A Novel Algorithm for Prediction of Protein Coding DNA from Non-coding DNA in Microbial Genomes Using Genomic Composition and Dinucleotide Compositional Skew

Baharak Goli, B.L. Aswathi, and Achuthsankar S. Nair

State Inter-University Centre for Bioinformatics,
University of Kerala,
Trivandrum 695581, India
baharak_goli@yahoo.com

Abstract. Accurate identification of genes encoding proteins in genome remains an open problem in computational biology that has been receiving increasing consideration with explosion in sequence data. This has inspired us to relook at this problem. In this study, we propose a novel gene finding algorithm which relies on the use of genomic composition and dinucleotide compositional skew information. In order to identify the most prominent features, two feature selection techniques widely used in data preprocessing for machine learning problems: CFS and ReliefF algorithm applied. The performance of two types of neural network such as multilayer perceptron and RBF network was evaluated with these filter approaches. Our proposed model led to successful prediction of protein coding from non-coding with 96.47% and 94.18 % accuracy for MLP and RBF Network respectively in case of CFS and 94.94 % and 92.46 % accuracy for MLP and RBF Network respectively in case of ReliefF algorithm.

Keywords: Identification of protein coding DNA, genomic composition, dinucleotide compositional skew, feature selection methods, machine learning.

1 Introduction

With the speedy development of genome sequencing technologies and databases the amount of genomic data has been increasing almost exponentially. Till date, 1646 microbial genomes have been sequenced successfully, while sequencing of more than 4900 microbial genomes are currently in progress. The most important biologically functional parts of DNA sequence of any organism are its genes. Genes control all major biological processes of an organism through the complex expression of their cognate gene products. Therefore, gene identification from genome sequences and

N. Meghanathan et al. (Eds.): CCSIT 2012, Part II, LNICST 85, pp. 535–542, 2012.
© Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2012

their biological importance in most of living organisms is a challenging issue in bioinformatics and computational biology [1]. A Large number of computational algorithms have been proposed to predict protein-coding DNAs over the last two decades [2], [3], [4], [5], [6]. These algorithms can be divided into two categories. The approaches in the first category, the ab initio programs, are based on various statistics of DNA sequences, which usually use training datasets from already known coding and non-coding sequences to find the discriminant function [7], [8]. The second category known as homology-based programs, includes algorithms that are based on similarity search in large databases of genomic information [9], [10], [11], [12], however these algorithms are not perfect due to lack of experimentally verified proteins in databases. Recently another type of gene prediction tools have developed which combine the result from two or more gene finding tool and have a higher performance[13]. In this study we have developed a statistics-based approach to discriminate protein coding DNA from non-coding DNA based on the use of genomic composition and dinucleotide compositional skew information.

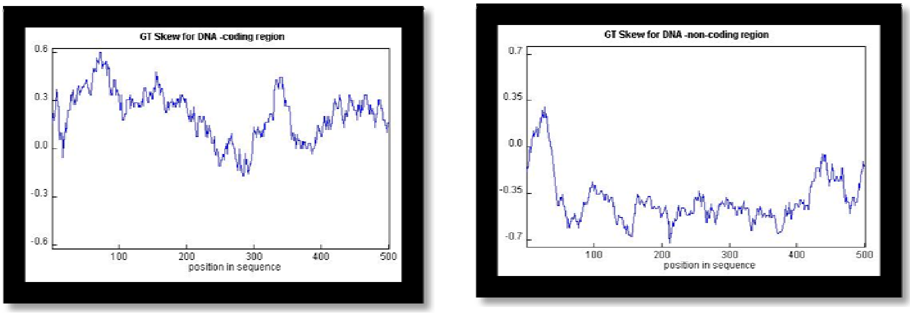
2 Materials and Methods

2.1 The Dataset

We selected Escherichia coli k12 MG1655 for our development available in the Integrated Microbial Genomes (IMG) database [14]. Protein coding sequences and non-coding sequences shorter than 100 were excluded. The final number of protein coding DNA was 4334 and non-coding was 1823.

2.2 Feature Transformation

The quantitative characteristics of DNA that we took into our consideration included nucleotide composition and nucleotide compositional skews. The nucleotide composition comprised of the frequencies of base (4 features), dimer (16 features), trimer (64 features), AG content, GC content, GT content and atomic composition (5 features). Nucleotide compositional skews included GC skew, AT skew, AG skew, TC skew, GT skew, Purine skew and Kito skew. The dinucleotide compositional skew indices are calculated as differences between single strand frequencies of certain nucleotides divided by the total of the frequencies. For example, the GT skew is computed as the value of ([G]-[T])/([G]+[T]) where brackets indicate the absolute values of correspondent nucleotides. The skew method plays an important role in studying the base composition bias related to both DNA replication [15] and transcription processes [16]. Graphical representation of AT skew for a protein coding DNA and non-coding of same size (500 nt) shows remarkable discriminative pattern (Figure 1).



GT skew for coding region
id |646311903 |
for Ecoli K-12 MG1655
Maximum value:0.792
Minimum value:-0.172

GT skew for non-coding region
id |NC_000913_4266331_4266831|
for Ecoli K-12 MG1655
Maximum value:0.32
Minimum value:-0.743

Fig. 1. Graphical representations of AT skew for a coding and non-coding DNA sequence

2.3 Feature Subset Selection Techniques

The performance of a classifier totally depends on the relations between a number of features, training set size and classifier complexity. Hence large number of features comprises an obstacle to efficiency of classification algorithms by increasing computation time and [17] over-fitting the training data set [18] a smallest subset of important and prominent features that attains maximal classification performance, faster classification models and smaller data bases should be retained. Feature selection is one of the important techniques in data preprocessing for machine learning and data mining problems. It trashes out irrelevant, noisy and redundant features, speeds up the data mining algorithm and improves prediction accuracy [19]. For this purpose we adopted two well-known feature selection techniques such as CFS (correlation-based feature selection) [20] and ReliefF feature selection algorithm [21] to select the appropriate discriminatory set of features. We briefly describe these feature selection algorithms below. In this study 96 features generated from the transformation step explained above and after feature selection a total of 19 features remained.

2.3.1 Correlation-Based Feature Selection Algorithm (CFS)

The correlation-based feature selection algorithm has been proved as a powerful technique in removing both unrelated and redundant features. It assesses the significance of subsets of features and uses a best first-search heuristic. This heuristic algorithm considers the relevance of individual features for predicting the class along with the level of correlation among them. The main logic in CFS is that good feature subsets include those features that are highly correlated with the target class and uncorrelated with each other. The CFS function is defined as follows:

$$M_s= k\ \bar{r}_{cf} / \sqrt{}\ (k+k(k-1)\ \bar{r}_{ff}) \qquad (1)$$

where $M_s$is the heuristic subsets evaluator function when the subset (s) containing k features, $\bar{r}_{cf}$ is the feature-classification correlative average value where (f∊S) and $\bar{r}_{ff}$is the feature-feature correlation average value.

2.3.2 Relief Feature Selection Algorithm (ReliefF)

ReliefF is a simple and powerful feature selection technique. It is an extension of Relief algorithm [22] developed to use in classification problems. It evaluates the relevance of features with strong dependencies between them. At each step of an iterative process, an instance k is selected randomly from the dataset and the weight for each feature is updated based on the distance of k to its NearHit (nearest neighbors from the same class) and NearMiss (nearest neighbors from each of the different classes). This process is iterated n times, where n is a predefined parameter. Generally n is equal to the number of samples in dataset. Finally the best subset includes those features with relevance above a chosen threshold t.

2.4 Building of Neural Networks

Artificial neural network is a supervised learning algorithm used commonly to solve classification problems. In this study, two types of neural networks configurations, multilayer perceptron trained by the back propagation algorithm and RBF network, were chosen. The weka suite, machine learning workbench developed in java programming language was used for implementation [23].

Back-propagation networks are apparently the most common and widely used algorithm for training supervised neural networks [24], [25], [26]. It has less memory requirements than most techniques and usually reaches an adequate error level significantly fast. It can be adopted on most types of networks, however it is most suitable for training multilayer perceptrons. RBF networks are supervised neural networks which are popular alternative to the MLPs which employ reasonably lesser number of locally tuned units and are adaptive in nature. They are widely used for pattern recognition and classification problems. RBF networks are suitable for modeling nonlinear data and can be trained in one phase instead of using an iterative process as in MLPs [27], [28].In this study, the training set consisting of 4334 coding and 1823 non-coding elements was given to the each network in the 10-fold cross-validation scheme. The accuracy of classification using each network was measured. For the comparison of the networks, the time taken by each network to build the model was also noted.

3 Results

3.1 Evaluation of Performance

The performance of our proposed models were estimated using standard 10-fold cross-validation in which the whole dataset is randomly partitioned into ten evenly-sized subsets. During each test, a neural network is trained on nine subsets and then tested on the tenth one. This method is repeated ten times so that each subset is used for both training and testing on each fold. Performance is measured for each test set, and the mean is reported as overall accuracy. Several measures were used to evaluate the performance of the neural networks (TP, TN, FP and FN representing true positive, true negative, false positive and false negative respectively).

Specificity=TN/ (TN+FP)*100
Sensitivity=TP/ (TP+FN)*100
Precision=TP/ (TP+FP)*100
Matthews correlation coefficient = (((TP*TN)-(FP*FN)))/(√(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN))
Accuracy= TP+TN/ (TP+TN+FP+TN)

Table 1. Performance of gene finding tools

Method	Sn (%)	Sp (%)	Pc (%)	MCC	Acc (%)
Multilayer perceptron + cfs	97.09	95.00	97.88	0.91	96.47
RBF Network + cfs	91.99	99.39	99.72	0.87	94.18
Multilayer perceptron + relief-f	95.29	94.13	97.47	0.88	94.94
RBF Network+relief-f	90.23	97.75	98.96	0.8378	92.46

The comparison of performances of different neural networks is shown in Table 1.Multilayer Perceptronin conjunction with correlation-based feature selection algorithm produced highest classification result. Time taken to build the models were 81.56 seconds for multilayer perceptron and 5.03 seconds for RBF network in case of correlation-based feature selection and 83.86 seconds for multilayer perceptron and 7.19 seconds for RBF network in case of relief feature selection algorithm in the same work station.

Self-consistency test and independent test (shown in Table 2) were also performed to evaluate the prediction model. Self-consistency test reflects the consistency of the developed model. It is an evaluation method to estimate the level of fitness of data in a developed method. In self-consistency test, observations of training datasets are predicted with decision rules acquired from the same dataset.The accuracy of self-consistency reveals the fitting ability of the rules obtained from the features of training sets. Since the prediction system parameters obtained by the self-consistency

Table 2. Accuracy of each classifier for self-consistency and independent data test

Method	Self-consistency test (%)	Independent data test(%)
Multilayer perceptron+cfs	97.09	95.61
RBF Network+cfs	94.15	93.83
Multilayer perceptron + relief-f	94.90	94.48
RBF Network+ relief-f	92.65	92.08

test are from the training dataset, the success rate is high. However poor result of self-consistency test reflects the inefficiency of classification method.In independent dataset the training set composed of two-thirds of protein coding DNA, and two-thirds of the non-coding sequences. The remaining sequences were used as the testing set.

4 Discussion

Existing gene finding tools employ different biological information for identification of protein coding regions. In this study, a novel gene finding algorithm which relies on the use of genomic composition and dinucleotide compositional skew information was proposed. Our results indicate that genomic composition and dinucleotide compositional skew in conjunction with two feature selection methods: Correlation-based feature selection and the Relief-F algorithm followed by two classification algorithms, multilayer perceptron and RBF Networks are significantly useful features in classification of protein coding DNA from non-coding. The ability of these discriminant features is evident from the above mentioned performance evaluation techniques.

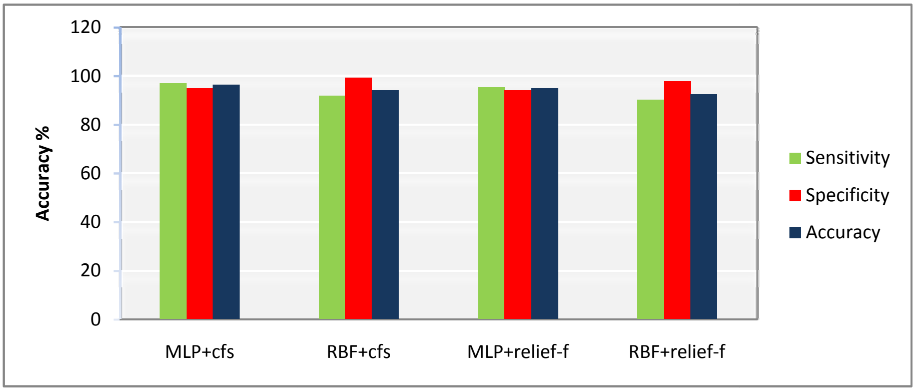


Fig. 2. Average Specificity, Sensitivity and Accuracy for the various methods

References

1. Baldi, P., Brunak, S.: Bioinformatics: The Machine Learning Approach. The MIT Press, Cambridge (1998)
2. Fickett, J.W.: The gene identification problem: an overview for developers. Comput. Che. 20, 103–118 (1996)
3. Mathé, C., Schiex, M.-F., Rouzé, P.: Current methods of gene prediction, their strength and weaknesses. Nucleic Acids Res. 30, 4103–4117 (2002)

4. Wang, Z., Chen, Y.Z., Li, Y.X.: A brief review of computational gene prediction methods. Geno. Prot. Bioinfo. 2, 216–221 (2004)
5. Do, J.H., Choi, D.K.: Computational approaches to gene prediction. Journal of Microbiology 44(2), 137–144 (2006)
6. Bandyopadhyay, S., Maulik, U., Roy, D.: Gene Identification: Classical and Computational Intelligence Approaches. IEEE Transactions on Systems, Man and Cybernetics, Part C 38(1), 55–68 (2008)
7. Delcher, A.L., Harmon, D., Kasif, S., White, O., Salzberg, S.L.: Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 27, 4636–4641 (1999)
8. Besemer, J., Lomsadze, A., Borodovsky, M.: GeneMarkS.:A self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res. 29, 2607–2618 (2001)
9. Gish, W., States, D.: Identification of protein encoding regions by database similarity search. Nature Genet. 3, 266–272 (1993)
10. Robison, K., Gilbert, W., Church, G.: Large-scale bacterial gene discovery by similaritysearch. Nat. Genet. 7, 205–214 (1994)
11. Frishman, D., Mironov, A., Mewes, H.W., Gelfand, M.: Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. Nucleic Acids Res. 26, 2941–2947 (1998)
12. Badger, J.H., Olsen, G.J.: CRITICA.:Coding region identification tool invoking comparative analysis. Mol. Biol. Evol. 16, 512–524 (1999)
13. Tech, M., Merkl, R.: YACOP: enhanced gene prediction obtained by a combination of existing methods. Silico Biol. 3, 441–451 (2004)
14. Markowitz, V.M., Korzeniewski, F., Palaniappan, K., Szeto, E., Werner, G., Padki, A., Zhao, X., Dubchak, I., Hugenholtz, P., Anderson, I., Mavromatis, K., Ivanova, N., Kyrpides, N.C.: The Integrated Microbial Genomes (IMG) system. Nucleic Acids Research 34, D344–D348 (2006)
15. Touchon, M., Nicolay, S., Audit, B., Brodie, B., Arneodo, A., d'Aubenton, C.Y., Thermes, C.: Replicationassociated strand asymmetries in mammalian genomes Toward detection of replication origins. PNAS 102(28), 9836–9841 (2005)
16. Fujimori, S.: GC–compositional strand bias around transcription start sites in plants and fungi. BMC Genomics 6(26), 1471, 2164/6/26 (2005)
17. Hall, M., Holmes, G.: Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. IEEE Trans. Knowl. Data Eng. 15, 1–16 (2003)
18. Wang, C., Ding, C., Meraz, R.F., Holbrook, S.R.: PSoL.: A positive sample only learning algorithm for finding non-coding RNA genes. Bioinformatics 22, 2590–2596 (2006)
19. Liu, H., Yu, L.: Towards integrating feature selection algorithms for classification and clustering. IEEE Transactions on Knowledge and Data Engineering 17(3), 1–12 (2005)
20. Hall, M.A.: Correlation based feature selection for machine learning. Doctoral dissertation, The University of Waikato, Dept of Comp. Sci. (1999)
21. Marko, R.S., Igor, K.: Theoretical and empirical analysis of relief and rreliefF. Machine Learning Journal 53, 23–69 (2003)
22. Kira, K., Rendell, L.A.: A practical approach to feature selection. In: Proceedings of the Ninth International Workshop on Machine Learning, pp. 249–256. Morgan Kaufmann Publishers Inc. (1992)
23. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA Data Mining Software: An Update. SIGKDD Explorations 11(1) (2009)
24. Werbos, P.J.: Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. PhD thesis, Harvard University (1974)

25. Parker, D.B.: Learning-logic. Technical report, TR-47, Sloan School of Management, MIT, Cambridge, Mass (1985)
26. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning internal representations by error propagation in Parallel distributed processing. Explorations in the Microstructure of Cognition, vol. I. Bradford Books, Cambridge (1986)
27. Moody, J., Darken, C.J.: Fast learning in networks of locallytuned processing units. Neural Computing 1, 281–294 (1989)
28. Broomhead, D.S., Lowe, D.: Multivariate functional interpolation and adaptive networks. Complex Syst. 2, 321–355 (1988)