Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 362738, 8 pages
http://dx.doi.org/10.1155/2014/362738

Review Article

A Survey on Evolutionary Algorithm Based Hybrid Intelligence in Bioinformatics

Shan Li,[1] Liying Kang,[1] and Xing-Ming Zhao[2]

[1] *Department of Mathematics, Shanghai University, Shanghai 200444, China*
[2] *Department of Computer Science, School of Electronics and Information Engineering, Tongji University, Shanghai 201804, China*

Correspondence should be addressed to Xing-Ming Zhao; zhaoxingming@gmail.com

Received 3 December 2013; Revised 29 January 2014; Accepted 29 January 2014; Published 6 March 2014

Academic Editor: Jean X. Gao

With the rapid advance in genomics, proteomics, metabolomics, and other types of omics technologies during the past decades, a tremendous amount of data related to molecular biology has been produced. It is becoming a big challenge for the bioinformatists to analyze and interpret these data with conventional intelligent techniques, for example, support vector machines. Recently, the hybrid intelligent methods, which integrate several standard intelligent approaches, are becoming more and more popular due to their robustness and efficiency. Specifically, the hybrid intelligent approaches based on evolutionary algorithms (EAs) are widely used in various fields due to the efficiency and robustness of EAs. In this review, we give an introduction about the applications of hybrid intelligent methods, in particular those based on evolutionary algorithm, in bioinformatics. In particular, we focus on their applications to three common problems that arise in bioinformatics, that is, feature selection, parameter estimation, and reconstruction of biological networks.

1. Introduction

During the past decade, large amounts of biological data have been generated thanks to the development of high-throughput technologies. For example, 1,010,482 samples were profiled and deposited in Gene Expression Omnibus (GEO) database [1] by the writing of this paper, where around thousands of genes on average were measured for each sample. The recently released pilot data from the 1000 genomes project indicate that there are 38 million SNPs (single-nucleotide polymorphism) and 1.4 million biallelic indels within the 14 populations investigated [2]. Beyond that, other large-scale omics data, for example, RNA sequencing and proteomics data, can be found in public databases and are being generated everyday around the world. Despite the invaluable knowledge hidden in the data, unfortunately, the analysis and interpretation of these data lag far behind data generation.

It has been a long history that intelligent methods from artificial intelligence were widely used in bioinformatics, where these approaches were utilized to analyze and interpret the big datasets that cannot be handled by biologists. For example, in their pioneering work, Golub et al. utilized self-organizing maps (SOMs) to discriminate acute myeloid leukemia (AML) from acute lymphoblastic leukemia (ALL) based only on gene expression profiles without any prior knowledge [3]. Later, support vector machine was employed to classify 14 tumor types based on microarray gene expression data [4]. Except for diagnosis, intelligent methods have been exploited to identify biomarkers [5], annotate gene functions [6], predict drug targets [7, 8], and reverse engineering signaling pathways [9], among others.

Despite the success achieved by standard intelligent methods, it is becoming evident that it is intractable to analyze the large-scale omics data with only single standard intelligent approaches. For example, when diagnosing cancers based on gene expression profiles, low accuracy is expected if a traditional classifier, for example, linear discriminant analysis (LDA), is employed to classify the samples based on all the genes measured. This phenomenon is caused due to the "large p small n" paradigm which arises in microarray data, where there are generally around 20 thousand of genes or variables that were measured for each sample while only tens or at most hundreds of samples were considered in each experiment. In other words, there are very few samples while a much larger number of variables are to be learned by the intelligent methods, that is, the curse of dimensionality problem. Therefore, it is necessary to employ other intelligent techniques to select a small number of informative features first, based on which a classifier can be constructed to achieve the desired prediction accuracy. Such hybrid intelligent methods, that is, the combination of several traditional intelligent approaches, are being proved useful in analyzing the big complex biological data and are therefore becoming more and more popular.

In this paper, we survey the applications of hybrid intelligent methods in bioinformatics, which can help the researchers from both fields to understand each other and boost their future collaborations. In particular, we focus on the hybrid methods based on evolutionary algorithm due to its popularity in bioinformatics. We introduce the applications of hybrid intelligent methods to three common problems that arise in bioinformatics, that is, feature selection, parameter estimation, and molecular network/pathway reconstruction.

2. Evolutionary Algorithm

In this section, we first briefly introduced evolutionary algorithm, which is actually a family of algorithms inspired by the evolutionary principles in nature. In the evolutionary algorithm family, there are various variants, such as genetic algorithm (GA) [10, 11], genetic programming (GP) [12], evolutionary strategies (ES) [13], evolutionary programming (EP) [14], and differential evolution (DE) [15]. However, the principle underlying all these algorithms is the same that tries to find the optimal solutions by the operations of reproduction, mutation, recombination, and natural selection on a population of candidate solutions. In the following parts, we will take genetic algorithm (GA) as an example to introduce the evolutionary algorithm.

Figure 1 presents a schematic flowchart of genetic algorithm. In genetic algorithm, each candidate solution should be represented in an appropriate way that can be handled by the algorithm. For example, given a pool of candidate solutions X of size M, $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M\}^T$, a candidate solution $\mathbf{x}_i$, that is, an individual, can be represented as a binary string $\mathbf{x}_i = [0, 0, 1, 0, \dots, 1]$. Take feature selection as an example; each individual represents a set of features to be selected, where element 1 in the individual means that the corresponding feature is selected and vice versa. After the representation of individuals is determined, a pool of initial solutions is generally randomly generated first.

To evaluate each individual in the candidate solution pool, a fitness function or evaluation function F is defined in the algorithm. The fitness function is generally defined by taking into account the domain knowledge and the optimal objective function to be solved. For instance, the prediction accuracy or classification error can be used as fitness function. If an individual leads to better fitness, it is a better solution and vice versa.

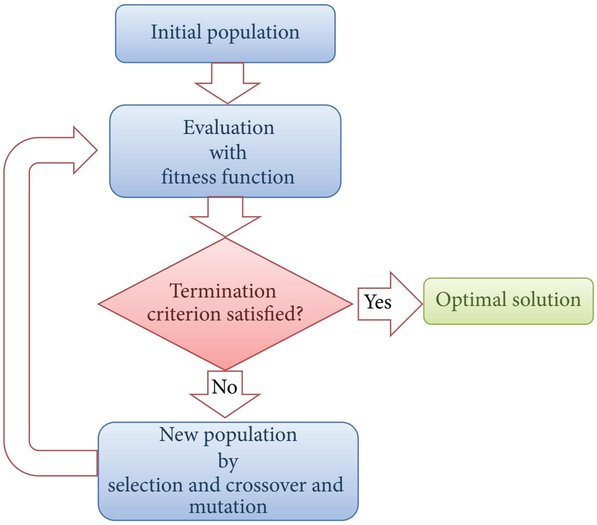

Initial population
Evaluation with fitness function
Termination criterion satisfied?
Yes
Optimal solution
No
New population by selection and crossover and mutation

Figure 1: The schematic flowchart of genetic algorithm.

Once the fitness function is determined, the current population will go through two steps: selection and crossover and mutation. In selection step, a subset of individual solutions will be selected generally based on certain probability, and the selected solutions will be used as parents to breed next generation. In the next step, a pair of parent solutions will be picked from the selected parents to generate a new solution with crossover operation; meanwhile, mutation(s) can be optionally applied to certain element(s) within a parent individual to generate a new one. The procedure of crossover and/or mutation continues until a new population of solutions of similar size is generated.

The genetic algorithm repeats the above procedure until certain criterion is met; that is, the preset optimal fitness is found or a fixed number of generations are reached. Despite the common principles underlying the evolutionary algorithm family, other variants of the algorithm may have implementation procedures that are different from the genetic algorithm. For example, in differential evolution, the individuals are selected based on greedy criterion to make sure that all individuals in the new generation are better than or at least as good as the corresponding ones in current population. Another alternative of the traditional genetic algorithm, namely, memetic algorithm (MA), utilizes a local search technique to improve the fitness of each individual and reduce the risk of premature convergence.

Since the evolutionary algorithm starts with a set of random candidate solutions and evaluates multiple individuals at the same time, the risk of getting stuck in a local optimum is reduced. Furthermore, the evolutionary algorithm can generally find optimal solutions within reasonable time, thereby becoming a popular technique in various fields.

3. Feature Selection in Bioinformatics

In bioinformatics, various problems are equivalent to feature selection problem. For example, in bioinformatics, biomarker discovery is one important and popular topic that tries to identify certain markers, for example, genes or mutations, which can be used for disease diagnosis. It is obvious that biomarker identification is equivalent to feature selection if we consider genes or mutations of interest as variables, where the informative genes or mutations are generally picked to discriminate disease samples from normal ones. However, it is not an easy task to select a few informative variables (generally <20) from thousands or even tens of thousands of features. Under the circumstances, the evolutionary algorithm has been widely adopted for identifying biomarkers along with other intelligent methods. Figure 2 depicts the procedure of feature selection with GA, where GA generally works together with a classifier as a wrapper method and the classifier is used to evaluate the selected features in each iteration. For example, Li et al. [16] utilized genetic algorithm and k-nearest neighbor (KNN) classifier to find discriminative genes that can separate tumors from normal samples based on gene expression data, and robust results were obtained by the hybrid GA/KNN method. Later, Jirapech-Umpai and Aitken [17] applied the GA/KNN approach to leukemia and NCI60 datasets, where the prediction results by the hybrid method are found to be consistent with clinical knowledge, indicating the effectiveness of the hybrid method. Since the simple genetic algorithm (SGA) often converges to a point in the search space, Goldberg and Holland adopted the speciated genetic algorithm, which controls the selection step by handling its fitness with the niching pressure, for gene selection along with artificial neural network (SGANN) [18]. Benchmark results show that SGANN reduces much more features than SGA and performs pretty well [19]. Recently, the hybrid approaches that, respectively, combined Pearson's correlation coefficient (CC) and Relief-F measures with GA were proposed by Chang et al. [20] to select the key features in oral cancer prognosis. These hybrid approaches outperform other popular techniques, such as adaptive neurofuzzy inference system (ANFIS), artificial neural network (ANN), and support vector machine (SVM). In addition to gene selection, the hybrid methods involving evolutionary algorithm have been successfully used to identify SNPs associated with diseases [21, 22] and peptides related to diseases from proteomic profiles [23–25].

Beyond biomarker identification, the evolutionary algorithm based hybrid intelligent methods have also been successfully applied to other feature selection problems in bioinformatics. For example, Zhao et al. [26] proposed a novel hybrid method based on GA and support vector machine (SVM) to select informative features from motif content and protein composition for protein classification, where the principal component analysis (PCA) was further used to reduce the dimensionality while GA was utilized to select a subset of features as well as optimize the regularization parameters of SVM at the same time. Results on benchmark datasets show that the hybrid method is really effective and robust. The hybrid method that integrates SVM and GA was also successfully used to select SNPs [27] and genes [28] associated with certain phenotypes and predict protein subnuclear localizations based on physicochemical composition features [29]. Recently, the hybrid SVM/GA approach was also utilized for selecting the optimum combinations of specific histone epigenetic marks to predict enhancers [30]. Saeys et al. predicted splice sites from nucleotide acid sequence by utilizing the hybrid method combining SVM and estimation of distribution algorithms (EDA) that is similar to GA [31]. Nemati et al. further combined GA and ant colony optimization (ACO) together for feature selection, and the hybrid method was found to outperform either GA or ACO alone when predicting protein functions [32]. In addition, Kamath et al. [33] proposed a feature generation with an evolutionary algorithm (FG-EA) approach, which employs a standard GP algorithm to explore the space of potentially useful features of sequence data. The features obtained from FG-EA enable the SVM classifier to get higher precision.

Feature selection is an important topic in bioinformatics and is involved in the analysis of various kinds of data. The hybrid methods that utilize the evolutionary algorithm have been proven useful for feature selection when handling the complex biological data due to their efficiency and robustness.

4. Parameter Estimation in Modeling Biological Systems

In bioinformatics, one biological system can be modeled as a set of ordinary differential equations (ODEs) so that the dynamics of the systems can be investigated and simulated. For example, Zhan and Yeung modeled a molecular pathway with the following ODEs [34]:

$$\begin{gathered}\dot{x}(t) = f(x(t), u(t), \theta), \\ x(t_0) = x_0, \\ y(t) = g(x(t)) + \eta(t),\end{gathered} \tag{1}$$

where $x \in R^n$ is the state vector of the system, $\theta \in R^k$ is a parameter vector, $u(t) \in R^p$ is the system's input, $y \in R^m$ is the measured data, $\eta(t) \sim N(0, \sigma^2)$ is the Gaussian white noise, and x_0 denotes the initial state. f is designed as a set of nonlinear transition functions to represent the dynamical properties of the biological system and g is a measurement function. It can be seen that, to make the model work, it is necessary to estimate the parameters in the model, which can be transformed into an optimization problem as follows:

$$P : \min_{\hat{\theta}, \hat{x}_0} \sum_{j=0}^{N-1} \sum_{i=1}^{n} w_{ij} \left\| y_i(t_j) - \hat{y}_i(t_j \mid \hat{\theta}) \right\|_l, \tag{2}$$

where $\hat{y}(t_j) = g(\hat{x}(t_j \mid \hat{\theta}))$, $\|\cdot\|_l$ denotes the l-norm, $\hat{x}(t_j \mid \hat{\theta})$ is the variable at time t_j with parameter $\hat{\theta}$, w_{ij} denotes the weight, and $\hat{y}$ means the estimated value. The problem P could be solved easily by employing the evolutionary algorithms [35–37]. For example, Katsuragi et al. [38] employed GA to estimate the parameters required by the simulation of dynamics of the metabolite concentrations, and Ueda et al. [39] applied the real-coded genetic algorithm to find the optimal values of the parameters. Recently, in order

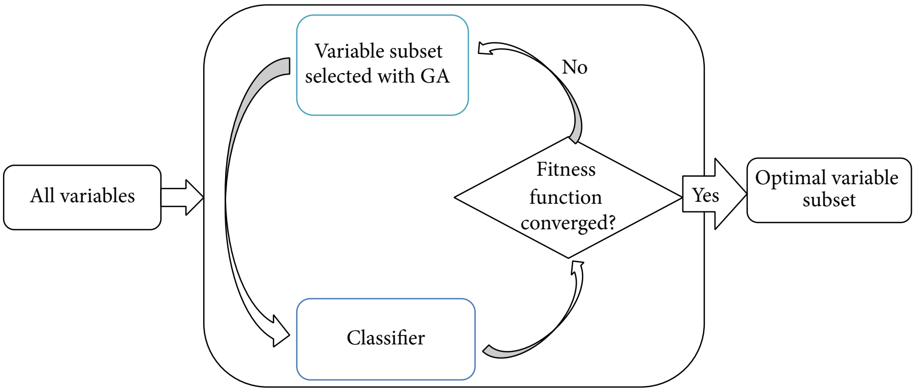


Figure 2: The flowchart of feature selection based on GA and classifier.

to improve the accuracy of parameter estimation, Abdullah et al. [40] proposed a novel approach that combines differential evolution (DE) with the firefly algorithm (FA), which outperformed other well-known approaches, such as particle swarm optimization (PSO) and Nelder-Mead algorithm.

In biological experiments, most data observed are measured at discrete time points while the traditional ODE model is a set of continuous equations, which makes it difficult to estimate the parameters in an accurate way. Therefore, the S-system, which is a type of power-law formalism and a particular type of ODE model, was widely used instead. For example, Savageau and Rosen [41] modeled the genetic network with the following S-system model:

$$\frac{dX_i}{dt} = \alpha_i \prod_{j=1}^{n+m} X_j^{g_{ij}} - \beta_i \prod_{j=1}^{n+m} X_j^{h_{ij}}, \qquad (3)$$

where X_i denotes the variable or reactant, n and m, respectively, denote the number of dependent and independent variables, α_i and β_i are nonnegative rate constants, and g_{ij} and h_{ij} are kinetic orders. Here, the parameters α_i, β_i, g_{ij}, and h_{ij} must be estimated. To optimize the parameters, Tominaga and Okamoto [42] utilized GA to approach the optimization problem with the following evaluation function E:

$$E = \sum_{i=1}^{n+m} \sum_{t=1}^{T} \left(\frac{X_i'(t) - X_i(t)}{X_i(t)} \right)^2, \qquad (4)$$

where T is the number of sampling points and $X_i(t)$ and $X_i'(t)$, respectively, denote experimentally observed and estimated value at time t for X_i. Later, Kikuchi et al. [43] found that it is difficult to estimate all the parameters from limited time-course data of metabolite concentrations. Hence, they changed the evaluation function E as follows:

$$\begin{aligned} E = {} & \sum_{i=1}^{n+m} \sum_{t=1}^{T} \left(\frac{X_i'(t) - X_i(t)}{X_i(t)} \right)^2 \\ & + c(n+m)T \left\{ \sum_{i,j} \left| g_{ij} \right| + \sum_{i,j,i \neq j} \left| h_{ij} \right| \right\}, \end{aligned} \qquad (5)$$

where c is a penalty constant that balances the two evaluation terms. Moreover, they adopted the simplex operations [44] instead of the random ones to accelerate the searching in GA. Considering only a few genes affecting both the synthesis and degradation processes of specific genes, Noman and Iba [45] further simplified the evaluation function as follows:

$$E_i = \sum_{t=1}^{T} \left(\frac{X_i'(t) - X_i(t)}{X_i(t)} \right)^2 + c \sum_{j=1}^{n+m-1} \left(\left| K_{i,j} \right| \right), \qquad (6)$$

where $K_{i,j}$ is the kinetic order of gene i. With this objective function, they adopted a novel hybrid evolutionary algorithm, namely, memetic algorithm (MA) [46], that combines global optimization and local search together to find the optimal solutions. Considering that the traditional S-system can only describe instantaneous interactions, Chowdhury et al. [47] introduced the time-delay parameters to represent the system dynamics and refined the evaluation function as follows:

$$E = \sum_{t=1}^{T} \left(\frac{X_i^{\text{cal}}(t) - X_i^{\text{exp}}(t)}{X_i^{\text{exp}}(t)} \right)^2 + B_i \times C_i \frac{2N}{2N - r_i}, \qquad (7)$$

where r_i is the number of all actual regulators, B_i is a balancing factor between the two terms, and C_i is the penalty factor for gene i. The trigonometric differential evolution (TDE) technique was adopted to estimate the set of parameters because of its better performance than other traditional evolutionary algorithms.

Parameter estimation is a key step in mathematical modeling of biological systems, which is however a nontrivial task considering the possible huge search space. Due to its excellent searching capability, the evolutionary algorithm is able to help determine the model parameters along with other intelligent approaches.

5. Molecular Network/Pathway Reconstruction

Recently, the network biology that represents a biological system as a molecular network or graph is attracting more and more attention. In the molecular network, the nodes denote the molecules, for example, proteins and metabolites, while edges denote the interactions/regulations or other functional links between nodes. Although it is easy to observe the activity of thousands of molecules at the same time with high-throughput screening, it is not possible to detect the potential interactions/regulations between molecules right now.

Under the circumstances, a lot of intelligent methods have been presented to reconstruct the molecular networks, such as Boolean network and Bayesian network. When reconstructing the molecular networks, one critical step is to determine the topology of the network to be modeled, based on which the interactions/regulations between molecules can be investigated. The topology determination problem can be treated as an optimization problem that is ready to be solved with the help of the evolutionary algorithm.

Take a gene regulatory network as an example; Figure 3 shows the flowchart of reconstructing the regulatory network based on gene expression data by utilizing Boolean network and evolutionary algorithm. In the example, we want to reconstruct the regulatory circuit that controls the gene expression of five genes. Since at least one edge exists while at most 10 edges exist in the network, the number of possible network structures will be $M = \sum_{i=1}^{10} C_{10}^{i} = 2^{10} - 1 \approx 2^{10}$. It is impossible to validate all network topologies by biologists in lab. With appropriate fitness function, the evolutionary algorithm is able to identify the optimal network structure that fits best the gene expression data, where the consistence between network topology and gene expression data is evaluated with Boolean network based on certain rules.

Repsilber et al. [48] modeled the gene regulatory network with a Boolean model as a directed acyclic graph $G = (V, F)$, where $V = \{x_1, x_2, \ldots, x_n\}$ denotes the set of genes in the regulatory network and $F = \{f_1, f_2, \ldots, f_n\}$ denotes the Boolean rules that describe the regulations between nodes (or genes). To determine the topology of the regulatory network that better fits the observed data, they employed GA with the following fitness function f:

$$f = \frac{1}{1 + (1/D) \sum_{ijk} \delta_{ijk}^{2}}, \tag{8}$$

where $\delta_{ijk} = (\text{sim\_data}_{ijk} - \text{network\_output}_{ijk})$ is the difference between the observed data and those estimated from the generated network. In this way, they successfully reconstructed the gene regulatory network that generates the expression profiles consistent with experiments.

Later, Mendoza and Bazzan [49] presented inconsistency ratio (IR) to evaluate each individual node in the network, where the IR is defined as follows:

$$\text{IR}_i = w^{-1} \sum_{k=1}^{2^K} \min \left(w_k(0), w_k(1) \right). \tag{9}$$

Here, $k = 1, 2, 3, \ldots, 2^K$ is the number of possible input combinations for a node, $w_k(0)$ denotes the weight of measurements with output of 0 while $w_k(1)$ denotes those with output of 1, and w is the sum of all weights. With the IR defined above, an evaluation function defined below was used to investigate the inconsistency between the network generated and the experimental data:

$$\phi = \frac{1}{1 + \left(\sum_{i=1}^{N} \text{IR}_i / (N \times 0.5) \right) + \left(\text{NP}/N^2 \right)}, \tag{10}$$

where $N \times 0.5$ denotes the maximum inconsistency to be generated by the network while (NP/N^2) is a penalty factor. With this evaluation function, the differential evolution (DE) approach was used to identify the optimal network structure [50].

Recently, to understand the signaling in distinct physiological situations, Terfve et al. [51] proposed a CellNOptR approach, which derives a Boolean logic model from a "prior knowledge network" and uses GA to search the optimal network structure that is consistent with the perturbation data. Later, Crespo et al. [52] employed Boolean logic model and genetic algorithm to predict missing gene expression values from experimental data and obtained promising results.

Although the Boolean network is simple and capable of handling large networks, it fails to provide quantitative information about regulations between molecules, which is however the key to understand the regulation process. In this case, the Bayesian network is widely adopted. Considering the expensive computation time required by Bayesian network, the evolutionary algorithm is widely used to determine the structures of the molecular networks modeled. In the Bayesian network, the molecular network is regarded as a directed acyclic graph described as follows:

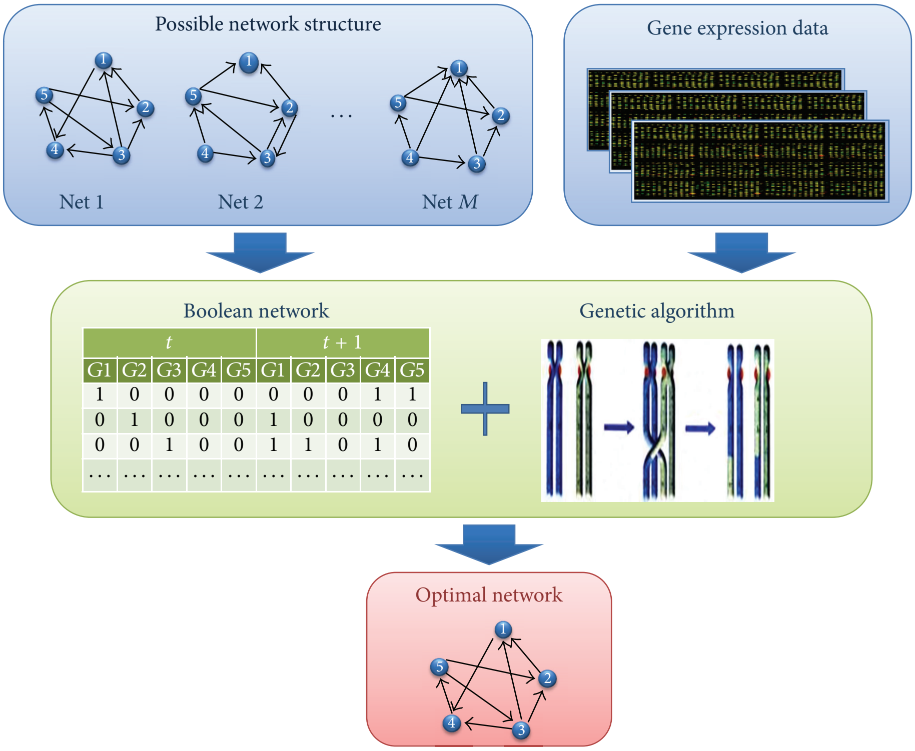
$$P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid \pi_i), \tag{11}$$

where x_i denotes node i in the set of variables, that is, the molecules considered, and π_i denotes the parent node of x_i. For example, Yu et al. [53] utilized GA to determine the optimal network structure consistent with experimental data along with the dynamic Bayesian network by defining an evaluation function based on Bayesian dirichlet equivalence (BDe) score and Bayesian information criterion (BIC) score. Later, Xing and Wu [54] employed the maximum likelihood (ML) score and the minimal description length (MDL) score as fitness values and determined the topology of gene regulatory networks with GA, where the regulatory network is modeled with Bayesian network. Recently, Li and Ngom [55] proposed a new high-order dynamic Bayesian network (HO-DBN) learning approach to identify genetic regulatory networks from gene expression time-series data and obtained the optimal structure of the networks with GA. In their method, the optimal structure $\widehat{S}$ was estimated by the maximum likelihood as follows:

$$\widehat{S} = \int_{\theta_s} P(X \mid \theta_s) P(\theta_s \mid S) \, d\theta_s, \tag{12}$$

where $X = \{x_1, x_2, \ldots, x_n\}$ and $\theta_s = \{\theta_1, \theta_2, \ldots, \theta_n\}$ is the parameter set.

In addition to Boolean and Bayesian networks, the Petri net [56] is also widely employed to reconstruct biological



t					$t+1$				
G1	G2	G3	G4	G5	G1	G2	G3	G4	G5
1	0	0	0	0	0	0	0	1	1
0	1	0	0	0	1	0	0	0	0
0	0	1	0	0	1	1	0	1	0
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯

Figure 3: The reconstruction of gene regulatory network based on gene expression with the hybrid method consisting of Boolean network and evolutionary algorithm.

networks. For example, in the Petri net model of metabolic networks, the nodes named places denote metabolites or products while transitions representing reactions are edges, where the values accompanying transitions denote rate constants. The input places for a transition denote the reaction's reactants while the output places denote its products, and the value of a place can be represented by its corresponding amount of substance. If a transition is deleted, a reaction happens, in which reactants are consumed and products are yielded. To find the optimal solutions, Nummela and Juistrom [57] defined a fitness function F as follows:

$$F = \sum \frac{\left|c_{mi} - c_{mi0}\right|}{n_m n_p} + 0.1 \times n_r, \qquad (13)$$

where c_{mi} means the computed concentration of the mth metabolite at time i, c_{mi0} is the corresponding target concentration, n_m means the number of metabolites, n_p is the number of time steps, and n_r is the number of reactions. With the hybrid method combining the Petri net and GA, they successfully identified a network that is consistent with the simulated data. Later, Koh et al. [58] have also successfully employed this hybrid method to model the AKt and MAPK signaling pathways.

The molecular networks enable one to investigate the biological systems from a systematic perspective, whereas the network topology is the key to construct and understand the network. Accumulating evidence demonstrates that the hybrid heuristic methods involving evolutionary algorithm are able to help determine the network topology consistent with experimental data in an accurate way due to its significant efficiency.

6. Conclusions

In this paper, we surveyed the applications of hybrid intelligent methods, which combine several traditional intelligent approaches together, in bioinformatics. Especially, we introduced the hybrid methods involving evolutionary algorithm and their applications in three common problems in bioinformatics, that is, feature selection, parameter estimation, and reconstruction of biological networks. The evolutionary algorithm was selected here due to its capability of finding global optimal solutions and its robustness. The hybrid intelligent approaches that combine evolutionary algorithm together with other standard intelligent approaches have been proved extremely useful in the above three topics. We hope this review can help the researchers from both bioinformatics and informatics to understand each other and boost their future collaborations. We believe that, with more effective hybrid intelligent methods introduced in the future, it will become relatively easier to analyze the ever-growing complex biological data.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was partly supported by the National Natural Science Foundation of China (91130032 and 61103075), the Innovation Program of Shanghai Municipal Education Commission (13ZZ072), and Shanghai Pujiang Program (13PJD032).

References

[1] T. Barrett and R. Edgar, "Gene expression omnibus: microarray data storage, submission, retrieval, and analysis," *Methods in Enzymology*, vol. 411, pp. 352–369, 2006.

[2] G. R. Abecasis, A. Auton, L. D. Brooks et al., "An integrated map of genetic variation from 1, 092 human genomes," *Nature*, vol. 491, no. 7422, pp. 56–65, 2012.

[3] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–527, 1999.

[4] S. Ramaswamy, P. Tamayo, R. Rifkin et al., "Multiclass cancer diagnosis using tumor gene expression signatures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 26, pp. 15149–15154, 2001.

[5] K. Q. Liu, Z. P. Liu, J. K. Hao et al., "Identifying dysregulated pathways in cancers from pathway interaction networks," *BMC Bioinformatics*, vol. 13, article 126, 2012.

[6] G. R. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble, "Kernel-based data fusion and its application to protein function prediction in yeast," *Pacific Symposium on Biocomputing*, pp. 300–311, 2004.

[7] D. Barh, K. Gupta, N. Jain et al., "Conserved host-pathogen PPIs. Globally conserved inter-species bacterial PPIs based conserved host-pathogen interactome derived novel target in C. pseudotuberculosis, C. diphtheriae, M. tuberculosis, C. ulcerans, Y. pestis, and E. coli targeted by Piper betel compounds," *Integrative Biology*, vol. 5, no. 3, pp. 495–509, 2013.

[8] M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen, and P. Bork, "Drug target identification using side-effect similarity," *Science*, vol. 321, no. 5886, pp. 263–266, 2008.

[9] X.-M. Zhao, R.-S. Wang, L. Chen, and K. Aihara, "Uncovering signal transduction networks from high-throughput data by integer linear programming," *Nucleic Acids Research*, vol. 36, no. 9, article e48, 2008.

[10] A. S. Fraser, "Simulation of genetic systems by automatic digital computers. I. Introduction," *Australian Journal of Biological Sciences*, vol. 10, pp. 484–491, 1957.

[11] J. H. Holland, *Adaptation in Natural and Artificial Systems: an Introductory analysis with Applications to Biology, Control, and Artificial Intelligence*, MIT Press, 1992.

[12] N. A. Barricell, "Numerical testing of evolution theories," *Journal of Statistical Computation and Simulation*, vol. 1, no. 2, pp. 97–127, 1972.

[13] I. Rechenberg, *Evolutionsstrategie: Optimierung Technischer Systeme Nach Prinzipien Der Biologischen Evolution*, Technical University of Berlin, 1971.

[14] L. J. Fogel, A. J. Owens, and M. J. Walsh, *Artificial Intelligence Through Simulated Evolution*, John Wiley & Sons, 1966.

[15] R. Storn and K. Price, "Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 11, no. 4, pp. 341–359, 1997.

[16] L. Li, C. R. Weinberg, T. A. Darden, and L. G. Pedersen, "Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method," *Bioinformatics*, vol. 17, no. 12, pp. 1131–1142, 2002.

[17] T. Jirapech-Umpai and S. Aitken, "Feature selection and classification for microarray data analysis: evolutionary methods for identifying predictive genes," *BMC Bioinformatics*, vol. 6, article 148, 2005.

[18] D. E. Goldberg and J. H. Holland, "Genetic algorithms and machine learning," *Machine Learning*, vol. 3, no. 2, pp. 95–99, 1988.

[19] J.-H. Hong and S.-B. Cho, "Efficient huge-scale feature selection with speciated genetic algorithm," *Pattern Recognition Letters*, vol. 27, no. 2, pp. 143–150, 2006.

[20] S. W. Chang, S. Abdul-Kareem, A. F. Merican et al., "Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods," *BMC Bioinformatics*, vol. 14, article 170, 2013.

[21] G. Mahdevar, J. Zahiri, M. Sadeghi, A. Nowzari-Dalini, and H. Ahrabian, "Tag SNP selection via a genetic algorithm," *Journal of Biomedical Informatics*, vol. 43, no. 5, pp. 800–804, 2010.

[22] C. S. Carlson, M. A. Eberle, M. J. Rieder, Q. Yi, L. Kruglyak, and D. A. Nickerson, "Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium," *American Journal of Human Genetics*, vol. 74, no. 1, pp. 106–120, 2004.

[23] K. A. Baggerly, J. S. Morris, J. Wang, D. Gold, L.-C. Xiao, and K. R. Coombes, "A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples," *Proteomics*, vol. 3, no. 9, pp. 1667–1672, 2003.

[24] E. F. Petricoin III, A. M. Ardekani, and B. A. Hitt, "Use of proteomic patterns in serum to identify ovarian cancer," *The Lancet*, vol. 359, pp. 572–577, 2002.

[25] L. Li, H. Tang, Z. Wu et al., "Data mining techniques for cancer detection using serum proteomic profiling," *Artificial Intelligence in Medicine*, vol. 32, no. 2, pp. 71–83, 2004.

[26] X.-M. Zhao, Y.-M. Cheung, and D.-S. Huang, "A novel approach to extracting features from motif content and protein composition for protein sequence classification," *Neural Networks*, vol. 18, no. 8, pp. 1019–1028, 2005.

[27] B. Gong, Z. Guo, J. Li et al., "Application of a genetic algorithm -Support vector machine hybrid for prediction of clinical phenotypes based on genome-wide SNP profiles of sib pairs," in *Proceedings of the 2nd International Confernce on Fuzzy Systems and Knowledge Discovery (FSKD '05)*, pp. 830–835, August 2005.

[28] L. Li, W. Jiang, X. Li et al., "A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset," *Genomics*, vol. 85, no. 1, pp. 16–23, 2005.

[29] W.-L. Huang, C.-W. Tung, H.-L. Huang, S.-F. Hwang, and S.-Y. Ho, "ProLoc: prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features," *BioSystems*, vol. 90, no. 2, pp. 573–581, 2007.

[30] M. Fernández and D. Miranda-Saavedra, "Genome-wide enhancer prediction from epigenetic signatures using genetic algorithm-optimized support vector machines," *Nucleic Acids Research*, vol. 40, no. 10, p. e77, 2012.

[31] Y. Saeys, S. Degroeve, D. Aeyels, P. Rouzé, and Y. Van de Peer, "Feature selection for splice site prediction: a new method using EDA-based feature ranking," *BMC Bioinformatics*, vol. 5, article 64, 2004.

[32] S. Nemati, M. E. Basiri, N. Ghasem-Aghaee, and M. H. Aghdam, "A novel ACO-GA hybrid algorithm for feature selection in protein function prediction," *Expert Systems with Applications*, vol. 36, no. 10, pp. 12086–12094, 2009.

[33] U. Kamath, J. Compton, R. Islamaj-Dogan, K. A. De Jong, and A. Shehu, "An evolutionary algorithm approach for feature generation from sequence data and its application to DNA splice site prediction," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 5, pp. 1387–1398, 2012.

[34] C. Zhan and L. F. Yeung, "Parameter estimation in systems biology models using spline approximation," *BMC Systems Biology*, vol. 5, article 14, 2011.

[35] M. Ashyraliyev, Y. Fomekong-Nanfack, J. A. Kaandorp, and J. G. Blom, "Systems biology: parameter estimation for biochemical models," *FEBS Journal*, vol. 276, no. 4, pp. 886–902, 2009.

[36] J. R. Banga and E. Balsa-Canto, "Parameter estimation and optimal experimental design," *Essays in Biochemistry*, vol. 45, pp. 195–209, 2008.

[37] C. G. Moles, P. Mendes, and J. R. Banga, "Parameter estimation in biochemical pathways: a comparison of global optimization methods," *Genome Research*, vol. 13, no. 11, pp. 2467–2474, 2003.

[38] T. Katsuragi, N. Ono, K. Yasumoto et al., "SS-mPMG and SS-GA: tools for finding pathways and dynamic simulation of metabolic networks," *Plant Cell Physiology*, vol. 54, no. 5, pp. 728–739, 2013.

[39] T. Ueda, D. Tominaga, N. Araki et al., "Estimate hidden dynamic profiles of siRNA effect on apoptosis," *BMC Bioinformatics*, vol. 14, article 97, 2013.

[40] A. Abdullah, S. Deris, S. Anwar, and S. N. Arjunan, "An evolutionary firefly algorithm for the estimation of nonlinear biological model parameters," *PloS One*, vol. 8, no. 3, Article ID e56310, 2013.

[41] M. A. Savageau and R. Rosen, *Biochemical Systems Analysis: A Study of Function and Design in molecular Biology (*, Addison-Wesley, 1976.

[42] D. Tominaga and M. Okamoto, "Design of canonical model describing complex nonlinear dynamics," in *Proceedings of the 7th International Conference on Computer Applications in Biotechnology*, pp. 85–90, 1998.

[43] S. Kikuchi, D. Tominaga, M. Arita, K. Takahashi, and M. Tomita, "Dynamics modeling of genetic networks using genetic algorithm and S-system," *Bioinformatics*, vol. 19, no. 5, pp. 643–650, 2003.

[44] S. Tsutsui, M. Yamamura, and T. Higuchi, "Multi-parent recombination with simplex crossover in real coded genetic algorithms," in *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 657–664, 1999.

[45] N. Noman and H. Iba, "Reverse engineering genetic networks using evolutionary computation," *Genome Informatics*, vol. 16, no. 2, pp. 205–214, 2005.

[46] P. Moscato, "On evolution, search, optimization, genetic algorithms and martial arts: towards memetic algorithms," Caltech Concurrent Computation Program 826, 1989.

[47] A. R. Chowdhury, M. Chetty, and N. X. Vinh, "Incorporating time-delays in S-System model for reverse engineering genetic networks," *BMC Bioinformatics*, vol. 14, article 196, 2013.

[48] D. Repsilber, H. Liljenström, and S. G. E. Andersson, "Reverse engineering of regulatory networks: simulation studies on a genetic algorithm approach for ranking hypotheses," *BioSystems*, vol. 66, no. 1-2, pp. 31–41, 2002.

[49] M. R. Mendoza and A. L. C. Bazzan, "Evolving random boolean networks with genetic algorithms for regulatory networks reconstruction," in *Proceedings of the 13th Annual Genetic and Evolutionary Computation Conference (GECCO '11)*, pp. 291–298, July 2011.

[50] A. Esmaeili and C. Jacob, "A multi-objective differential evolutionary approach toward more stable gene regulatory networks," *BioSystems*, vol. 98, no. 3, pp. 127–136, 2009.

[51] C. Terfve, T. Cokelaer, D. Henriques et al., "CellNOptR: a flexible toolkit to train protein signaling networks to data using multiple logic formalisms," *BMC Systems Biology*, vol. 6, article 133, 2012.

[52] I. Crespo, A. Krishna, A. Le Bechec, and A. del Sol, "Predicting missing expression values in gene regulatory networks using a discrete logic modeling optimization guided by network stable states," *Nucleic Acids Research*, vol. 41, no. 1, article e8, 2013.

[53] J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink, and E. D. Jarvis, "Advances to Bayesian network inference for generating causal networks from observational biological data," *Bioinformatics*, vol. 20, no. 18, pp. 3594–3603, 2004.

[54] Z. Xing and D. Wu, "Modeling multiple time units delayed gene regulatory network using dynamic Bayesian network," in *Proceedings of the 6th IEEE International Conference on Data Mining—Workshops (ICDM '06)*, pp. 190–195, December 2006.

[55] Y. Li and A. Ngom, "The max-min high-order dynamic Bayesian network learning for identifying gene regulatory networks from time-series microarray data," in *Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB '13)*, pp. 83–90, 2013.

[56] G. Rozenberg and E. Engelfriet, "Elementary net systems," in *Lectures on Petri Nets I: Basic Models*, vol. 1497, pp. 12–121, 1998.

[57] J. Nummela and B. A. Juistrom, "Evolving, petri nets to represent metabolic pathways," in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '05)*, pp. 2133–2139, June 2005.

[58] G. Koh, H. F. C. Teong, M.-V. Clément, D. Hsu, and P. S. Thiagarajan, "A decompositional approach to parameter estimation in pathway modeling: a case study of the Akt and MAPK pathways and their crosstalk," *Bioinformatics*, vol. 22, no. 14, pp. e271–e280, 2006.