Biochemical and Biophysical Research Communications 422 (2012) 36–41

Contents lists available at SciVerse ScienceDirect

Biochemical and Biophysical Research Communications

journal homepage: www.elsevier.com/locate/ybbrc

The elusive short gene – an ensemble method for recognition for prokaryotic genome

Baharak Goli *, Achuthsankar S. Nair

Department of Computational Biology and Bioinformatics, University of Kerala, Trivandrum 695581, India

ARTICLE INFO

Article history:
Received 13 April 2012
Available online 25 April 2012

Keywords:
Computational gene finding
Short gene prediction
Ensemble classifier
Feature selection
Adaboost.M1
Random forests

ABSTRACT

Accurate prediction of short protein coding DNA from genome sequence information remains an unsolved problem in DNA sequence analysis. Popular gene finding tools show drastic reduction in accuracy while attempting to predict genes of length less than 400 nt, a length we define as short. This study performs a quantitative evaluation of a set of selected coding measures in terms of their discriminative power in recognizing short genes in prokaryotic genomes. By performing Fast Correlation Based Feature Selection (FCBF) technique, we identified a subset of coding measures with high discriminative power. Using the measures identified thus, we present a novel approach for short genes recognition. A short-gene predictor employing AdaBoost.M1 in conjunction with random forests as the base classifier gives 92.74% accuracy, 94.77% sensitivity and 90.06% specificity on short genes.

1. Introduction

Small proteins have been discovered to be vital functional elements in cellular biology. They have shown to play important roles in several functions including the regulation of amino acid metabolism [1], antimicrobial activity [2], stabilizing factors for larger protein complexes [3], iron-homeostasis [4], acting as chaperones of metals and nucleic acids [5,6] and so on. The most widely used gene prediction algorithms in prokaryotic genomes often fail to identify short protein coding DNA due to limitation in coding measures in short nucleotide sequences, resulting in incomplete or wrong annotations [7]. Hence, study of coding measures in short DNA sequences is of major importance in gene prediction and genome annotation.

Performance of gene finding algorithms is highly dependent on the coding measures that are used to annotate the sequence. Coding measures describe the likelihood that a DNA sequence is coding for a protein or at least part of it. The ability of many coding measures for classification of longer sequences have been exploited in the past two decades while little is reported about their significance on shorter sequences. Notable discriminant features for longer sequences are given in Table 1. These are reported as effective in gene finding at various accuracies. However, no literature refers to the suitability of these or other features in the case of short genes.

We define short genes as fragments less than 400 nt in length, since many of the popular gene prediction tools drastically drop in accuracy for these fragment sizes. We analyzed the length distribution of entire protein coding sequences of four prokaryotic species- *Escherichia coli K-12 MG1655, Klebsiella pneumoniae 342, Yersinia pestis KIM 10, Enterobacter 638* and found short genes comprise a large percentage of the coding sequences. The results of this analysis are shown in Fig. 1.

To test the power of available gene finding tools to detect short genes, protein coding sequences of *E. coli K-12 MG1655* with two widely used prokaryotic gene finding tool-FrameD and GeneMark were analyzed. The results shown in Fig. 2 clearly illustrate that prediction of significant fraction of short genes are beyond the detection ability of these tools.

To overcome the limitations of available gene finding tools in prediction of short genes, we scrutinized the discriminative power of 177 coding features extracted from six prokaryotic organisms. We combined various types of features hypothesizing that the combined strength of these sequence parameters would result in a better recognition of short coding sequences.

2. Materials and methods

2.1. Dataset construction

Coding and non-coding sequences of the following organisms were chosen for this study (i) two unique *E. coli* strains: *E. coli K-12 MG1655* and *E. coli UTI89 (UPEC)* (ii) four *Enterobacteriaceae* strains: *Buchnera aphidicola 5A*, *Enterobacter 638*, *K. pneumoniae 342* and *Y. pestis KIM 10*. These were obtained from Integrated Microbial Genome (IMG) database [25]. To build a general dataset, both closely and distantly related spices were included. These

* Corresponding author.
E-mail address: baharak_goli@yahoo.com (B. Goli).

0006-291X/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.bbrc.2012.04.090

Table 1
Discriminant features for longer sequences.

Coding measure	Reference
Base compositional bias	[8]
Codon usage bias	[9]
Base compositional bias between codon positions	[10]
Hydrophobicity	[11]
Mono and diamino acid usage	[12]
Asymmetry in the codon positions with respect to purine/pyrimidine, amino/keto and strong/weak hydrogen bonding nature of nucleotides represented by Z curve	[13,14]
Global features obtained from multi-fractal analysis of DNA sequence	[15]
Local singularity density distribution in the coding and non-coding sequences estimated by wavelet transform modulus maxima methodology	[16]
G + C content	[17]
Oligonucleotide compositions and codon usage	[18]
Codon frequency	[19,20]
Hexamer usage	[21]
Oligomer frequencies	[22]
Entropy distance profile	[23]
GC bias in the 1st, 2nd, and 3rd positions of each codon	[24]

organisms were opted based on their relation to *E.* coli K12 MG1655. Short coding and non-coding sequences were extracted. The training set was formed by taking two-thirds of coding and non-coding sequences from each organisms and the remaining one-third was allotted to the test set. Training dataset comprised of 3270 coding and 2489 non-coding sequences. Testing dataset comprised of 1637 coding and 1247 non-coding sequences.

2.2. Feature vector formation

We now describe 6 sets of features we extracted, namely physicochemical and conformational properties (65 features), *k*-mer frequencies (84 features), GC content and its fraction at different codon positions (4 features), Codon usage bias (6 features), Amino acid properties (2 features) and Rho statistic (16 features).

2.3. Physicochemical and conformational properties

A set of 16 diverse physicochemical and 49 conformational DNA dinucleotide properties were obtained from the DiProDB database [26]. The considered dinucleotide properties are given in Tables 2 and 3. Our approach was to consider the property values across the length of the sequence on a sliding window mode, with a window size of 40 and step size of one. These window dimensions were chosen empirically during testing of the developed tool, based on the best accuracy achieved. We computed 65 features as follows. Let $p_1, p_2 \ldots p_{65}$ be the 65 chosen physicochemical and conformational properties and $d_1, d_2 \ldots d_{16}$ be the dinucleotide frequencies in a particular sequence under study. Feature value f_i corresponding to property p_i was calculated for window w as: $f_{iw} = \sum_{k=1}^{16} d_k \cdot p_i/n$, where n is number of dinucleotides in the considered window. The feature element of the whole sequence was taken as the average over all windows: $f_i = \sum_{w=1}^{n} f_{iw}/N$, where N is number of windows of size 40 in the considered sequence. The f_i values were computed for all the 65 parameters to form a feature vector $F_1 = [f_1, f_2 \ldots f_{65}]$.

2.4. k-mer frequencies

The k-mer frequencies of nucleotides have been widely used for gene prediction [41] and biosequence characterization [42]. We computed *k*-mer frequencies for $k = 1, 2$ and 3 for each sequence, resulting in total of 84 features (4 frequencies for $k = 1$, 16 for $k = 2$ and 64 for $k = 3$). The feature vector formed by *k*-mer frequencies is $F_2 = [f_{66}, f_{67} \ldots f_{149}]$.

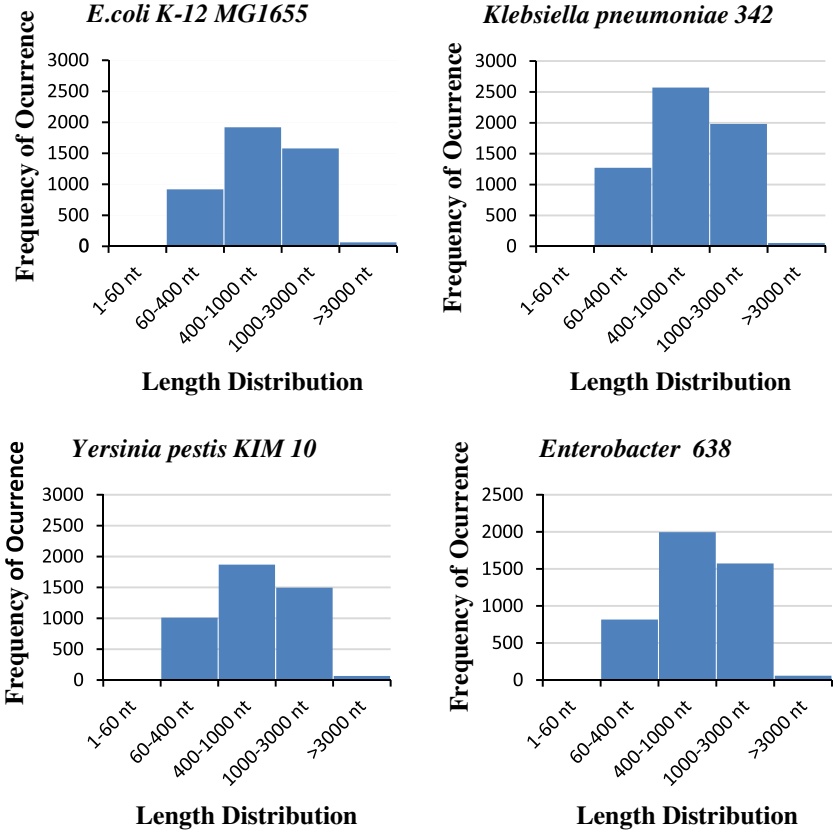

Fig. 1. Length distribution of entire protein coding sequences of *E. coli K-12 MG1655, Klebsiella pneumoniae 342, Yersinia pestis KIM 10* and *Enterobacter 638*.

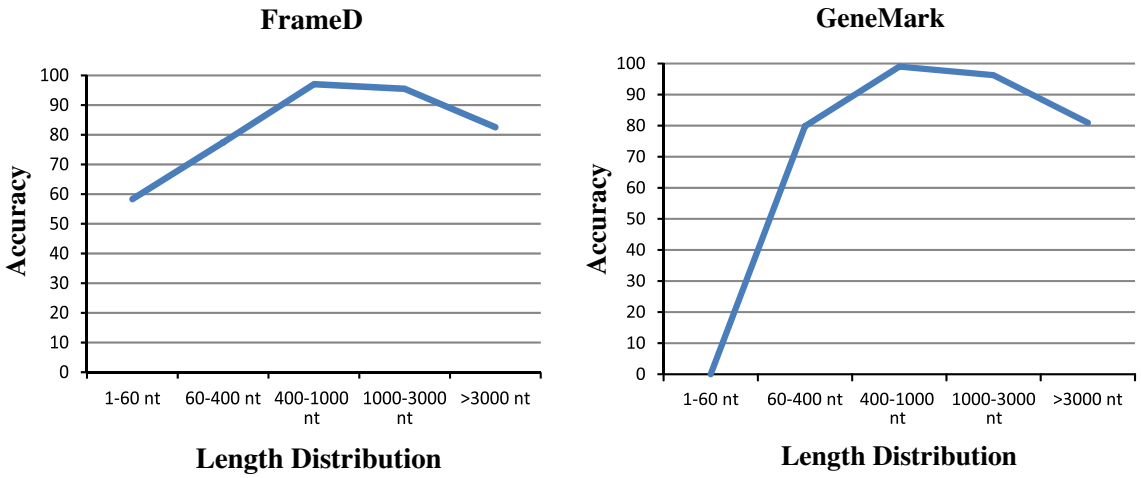

Fig. 2. Accuracy of FrameD and GeneMark to detect short genes.

Table 2
Physicochemical properties of DNA.

Physicochemical properties	Reference
Stacking energy	[27]
Melting temperature	[28]
Probability contacting nucleosome core	[29]
Mobility to bend towards major groove	[30]
Mobility to bend towards minor groove	[30]
Enthalpy	[31]
Entropy	[31]
Free energy	[32]
Slide stiffness	[33]
Shift stiffness	[33]
Roll stiffness	[33]
Tilt stiffness	[33]
Twist stiffness	[33]
Rise stiffness	[33]
Flexibility_slide	[34]
Flexibility_shift	[34]

Table 3
Conformational properties of DNA.

Conformational properties	Reference
Bend	[35]
Tip	[35]
Inclination	[35]
Major groove width	[35]
Major groove depth	[35]
Major groove size	[36]
Major groove distance	[36]
Minor groove width	[35]
Minor groove depth	[35]
Minor groove size	[36]
Minor groove distance	[36]
Persistance length	[29]
Propeller twist	[36]
Clash strength	[36]
Twist(DNA–protein complex)	[37]
Twist_twist	[38]
Tilt_tilt	[38]
Roll_roll	[38]
Twist_tilt	[38]
Twist_roll	[38]
Tilt_roll	[38]
Shift_shift	[38]
Slide_slide	[38]
Rise_rise	[38]
Shift_slide	[38]
Shift_rise	[38]
Slide_rise	[38]
Twist_shift	[38]
Twist_slide	[38]
Twist_rise	[38]
Tilt_shift	[38]
Tilt_slide	[38]
Tilt_rise	[39]
Roll_shift	[38]
Roll_slide	[38]
Roll_rise	[38]
Tilt(DNA–protein complex)	[39]
Roll(DNA–protein complex)	[39]
Shift(DNA–protein complex)	[39]
Slide(DNA–protein complex)	[39]
Rise(DNA–protein complex)	[39]
Twist	[36]
Tilt	[36]
Roll	[36]
Slide	[36]
Shift	[37]
Rise	[37]
Wedge	[40]
Direction	[40]

2.5. GC content and its fraction at different codon positions

Genome composition in terms of GC content and its fraction at different codon positions have been widely used in many bio-sequence classifications and analyses [43,44]. We have chosen the following GC measures as feature vector elements:

f_{150} = GC content (computed as number of GC bases/sequence length).
f_{151} = GC_1 (computed as number of GC bases at 1st codon position/number of codons).
f_{152} = GC_2 (computed as number of GC bases at 2nd codon position/number of codons).
f_{153} = GC_3 (computed as number of GC bases at 3rd codon position/number of codons).

The feature vector incorporating the above elements is given by $F_3 = [f_{150}, f_{151}, f_{152}, f_{153}]$.

2.6. Codon usage bias

Codon usage bias (CUB) is defined as a deviation from regular codon usage in the genic regions. Codon usage bias has numerous applications, such as gene prediction and recognition of laterally transferred genes [45,46]. Measures such as the 'frequency of

optimal codons' (F_{op}) [47] and *CAI* (Codon Adaptation Index) [48] have been exploited for estimating and analyzing codon usage bias [49]. We chose these two features and also four base compositions at synonymous third codon positions to form a feature vector. F_{op} is the ratio of optimal codons to synonymous codons. F_{op} is defined as: $X_{op}/X_{op}+X_{non}$ where X_{op} represents the number of 'optimal' codons and X_{non} represents the number of 'nonoptimal' codons. *CAI* estimates the deviation of a given protein coding gene sequence concerning a reference set of known genes and is based on 'relative adaptedness' value which is noted by W_i and defined for codon i as: $RSCU_i/RSCU_{\max} = X_i/X_{\max}$ where $RSCU_{max}$ is relative synonymous codon usage value for the most frequently used codon for an amino acid. The codon adaptation index is defined as: $CAI = (\exp(\frac{1}{L}\sum_{i=1}^{L}\ln(W_i)))$, where L is the number of codons in the sequence. F_{op}, *CAI* and four base compositions at synonymous third codon positions were taken as feature elements and a feature vector is formed as $F_4 = [f_{154}, f_{155}\ldots f_{159}]$.

2.7. Amino acid properties

Physiochemical and thermodynamic properties of the amino acids are of great relevance in gene prediction [11]. Two of the most widely used physiochemical parameters in different biosequence classification applications are hydrophobicity and aromaticity. We converted the nucleotide sequences to amino acid sequences using codon table specific to the concerned species. Hydrophobicity was computed as the average of the hydrophobic indices of each amino acid [50]. Aromaticity was calculated as the number of aromatic amino acids (Phe, Tyr, Trp) in each sequence. The above two elements were taken as f_{160}, f_{161} in the feature vector $F_5 = [f_{160}, f_{161}]$.

2.8. Rho statistic

Rho is a nonparametric measure which is used to estimate the dinucleotide representation [51]. It is computed using the formulae $rho_{xy} = f_{xy}/f_x \times f_y$, where f_{xy} is the frequency of dinucleotide xy, f_x is frequency of nucleotide x and f_y is frequency of nucleotide y. Rho statistic is equal to 1.00 when dinucleotide xy is constructed by pure chance, it is superior to 1.00 when dinucleotide xy is over-represented and it is inferior to 1.00 when dinucleotide xy is under-represented. We computed 16 dinucleotide rho statistic and formed the feature vector $F_6 = [f_{162}, f_{163}\ldots f_{177}]$.

The combined feature vector made out of the 6 partial vectors was formed for each sequence in the dataset. The combined feature vector has 177 elements and is given by $F = [F_1|\ F_2\ |\ F_3\ |F_4|\ F_5|\ F_6]$. The values were further normalized using z-score into the range $[-1, 1]$ to remove the overbearing effect of any feature due to its large range of values.

3. Algorithm and implementation

3.1. Pre-processing step: feature subset selection

The performance of a classifier is highly dependent on the feature vector size, the training sample size and classifier complexity. As the number of features increase, dependability of the parameter estimates reduces and the performance of the classifier may degrade [52]. Feature selection is an important preprocessing step to machine learning and data mining problems. Feature selection methods aim at separating out features with high impact. It reduces dimensionality, removes all weaker features, decreases computational time and increases learning accuracy [53]. In this study, Fast Correlation Based Feature Selection (FCBF) algorithm [54] was adopted to identify a subset of discriminant features. FCBF is an entropy-based feature selection method. The entropy of the particular feature X is defined as: $H(X) = -\sum_i P(x_i)\log_2(P(x_i))$, where $P(x_i)$ is the prior probability for a finite set of values of X. The conditional entropy of a variable X, given another variable Y is defined as: $H(X|Y) = -\sum_j P(y_i)\sum_i P(x_i|y_i)\log_2(P(x_i|y_i))$, where $P(x_i|y_i)$ is the posterior probabilities of X given the values of Y. The amount by which the entropy of X reduces returns supplementary information about X provided by Y and is known as information gain and it is defined as: $IG(X|Y) = H(X) - H(X|Y)$. Feature Y is considered as more highly correlated to feature X than to feature Z, if $IG(X|Y) > IG(Z|Y)$.

With the use of FCBF, we reduced the original set of 177 features to 22 prominent features (shown in Table 4). The experimental results show that F2 has more representations and F5 was totally excluded from our feature set.

Table 4
Selected prominent features.

Feature type	Selected features using FCBF
Codon usage bias	CAI
GC content and its fraction at different codon positions	GC1,GC2
k-mer frequencies	ACC, ATC, ATG, CAG, CTG, GAA, GAC, GCA, GCT, GGT, GTC, TAA, TAG, TGA
Physicochemical and conformational properties	Twist_rise, Stacking energy
Rho statistic	rho_{tc}, rho_{tg}
Amino acid properties	–

3.2. Ensemble classifier

Accuracy and computational cost have always been matter of concern for machine learning systems. There have recently been many attempts in designing new learning algorithms with higher predictive accuracy by generating and ensembling multiple learners. Ensemble learning [55] deals with multiple classifiers to solve a pattern classification problem. Two effective ensemble methods are bagging [56] and boosting [57]. Bagging generates new training sets by sampling with substitution from the training data while boosting adopts adaptive sampling by using all instances at each iteration. Each instance is typically weighted in the training data. Weights are optimized then causing the learner to concentrate on different instances which consequently lead to multiple classifiers. In both methods the multiple classifiers are then combined using simple voting system to make a meta classifier. In bagging, each classifier has the same vote whereas boosting assigns different voting strengths to classifiers based on their accuracy. Within the classifier ensemble models, AdaBoost is considered to be one of the best since it is built on a solid theoretical foundation,

Table 5
The resulting performance based on standard 5-fold cv.

	Specificity	Sensitivity	Precision	MCC	5-fold CV
All features	89.87	93.11	92.35	0.83	91.71
Selected features using FCBF	90.06	94.77	92.61	0.85	92.74

Table 6
The resulting performance based on test set.

	Specificity	Sensitivity	Precision	MCC	Test set accuracy
All features	89.01	90.77	91.55	0.79	90.01
Selected features using FCBF	89.89	91.26	92.22	0.81	90.67

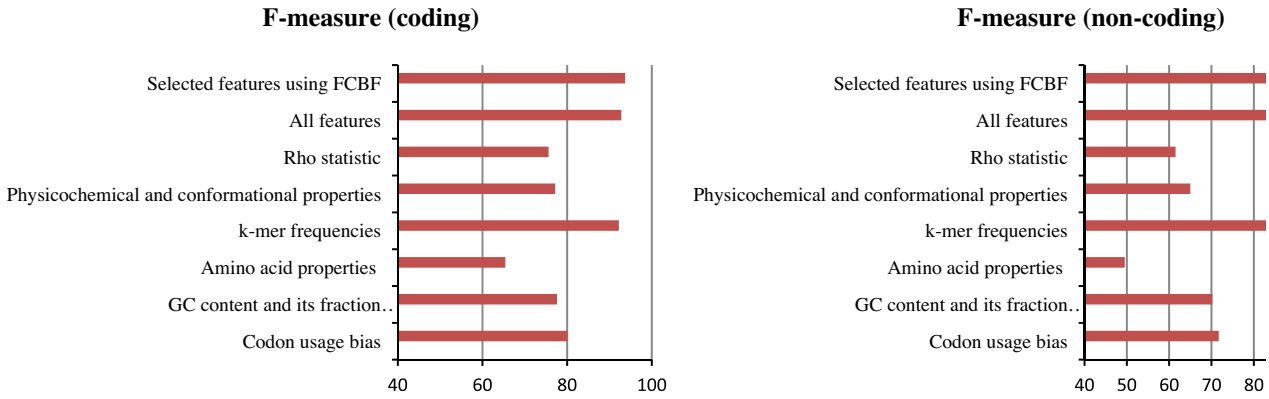

Fig. 3. F-measures of our proposed algorithm with different input and conformational features.

produces very accurate prediction and smaller classification error. In this study we employed AdaBoost.M1 [58] which is an extension of Adaboost.M1 algorithm. We carried out experiments with Adaboost.M1algorithm in conjunction with random forests as the base learner. Random forests was chosen after empirical evaluation of 5 base classifiers: SVM with RBF kernel, decision stump, RBF network, random forests and nearest neighbor classifiers.

4. Result

Algorithms were implemented in data mining toolkit Weka version 3.6.6 [59] on a PC with a 2.13 GHz Intel CPU and 4 GB RAM, using Windows 7. The capability of this hybrid model was evaluated using basic performance measures of accuracy, sensitivity and specificity based on standard 5-fold cross validation and test set. By applying Fast Correlation Based Feature Selection, the resulting performance improved by more than one percentage point (shown in Tables 5 and 6). To analyze the impact of feature selection on performance, F-measure (the harmonic mean of sensitivity and specificity) was computed for both coding and non-coding regions using different subsets of features. Combining the features improved the F-measure to 92.8 for coding set and 90.3 for non-coding set. The results are given in Fig. 3. The obtained error rates of proposed model are shown in Table 7.

Self-consistency test was also carried out to evaluate the performance of our algorithm. Self-consistency test estimates the level of fitness of data in a constructed model. In this test, observations of training datasets are predicted with decision rules acquired from the same dataset. Since the prediction system parameters obtained by the self-consistency test are from the training dataset, the success rate is high. Low accuracy of self-consistency test indicates low efficiency classifier. Self-consistency test gave an accuracy of 100% and a MCC of 1.

5. Discussion

We checked the performance of a standard gene finding tool FrameD on the same test set of genomic sequences that we used and the accuracy was noted as 54.97%. This emboldens us to report our new method as a remarkable improvement in short gene finding, an area that has been more or less neglected in computational gene finding. Our results demonstrate that the trimer frequency in DNA sequences contributes most to the prediction accuracy, followed by Condon adaptation index and GC content at first and second codon positions. Among physiochemical and conformational properties, twist_rise and stacking energy are the most important discriminating features for short gene recognition.

Our investigations raise many new research questions. Primarily we feel that the correlation of prediction strength of various coding measures with length of query sequences is to be investigated further. If striking correlations are found in the case of majority of features, then separate recognition tools can be arrayed to specialize on different sequence length, or range of lengths. In other words, from the specialized case of short versus long genes, we need to address the general case of various gene lengths or range of lengths. This is analogous to species-specific gene finding tools. While the above are rather computational issues, the more interesting question relates to establishing biological significance of the features found to be of computational predictive power.

Table 7
The obtained error rates of proposed algorithm.

	Mean absolute error (%)	Root mean squared error (%)	Relative absolute error (%)	Root relative squared error (%)
All features	0.08	0.28	16.85	56.74
Selected features using FCBF	0.07	0.26	14.78	53.66

References

[1] C. Yanofsky, Transcription attenuation: once viewed as a novel regulatory strategy, Journal of Bacteriology 182 (2000) 1–8.
[2] R.L. Gallo, V. Nizet, Endogenous production of antimicrobial peptides in innate immunity and human disease, Current Allergy and Asthma Reports 3 (2003) 402–409.
[3] D. Schneider, T. Volkmer, M. Rogner, PetG and PetN, but not PetL, are essential subunits of the cytochrome b6f complex from Synechocystis PCC 6803, Research in Microbiology 158 (2007) 45–50.
[4] B.A. Sela, Hepcidin the discovery of a small protein with a pivotal role in iron homeostasis, Harefuah 147 (2008) 261–266, 276.
[5] M.R. Hemm, B.J. Paul, T.D. Schneider, G. Storz, K.E. Rudd, Small membrane proteins found by comparative genomics and ribosome binding site models, Molecular Microbiology 70 (2008) 1487–1501.
[6] A. Gaballa, H. Antelmann, C. Aguilar, S.K. Khakh, K.B. Song, G.T. Smaldone, J.D. Helmann, The *Bacillus subtilis* iron-sparing response is mediated by a Fur-regulated small RNA and three small, basic proteins, Proceedings of the National Academy of Sciences of the United States of America 105 (2008) 11927–11932.
[7] M.R. Brent, R. Guigo, Recent advances in gene structure prediction, Current Opinion in Structural Biology 14 (2004) 264–272.
[8] J.C. Shepherd, Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification, Proceedings of the National Academy of Sciences of the United States of America 78 (1981) 1596–1600.
[9] R. Staden, A.D. McLachlan, Codon preference and its use in identifying protein coding regions in long DNA sequences, Nucleic Acids Research 10 (1982) 141–156.
[10] J.W. Fickett, Recognition of protein coding regions in DNA sequences, Nucleic Acids Research 10 (1982) 5303–5318.
[11] A. Tramontano, M.F. Macchiato, Probability of coding of a DNA sequence: an algorithm to predict translated reading frames from their thermodynamic characteristics, Nucleic Acids Research 14 (1986) 127–135.

[12] P. McCaldon, P. Argos, Oligopeptide biases in protein sequences and their use in predicting protein coding regions in nucleotide sequences, Proteins 4 (1988) 99–122.
[13] C.T. Zhang, J. Wang, Recognition of protein coding genes in the yeast genome at better than 95% accuracy based on the Z curve, Nucleic Acids Research 28 (2000) 2804–2814.
[14] J. Wang, C.T. Zhang, Identification of protein-coding genes in the genome of *Vibrio cholerae* with more than 98% accuracy using occurrence frequencies of single nucleotides, European Journal of Biochemistry 268 (2001) 4261–4268.
[15] L.Q. Zhou, Z.G. Yu, J.Q. Deng, V. Anh, S.C. Long, A fractal method to distinguish coding and non-coding sequences in a complete genome based on a number sequence representation, Journal of Theoretical Biology 232 (2005) 559–567.
[16] O.C. Kulkarni, R. Vigneshwar, V.K. Jayaraman, B.D. Kulkarni, Identification of coding and non-coding sequences using local Holder exponent formalism, Bioinformatics 21 (2005) 3818–3823.
[17] A.V. Lukashin, M. Borodovsky, GeneMark.hmm: new solutions for gene finding, Nucleic Acids Research 26 (1998) 1107–1115.
[18] W.S. Hayes, M. Borodovsky, How to interpret an anonymous bacterial genome: machine learning approach to gene identification, Genome Research 8 (1998) 1154–1171.
[19] D. Frishman, A. Mironov, H.W. Mewes, M. Gelfand, Combining diverse evidence for gene recognition in completely sequenced bacterial genomes, Nucleic Acids Research 26 (1998) 2941–2947.
[20] J. Besemer, A. Lomsadze, M. Borodovsky, GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions, Nucleic Acids Research 29 (2001) 2607–2618.
[21] G.B. Hutchinson, M.R. Hayden, The prediction of exons through an analysis of spliceable open reading frames, Nucleic Acids Research 20 (1992) 3453–3462.
[22] A.L. Delcher, D. Harmon, S. Kasif, O. White, S.L. Salzberg, Improved microbial gene identification with GLIMMER, Nucleic Acids Research 27 (1999) 4636–4641.
[23] H. Zhu, G.Q. Hu, Y.F. Yang, J. Wang, Z.S. She, MED: a new non-supervised gene prediction algorithm for bacterial and archaeal genomes, BMC Bioinformatics 8 (2007) 97.
[24] D. Hyatt, G.L. Chen, P.F. Locascio, M.L. Land, F.W. Larimer, L.J. Hauser, Prodigal: prokaryotic gene recognition and translation initiation site identification, BMC Bioinformatics 11 (2010) 119.
[25] V.M. Markowitz, F. Korzeniewski, K. Palaniappan, E. Szeto, G. Werner, A. Padki, X. Zhao, I. Dubchak, P. Hugenholtz, I. Anderson, A. Lykidis, K. Mavromatis, N. Ivanova, N.C. Kyrpides, The integrated microbial genomes (IMG) system, Nucleic Acids Research 34 (2006) D344–348.
[26] M. Friedel, S. Nikolajewa, J. Suhnel, T. Wilhelm, DiProDB: a database for dinucleotide properties, Nucleic Acids Research 37 (2009) D37–40.
[27] J. Sponer, H.A. Gabb, J. Leszczynski, P. Hobza, Base-base and deoxyribose-base stacking interactions in B-DNA and Z-DNA: a quantum-chemical study, Biophysical Journal 73 (1997) 76–87.
[28] O. Gotoh, Y. Tagashira, Stabilities of nearest-neighbor doublets in double-helical DNA determined by fitting calculated melting profiles to observed profiles, Biopolymers 20 (1981) 1033–1042.
[29] M.E. Hogan, R.H. Austin, Importance of DNA stiffness in protein-DNA binding specificity, Nature 329 (1987) 263–266.
[30] M.R. Gartenberg, D.M. Crothers, DNA sequence determinants of CAP-induced bending and protein binding affinity, Nature 333 (1988) 824–829.
[31] N. Sugimoto, S. Nakano, M. Yoneyama, K. Honda, Improved thermodynamic parameters and helix initiation factor to predict stability of DNA duplexes, Nucleic Acids Research 24 (1996) 4501–4505.
[32] K.J. Breslauer, R. Frank, H. Blocker, L.A. Marky, Predicting DNA duplex stability from the base sequence, Proceedings of the National Academy of Sciences of the United States of America 83 (1986) 3746–3750.
[33] J.R. Goni, A. Perez, D. Torrents, M. Orozco, Determining promoter location based on DNA structure first-principles calculations, Genome Biology 8 (2007) R263.
[34] M.J. Packer, M.P. Dauncey, C.A. Hunter, Sequence-dependent DNA structure: dinucleotide conformational maps, Journal of Molecular Biology 295 (2000) 71–83.
[35] H. Karas, R. Knuppel, W. Schulz, H. Sklenar, E. Wingender, Combining structural analysis of DNA with search routines for the detection of transcription regulatory elements, Computer Applications in the Biosciences 12 (1996) 441–446.
[36] A.A. Gorin, V.B. Zhurkin, W.K. Olson, B-DNA twisting correlates with base-pair morphology, Journal of Molecular Biology 247 (1995) 34–48.
[37] M. Suzuki, N. Yagi, J.T. Finch, Role of base-backbone and base-base interactions in alternating DNA conformations, FEBS Letters 379 (1996) 148–152.
[38] F. Lankas, J. Sponer, J. Langowski, T.E. Cheatham 3rd, DNA basepair step deformability inferred from molecular dynamics simulations, Biophysical Journal 85 (2003) 2872–2883.
[39] W.K. Olson, A.A. Gorin, X.J. Lu, L.M. Hock, V.B. Zhurkin, DNA sequence-dependent deformability deduced from protein-DNA crystal complexes, Proceedings of the National Academy of Sciences of the United States of America 95 (1998) 11163–11168.
[40] E.S. Shpigelman, E.N. Trifonov, A. Bolshoy, CURVATURE: software for the analysis of curved DNA, Computer applications in the biosciences: CABIOS 9 (1993) 435–440.
[41] Y. Saeys, P. Rouze, Y. Van De Peer, In search of the small ones: improved prediction of short exons in vertebrates, plants, fungi and protists, Bioinformatics 23 (2006) 414–420.
[42] C. Burge, A.M. Campbell, S. Karlin, Over- and under-representation of short oligonucleotides in DNA sequences, Proceedings of the National Academy of Sciences of the United States of America 89 (1992) 1358–1362.
[43] A.S. Warren, J. Archuleta, W.C. Feng, J.C. Setubal, Missing genes in the annotation of prokaryotic genomes, BMC Bioinformatics 11 (2010) 131.
[44] M. Semon, D. Mouchiroud, L. Duret, Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance, Human Molecular Genetics 14 (2005) 421–427.
[45] C. Burge, S. Karlin, Prediction of complete gene structures in human genomic DNA, Journal of Molecular Biology 268 (1997) 78–94.
[46] D. Cortez, L. Delaye, A. Lazcano, A. Becerra, Composition-based methods to identify horizontal gene transfer, Methods in Molecular Biology 532 (2009) 215–225.
[47] T. Ikemura, Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system, Journal of Molecular Biology 151 (1981) 389–409.
[48] P.M. Sharp, W.H. Li, The codon Adaptation Index–a measure of directional synonymous codon usage bias, and its potential applications, Nucleic Acids Research 15 (1987) 1281–1295.
[49] A. Carbone, A. Zinovyev, F. Kepes, Codon adaptation index as a measure of dominating codon bias, Bioinformatics 19 (2003) 2005–2015.
[50] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, Journal of Molecular Biology 157 (1982) 105–132.
[51] S. Karlin, L.R. Cardon, Computational DNA sequence analysis, Annual Review of Microbiology 48 (1994) 619–654.
[52] S.J. Raudys, A.K. Jain, Small sample size effects in statistical pattern recognition: recommendations for practitioners, IEEE Transactions on Pattern Analysis and Machine Intelligence 13 (1991) 252–264.
[53] Y. Saeys, S. Degroeve, D. Aeyels, P. Rouze, Y. Van de Peer, Feature selection for splice site prediction: a new method using EDA-based feature ranking, BMC Bioinformatics 5 (2004) 64.
[54] L. Yu, H. Liu, Feature selection for high-dimensional data: a fast correlation-based filter solution, ICML (2003) 856–863.
[55] T. Dietterich, Machine-learning research: four current directions, AI Magazine 18 (1997) 97–136.
[56] L. Breiman, Bagging Predictors, Machine Learning 24 (1996) 123–140.
[57] Y. Freund, R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, Proceedings of the Second European Conference on Computational Learning Theory, Springer-Verlag, 1995.
[58] Y. Freund, R. Schapire, Experiments with a New Boosting Algorithm, International Conference on Machine Learning, 1996, pp. 148-156.
[59] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA data mining software: an update, SIGKDD Exploration Newsletter 11 (2009) 10–18.