Preface

Computational methods are essential for analysing biological data because major developments in molecular biology and advances in high-throughput genomic technologies have led to an explosive growth in the amount and complexity of information that is routinely collected. Computational biology, bioinformatics and systems biology have evolved in response to this challenge, to cater to various tasks related to exploration, and thereby to generate knowledge or hypotheses.

Many problems in the above areas are closely related to different tasks of pattern recognition and machine learning. Much of biological data is noisy and has missing values. Data-cleaning and missing-value estimation are essential in such situations. Gene expression data need efficient feature selection methods for identifying a few genes that are of interest, thereby reducing the problems of dimensionality. For the purpose of decision-making, classification, clustering and prediction methodologies are necessary. Examples include gene function prediction, protein classification and microRNA target prediction. Clustering is used as one of the basic exploratory data-processing methods for problems such as sequence grouping, identification of co-expressed genes, and protein module extraction. Similarly, there are optimization problems galore in drug design and many other areas.

In this special issue, we have focussed on the design and application of new and improved techniques of pattern recognition and machine learning. They are important for gaining deeper biological insights from the huge amount of data collected. The issue provides a wealth of information for academicians, practitioners and students working in computational biology and bioinformatics, systems biology, pattern recognition, and machine learning. Extended versions of some selected articles of PReMI-2013 (the 5th International Conference on Pattern Recognition and Machine Intelligence, Kolkata, December 2013) have been considered for review, in addition to other invited ones. Topics considered include pattern recognition and machine-learning approaches for sequence analysis, microarray data analysis, biochemical pathway analysis, NGS data analysis, microRNA data analysis, classification of diseases and analysis of comorbid diseases and of data related to evolutionary biology. All submissions have undergone the journal's peer-review procedures.

There are 13 papers in this special issue. In the first paper titled 'pubmed.mineR: An R package with text-mining algorithms to analyse PubMed abstracts', the authors have developed an R package, pubmed.mineR, where they have combined the advantages of existing algorithms, overcome their limitations, offered user flexibility and linked with other packages in Bioconductor and The Comprehensive R Network (CRAN) in order to expand the user capabilities for executing multifaceted approaches. Three case studies are presented, namely, evolving role of diabetes educators, Cancer Risk Assessment and dynamic concepts on disease and comorbidity to illustrate the use of pubmed.mineR. This work is helpful to mine the PubMed literature database, a valuable source of information for scientific research.

The representation of proteins as networks of interacting amino acids, referred to as protein contact networks (PCN), and their subsequent analysis using graph theoretic tools, can provide novel insights into the key functional roles of specific groups of residues. In the second article titled 'Analysis of core–periphery organization in protein contact networks reveals groups of structurally and functionally critical residues', the authors have characterized the networks corresponding to the native states of 66 proteins (belonging to different families) in terms of their core–periphery organization. The resulting hierarchical classification of the amino acid constituents of a protein arranges the residues into successive layers – having higher *core order* – with increasing connection density, ranging from a sparsely linked *periphery* to a densely intra-connected *core* (distinct from the earlier concept of protein core defined in terms of the three-dimensional geometry of the native state).

The third paper titled 'MIPCE: An MI-based protein complex extraction technique' deals with protein–protein interactions. Identifying protein complexes is of great importance for understanding cellular organization and functions of organisms. In this article, a method, called MIPCE, has been proposed to identify protein complexes in a PPI network, based on mutual information (MI). MIPCE has been biologically validated by GO-based score, and satisfactory results have been obtained.

A canonical correlation analysis (CCA) has been done, as described in the fourth paper titled 'DNA pattern recognition using canonical correlation algorithm', to describe related views of the same semantic object for identifying patterns. The method determines

Published online: 28 September 2015

the required genetic code in the DNA sequence. CCA finds correlations between two observations of the same semantic pattern and test sequence. It is concluded that a relationship possesses maximum value in the position where the pattern exists. As a case study, the potential of CCA is demonstrated on the sequence found from HIV-1 preferred integration sites.

Reduction of dimensionality is a routine process to get rid of curse of dimensionality, and has emerged as a task involved in modelling complex biological systems. In the fifth article titled 'Feature selection using feature dissimilarity measure and density-based clustering: Application to biological data', an unsupervised feature selection technique is proposed, using maximum information compression index as the dissimilarity measure and the well-known DBSCAN, a density-based cluster identification technique, for identifying the largest natural group of dissimilar features. The algorithm is fast and less sensitive to user-supplied parameters. Moreover, the method automatically determines the required number of features and identifies them.

Use of computational methods to predict gene regulatory networks (GRNs) from gene expression data is a challenging task. In the sixth paper titled 'Semi-supervised prediction of gene regulatory networks using machine learning algorithms', a semi-supervised method has been developed for GRN prediction by utilizing two machine learning algorithms, viz., support vector machines (SVM) and random forests (RF). Both inductive and transductive learning approaches have been adopted to obtain reliable negative training data from the unlabelled data. The proposed semi-supervised methods have been applied to gene expression data of *Escherichia coli* and *Saccharomyces cerevisiae*, and their performance evaluated.

The seventh paper, titled 'Identification of certain cancer-mediating genes using Gaussian fuzzy cluster validity index', describes a novel cluster validity index, developed based on the notion of fuzzy sets, for validating the clusters obtained by a clustering algorithm applied on cancer gene expression data. Gaussian fuzzy index (GFI) has then been used for the identification of genes that have altered quite significantly from normal state to carcinogenic state with respect to their mRNA expression patterns. The effectiveness of the methodology has been demonstrated on three gene expression cancer datasets dealing with human lung, colon and leukaemia. The performance of GFI has been compared with many existing cluster validity indices. The results are appropriately validated biologically and statistically.

Selection of informative genes, from microarray gene expression profiles, is an important data analysis step to identify a set of genes which can further help in finding the biological information embedded in microarray data, and thus assists in diagnosis, prognosis and treatment of the disease. In the eighth paper titled 'Graph-based unsupervised feature selection and multiview clustering for microarray data', the authors have presented an unsupervised feature selection technique which attempts to address the goal of explorative data analysis, unfolding the multifaceted nature of data. It focuses on extracting multiple clustering views considering the diversity of each view from high-dimensional data. The effectiveness of the technique has been demonstrated on several benchmark datasets.

Pathway analysis forms a key task in the area of molecular biology, under the framework of systems biology. Various T-cell co-receptor molecules and calcium channel CRAC play pivotal roles in the maintenance of cell's functional responses by regulating the production of effector molecules (mostly cytokines) that aid in immune clearance and also in maintaining the cell in a functionally active state. Any defect in these co-receptor signalling pathways may lead to an altered expression pattern of the effector molecules. To study the propagation of such defects with time and their effect on the intracellular protein expression patterns, a comprehensive and largest pathway map of T-cell activation network has been reconstructed manually, as described in the ninth article titled 'Temporal protein expression pattern in intracellular signalling cascade during T-cell activation: A computational study'. The entire pathway reactions are then translated using logical equations and simulated using the published time series microarray expression data as inputs. After validating the model, the effect of *in silico* knock-down of co-receptor molecules on the expression patterns of their downstream proteins has been studied and simultaneously the changes in the phenotypic behaviours of the T-cell population have been predicted, which shows significant variations among the proteins expression and the signalling routes through which the response is propagated in the cytoplasm.

MicroRNAs are a class of important post-transcriptional regulators. Genetic and somatic mutations in miRNAs, especially those in the seed regions, have profound and broad impacts on gene expression and physiological and pathological processes. Over 500 SNPs have been mapped to the miRNA seeds, which are located at position 2–8 of the mature miRNA sequences. The authors of the tenth paper titled 'Knowledge-based analysis of functional impacts of mutations in microRNA seed regions' have developed a knowledge-based method to analyse the functional impacts of mutations in miRNA seed regions. The gene ontology-based similarity score (GOSS) and the GOSS percentile score have been determined for all 517 SNPs in miRNA seeds. In addition to the annotation of SNPs for their functional effects, a detailed analysis pipeline has been developed for finding the key functional changes for seed SNPs. A detailed gene ontology graph-based analysis of enriched functional categories for miRNA target gene sets has been performed in the article.

In 'Phylogeny of metabolic networks: A spectral graph theoretical approach', the eleventh article, the authors have constructed metabolic networks of 79 fully sequenced organisms and compared their architectures. They have used spectral density of

normalized Laplacian matrix for comparing the structure of networks. The eigenvalues of this matrix reflect not only the global architecture of a network but also the local topologies that are produced by different graph evolutionary processes, like motif duplication or joining. A divergence measure on spectral densities has been used to quantify the distances between various metabolic networks, and a split network has been constructed to analyse the phylogeny from these distances. They have focused on the species that belong to different classes but appear more related to each other in the phylogeny.

Protein–protein interaction (PPI) site prediction aids in ascertaining the interface residues that participate in interaction processes. A fuzzy support vector machine (F-SVM)-based method has been developed, as described in the twelfth paper titled 'Protein–protein interaction site prediction in *Homo sapiens* and *E. coli* using an interaction-affinity based membership function in fuzzy SVM', as an effective method to predict PPI sites. The performances of both SVM and F-SVM on the PPI databases of the *H. sapiens* and *E. coli* organisms have been evaluated and the statistical significance has been estimated for the proposed method over classical SVM and other fuzzy-membership-based SVM methods available in the literature.

A systematic understanding of the rice cellular metabolism, for designing efficient stress-tolerant, more nutritious, high-yielding rice varieties, is essential. The thirteenth article, titled 'Flux balance analysis of genome scale metabolic model of rice (*Oryza sativa*): Aiming to increase biomass', describes an analysis of a genome-scale metabolic model of rice leaf using Flux Balance Analysis to investigate whether it has potential metabolic flexibility to increase the biosynthesis of any of the biomass components. The authors have initially simulated the metabolic responses with the objective to maximize the biomass components. Using the estimated maximum value of biomass synthesis as a constraint, they have further simulated the metabolic responses optimizing the cellular economy. In order to mimic this physiological state, they have randomly varied the ICTs' transport capacities and investigated their effects.

We take this opportunity to thank the Editor-in-Chief, *Journal of Biosciences*, for giving us an opportunity to highlight, in this special issue, the effectiveness and methodologies of pattern recognition and machine intelligence for solving a wide range of problems in molecular biology. We also thank the reviewers for their critical and insightful comments on the manuscripts, which have enabled the authors to improve the quality of their articles. It is our pleasure to thank the authors for their contributions, without which this special issue could not have come into existence. We hope that the articles in this special issue will not only help the readers appreciate the importance of pattern recognition and machine intelligence approaches for solving biological problems, but will also inspire them to come up with novel algorithms and approaches.

Sanghamitra Bandyopadhyay* and Rajat K De
India Statistical Institute, Kolkata, India
**Corresponding author (Email, sanghami@isical.ac.in)*