Agent-based Optimization of Biological Response Networks

Chien-Feng Huang
Department of Computer Science and Information Engineering
National University of Kaohsiung
Kaohsiung, 811, Taiwan
cfhuang15@nuk.edu.tw

Yu-Feng Lin
Department of Computer Science and Information Engineering
National Cheng Kung University
Tainan, 701, Taiwan
p78991113@mail.ncku.edu.tw

Vincent S. Tseng
Department of Computer Science and Information Engineering
National Cheng Kung University
Tainan, 701, Taiwan
tsegsm@mail.ncku.edu.tw

Abstract—One of the major challenges in systems biology today is to devise robust methods of interpreting data concerning the expression levels of the genes in an organism in a way that will shed light on the collective interactions between multiple genes and their products. The ability to better understand and predict the structures and actions of complex biological systems is of significant importance to modern drug discovery as well as our understanding of the mechanisms behind an organism’s ability to react to its environment. In this paper we present a study for robust biological pathway construction through an agent-based methodology—probability collectives multi-agent systems (PCMAS). This technique relies on the search of particular seed nodes by probability collectives to construct biological networks based on various sets of interaction information and gene expression data. As an application, expression data of ofloxacin response in M. tuberculosis is used to build response networks. We then demonstrate how this approach provides robust prediction of response networks to facilitate drug target identification on systems-level.

Keywords- Biological networks; Mycobacterium tuberculosis probability collectives multi-agent systems; robust prediction

I. INTRODUCTION

Two dramatic revolutions during the past thirty years in the life sciences – molecular biology (in the 1970s) and genomics (in the 1990s) – have enhanced the knowledge on which modern medical science and pharmaceutical discovery are based. One of the most important challenges in the post-genome research is to move toward a new view of biology – a systems level approach. It is currently estimated that there are 20,000–25,000 human protein-coding genes. Even with the discovery of RNA genes and microRNAs, this surprisingly low estimate in the human genome strongly indicates that functional complexity may originate in locations and processes beyond the identification of particular genes. Rather than looking at biological elements in isolation, Systems Biology intends to examine the collective behavior and relationships of all of the components in a biological system and attempt to model and to control its dynamical behavior.

One of the major challenges in systems biology today is to devise generally robust methods of interpreting this data in a way that will shed light on the complex relationships between multiple genes and their products. Such methodology is important both as an aid to arenas of research like drug discovery and design, but also as a powerful tool toward deeper understanding of the interactions among all the components in a biological system as well as their collective behavior.

In particular, one crucial goal within systems biology is to develop the capability for analyzing biological interaction networks as they respond to different environments. Groundwork with respect to this goal has been laid in seminal contributions from Ideker et al. [1], [2] and Zien et al. [3]. Such a capability is of significant importance for the understanding of biological activities on a systems level that eventually leads to the identification of risk factors of genetic disorders as well as the prevention of infection from bacterial pathogens. More recent work, including [4] and [5], followed this line of research to reveal the role of transcription factors in biological systems and/or environments. In addition, Desmeulles et al. [6] studied generic models by numerical simulations for systems biology, in order for different biologists to work around an interdisciplinary subject and also permit to embed different expert system modeling methods like fuzzy cognitive maps. Here we review a method for analyzing biological response data from gene-expression arrays by combining it with computationally derived network information. Then we show how the PCMAS (Probability Collectives Multi-agent Systems) can be used as an optimization tool for network construction, thereby shedding light on discovering crucial components in specific drug response networks.

The organization of this paper is as follows. Section 2 presents the preliminaries for biological network construction. In Section 3, a review of PC is provided. In Section 4 we present empirical results and discussions for optimization of biological network construction of Mycobacterium tuberculosis (M. tuberculosis) using PCMAS. We then conclude this paper in Section 5.

II. PRIOR WORK

In this section we review the method of the response network developed by [8], which is comprised of various steps, including, (i) M. tuberculosis network construction by computational methods, (ii) network filtering and response
network identification by superimposing experimental gene expression data upon the computationally derived *M. tuberculosis* network and (iii) differential network expression analysis.

In the case of biological networks, depending on the particular network representation, node properties can include gene, protein or chemical names, and edge properties may refer to specific interactions, such as binding or catalysis.

In order to construct the *M. tuberculosis* network, Cabusora et al. [8] identified three types of interactions as relevant for such a network: (i) protein interaction, (ii) metabolic reactions and (iii) co-expression in regulons. For protein interaction data, Cabusora et al. used identified component genes fused together into one single gene (with multiple functions according to the functions of the component genes) according to Enright et al. [9],[10]. This method for recognizing possible interacting proteins is called the Rosetta Stone approach [10], which is based on the observation that individual genes in one organism that are fused into a single chain in another organism are likely to interact physically with each other.

In this study, the gene expression information from *M. tuberculosis* (H37Rv) was obtained after its growth in Middlebrook 7H9 supplemented with albumin/dextrose/NaCl/glycerol, Dubos medium or defined minimal medium [11]. Cultures were grown before adding either drug or solvent, and RNA was isolated at selected intervals thereafter. For the drug-treated culture, a parallel culture was treated with an equivalent amount of solvent (DMSO, ethanol or water) for the same amount of time. RNA from the latter culture was used as the reference sample to which the drug-treated sample was compared. The gene expression data is accessible through the Gene Expression Omnibus at NCBI (GEO; http://www.ncbi.nlm.nih.gov/geo) with GEO platform accession number GPL1396.

A. Algorithm and Implementation

The algorithm employed by [8] is implemented as follows. After the large biological network is constructed, it is stored as a Petri-Net [12], converted into a node graph, stripped of ubiquitous chemicals according to a maximal permissible node degree and loaded with gene-expression values. In addition, input parameters, such as the k-value for k-shortest path calculation and the maximal path-length l, are set. In the next step, the network is filtered according to the parameters and a set of seed nodes. The core of the algorithm is the component that scores particular sub-networks, yielding a sub-network with the highest score. Further components of the algorithm include a statistical analysis routine for sub-network score validation and a network algebra component for differential network analysis. A flow diagram of the corresponding processes is shown in Figure 1.

Figure 1. Global Architecture of the algorithm

B. Sub-Network Filtering and Scoring

A first-pass analysis of the iterative process of *Network Scoring* is performed during Sub-Network Filtering. Here, particular nodes are tagged as seed nodes, and shortest pathways between them are identified by Dijkstra’s algorithm [13]. Successively longer pathways are identified by Yen’s algorithm for the k-shortest simple paths problem, using an implementation by Hershberger et al. [14].

The algorithm performs the following functions: (i) compilation of a list of seed nodes, (ii) computation of all possible pairs of seed nodes from this list, (iii) calculation of shortest and k-shortest paths (with maximal path-length l) between each pair of seed nodes using Dijkstra’s and Yen’s algorithms, (iv) recording of all nodes and edges on identified paths, (v) filtering (deletion, hiding) of all other nodes and edges that are not on the selected paths, (vi) sub-network scoring, (vii) reiteration.

*Network Scoring* uses expression values as metrics for weighted edges in the network. In the graph, genes, proteins and other cellular components are coded as nodes which are connected by edges in the biological network. Because the sub-network filtering step assumes weights on edges for scoring, such edge weights must be calculated from node scores, i.e., gene expression levels.

**Bioconductor** modules were then employed in combination with the R package to analyze the gene-expression data and to extract p-values p_m for each expressed gene m. The p_m is converted into a Z-score Z_m = Φ^{-1}(1 - p_m), where Φ^{-1} denotes the inverse normal distribution function.  

To calculate a total score of the sub-network N, the Z_m’s are summed over all m, given the constraint of the k-shortest paths with maximal path-length / between each two seed nodes in N : Z_N = \frac{1}{\sqrt{m}} \sum_{m \in N(k,l)} z_m , where N(k, l) denotes


2. For the edge scores, one can use either the product probability p_i · p_j or the correlation coefficient p_ij for an edge m with origin node i and terminus j, yielding for example z_m = Φ^{-1}(1 - p_i · p_j).
shortest $N(k, l)$. Given a particular set of seed nodes, the shortest path approach already guarantees a best scored sub-network. But to obtain an optimal set of seed nodes for a better scoring sub-network, one has to search through the network. To this end, we employ the PCMAS for the search of seed nodes in order to optimize network construction. After the search process converges or the iteration reaches the terminating condition, the algorithm stops and outputs the highest scoring sub-networks.

III. PROBABILITY COLLECTIVES MULTI-AGENT SYSTEMS

Biologically-inspired algorithms, such as Genetic algorithms (GA) [15], Ant Colony Optimization (ACO) [16], Particle Swarm Optimization (PSO) [17], have been used as computational models to mimic evolutionary and social learning systems and as adaptive algorithms to solve complex problems. The core component of this class of optimization algorithms is a population of solutions that are employed to search for optimal solutions to the problem at hand. In the past decade, the research on Multi-agent Systems (MAS) has advanced the population-based algorithms with theoretical principles that shed light on the mechanism of system’s decomposition and interacting agents’ coordinating behavior in complex systems. As the complexity of a system grow, a generally more effective way to handle the system is through decomposition of the system into distributed and decentralized sub-systems and a given optimization task can be accomplished collectively by the sub-systems. In this scenario, the smaller subsystems can be regarded as a group of learning agents. These agents are self-interested and are dedicated to optimizing their individual rewards or payoffs that in turn collectively optimize the global goal on the systems level.

Probability Collectives (PC) theory [18] refers to the systems-level objective as the world utility, which measures the performance of the whole system. PC is a broad framework for modeling and controlling distributed systems, and has deep connections to game theory, statistical physics, distributed control and optimization. Typically the search of adaptive, distributed agent-based algorithms is conducted by having each agent run its own reinforcement learning algorithm [19]. In this methodology the global utility function $G(x)$ in the system maps a joint move of the agents, $x \in X$, to the performance of the overall system. Moreover, in practice the agents in a MAS are bounded rational. The equilibrium they reach typically involves mixed strategies rather than pure strategies; i.e., they don’t settle on a single point $x$ that optimizes $G(x)$. This suggests formulating a framework to explicitly account for the bounded rational, mixed strategy characteristics of the agents. PC adopts this perspective to show that the equilibrium of a MAS is the minimizer of a Lagrangian $L(P)$ (derived using information theory) that quantifies the expected value of $G$ for the joint distribution $P(x_1, x_2, \ldots, x_N)$ [18], [20].

The PC\(^3\) approach differs from traditional optimization methods such as gradient descent, GA and PSO that concentrate on a specific choice for the design variables (i.e., pure strategies) and on how to update that choice. Since the PC approach operates directly on probability distributions that optimize an associated Lagrangian, it offers a direct treatment for incorporating uncertainty, which is also represented through probabilities [21]. This is the most salient feature that this class of algorithms possesses – the search course is guided by a probability distribution over $x$, rather than a single value of $x$. By building such a probabilistic model of promising solutions and sampling the built model to generate new candidate solutions, PC allows the agents to significantly expand the range of exploration of the search space, and simultaneously focus on promising areas of solutions. As a result, this class of algorithms may provide a more robust and scalable solution to many important classes of optimization problems.

In [22] several benchmark testbeds were first studied to compare PCMAS with GA, which bear important characteristics of search space such as multimodality, nonlinearity and non-separability. Using these problems Huang et al. [22] showed that PC can outperform GA in the rate of descent, trapping in false minima and long term optimization. More recently, Huang and Chang [23] presented a study on PCMAS and concluded the search by the PCMAS is more robust than that of the GA.

Also recently, the summary for PC by kulkarni and Tai [7] indicated several advantages of PC that make it a competitive choice over other algorithms. In a nutshell, PC has the following advantages that can facilitate optimization tasks:

1. PC is a distributed solution approach in which each agent independently updates its probability distribution at any instant and can be applied to continuous, discrete or mixed variables, etc.
2. In PC the probability of the strategy set is always a vector of real numbers, thereby allowing the techniques of the straightforward optimization for Euclidean vectors, such as gradient descent, to be applicable.
3. PC is robust in the sense that the cost function can be irregular or noisy. PC also provides the sensitivity information about the problem in the sense that a variable with a peaky distribution plays a more important role in the optimization task than a variable with a broad distribution. In essence, peaky distribution provides the best choice of action that contributes most to the optimization of the global utility.
4. The computational and communication load is marginally less and equally distributed among all the agents. Thus the framework of PC enables an efficient

\(^3\)PCMAS is used in this paper to reflect the nature of PC in the context of multi-agent systems. Thus, PC and PCMAS are two exchangeable terms here.
A way to handle large-scale problems with many variables.

Since PCMAS has been proven as an effective methodology in solving combinatorial optimization, here we naturally intend to investigate how this approach can be employed to study complex biological problems, in particular, to solve the NP-hard combinatorial optimization problem for biological network construction.

In a nutshell, the core of PCMAS is an update process of each agent’s probability distribution $q_i(x_i)$ that prescribes a mixed strategy over player $i$’s possible pure strategies. Using Newton updating and enforcing the constraint on total probability, the following update rule at each iteration is obtained [23]:

$$q_i(x_i) \rightarrow q_i(x_i) - \alpha q_i(x_i) \times \left\{ (E[G | x_i] - E[G]) / T + S(q) + \ln q_i(x_i) \right\},$$  \hspace{1cm} (1)

where $\alpha$ plays the role of a step size. The step size is required since the expectations result from the current probability distributions of all the agents. To perform the gradient descent in probability space each agent must estimate the expected value of any of its actions, $E[G|x_i]$, from Monte-Carlo samples. Briefly, optimization proceeds in alternating rounds of Monte-Carlo sampling blocks, and updates to the agent’s probability distribution over the parameter value. To draw a Monte-Carlo sample each agent chooses the value for its parameter $x_i$ from its current probability distribution, and the world cost function $G(x)$ is evaluated.

The number of samples in each Monte-Carlo block determines accuracy of the expected cost estimate. If sampling the objective function is costly, one may wish to gain the most information from the least number of samples. The kernel density estimation implies and exploits weak prior knowledge about smooth interpolation between the sample points. Additionally, samples from the previous iterations may be re-used by geometrically weighting them according to their “age” in iterations. The imperfections that these augmentations introduce can be considered as another contribution to the bounded rationality term that broadens the probability distribution. The primary free parameters in the optimization are the Gaussian kernel width ($\tau$), the rate of cooling ($\delta T/T$), the number of Monte-Carlo samples per iteration, the proportional step size in the gradient descent ($\alpha$), and data-aging rate ($\gamma$).

The procedure for updating temperature $T$ in PCMAS, referred to as the annealing schedule, plays an important role in the efficiency and reliability of the approach [24]. Typically, a geometric schedule is applied, which involves multiplying the temperature by some fixed factor every several iterations. The detailed steps of the PCMAS are described in the following [24]:

1. **Initialize**
   - **A.** Set the initial parameters, including temperature $T$ and assign the annealing schedule for $T$ as a function of iteration number.
   - **B.** Specify the number of Monte-Carlo samples $N$ for each iteration.

2. **Minimize the lagrangian**
   - **A.** Monte-Carlo sample the agent probabilities to obtain $N$ joint actions.
   - **B.** For each sample, evaluate the objective function and compute the private utility for each agent.
   - **C.** Compute the expected utilities for the agents for each of their possible moves using a regression.
   - **D.** Update the probability distributions using Eq. (1).
   - **E.** Update the parameters. Use the assigned annealing schedule for $T$.

3. **Evaluate**
   - **A.** Determine the highest probability action for each agent.
   - **B.** Evaluate the objective function with this set of actions.

IV. **NETWORK OPTIMIZATION THROUGH PROBABILITY COLLECTIVES MULTI-AGENT SYSTEMS**

As discussed previously, PCMAS can be used as computational tools for solving optimization problems, including virtually any types of combinatorial optimization problems. Here PCMAS is employed to search for optimal combinations of various seed nodes in order to build optimal response networks for the case study on $M. tuberculosis$.

The study of optimization usually involves defining a performance measure that captures the idea of rate of improvement in solution quality, so that its change over time can be monitored for investigation. In many practical problems, a traditional performance metric is the “best-so-far” curve that plots the score of the best candidate solution that has been acquired thus far by function evaluation $N$.

In this section, the optimization process is conducted using a PCMAS with $n$ agents designated to search for $n$ seed nodes that optimize the sub-network scores. Note that in this case a function evaluation stands for the evaluation of a sub-network expanded by the seed nodes. We then conducted 50 runs for the PCMAS, with each run consisting of 6000 function evaluations in order to show the convergence of the search process. Figure 2 displays the results on averaged best-so-far score for which the ofloxacin (a drug treating $M. tuberculosis$ infection discussed in more detailed shortly) expression profile was obtained by exposing $M. tuberculosis$ to 10ug/ml ofloxacin for 6h against a reference of $M. tuberculosis$ exposure to DMSO (Dimethyl Sulfoxide).

The value of the averaged best-so-far score is calculated by averaging the best-so-far scores obtained at each 1000 function evaluations for all 50 runs. The vertical bars overlaying the metric curves in Figure 2 represent the 95-percent confidence intervals, i.e., the intervals within which
the actual mean best-so-far would reside with probability 0.95.

The range of the 95% confidence intervals indicates the robustness of the search algorithm, which tells us about how stable the estimate is. Wider confidence intervals in relation to the estimate itself indicate instability. On the other hand, narrow confidence intervals in relation to the point estimate shows that the estimated value is relatively stable. Therefore, the narrow confidence intervals in Figure 2 show that repeated PCMAS experiments would yield reliable models for the optimal (or suboptimal) construction of the response networks.

Figure 2. Best-so-far scores of Ofloxacin drug response optimized by PCMAS

Here we examine the drug targeting mechanism of *M. tuberculosis*. As an example we use ofloxacin, one of the class of drugs known as fluoroquinolones that trigger a DNA repair response [25]. Ofloxacin is an antibiotic that stops bacteria multiplication by inhibiting the reproduction and repair of their genetic material (DNA). Ofloxacin is used to treat pneumonia and bronchitis caused by Haemophilus influenzae and Streptococcus pneumoniae. Ofloxacin is known to have significant bactericidal activity against *M. tuberculosis* showing good penetration in macrophages, which are the survival and multiplication site for mycobacteria, and high tissue penetration [26]. Ofloxacin has been used in a variety of settings for the management of tuberculosis, but the precise mechanism by which any of the fluoroquinolones interfere with DNA synthesis has not yet been elucidated.

The response network generated by the PCMAS is shown in Figure 3 (using the same gene expression profile for the results shown in Figure 2). From the figure, it can be seen that the two important genes associated with the SOS gene repair response, *uvrB* and *recA*, indeed show up in the network. These results further confirms the prediction by Huang et al. [27]. In addition, this response network also displays the most notable hub in the network, *ltp1* (a hypothetical lipid-transfer protein), which can further pursued as a highly plausible drug target. Among the nodes that connect directly to *ltp1*, the appearance of *fadE27* also implies that the lipid degradation pathway is involved in the treatment of ofloxacin. The response network approach in conjunction with the optimization through PCMAS thus enables us to discern immediately the key nodes and the nature of different drug treatment.

V. CONCLUSIONS

In this study we demonstrated a robust combinational optimization method for biological network construction using the Probability Collectives Multi-agent Systems to examine and to interpret expression data in the context of network connectivity. The computational means is flexible enough to accommodate new information about protein or gene interactions, which facilitates hypotheses about biological problems on a systems level.

This approach to constructing response networks and optimization enables prediction and easy identification of crucial sub-networks in biological networks under the presence of specific drugs. Such comparison of the response networks thus facilitates the identification of important nodes for potential drug targets. In this paper, we employ the 95% confidence intervals as the measurement of robustness of the search algorithms. We show that the search of the PCMAS generates narrow confidence intervals and thus yields reliable models for the construction of the response networks. There are other statistics approaches that one can use to this end, such as cross-validation and bootstrapping [28]. This line of research is our proposed future work.

In summary, the PCMAS methodology provides an excellent systems biology tool for research on biological networks. In particular, the case study on ofloxacin drug response of *M. tuberculosis* highlights the interconnectivity of cellular processes and thus sheds light into target identification for future drug discovery efforts.

ACKNOWLEDGMENT

This work was supported by a grant (NSC98-2218-E-390-004) from the National Science Council of Taiwan.

REFERENCES


http://www.medicinenet.com/ofloxacin/article.htm


Figure 3. Ofloxacin response sub-network