Swarm Intelligence based Gene Classification

Manasi M. Jahagirdar[*], **Prof. S. M. Kamalapur**[**]

[*] Department of Computer Engineering, KKWIEER, Nashik-422003, University of Pune, Maharashtra, India.
[**]Associate Professor, Department of Computer Engineering, KKWIEER, Nashik-422003, University of Pune, Maharashtra, India.

Abstract- The classification of genes is quite important in the understanding of gene regulation. The genes are grouped into transcription units for the purpose of construction and regulation of gene expression and synthesis of proteins. This knowledge further contributes as essential information for the process of drug design and to determine the protein functions of newly sequenced genomes. It is possible to use the diverse biological information across multiple genomes as an input to the classification problem. The purpose of this work is to show that Particle Swarm Optimization may improve the results of classification as compared to other algorithms. To validate the approach E.Coli complete genome is taken as the benchmark genome.

Index Terms- Classification, Drug Design, Protein Synthesis, Particle Swarm Optimization, Transcription Units.

General Terms- Algorithms.

I. Introduction

The developments in Bio-technological studies, have led to the design and development of large number of computer algorithms for bio-synthesis. Amongst which Gene synthesis constitutes of a major portion of research. The genomic era has opened up opportunities for analysis of complete gene organization, especially in bacteria. These have led to interesting conclusions about tendencies of genes and related functions. Data clustering is the process of grouping together similar multi-dimensional data into number of clusters or bins. Clustering algorithms have been applied to a wide range of problems, of which Gene analysis or Computational Biology form a considerable part.

The availability of complete genome sequences give rise to the need for more computational methods for discovering the regulation and synthesis of genomes. Classifying genes into different clusters or groups can thus enhance the knowledge of gene function. The approach takes into account several data sources including gene co-ordinates, regulatory control signals etc. Knowledge of gene organization is becoming increasingly important in the search for novel antibacterial targets and for understanding the processes involved in bacterial pathogenesis. Altogether, these facts point to the critical need for gene classification in targeted organisms.

Based on the sequence and annotations of the *E. coli* genome, the common features shared among pairs of adjacent genes within operons are analyzed against pairs of adjacent genes positioned at the boundaries of transcription units, but transcribed in the same direction. Their differences in terms of distances between genes, measured in base pairs, and in terms of functional relationships are evaluated. It is also shown that such differences can be used to develop a method to cluster genes in the whole genome sequence. This method might help the identification of transcription unit boundaries in other prokaryotic genomes.

In this paper, an approach based on Swarm Intelligence is presented to classify genes from target genome. The next section contains the survey of the similar work done before. Section III comprises of the proposed system description, followed by Datasets used, Results and Conclusions.

II. LITERATURE REVIEW

Several computational methods have been devised to cluster genes and group into a few general categories [9]. The first one being clustering by detecting Promoters and Transcription Terminators. A transcription unit can be identified if the promoter and the terminator genes of a gene sequence are identified [9]. Several algorithms have been developed to predict rho-independent transcription terminators [2, 3] efficient prokaryotic promoter-searching algorithm is not available as yet, even for the model organism E.coli [3].

The drawbacks of the method mentioned above can be overcome by the next method Construction of Hidden Markov Model (HMM). This method was reported to classify 60% of known genes in E.coli [4]. However, this method is difficult to apply in organisms where promoters and terminators are not as well characterized. The third method Probabilistic Machine Learning Approach using Variety of Data, estimates the probability of any consecutive sequence of genes on the same strand to be in a transcription unit and yielded 67% accuracy in E.coli [5]. With the generation of a large amount of gene expression data, co-expression pattern has been used as a tool to improve gene classification [6].

Bockhorst et al. [7] developed a Bayesian network approach to cluster and showed the method was able to predict 78% of E.coli transcription units with 10% false positives. However, these methods again are only applicable to organisms in which vast amounts of experimental data are available. The fourth category of methods proposed using artificial intelligence and genetic algorithms. This method was reported to have a maximum of 88% accuracy in identification of adjacent gene pairs to be in a transcription unit and found 75% of known transcription units in E.coli. This method has opened the possibility of transcription unit identification in bacterial genomes other than E.Coli.

III. DOMAIN CONCEPTS

Transcription Units

Transcription units are genetic regulatory system found in the organisms in which genes for functionally related proteins is clustered along a DNA. This feature allows protein synthesis to be controlled and coordinated in response to the needs of the cell. By generating proteins only as and when required, operons allows the cells to conserve energy. The part of the chromosome containing genes under consideration can be categorized into two regions: one that includes structural genes (i.e. genes that code for protein structure) and other is the regulatory region. This overall unit is known as an operon.

The gene pairs can be categorized as (i) WO (Within Operon) pair and (ii) TUB(Transcription Unit Border) pair. Adjacent genes that fall into the same transcription unit can be termed as WO gene pair. Whereas, the gene pair that lies at the borders of the transcription units are termed as TUB pairs.

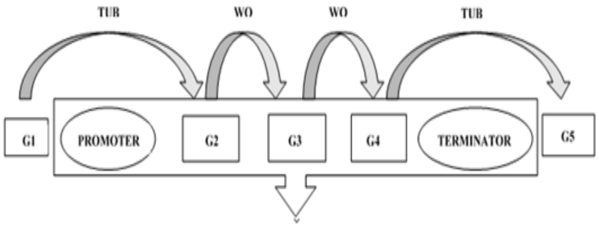

Figure 1. WO and TUB gene Pairs

Features for Gene Classification

Five properties were originally considered for the prediction of operons: (i) the intergenic distance, (ii) the metabolic pathway, (iii) the COG gene function, (iv) the operon length.

However, the gene length ratio and the operon length are not as suitable for operon prediction as the other three features. Thus the intergenic distance, the metabolic pathway, and the COG gene function are generally selected to predict operons. The intergenic distance property not only plays an important role in the initial step, but also yields good prediction results [20]. This property can be used to universally predict operons in bacterial genomes with a completed chromosomal sequence.

Intergenic Distance:

This property is defined as the distance (in bp i.e. base pairs) between two ORF's (Open Reading Frames). A drawback with intergenic distance is the fact that every species has different spacing. Also, some highly expressed operons are exceptions to this rule, which can also lead to correct identification of transcription units.

$$\text{Distance} = \text{Gene}_2\text{\_start-}(\text{Gene}_1\text{\_end}+1) \qquad (1)$$

Functional Relationship:

Operon contains genes that are often functionally related. The Clusters of Orthologous Groups (COG) and Metabolic Pathway are the most representative of the functional relationship category. The proteins that are produced are often present in the same pathway, or are a part of the same complex. Improved clustering is expected when incorporating this knowledge into the process.

Transcription Unit Length:

The length of a transcription unit is given by the number of genes within that unit. If it contains of just one single gene, then it is known as a singleton unit.

IV. IMPLEMENTATION DETAILS

4.1 Calculation of Pair Score

The properties used in this study are the intergenic distance, the metabolic pathway, and the COG gene function. The fitness values of the three properties are calculated based on the log-likelihood method as shown below.

Intergenic distance:

As shown, the equation given below is used to calculate the pair-score of intergenic distance [20].

$$\text{LLProperty}(\text{genei}, \text{genej}) = \ln\left(\frac{\frac{\text{Nwo(Property)}}{\text{TNWO}}}{\frac{\text{NTUB(Property)}}{\text{TNTUB}}}\right) \qquad (2)$$

Where, $N_{WO}(property)$ and $N_{TUB}(property)$ correspond to the number of WO and TUB pairs in the interval distance (10, 20, 30…). TN_{WO} and TN_{TUB} are the total pair numbers within WO and TUB, respectively.

Metabolic pathways:

The pathway pair-score is only taken into account when the two adjacent genes have the same pathway. Equation mentioned above is used to calculate the pathway pair-score.

COG gene function:

Equation mentioned above along with the following equation are used to calculate the COG pair-score [1].

$$LLCOGd(genei, genej) = \ln\left(\frac{1-\frac{Nwo(COG)}{TNWO}}{1-\frac{NTUB(COG)}{TNTUB}}\right) \qquad (3)$$

where LL_{COGd} ($gene_i$, $gene_j$) represents the pair-score of adjacent genes with a different COG gene function.

Fitness Calculation

Calculation of operon fitness value

While the pair-scores of each particle are calculated based on the metabolic pathway and the COG function, the fitness value of the operon in BPSO is calculated by multiplying the pair-score average with the gene number in the same operon.

Calculation of particle fitness value

Finally, the fitness value of a particle is calculated as the sum of the fitness values from all putative operons in the particle.

Particle Updating

Each particle is updated through an individual best ($pbest_i$), a global best (*gbest*) value, as well as other parameters. The *pbesti* value represents the position of the i^{th} particle with the highest fitness value at a given iteration, and *gbest* represents the best position of all *pbest* particles.

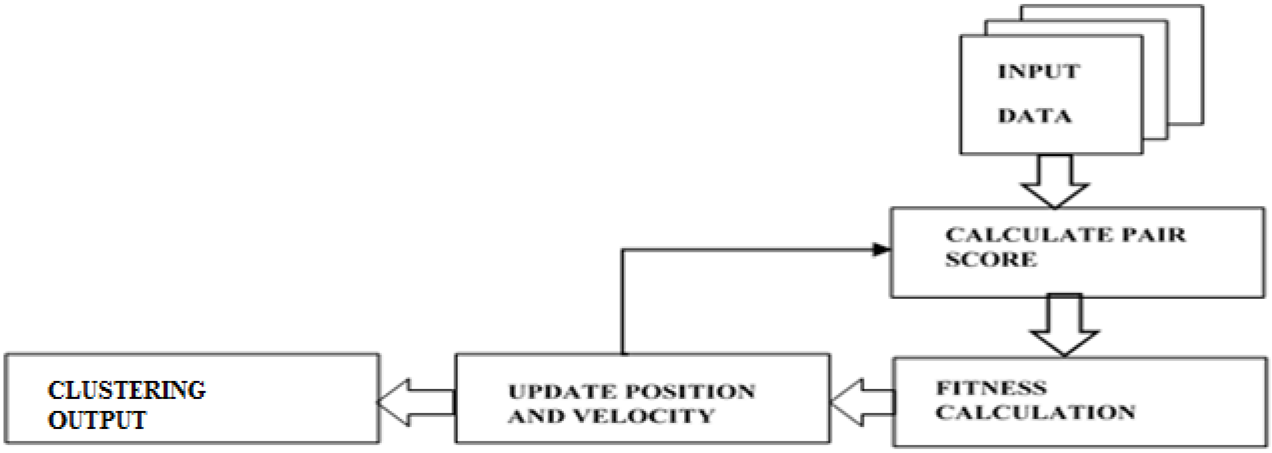

Figure 2: Block Diagram

V. PARTICLE SWARM OPTIMIZATION

Particle swarm optimization (PSO) is a population-based stochastic optimization technique developed by Kennedy and Eberhart in 1995 [4]. PSO has been developed through simulation of the social behavior of organisms, such as the social behavior observed of birds in a flock or fish in a school.

It describes an automatically evolving system. In PSO, each single solution is known as particle in the search space. Each particle uses their memory and knowledge gained by the swarm as a whole to find the optimal solution. The fitness value of each particle is evaluated by an optimized fitness function, and the particle velocity directs the movement of the particles.

Each particle adjusts its position according to its own experience during movement. In addition, each particle also searches for the optimal solution in a search space based on the experience of a neighboring particle, thus making use of the best position encountered by itself and its neighbor.

The entire process is reiterated a predefined number of times or until a minimum error is achieved. PSO has been successfully employed to many application areas; it obtains better results quickly and has a lower cost compared to other methods. However, PSO is not suitable for optimization problems in a discrete feature space. Hence, Kenney and Eberhart developed binary PSO (BPSO) to overcome this problem [20].

The basic elements of PSO are briefly introduced below:

(i) *Population:* A swarm (population) consists of N particles.

(ii) *Particle position, x_i:* Each candidate solution can be represented by a D-dimensional vector; the ith particle can be described as $x_i = (x_{i1},x_{i2}, . . . , x_{iD})$, where x_{iD} is the position of the ith particle with respect to the Dth dimension.

(iii) *Particle velocity, v_i:* The velocity of the ith particle is represented by $v_i = (v_{i1},v_{i2}, . . . ,v_{iD})$, where v_{iD} is the velocity of the ith particle with respect to the Dth dimension. In addition, the velocity of a particle is limited within $[V_{min},V_{max}]^D$.

(iv) *Inertia weight, w:* The inertia weight is used to control the impact of the previous velocity of a particle on the current velocity.

(v) *Individual best, $pbest_i$:* $pbest_i$ is the position of the ith particle with the highest fitness value at a given iteration.

(vi) *Global best, gbest:* The best position of all pbest particles is called global best.

(vii) *Stopping criteria:* The process is stopped after the maximum allowed number of iterations is reached.

In the PSO algorithm, each particle represents a candidate solution to the problem, and a swarm consists of N particles moving around a D-dimension search space until the computational limitations are reached.

VI. RESULTS AND DISCUSSION

Data set Preparation

The entire microbial genome data were downloaded from the GenBank database (http://www.ncbi.nlm.nih.gov/). The related genomic information contains the gene name, the gene ID, the position, the strand, and the product. The experimental operon data set of the E. coli genome was obtained from RegulonDB (http://regulondb.ccg.unam.mx/) [1], which contains highly reliable data of validated experimental operons of the E. coli genome. The metabolic pathway and COG data of the genomes were obtained from KEGG (http://www.genome.ad.jp/kegg/pathway.html) and NCBI (http://www.ncbi.nlm.nih.gov/COG/), respectively.

Result Set

The result set shows the error values for the clustering on various metabolic pathway datasets. The table shows the minimum amount of error value occurs if PSO algorithm is applied.

Table 1. Cluster Results

Dataset	Dimensions	Instances	Err. value
Metabolic Pathway data(Directed)	24	53414	1.813E-5
Metabolic Pathway data(Undirected)	29	65554	1.362E-4

VII. CONCLUSION

The analysis of gene data is gaining increasing importance. This study proposes a method to identify operons at complete genome level. The gene features like Intergenic distance, Metabolic pathways, gene clusters of orthologous groups make feasible input parameters for identification process. This identification can be very valuable contribution for various genetic applications.

ACKNOWLEDGMENTS

We wish to express our gratitude to Prof. Swati Bhavsar, Dept. of Microbiology, R. Y. K. college of science Nashik for her valuable help in the data interpretation. Also sincere thanks to Dr. Neelima Kulkarni for domain concepts.

REFERENCES

[1] Bockhorst,J., Craven,M., Page,D., Shavlik,J. and Glasner,J. "A Bayesian network approach to operon prediction." Bioinformatics, 19, 1227-1135(2003).

[2] Brendel,V. and Trifonov,E.N. "A computer algorithm for testing potential prokaryotic terminators." Nucleic Acids Res., 12, 4411-4427.(1984)

[3] Brendel,V. and Trifonov,E.N. "Computer-aided mapping of DNA-protein interaction sites." Proceedings of the Ninth International CODATA Conference, Jerusalem, Israel, pp. 17-20, 115-118.(1984)

[4] Craven,M., Page,D., Shavlik,J., Bockhorst,J. and Glasner,J. "A probabilistic learning approach to whole-genome operon prediction." Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology, La Jolla, CA, pp.116-127(2000).

[5] Ermolaeva,M.D., Khalak,H.G., White,O., Smith,H.O. and Salzberg,S.L. "Prediction of transcription terminators in bacterial genomes." J. Mol. Biol., 301, 27-33(2000).

[6] Ermolaeva,M., White,O. and Salzberg,S.L. "Prediction of operons in microbial genomes." Nucleic Acids Res., 29, 1216-1221(2001).

[7] J. Kennedy and R. Eberhart, "Particle swarm optimization," in IEEE International Joint Conference on Neural Network. vol. 4 Perth, Australia, 1995, pp. 1942-1948.

[8] Li-Yeh Chuang, Cheng-Huei Yang, Jui-Hung Tsai, and Cheng-Hong Yang," Operon Prediction using Chaos Embedded Particle Swarm Optimization", IEEE-ACM TRANS- ACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS(2013).

[9] L. Wang, J. D. Trawick, R. Yamamoto, and C. Zamudio, "Genome-wide operon prediction in Staphylococcus aureus," Nucleic Acids Res., vol. 32, pp. 3689-702, 2004.

[10] L. Y. Chuang, J. H. Tsai, and C. H. Yang, "Binary particle swarm optimization for operon prediction," Nucleic acids research, vol. 38, p. e128(2010).

[11] L. Y. Chuang, J. H. Tsai, and C. H. Yang, "Complementary Binary particle swarm optimization for operon prediction," Nucleic acids research, (2010).

[12] Mironov,A.A., Koonin,E.V., Roytberg,M.A. and Gelfand,M.S. "Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes." Nucleic Acids Res., 27, 2981-2989(1999).

[13] Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. "The use of gene clusters to infer functional coupling." Proc. Natl Acad. Sci. USA, 96, 2896-2901(2002).

[14] Ozoline,O.N., Deev,A.A. and Arkhipova,M.V. "Non-canonical sequence elements in the promoter structure. Cluster analysis of promoters recognized by Escherichia coli RNA polymerase." Nucleic Acids Res., 23, 4703 4709(1997).

[15] Sabatti,C., Rohlin,L., Oh,M. and Liao,J.C. "Co-expression pattern from DNA microarray experiments as a tool for operon prediction." Nucleic Acids Res., 30, 2886-2893(2002).

[16] Salgado,H., Moreno-Hagelsieb,G., Smith,T.F. and Collado-Vides,J. "Operons in Escherichia coli: genomic analyses and predictions." Proc. Natl Acad. Sci. USA, 97, 6652-6657(2000).

[17] Unniraman,S., Prakash,R. and Nagaraja,V. "Conserved economics of transcription termination in eubacteria." Nucleic Acids Res., 30, 675-684(2002).

[18] Vitreschak,A.G, Rodionov,D.A., Mironov,A.A. and Gelfand,M.S. "Regulation of riboavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation." Nucleic Acids Res., 30, 3141-3151(2002)

[19] Wolf,Y.I., Rogozin,I.B., Kondrashov,A.S. and Koonin,E.V. "Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context." Genome Res.,11, 356-372(2002).

[20] Yada,T., Nakao,M., Totoki,Y. and Nakai,K. "Modeling and predicting transcriptional units of Escherichia coli genes using hidden Markov models". Bioinformatics, 15, 987-993(1999).

[21] Zheng,Y., Szustakowski,J.D., Fortnow,L., Roberts,R.J. and Kasif,S. "Computational identi_cation of operons in microbial genomes". Genome Res., 12, 1221-1230(2002).

AUTHORS

First Author- M. M. Jahagirdar, Post Graduate Student, was with Pune University, Maharashtra, India. She is now with the Department of Computer Engineering, KKWIEER, Pune University, Maharashtra, India. (e-mail: jmanasi04@yahoo.com).
Second Author - Prof. S. M. Kamalapur is with the Computer Engineering Department, University of Pune, Maharashtra,India.(e-mail: snehal_kamalapur@yahoo.com).