the authors investigate the generalized assignment problem with the objective of finding a minimum-cost assignment of jobs to agents subject to capacity constraints. A complicating feature of the model is that the coefficients for resource consumption and capacity are random. The problem is formulated as a stochastic integer program with a penalty term associated with the violation of the resource constraints and is solved with a branch-and-price algorithm that combines column generation with branch and bound. To speed convergence, a stabilization procedure is included. The performance of the methodology was tested on four classes of randomly-generated instances. The principal results showed that the value of the stochastic solution (VSS), i.e., the gap between the stochastic solution and the expected value solution, was 35.5% on average. At the root node of the search tree, it was found that the linear programming relaxation of the master problem associated with column generation provided a much tighter lower bound than the relaxation of the original constraint-based formulation. In fact, two thirds of the test problems evidenced no gap between the optimal integer solution and the relaxed master problem solution. Additional testing showed that (1) initializing the master problem with a feasible solution out performs the classical big-M approach; (2) SOS type 1 branching is superior to single-variable branching; and, (3) variable fixing based on reduced costs provides only a slight decrease in runtimes.

Chapter 4 - The analysis of the evolutionary relationships among biological sequences, known as phylogenetics, constitutes one of the most powerful tools of computational biology. Besides its classical use to ascertain the evolution of a group of species, phylogenetics has many other applications such as the prediction of the function of a protein and the detection of genes under specific selective constrains. The advent of the genome era has brought about the possibility of extending such analyses to larger sets comprising thousands of sequences from complete genomes. The use of whole genomes, rather than that of reduced sets of genes or proteins, opens the door to a wide range of new possibilities. On the other hand, however, it poses many conceptual and technical challenges that require the development of new algorithms to interpret and manipulate large-scale phylogenetic data. Here the authors survey recent progress in the development of automated pipelines to reconstruct and analyze large collections of phylogenetic trees and provide some examples of how they have been used to address important biological questions.

Chapter 5 - A thorough understanding of electrostatic, elastic and topological behaviour of DNA has provided some relevant mechanistic insights into the regulation of genetic expression. Although this approach has proved valuable for the study of many biological processes, it is limited by the simple description level represented by DNA. Indeed, genomic DNA in eukaryotic cells is basically divided into chromosomes, each consisting in a single huge chromosomal fiber hierarchically supercoiled. Since this organisation plays a critical role in all processes involved in DNA metabolism, tremendous efforts have been done to build relevant models of chromatin structure and dynamics. Namely, by shifting from a DNA (as a simple molecular polyelectrolyte) point of view to a chromatin (as a polymorph supramolecular nucleoprotein complex) point of view, we should go towards more efficient mechanistic framework in which the control of genetic expression and other DNA metabolism processes could be interpreted. This review gives an historical overview of the progresses that have been done during the last 30 years in this field, and discusses what the most challenging outcomes are now.

Chapter 6 – The authors present an algorithm to reconstruct protein C_α traces from 4-class distance maps, and benchmark it on a non-redundant set of 258 proteins of length between 51 and 200 residues. They first represent proteins as contact maps, and show that even when exact maps are available, only low-quality models can often be obtained. The authors then adopt a more powerful simplification of distance maps: multi-class contact maps. They show that the reconstructions based on 4-class native maps are significantly better than those from binary maps. Furthermore, the authors build two predictors of 4-class maps based on recursive neural networks: one ab initio, or relying on the sequence and on evolutionary information; one in which homology information is provided as a further input, showing that even very low sequence similarity to PDB templates yields more accurate maps than the ab initio predictor. They reconstruct C_α traces based on both ab initio and homology-based 4-class map predictions. The authors show that homology-based predictions are generally more accurate than ab initio ones even when homology is dubious.

Chapter 7 - Understanding the protein mechanics is *a priori* requisite for gaining insight into protein's biological functions, since most protein performs its function through the structural deformation renowned as conformational change. Such conformational change has been computationally delineated by atomistic simulations, albeit the mechanics of large protein structure is computationally inaccessible with atomistic simulation such as molecular dynamics simulation. In a recent decade, normal mode analysis with coarse-grained modeling of protein structures has been a computational alternative to atomistic simulations for understanding large protein mechanics. In this review, the authors delineate the current state-of-art in coarse-grained modeling of proteins for normal mode analysis. Specifically, the pioneered coarse-grained models such as Go model and elastic network model as well as recently developed coarse-grained elastic network model are summarized and discussed for understanding large protein mechanics.

Chapter 8 - The aim of this study was to differentiate between superficial and advanced bladder cancers by analyzing the gene expression profiles of these tumors by using the self-organizing maps (SOMs). The authors also used the GoMiner software for the biological interpretation of 473 interesting genes.

Materials and Methods: Between December 2003 and November 2004, 17 patients with urothelial bladder cancers who were admitted to the Chang Gung Memorial Hospital for

transurethral resection of the tumor were included in this study. The gene expression data included 7400 cDNAs in 17 arrays. The software, GeneCluster 2.0, was used for analyzing gene expression data by using SOMs. The authors used a 2-cluster SOM to cluster automatically a set of 17 tissues samples into superficial and advanced bladder cancers based on the gene expression patterns. The authors also used the GoMiner software for the biological interpretation of top 473 interesting genes.

Results: Patients included 11 males and 6 females. Pathological studies confirmed the presence of superficial tumors in 9 patients and advanced tumors in 8 patients. Of the 7400 genes analyzed, 473 genes showed significant changes in their expression. Of these 268 were up-regulated, and 205 were down-regulated. Using the top 473 genes, SOMs were used to differentiate between the gene expression patterns of superficial and advanced bladder cancer tissue samples. The patient tissue samples were clustered into 2 groups, namely, superficial and advanced bladder cancers, comprising 10 and 7 samples, respectively. Only one patient tissue sample with advanced bladder cancer was clustered into the superficial bladder cancer group. This analysis had a high accuracy rate of 94% (16/17). The top 473 genes also were classified into biologically coherent categories by the GoMiner software. The results revealed that 452, 435, and 452 genes were associated with biological processes, cellular components, and molecular functions, respectively.

Conclusion: Based on the authors' results, they believe that superficial and advanced urothelial bladder cancers can be differentiated by their gene expression profiles analyzed by SOMs. The SOM method may be used on microarray data analysis to distinguish tumor stages and predict clinical outcomes. The genes that are uniquely expressed in either stage of bladder cancer can be considered as possible candidates for biomarkers.

Chapter 9 - New technologies for collecting genotypic data from natural populations open the possibilities of investigating many fundamentalbiological phenomena, including behavior, mating systems, heritabilities of adaptive traits, kin selection, and dispersal patterns. The power and potential of genotypic information often rests in the ability to reconstruct genealogical relationships among individuals. These relationships include parentage, full and half-sibships, and higher order aspects of pedigrees. Some areas of genealogical inference, such as parentage, have been studied extensively. Although methods for pedigree inference and kinship analysis exist, most make assumptions that do not hold for wild populations of animals and plants. In this chapter, the authors focus on the full sibling relationship and first review existing methods for full sibship reconstructions from microsatellite genetic markers. The authors then describe our new combinatorial methods for sibling reconstruction based on simple Mendelian laws and its extension even in the presence of errors in the data. They also describe a generic consensus method for combining sibling reconstruction results from other methods. They present experimental comparison of the best existing approaches on both biological and simulated data. The authors discuss relative merits and drawbacks of existing methods and suggest a practical approach for reconstructing sibling relationships in wild populations.

Chapter 10 - Microarray gene expression profiling, which monitors the expression of tens of thousands of genes simultaneously, is a promising tool for developing prognostic markers for cancer patients. Many researchers have applied microarray gene expression profiling in order to develop better prognostic markers, and have demonstrated promising results in many types of cancer. Unfortunately, there are concerns regarding the premature clinical use of newly-developed prognostic gene signatures, as problems associated with their application

remain unresolved, diminishing the reliability of their intended results. This review first discusses these presently unsolved problems in the development of prognostic gene signatures. Recent computational approaches to circumventing these problems are then presented, and therein the authors discuss these approaches in the categorized framework of mechanism-derived bottom-up approaches, meta-analytic approaches, integrative approaches that combine genomics and clinical data, and sub-type-specific analysis approaches. The authors believe that recent bioinformatics approaches, which integrate rapidly accumulating genomics, clinical, and other forms of data, will help overcome current problems, and will help realize the successful application of prognostic gene signatures in personalized medicine.

Chapter 11 - The authors calculate time of folding and explore the transition state ensembles for ten proteins with known experimental data at the point of thermodynamic equilibrium between unfolded and native state using a Monte Carlo Gō model and Dynamic programming where each residue is considered to be either folded as in the native state or completely disordered. The order of events in folding simulations has been explored in detail for each of the proteins. The times of folding for ten proteins which reach the native state within a limit of 10^8 Monte Carlo steps are in a good correlation with experimentally measured folding time at mid-transition point (the correlation coefficient is 0.71). A lower correlation was obtained if to use Dynamic programming approach (the correlation coefficient is 0.53). Moreover, Φ-values calculated from the Monte Carlo simulations for ten proteins correlate with experimental data (the correlation coefficient is 0.41) practically at the same level as Φ-values calculated from Dynamic programming approach (the correlation coefficient is 0.48). The model provides good prediction of folding nuclei for proteins whose 3D structures have been determined by X-ray, and exhibits a more limited success for proteins whose structures have been determined by NMR.

Chapter 12 - The structural class of a given protein represents the first level in the hierarchical structural classification. Its knowledge starts the progressive identification of the next levels, which allows to relate the protein to a family, in evolutionary as well as functional terms. A number of computational methods have been proposed to predict the structural class based on primary sequences. Most of the prediction methods use simple sequence representations such as composition vectors and polypeptide composition or also more advanced representations that combine physico-chemical properties and sequence composition. Moreover, different classification algorithms, including neural network, rough sets and logistic regression, and the application of complex classification models, such as ensembles, bagging and boosting, have been recently used.

However, the accuracy of all these methods is strongly affected by sequence similarity. Some algorithms were tested on small datasets with high sequence identity percentage, which results in an overestimated prediction accuracy. On the other hand, low similarity sequences pose a substantial challenge.

The main aim of this paper is to present the state of the art in this field, describing some methods developed in the last years for the prediction of the protein structural class and underlining the need of using protein datasets of varying similarity and new testing procedures in order to evaluate correctly the quality and the accuracy of new prediction methods.

Chapter 13 - The computational process is based on the fundamental semiotic activity linking mathematical equations to the materialized physical world. Its limits are defined by the set of imposed physical values constituting the background structure of the Universe.

Computation becomes a direct consequence of semiosis, in the case where the arbitrariness of semiotic signs appears strictly defined in the semiotic context. This results in emergence of an internal formal structure of the system that can be modeled and computed. The authors consider formalization of the Peircean semiotics in the framework of Peirce algebra as a prerequisite of the understanding of fundamentals of computation. In reproducible semiotic structures such as biological entities, the factor that makes these systems "closed to efficient causation" (in Robert Rosen's sense) is a basic element of the Peirce algebra that provides semantic closure to the system via introduction of the parameter of organizational invariance. Approaches to define this factor as a set of dual components, related both to relations and sets, are discussed in the frames of the semiotic interpretation of quantum mechanics where actualization is explained as a signification process within the network of quantum measurements. In this framework, enzymes are molecular automata, the set of which maintains highly ordered robust coherent state of the quantum computer, and genome concatenates error-correcting codes into a single reflective set that can be described by the Peirce algebra. The biological evolution can be viewed as a functional unfolding of the organizational invariance constraints in which new limits of iteration emerge, possessing criteria of perfection and having selective values.

Chapter 14 - Extraction of functional sequences from promoters has been achieved with alignment-based motif search algorithms, but recently, several new approaches have been developed. In this review, I would like to introduce a methodology called as the LDSS (Local Distribution of Short Sequences) analysis. This approach evaluates distribution profiles of short sequences along the promoter region, and sequences that preferentially appear at specific promoter regions are extracted. Application of this strategy to human, mouse, *Drosophila*, *C. elegans*, *Arabidopsis*, rice, and also yeast resulted in successful extraction of both well known promoter elements and also novel putative elements. This method is so sensitive that various kinds of minor elements can be simultaneously detected by analysis of all the promoters of a genome as one batch. However, the LDSS analysis does not detect all the elements involved in transcriptional regulation, but position-insensitive elements are out of range of the analysis. No need of microarray data for the analysis enables its application to wide range of species beyond the model species.

Chapter 15 - Combinatorial peptide library technology has been proven to be a powerful approach to both T-cell epitope determination and analysis of TCR specificity and degeneracy. During the past ten years, the authors have developed mathematical models and bioinformatics approaches for deconvolution of positional scanning synthetic combinatorial libraries (PS-SCL). PS-SCL are composed of trillions of peptides systematically arranged in mixtures with defined amino acids at each position. Starting from the mathematical model building to deconvolute the spectrum of PS-SCL, the authors proposed a biometrical approach using the score matrix to systematically search the protein databases for putative antigenic peptide candidates. The authors then evaluated of the assumption of independent contribution of the side chains of the amino acids in peptides and applied more sophisticated machine learning algorithms to improve the prediction accuracy based on synthesized peptide data. Finally, they implemented the above approach into a web based tool for searching protein database for T-cell epitopes based on experimental data from PS-SCL with the website employed a strong statistical analysis package, relational database and Java applets. The authors' work has provided a sound basis for PS-SCL data analysis and has been proven efficient and successful for identifying target antigens and highly active peptide mimics.

Chapter 16 - To improve the flexibility and extensibility of application programs, a method to support scripting function by embedding a Lua language interpreter is described. Using this approach, variations of input data and parameters for calculations are supported by a script file without rewriting the application program. For instance, the script information is supplied to the program and internal data can also be exported to a script for extended calculations. This chapter summarizes the basic framework of embedding this scripting language to interact with existing codes using the application programming interface provided by Lua.

The method was applied to the molecular structure viewer program MOSBY to support additional visualizations and calculations from atomic coordinate data by script programs. Atomic structure data in the original C structure were mapped to a Lua script by using a mechanism called "metamethod" in Lua. In addition, the table data type in Lua provides a simple database useful for a configuration of molecular graphics. The design of a "domain-specific language" in biocomputing is discussed with reference to other scripting languages.

Chapter 17 - At present, the third wave of medical experiments, in silico or computational simulating, is accepted as a powerful tool to drive the medical society into the new post genomics phase. Computational biology research is an important facet of bioinformatics. In this article, the author presents the concept and shares the experience in computational hematology research. Cases on hemoglobin and prothrombin disorders are demonstrated. Briefly, the computational research can help understand the genome, proteome and expression of hemoglobin and prothrombin disorders.