INSILICO PROMOTER PREDICTION USING GREY RELATIONAL ANALYSIS

UMA DEVI TATAVARTHI[1], VENKATA NAGESWARA RAO PADMANABHUNI[1], APPA RAO ALLAM[2], RAMACHANDRA SRIDHAR GUMPENY[3]

[1]GITAM University, Visakhapatnam,
[2]JNTUK, Kakinada, [3]Endocrine and Diabetes Centre, Visakhapatnam
E-mail of corresponding auhor: uma@gitam.edu[1]

ABSTRACT

In machine learning, multiclass or multi-label classification is the special case within statistical classification of assigning one of several class labels to an input object. The multiclass problem is more complex than binary classification and less researched problem. In biology promoter is the DNA region where the transcription initiation takes place. Reliable recognition of promoter region is essential for understanding biological mechanical of the gene. This study proposes a new approach for predicting the promoter from the DNA sequence based on the modeling of Grey Relational Analysis (GRA). In order to construct a promoter prediction system, GRA approach is developed and applied to the real data set with 2111 samples of promoters and non-promoters of 4 species. The results of the current model are compared to those of traditional ones, logistic regression and back-propagation neural network. The results illustrate that the prediction of the proposed GRA model demonstrates better prediction accuracy than the conventional ones. The current results show that the proposed GRA provides a novel approach in predicting the promoter from a genome.

Key words: *Promoter Prediction, Grey Relational Analysis, Grey Systems, Classification*

1. INTRODUCTION

Promoters are modular DNA structures containing complex regulatory elements required for gene transcription initiation (Lin and Li, 2010). In genetics, a promoter is a region of DNA that facilitates the transcription of a particular gene. Promoters are typically located near the genes they regulate, on the same strand and upstream (towards the 5' region of the sense strand) (Promoter (Biology), 2007). Hence, the identification of promoters using machine learning approach (*insilico*) is very important for improving genome annotation and understanding transcriptional regulation. In recent years, many methods have been proposed for the prediction of eukaryotic and prokaryotic promoters. However, the performances of these methods are still far from being satisfactory.

Recent availability of several genome sequences has allowed whole genome analyses to unravel their functional properties. One of the challenges of the genomics is to understand how genomes are transcribed. Specifically, the genes are small sequences spread out along the genomes which, after being transcribed in mRNA, are translated in proteins turning out to be functional units of the cells. Although several sequences located within the genes or in their closed vicinity control their specificity of expression there are typical regions, the promoters, that define the *Transcription Start Site* (TSS) of the genes.

In this paper, a new approach is proposed for predicting the promoter using machine learning algorithm: Grey Relational Analysis. By applying the proposed method to the promoter and non-promoter sequences of Homo Sapiens, Drosophila Melanogaster, Escherichia Coli(commonly abbreviated as E. Coli.) and Saccharomyces Cerevisiae, the sensitivities and specificities obtained are: 94.9% and 96.54% for Homo Sapiens, 96.4% and 100% for Drosophila Melanogaster, 93.2% and 100% for Escherichia Coli and 98.3% and 94.14% for Saccharomyces Cerevisiae. The high accuracies indicate that this method can be used as an efficient method for the identification of eukaryotic and prokaryotic promoters. This approach can also be extended to

predict other species promoters at genome level also.

2. RELATED WORKS:

To accurately predict promoter regions, finding discriminative and informative features is the first and key step. As far as feature choice is concerned, there are two distinct types of features used in the area of promoter prediction: signal and context structure features. The most important signal features include CpG islands, transcription factor binding sites (TFBSs) such as TATA-box and CAAT-box, and initiator(Inr). PWM (Bucher, 1990) derives four weight matrices of TATA-box, cap signal, CCAAT-box and GC-box respectively.

PromoterScan (Prestridge, 1995) uses a weight matrix to score TATA-box. A weight matrix is a simple generative model for a short, ungapped sequence motif (Down and Hubbard, 2002). PWM is used extensively in signal feature extraction processing, as it can create a profile that represents the common feature across the training sequence. This profile can be used to scan new sequences and make a decision as to whether these sequences are related to the training group (Raychaudhuri, 2006).

Hidden Markov Model (HMM) (Krogh and Brown, 1994) is a more sophisticated method for feature extraction from sequences compared to PWM. HMM can represent spacer-included motifs (Murakami *et al*., 2000) of a sequence family. Generalized Hidden Markov Model (GHMM) (Stormo and Haussler, 1994) is used for generating multi-symbol strings in gene finding systems (Kulp *et al*., 1996). The Pol II promoter prediction program (Murakami, *et al.,* 2000) is built based on PromFD (Chen *et al.*, 1997) and utilizes HMM to acquire additional motifs.

McPromoter is developed based on GenScan (Burge and Karlin, 1997), and uses stochastic segment models (SSMs) (Ostendorf *et al*., 1995) which is a generalization of HMM to represent six segments of the promoter sequence from -250 to +50bp: upstream 1 and 2, TATA box, spacer, initiator and downstream(Uwe Ohler, 2006).

3. MATERIAL AND METHOD:

In order to accomplish the task of promoter prediction, positive datasets and negative datasets were taken from different curated databases available on the *World Wide Web*. These databases contain non-redundant collection of promoters, for which the transcription start site has been determined experimentally. Table. 1 describes the database name, species and number of sequences in that class, that are used in this study.

Class	Database	Taxonomic group/organism	No. of sequences
1	EPD	Homo Sapiens	608
2	DCPD	Drosophila Melanogaster	192
3	PromEC	E.Coli.	471
4	SCPD	Saccharomyces Cerevisiae	232
5	Essential Genes	Homo Sapiens	118
6	Essential Genes	Drosophila Melanogaster	100
7	Essential Genes	E.Coli	300
8	Essential Genes	Saccharomyces Cerevisiae	90
		Total	2111

Table: 1. *Class Label, Database, Taxonomic Group/organism, Number of Sequences considered*

4. Feature Extraction from DNA sequence:

Good input representations make it easier for the classifier to recognize underlying regularities. Therefore, good input representations are crucial to the success of classifier learning. The sequence or primary structure of a nucleic acid is the exact specification of its atomic composition and the chemical bonds connecting those atoms. From a one dimensional point of view, a DNA sequence contains characters from the 4-letter nucleic acid alphabet A,C,G or T. Let a given DNA sequence either promoter or non-promoter be: $S = s_1,s_2,s_3,\ldots\ldots s_{L-1}, s_L$, where $s_i \in \{A,C,G,T\}$, $1 \le i \le L$. The feature values of the sequence are calculated on the composition of the tri-mers. i.e., $\Delta 3 \equiv \{AAA, AAC, AAG, AAT, ACA, ACC, \ldots, TTT\}$. The 64 compositional frequencies are calculated as:

$$v_i = \frac{f_i}{|s|-2}, \qquad 1<=i<=64$$

f_i denote the frequency of occurrence of the i^{th} feature and $|S|$ denote the length of the sequence. The feature values v_i are normalized frequency counts(Nageswara Rao et al., 2008). Feature extraction is transforming the data in the high-dimensional space to a space of fewer dimensions. By applying Principal Component Analysis on the 2111x64 matrix it is reduced to 2111x41. The

obtained matrix is considered as input for the GRA Classifier.

5. Grey Relational Procedure:

Grey Relational Analysis (GRA) has been one of the most practical analytical tools (Deng, 1988; Liu and Lin, 2005; Nagai and Yamaguchi, 2004; Wen, 2004; Wen *et al*., 2006; Yamaguchi *et al*., 2006b). Several GRA models are developed and well summarized in (Liu and Lin, 2005; Wen *et al*., 2006; Yamaguchi *et al*., 2007; You *et al*., 2006). The GRA models provide appropriate tools for examining a rank of order of multiple objects with resemblance from an objective. In the recent years, GRA models have been applied to a lot of applications, such as decision making in computer science (Akabane et al., 2005; Yamaguchi et al., 2006a), system modeling, social science, geometry, chemistry, management(Lin et al., 2009; Nagai et al., 2005; Rui and Wunshch, 2005; Yamaguchi *et al*., 2005; Yamaguchi et al. 2006a), economics, marketing research (Yamaguchi *et al*., 2004).

Owing to the usefulness and robustness of GRA, a new approach for predicting the promoter is proposed, based on GRA. This study applies GRA model to predict promoter from the data set of 2111 samples from 4 species. The results of the current model are compared to those of traditional and hybrid models, which include conventional logistic regression, logarithm logistic regression and ANN approaches. The result shows that in predicting the promoter, GRA model gave better performance and demonstrates stronger prediction power than conventional logistic regression and ANN approaches. A system that has no information is defined as a black system, while a system that is full of information is called white. Thus, when the information of a system is either incomplete or undetermined, it is defined as grey system. The grey number in grey system represents a number with incomplete information. The grey element represents an element with incomplete information. The grey relation is the relation with incomplete information. This section describes the basic definitions of grey relational analysis, GRA. The inner product and metric of two vectors are first defined. What follows are properties of norm space, grey relational space, grey relational grade for both globalized and localized grey relationships.

The GRA includes local relation and global relation analysis. Grey relational-based classifier is used to classify promoter/non-promoters. GRA is a method to determine the relation of a discrete data to other sequence data(Chang, 2000; Wu and Chen, 1999). The novel grey relational procedure is introduced. Suppose the test sequence $\varphi_i(0)$, i=1,2,3,…,n, and K comparative sequences $\Phi(k)=[\varphi_1(k), \varphi_2(k), \varphi_3(k),\ldots, \varphi_i(k),\ldots, \varphi_n(k)]$, k=1, 2, 3, …, K, can be represented as

$$\Phi_{test}=[\varphi_1(0)\varphi_2(0)\varphi_3(0)\ldots.\varphi_i(0)\ldots.\varphi_n(0)]$$

$$\Phi_{comp}=\begin{bmatrix}\Phi(1)\\ \Phi(2)\\ .\\ .\\ .\\ \Phi(k)\\ .\\ .\\ .\\ \Phi(K)\end{bmatrix} = \begin{bmatrix}\varphi_1(1)\varphi_2(1)\varphi_3(1)\ldots.\varphi_i(1)\ldots.\varphi_n(1)\\ \varphi_1(2)\varphi_2(2)\varphi_3(2)\ldots.\varphi_i(2)\ldots.\varphi_n(2)\\ .\\ .\\ .\\ .\\ .\\ \varphi_1(k)\varphi_2(k)\varphi_3(k)\ldots.\varphi_i(k)\ldots.\varphi_n(k)\\ .\\ .\\ .\\ \varphi_1(K)\varphi_2(K)\varphi_3(K)\ldots.\varphi_i(K)\ldots.\varphi_n(K)\end{bmatrix}$$

Compute the absolute deviation of the test sequence Φ_{test} and k comparative sequence $\Phi(k)$ by

$$\Delta\varphi_i(k) = [\varphi_i(0) - \varphi_i(k)]$$

The deviation matrix $\Delta\Phi$ can be represented as

$$\Delta\Phi=\begin{bmatrix}\Delta\varphi_1(1)\Delta\varphi_2(1)\ldots.\Delta\varphi_i(1)\ldots\Delta\varphi_n(1)\\ \Delta\varphi_1(2)\Delta\varphi_2(2)\ldots.\Delta\varphi_i(2)\ldots\Delta\varphi_n(2)\\ \Delta\varphi_1(3)\Delta\varphi_2(3)\ldots.\Delta\varphi_i(3)\ldots\Delta\varphi_n(3)\\ .\\ .\\ .\\ \Delta\varphi_1(k)\Delta\varphi_2(k)\ldots.\Delta\varphi_i(k)\ldots\Delta\varphi_n(k)\\ .\\ .\\ .\\ \Delta\varphi_1(K)\Delta\varphi_2(K)\ldots.\Delta\varphi_i(K)\ldots\Delta\varphi_n(K)\end{bmatrix}$$

The grey relational grades $r(k)$ can be calculated as:

$$r(k) = \exp\left[-\xi\left(\frac{ED(k)}{\Delta\varphi_{max}-\Delta\varphi_{min}}\right)^2\right],\ \xi\in(0,25)$$

$$= \exp\left[-\xi\left(\frac{\sqrt{\Sigma_{i=1}^{n}(\Delta\varphi_i(k))^2}}{\Delta\varphi_{max}-\Delta\varphi_{min}}\right)^2\right]$$

$\Delta\varphi_{max}=\max_{\forall k}[\max_{\forall i}\Delta\varphi_i(k)]$

$\Delta\varphi_{min}=\min_{\forall k}[\min_{\forall i}\Delta\varphi_i(k)]$

Where ED(k) is the Euclidean Distance(ED) between vector Φ_{test} and vector $\Phi(k)$; $\Delta\varphi_{min}$ and $\Delta\varphi_{max}$ are the minimum and maximum values of the matrix $\Delta\Phi$, respectively; ξ is a recognition coefficient with parameter interval (0,25), ξ=15 was chosen in this study. The grey relational grades $r(k)$ are inversely proportional to the distances. If the test vector Φ_{test} is similar to any comparative vector $\Phi(k)$, the grade $r(k)$ will be a maximum value. GRA uses the grey relational grade to measure the relationship between the reference sequence data and comparative sequences data. The dimension of grey relational vector $\Gamma=[r(1),r(2),\ldots..,r(k),\ldots..,r(K)]$ can be reduced from K-dimension to m-dimension by

$\gamma_j=\sum_{k=1}^{K} r(k)w_{kj}$, j= 1,2,3,…m

$$w_{kj}=\begin{cases}1, & k\epsilon\ Class\ j,\\ 0, & k\notin Class\ j.\end{cases}$$

The final grey grade g_j that an unknown vector Φ_{test} belongs to Class j can be derived from the following equation:

$$g_j=\frac{\gamma_j}{\sum_{j=1}^{m}\gamma_j}\ ,j=1,2,3\ldots.,m$$

which defines the decision for classifying an unknown vector Φ_{test}.

6. RESULTS AND DISCUSSION:

The following parameters: sensitivity (Sn), specificity (Sp), and Precision(Pr) are used to evaluate the predictive performance of the classifier. Let TP Number of true positive instances, FN Number of false negative instances, FP Number of false positive instances and TN Number of true negative instances.

True Positive Rate(TPR)/Sensitivity(Sn)/ Recall = TP/(TP+FN)

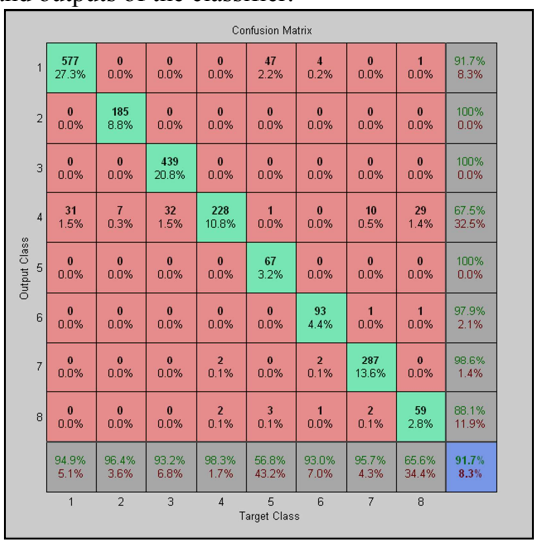
False Negative Ration(FNR)/Miss = FN/(TP+FN)

True Negative Rate(TNR)/ Specificity(Sp) = TN/(TN+FP)

False Positive Rate(FPR)/Fall = FP/(TN+FP)

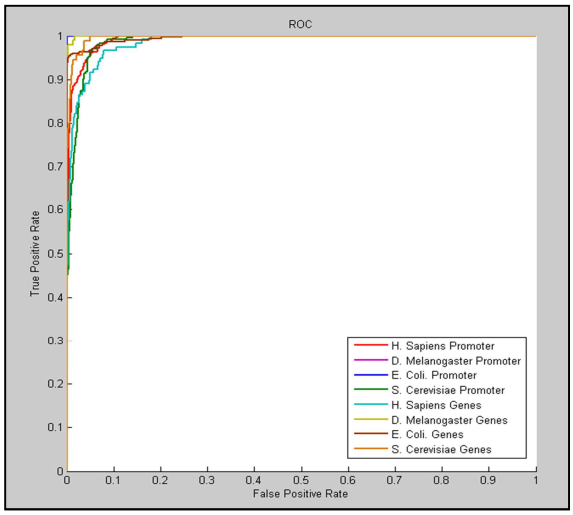
Positive Predictive Value(PPV)/ Precision(Pr) = TP/(TP+FP)

A confusion matrix is drawn between the targets and outputs of the classifier.



The diagonal cells show the number of promoters/non-promoters that were correctly classified for each class. The off-diagonal cells show the number of promoters/non-promoters that were misclassified (e.g. promoters of Homo Sapiens were classified as Promoters of E.Coli etc.). The True Positive Rate/Sensitivity and False Negative Rate/Miss for each class are presented in the last row in green and red colors respectively. The Precision of each class is indicated in the last column(in green). The blue cell shows the total percentage of correctly predicted promoters/non-promoters (in green) and the total percentage of incorrectly predicted promoters/non-promoters(in red).

A Receiver Operating Characteristic (ROC) curve, a plot of the true positive rate (sensitivity) versus the false positive rate(1-specificity) is also drawn. As the curves are towards the Y-Axis and away from the X-Axis, the performance of the classifier can be considered as good.



7. CONCLUSION

The successful prediction of promoters with high accuracy using Grey Relational Analysis clearly indicates that the novel method has a promise as an approach for successful Prokaryotic and Eukaryotic promoter prediction. The experience gained from the above example shows that *n-mer* frequencies and Grey Relational Analysis is quite suitable to classify between promoter and non promoter regions. The main aim of this paper is to develop an efficient tool that can discriminate between promoter and non promoter in a given sequence with high accuracy. High result accuracy of the program indicated that the novel approach can be further successfully used for the prediction of Eukaryotic promoters in entire chromosome. This method is currently applied for estimating the number of promoters in different chromosomes of the human genome. Another challenge being addressed is the localization of promoters rather than a simple classification similar to the one at present. It is expected that the promising results using GRA will improve the performance of bio-molecular sequence analysis and promoter prediction in particular.

8. ACKNOWLEDGEMENTS

The authors would like to thank GITAM University, JNTUK, Kakinada and Acharya Nagarjuna University for providing computational facility and access to e-journals to carry out this research.

REFERENCES:

Akabane, T., Yamaguchi, D., Li, G. D., Mizutani, K., & Nagai, M. M. (2005). Kansei information processing model applied multi-agent systems based on grey theory. *Japanese Journal of Japan Society of Kansei Engineering*, 5(4), 73–80.

Bucher, P. (1990). Weight matrix descriptions of four eukaryotic rna Polymerase II promoter elements derived from 502 unrelated promoter sequences. *Molecular Biology*, 212: 563-589.

Burge, C. and Karlin, S. (1997). Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology*. 268: 78-94.

Chang W-C (2000) A comprehensive study of grey relational generating. *J Chinese Grey Assoc.,* 1:53–62

Chen, Q. K., Hertz,G.Z. and Stormo, G.D.(1997). PromFD 1.0: a computer program that predicts eukarotic pol II promoters using strings and IMD matrices. *Comput. Applic. Biosci*. 13(1): 29-35.

Deng, J. L. (1988). *Grey system*. Beijing: China Oceans Press.

Down, T. A. and T. J. P. Hubbard (2002). Computational Detection and Location of Transcription Start Sites in Mammalian Genomic DNA. *Genome Research*,12: 458-461.

Krogh, A. and M. Brown (1994). Hidden Markov models in Computational biology applications to protein modeling. *Journal of Molecular Biology*, 235(5): 1501-1531.

Kulp, D., Haussler, D., and Eeckman, F.H.(1996). A Generalized Hidden Markov Model for the Recognition of Human Genes in DNA. *Proc Int Cong Intell Syst Mol Biol*. 4**:** 134-142.

Lin, H. and Li, Q.Z.(2010). Eukaryotic and prokaryotic promoter prediction using hybrid approach. *Theory of Biosci*: Springer-Verlag. DOI 10.1007/s12064-010-0114-8.

Lin, S.L., Wu, S.J., Ma, H.L., & Wu, D.B. (2009). Development of credit risk model in banking industry based on GRA. *In Proceedings of the eighth international conference on machine learning and cybernetics*, Baoding, (pp. 2903–2909), 12–15 July.

Liu, S. F., & Lin, Y. (2005). *Grey information: Theory and practical applications*. London: Springer.

Murakami, K.,Ohta, Y. and Tanikawa, K.(2000). A Transcription Regulatory Region Analysis System. *Genome Informatics*.11:297-297).

Nagai, M., & Yamaguchi, D. (2004). *Elements on grey*

system theory and applications. Tokyo: Kyoritsu-Shuppan.

Nagai, M., Yamaguchi, D., & Li, G. D. (2005). Grey structural modeling. *Journal of Grey System*, 8(2), 119–130.

Nageswara Rao, P.V., Uma Devi, T., Kaladhar, DSVGK., Sridhar, G.R. and Appa Rao, Allam.(2009). A Probabilistic Neural Network Approach for Protein Superfamily Classification. *Journal of Applied and Theoretical Information Technology*, Vol.6. No.1. pp 101-105.

Ostendorf, M., V. Digalakis, et al. (1995). From HMMs to segment Models: a unified view of stochastic modeling for speech recognition. *IEEE Transactions on Speech and Audio Processing*. 4(360-378).

Prestridge, D. S. (1995). Predicting Pol II promoter Sequences using Transcription Factor Binding Sites. *Journal of Molecular Biology*, 249: 923-932.

Promoter(Biology). (2007). Retrieved from wikipedia.org: http://en.wikipedia.org/wiki/Promoter_%28biology%29

Raychaudhuri, S., Ed. (2006). *Computational Text Analysis for Functional Genomics and Bioinformatics*, Oxford University Press Inc., New York.

Rui, X., & Wunshch, D. C. II, (2005). Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3), 645–678.

Stormo, G. D. and D. Haussler (1994). Optimally parsing a sequence into different classes based on multiple types of evidence. *Int Conf Intell Syst Mol Biol*. , Stanford, California, USA.

Uwe Ohler. (2006). Identification of core promoter modules in Drosphila and their application in accurate transcription start site prediction. *Nucleic Acids Research*, Vol.00, No.00, 1-8.

Wen, K. L. (2004). *Grey systems: Modeling and prediction*. Tucson: Yang's Scientific Research Institute.

Wen, K. L., Changchien, S. K., Yeh, C. K., Wang, C. W., & Lin, H. S. (2006). *Apply MATLAB in grey system theory*. Taipei: CHWA Publisher.

Wu JH, Chen C-B (1999) An alternative form for grey relational grades. *J Grey Systems,* 1:7–12

Yamaguchi, D., Li, G. D., Mizutani, K., Akabane, T., Nagai, M., & Kitaoka, M. (2006a). A k-means clustering approach based on grey theory. *In The 2006 IEEE international conference on systems, man, and cybernetics*, Taipei (Vol. 00137, pp. 2291–2296).

Yamaguchi, D., Kobayashi, T., Mizutani, K., Akabane, T., & Nagai, M. (2004). Marketing research method based on grey theory considering with consumer's kansei. Japanese *Journal of Japan Society of Kansei Engineering,* 4(2), 101–106.

Yamaguchi, D., Li, G. D., Mizutani, K., Akabane, T., Nagai, M., & Kitaoka, M. (2006b). A realization algorithm of grey structural modeling with MATLAB. *Proceeding 2006 IEEE international conference on cybernetics & intelligent systems*, 528–533.

Yamaguchi, D., Li, G. D., & Nagai, M. (2005). New grey relational analysis for finding the invariable structure and its applications. *Journal of Grey System*, 8(2), 167–178.

Yamaguchi, D., Li, G. D., & Nagai, M. (2007). Verification of effectiveness for grey relational analysis models. *Journal of Grey System*, 10(3), 169–182.

You, M. L., Wang, C. W., & Yeh, C. K. (2006). The development of completed grey relational analysis toolbox via Matlab. *Journal of Grey System*, 9(1), 57.