Co-evolutionary Algorithm for Analyzing Gene Expression Data

Jimbo H. Claver[1*] and Isidore. S. Ngongo[2]
[1]American University of Afghanistan, Department of Mathematics & Statistics, Faculty Building 1, Office 15, Darul Aman Road P.O.Box 458, Kabul
[2]University of Paris 1, Pantheon - Sorbonne, Department of Applied Mathematics, 12 75231 Paris CEDEX 05, France
[*]Corresponding author

***Abstract*—We investigate the employment of the co-evolutionary genetic algorithm (CoGA) as a search mechanism in a support system for designing the prediction, functionality and interaction of expression level in population of gene expressions. To correctly identify interactions between various experimental conditions or expression levels, we proposed a fitness function which is a metric on two randomly chosen expressed populations and integrated for the whole population. Our result indicates that expression level variability is not simply the manifestation of noise in the system, but instead it is probably the results of processes involving stochastic state transitions. These results finally suggest that even if genes in a population are independent, the number of proteins (and/or mRNAs) more likely co-regulated.**

Keywords-co-evolution; genetic algorithm; gene expression data; constrained optimization; nonlinear modeling; and computational biology; biostatistics and data analysis

I. INTRODUCTION

Gene expression refers to the sum of processes that result in a particular level of a specified mRNA and protein in the cell. In many studies of cell biodynamic, gene expression is the starting point for elucidation of mechanism at the microscopic and molecular levels. While gene expression profiles are governed by very unclear rules, describing the coordination of gene expressions is therefore a central step towards understanding cellular systems (Raj et al, 2008). A comprehensive understanding of mostly subtle differences in expression levels of genes is crucial for elucidating the molecular mechanism of certain diseases and their treatment (Rainer Spang et al., 2010). It is known that gene regulatory interactions are context-dependent, active in some cellular states and not in others. Thus correlations in gene expression levels provide automatically noninvasive means to probe the state of small codependence of expression levels due to active regulatory connections between genes. Recent works have shown that gene expression is subject to stochastic fluctuations causing substantial cell-to-cell variability (Pedraza et. al, 2005), and that fluctuation in the concentration of regulatory proteins can cause corresponding fluctuation in the expression of a target only when the regulatory link is active (Rosenfeld *et. al*, 2005). Thus gene-to-gene correlation in expression may provide information about the activity state of regulatory connections without explicit perturbation of cellular components. However, such analysis may be complicated since gene expression level correlations arise not only from regulations but also from global variations in the overall rate of expression of all genes (Elowitz *at. al*, 2002). It have been reported that quantitative microarrays show that expression levels for more than 80% of genes are very low, with fewer than two mRNA copies per cell (Holstege et. al. 1998). At the current state of technology, expression levels for a substantial fraction of human genes can be assessed, and in the near future it is likely that the same analysis will be available genome-wide. The technology to generate large amounts of gene expression data is already available and will likely improve within the coming years. The bottleneck in dealing cogently with the upcoming data explosion is clearly focused on the development of data analysis tools that i) identify the difference in gene expression profiles and ii) capture or predict the mutual or co-dependence of expression levels of pairs of genes. To address the first question (i), statistical approaches focusing mainly on unsupervised learning procedures (West, 2000) have been used. Other methods such as hierarchical average linkage clustering (Eisen *et. al*, 1998), deterministic annealing based clustering (Alon et, 1998), self-organizing maps (Tamayo, 1989), Principal Component Analysis (Hilsenbeck, 1999) and Singular Value Decomposition (Alter *at. al*, 2000) have been used and have provided very broad overviews of the internal structure of the data, but the obvious shortcoming of these approaches is that information on the available data may be incomplete which may cause some biases. Addressing the second question (ii), Support Vector Machines (Brown et al, 2000) and the Bayesian regression approaches (West et. al, 2000) have been successfully applied. However it is clear that gene co-expression analysis goes beyond those methods since conflict between expression levels are expected and actually observed in practice. It is then important to look for new tools to detect and explore such conflicts, so as to generate scientific understanding. To further understand the noise correlations or simply correlations in expression levels of genes, we implement a new mathematical model and apply a co-evolutionary Genetic Algorithm to the gene expression dataset of (Beer-Tavazoie) (Beer and Tavarzoie, (2000) to validate it. This approach is important in finding the co-dependence in gene expression levels over time.

Such searches are often performed automatically, and solutions to the problem may be explored through Genetic Algorithms (GAs) as we assume that expression levels are correlated if and only if genes are co-regulated and vice versa. However it should be stressed that gene expression data is vast

and noisy, and techniques related to traditional statistics and other machine learning based approaches appear to be limited. Evolutionary approaches, and in this context co-evolutionary GAs, could be more appropriate to the task as they use pair genes and elucidate via a random search technique the level of co-expression. This approach could also be used in many other complex biological problems. This paper is organized as follows: In Section 2 we formulate our problem, and in Section 3 we present our co-evolutionary approach and results for estimating expression levels. Finally we end our work in Section 4 with a discussion and short conclusion.

II. Problem Specification

A. Co-evolutionary Genetic Algorithm (CoGA) for Beer and Tavazoie dataSet

Since the 1990s, various technical and numerical methods have had considerable success in Biology. GAs (Goldberg, 1989), a biologically-inspired technology, are randomized search and optimization techniques guided by the principle of evolution and natural genetics. They are efficient, adaptive and robust search process, producing near optimal solutions and have a large degree of parallelism (Rabindra Ku. Jena *et. al.*,2009). In general performing evolutionary algorithms for solving certain problems in quantitative biology which need optimization techniques for robust, fast and closed approximate solutions appear to be appropriate and natural (Rabindra Ku. Jena *et. al.*, 2009). Usually GA works with one set of variables or population but in case we have a pair of populations, we can adjust the GA to co-evolutionary GA. This is the reason why we particularly use a co-evolutionary GA rather than a standard GA in this context as we are looking not at one gene but pair of gene co-expressing. Moreover, the error generated in experiments with gene expression data can be handled with the robust characteristics of the algorithm, but to some extent, such errors may be regarded as contributions to expression diversity, which is a desirable property. Moreover, using co-evolutionary Genetic Algorithms to capture the co-regulation of genes through some minimal correlation of expression levels in the population may appear as a promising area of research. To the best of our knowledge there is no reported work in this direction at the moment. Co-evolutionary GAs are executed iteratively on a set of coded solutions, called populations, with three basic operators: selection, crossover and mutation. They use only so-called *fitness* information and probabilistic transition rules for moving to the next iteration. Because co-evolutionary GAs are based on manipulating populations of bit-strings using both crossover and point-wise mutation, it seems particularly suitable for our task as we can easily represent the gene expression as a string of real numbers same as in (Rabindra Ku. Jena *et. al.*, 2009). In this work, we suggest that both optimization and probabilistic approaches are necessary for developing a gene expression level "oriented" pattern searching algorithm. The presentation of the problem formulation is done in the section that follows.

B. Problem Setup

Given an optimization problem the most general formulation of constraints is:

1. Cost

$$f(v_k)=\begin{cases} f_1(v_k); & v_k \in P_1 \\ f_2(v_k); & v_k \in P_2 \end{cases} \quad (1)$$

2. Subject to: $g(v_k) \le \alpha_k$;

$$h(v_k) \le \delta_k \;\; {}_{\alpha\nu\delta} \;\; v_k \in A \quad (2)$$

where A is the feasible region, the optimization problem is subject to some inequality and equality constraints. This general setup could be reformulated in our context as:

i) Optimize $f(.)$:

$$\min_{p_1,\ldots,p_n} f(P_1,P_2)=\begin{cases} \min\limits_{p_{1,1},\ldots,p_{1,n}} f_1(P_1); & x_k, x_l \in P_1 \\ \min\limits_{p_{2,1},\ldots,p_{2,n}} f_2(P_2); & x_k, x_l \in P_2 \end{cases} \quad (3)$$

ii) Subject to:

$$\sum_{i=1}^{n} p_{1,i} x_{1,k}^2 + \sum_{j=1}^{n} p_{2,j} x_{2,l}^2 \le \sum_{k=1}^{n} x_{1,k}^2 + \sum_{l=1}^{n} x_{2,l}^2$$

Where the above part (i) is the idea that the expression levels in both genes must be down regulated at the same time and part (ii) is the expression level constraint showing that the overall perturbed expression levels must be less than the unperturbed expression levels.

iii) Feasibility constraints:

$$\sum_{i=1}^{n} p_{1,i} x_{1,k} \le 50; \quad \sum_{i=1}^{n} p_{2,j} x_{2,l} \le 25; \quad \sum_{i=1}^{n}\sum_{j=1}^{n} (p_{1,i} + p_{2,j}) \le 1 \quad (4)$$

Feasibility constraints set of controlled conditions on the randomly perturbed expression levels. The numbers 50 and 25 are some fractions of the total number of expressions, but are flexible and may be changed in order to calibrate the solution to the problem. As we aim in this work is to explain how the co-expression of genes affect their co-regulation, we must be aware that existing gene-gene interactions are random and could have some hierarchical organization representing some relevant genetic functions. We will next practically implement this concept by choosing two rows (genes) at random from a data set (Beer-Tavazoie, 2004). These two genes have a list of expression levels for factors $i = 1, 2$ and $k = 1, \ldots\ldots, n$ genes, which mathematically are represented by:

$$X_1 = \{x_{1,1}, x_{1,2}, \mathrm{K}\ , x_{1,n}\}; \quad X_2 = \{x_{2,1}, x_{2,2}, \mathrm{K}\ , x_{2,n}\}$$

Corresponding to both populations P_1 and P_2, the change in a population is closely related to the change in the probability affected to each component. Each element in each population will be evaluated by a fitness function, in order to produce in total two vectors:

$$f(P_1)=\{f(x_{1,1}), f(x_{1,2}), \mathrm{K}, f(x_{1,n})\}; \quad f(P_2)=\{f(x_{2,1}), f(x_{2,2}), \mathrm{K}, f(x_{2,n})\}$$

with

$$f\left(P_1, P_2\right) = \begin{cases} f_1 = p_{1,k} x_{1,k}^2 + p_{2,l} x_{2,l}; & x_k, x_l \in P_1 \\ \\ f_2 = p_{2,l} x_{2,l}^2 - p_{1,k} x_{1,k}; & x_k, x_l \in P_2 \end{cases} \qquad (5)$$

Equations 1 to 5 lead us to the following remark:

Remark 1. Two genes X_1 and X_2 with respective expression levels $\left(x_{1,1}, x_{1,2}, \mathrm{K}\ , x_{1,n}\right)$ and $\left(x_{2,1}, x_{2,2}, \mathrm{K}\ , x_{2,n}\right)$ are co-regulated if they are co-expressed and co-evolved and vice versa.

III. Gene Expression Data

Systematic experimentation has acquired a detailed understanding of the mechanisms of transcriptional regulations for a handful of well-studied genes, but we lack tools to understand how the dynamic noise correlation in may lead to gene expression level correlation remains unclear (Beer and Tavazoie, 2004) . Here as a formal step in this direction, we quantify the random codependence of two genes by using a complementary to classical genetic algorithm to find the optimal set of co-expressed genes. While our random search approach is applicable to any microarray expression datasets, here we use the Beer and Tavarzoie dataset with 255 total conditions. This data explores a diverse set of experimental conditions, and the significant redundancy improves signal noise. Noise in the expression data may present some limitations on our ability to predict gene expression and impose certain constraints on our approach (Beer and Tavazoie, 2004). We must deal with the fact that under each condition the measured gene expression level may be significantly different from than the actual expression. The degree of which co-regulated genes are actually co-expressed is important in biology (Kim et al., 2001). It was also found that within the large set of genes there are smaller group of genes with tighter co-expressions (Davidson *et al.*, 2000). In this work Beer and Tavarzoie gene expression profiles measured under 6015 stress conditions with 255 genes is employed for our analysis, the ultimate goal again is to predict the level of co-expression in population of genes. In the next section, we will present and apply our co-evolutionary genetic algorithm *(CoGA*).

IV. Co-Evolutionary Genetic Algorithm

In this section, we present and explain the flow of our algorithm. We start with two populations *P1, P2* of probabilities associated gene expression of *size 255*.

A. CoGA and Chart

Parameters: population size *n = 255*, number of selections $n_{sel} = 50$, number of crossovers $n_{cro} = 126$, number of mutations $n_{mut} = 79$, maximum number of iterations =100.

- Load the data D, consisting of a matrix with 6189 rows and n columns ;
- Normalize D;
- Choose two rows of *D* at random. Call them X_1, X_2. These are constant and do not change throughout the algorithm;

Initialization:

- Initialize $P_1(1), P_2(1)$ as populations of *n* probabilities
  $P_1(1) = \{p_{1,1}(1), p_{1,2}(1), p_{1,3}(1), \mathrm{K}\ , p_{1,n}(1)\}$
  $P_2(1) = \{p_{2,1}(1), p_{2,2}(1), p_{2,3}(1), \mathrm{K}\ , p_{2,n}(1)\}$;
- Evaluate $f(P(1)), f(P(2))$ across all elements of both populations;
- Rank the populations $P_1(1), P_2(1)$ in ascending numerical order with respect to the fitness function *f*.

Main loop:

- For j = 2 : m_p, do
  For r=1,2 do
  a) Selection: take the top n_{sel} population members from $\mathrm{P_r}(j-1)$ and copy to $\mathrm{P_r}(j)$;
  b) Crossover: at random from the top 50% of $\mathrm{P_r}(j-1)$ select two parents and "interlace" them to produce two children, and select one child. Check if each child is feasible; if not, copy a random volatility from $\mathrm{P_r}(j-1)$ to $\mathrm{P_r}(j)$ for each child. In this way produce n_{cro} members for population $\mathrm{P_r}(j)$;
  c) Mutation: from the top 10% of population $\mathrm{P_r}(j-1)$, select a volatility and mutate. Check if the new volatility is feasible. If so, then copy to population $\mathrm{P_r}(j)$; if not, then repeat step (c). Repeat this a total of n_{mut} times for both the population $\mathrm{P_r}(j-1)$;

d) Check that the sum of population $P_r(j)$ is 1. If it is not, then perturb those elements of the population given by crossover until the population sum is 1.

e) For all elements in the population $P_r(j)$, find their fitness values;

f) Rank the population in ascending numerical order with respect to the fitness function.

- Output $P_1(m_p), P_2(m_p)$, the final populations.

Remark 2. The parameters for *crossover*, *mutation* and *selection* are chosen to enable fast and simple computation. Each member of our two populations is a real number between 0 and 1, with the condition that the population members in each population sum to 1. Hence the "interlacing" from step 7(b) is performed by taking alternate digits from the respective decimal expansions of the two parents. Now, we will present the plots of whole populations which are typical outputs from runs of our algorithm.

B. *Outputs*

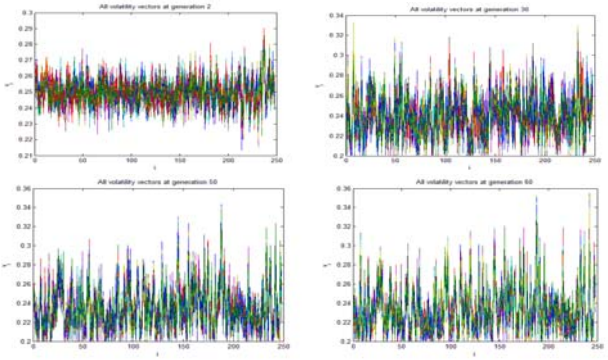

FIGURE I. ALL EXPRESSION LEVELS AT GENERATIONS 2, 30, 50 AND 60

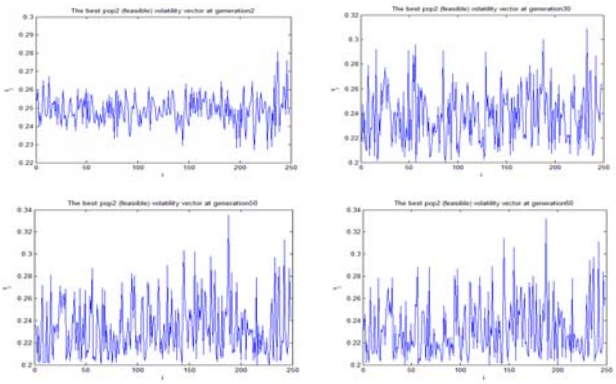

FIGURE II. THE OPTIMAL EXP. LEV FOR THE GENE POPULATION OVER GENERATIONS 2, 30, 50 AND 60. OBSERVE THAT AS THE GENERATION TIME INCREASES, WE ARE MORE LIKELY TO OBTAIN STABLE SOLUTIONS

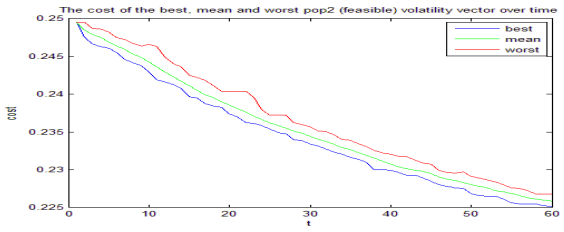

FIGURE III. PLOTS OF COST OF BEST, MEAN AND WORST POPULATION EXPRESSION LEVELS (EXP. LEV) OVER TIME. ACCORDING TO THE FITNESS FUNCTION F, THE BEST EXP. LEV PRESENTS A HIGHER CONVEXITY AFTER 60 GENERATIONS

Remark 3. Gene expression levels in our context are stochastic processes and are then by definition are *suboptimal*; therefore we do not need any concept of *global minimum* for the search.

V. CONCLUSION

Optimal gene expression levels can effectively be produced using a co-evolutionary genetic algorithm approach. The proposed algorithm begins with an initial population of expression levels obtained from a dataset (For example Beer-Tavazoie, 2000). Random fluctuations on *p* are applied and iterated to produce the next generation of expression levels. Relevant operators such as selection, crossover and mutation were also applied. The overall performance of the co-evolutionary GA for capturing optimal expression levels is satisfactory. Our innovative computational approach allows to efficiently capture the optimal co-expressed levels at each generation of the *CoGA*, which turns out to be important in practice, since it accounts for allow a better classification of gene function and therefore the basis of certain diseases. Our result also indicates that variability in the expression level is not simply the manifestation of noise in the system, but instead, it is probably the results of processes involving stochastic state transition. These results finally suggest that even if genes in a population are independent, their expression level could be co-regulated and as such optimized.

ACKNOWLEDGMENT

We thank NAIST, ISM, JMS, JSS, for the generous research support during the preparation of this work. We would also ike to express our gratitude to AUAF and the University of Paris 1 for providing us with some research facilities and software.

REFERENCES

[1] O. Alter, P. Brown, D. Botstein,, Singular Value Decomposition for Genome-wide Expression Data Processing and Modelling, PNAS 97 (18) (2000), 10101-10106.

[2] M.A. Beer and S Tavazoie, Predicting gene expression from sequence, Cell, vol. 117(2004), 185-198

[3] M. P. Brown et al., (2000), Knowledge Based Analysis of Microarray Gene Expression Data by using Support Vector Machines, PNAS 97. (1) (2000), 262-267.

[4] E. H. Davidson, D. R. McClay, L. Hood, Regulatory Gene Networks and the Properties of the Developmental Process, PNAS 100 (4), 1475-1480 (2003).

[5] M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, Cluster Analysis and Display of Genome-Wide Expression Patterns, PNAS 95, 14863-14969 (1998).

[6] M. B. Elowitz, A. J. Levine, E. D. Siggia, P. S. Swain, Stochastic Gene Expression in a Single Cell, Science 297 (2002), 1183-1186.

[7] D. E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning, Addison-Wesley (1989).

[8] S. Hilsenbeck, W. Friedrichs, R. Schiff, P. O'Connell, R. Hansen, K. Osborne, S. Fuqua, Statistical Analysis of Array Expression Data as Applied to the Problem of Tamoxifen Resistance, J. Natl. Cancer Inst. 91 (5) (1999), 453-459.

[9] F. Holstege et al., Dissecting the Regulatory Circuitry of a Eukaryotic Genome, Cell 95 (5) (1998), 717-728.

[10] A. Homaifar, C. X. Qi, S. H. Lai, Constrained Optimization via Genetic Algorithms, Simulation 62(4) (1994), 242-253.

[11] R. K. Jena, M. M. Aqel, P. Srivastava, P. K. Mahanti,, Soft Computing Methodology in Bioinformatics, Euro. J. Scientific Research 26 (2) (2009), 180—203.

[12] S. Kim et al, A Gene Expression Map for Caenorhabditis Elegans, Science 293 (2001), 2087-2092.

[13] J. M. Pedraza, A. van Oudenaarden, Noise Propagation in Genetic Networks, Science 307 (2005), 1965-1969.

[14] A. Raj, A. van Oudenaarden , Nature, Nurture, or Chance: Stochastic Gene Expression and its Consequences. Cell 135 (2) (2008, 216 – 226.

[15] J. M. Raser, E. K. O'Shea, Noise in Gene Expression: Origins, Consequences and Control, Science 309 (2005), 2010-2013.

[16] K. J. Rabindra, , Soft Computing methodology in bioinformatics, European Journal of Scientific Research vol 2, 2 (2009), 189-203

[17] Rainer *et al.*, Prediction and uncertainty in the analysis of gene expression profiles, (2010) http://www.bioinfo.de/talks/spang/main.html

[18] C. R. Reeves, Modern Heuristic Techniques for Combinatorial Problems, Wiley (1993).

[19] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. Lander, T. Golub, Interpreting Patterns in Gene Expression with Self-organizing Maps: Methods and Application to Hematopoietic Differentiation. PNAS 96 (6) (1999), 2907-2912.

[20] M. West, J. Nevins, J. Marks, R. Spang, C. Blanchette, H. Zuzan, DNA Microarray Data Analysis and Regression Modeling for Genetic Expression Profiling, ISDS Discussion (2000). Available at http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.8604