IRSTI 34.27.29

A COMPARISON OF METHODS FOR CLASSIFYING PROMOTER REGIONS IN E. COLI BASED ON STRUCTURAL PROPERTIES OF DNA

***Carmen Wright* [1], *J. Kaur* [2], *A.S. Newsome* [3], *Charles Bland* [4]**
[1] Ph.D, Assistant Professor,
[2] Graduate Student, Bioinformatics Program,
[3] Ph.D, Director of Bioinformatics and Associate Professor,
[4]Ph.D, Assistant Professor,
[1] Department of Mathematics, Jackson State University, Jackson, MS USA
[2,3,4] Mississippi Valley State University,
Itta Bena, MS, USA, asnewsome@mvsu.edu, (662)254-3395

One of the major challenges in biology is the correct identification of promoter regions. Computational methods based on motif searching have been the traditional approach taken. Studies have shown that DNA structural properties, such as free energy, curvature, and stress-induced duplex destabilization (SIDD) are useful in promoter classification, as well. In this paper, these properties were compared for their effectiveness in correctly classifying promoters. When using a decision tree for promoter classification based on DNA structural properties, SIDD showed a slight improvement over free energy and curvature, with f-score values 70.9%, 67.1%, and 61.5%, respectively.

Keywords: *promoter classification, DNA curvature, SIDD, free energy*

Identification of promoters is an important issue in biology, given that they are central in understanding the process by which genes are regulated. Wet-lab methods for promoter identification provide accuracy but suffer from being time-consuming. To facilitate faster processing, computational methods are required. Although far from perfect, they do provide a means for quickly identifying potential targets for experimental validation.

Several computational methods for promoter classification have been proposed. Most include some analysis of sequence patterns commonly found in promoter regions, such as -10 and -35 motifs [1, 2]. However, these patterns are not always sufficiently conserved to allow for adequate classification. Furthermore, there are clearly other factors not directly related to sequence motifs that are closely associated with promoter regions.

Promoter regions have unique characteristics in their physical structure that play major roles in transcription by facilitating protein-DNA interactions. Some of these properties include GC skew, bendability, free energy, curvature, base stacking, and stress-induced duplex destabilization (SIDD). Studies have reported impressive results using DNA structural properties for identifying promoter regions [3, 4, 5, 6]. This study assesses the feasibility of a computer-based classification approach for promoter identification in prokaryotes based on DNA free energy [7], curvature [8], and SIDD [9].

Analysis was performed on the genome of *E. coli K12*. Each sequence value was converted to its corresponding numeric structural property value.

Dataset.

The whole genome of *E. coli K12* was downloaded from NCBI. Experimentally verified transcription start sites were obtained from the Regulon database (Release: 6.4) [10]. This database release provided a compilation of 1771 promoter sequences. The dataset was filtered for unique promoters with known TSS locations, resulting in 1648 records.

Structural profiles were computed from the sequence data. The SIDD profile computations were obtained from Benham [5]. The free energy profile was computed using the nearest-neighbor thermodynamic parameters of base pairings described in [11]. The curvature profile was computed using the CURVATURE program [12, 13].

Classification

The training and testing datasets were constructed from the *E. coli K12* structural profile data. Positive instances (promoters) were defined as the 500 bp region from -400 to +100, with respect to TSSs. This dataset was composed of 1648 positive instances and 4944 negative instances, which represents a 3:1 ratio of negatives and positives. A randomly selected two-third and one-third split was used for training and testing data, respectively. The Weka data mining suite [14] was used to perform the classifications using its J48 decision tree.

Evaluation Measures

Classification results were used to evaluate the predictability of the structural properties. In order to compare predictions using a one-dimensional performance measure, the weighted average of the precision and recall (known as f-score) was computed for curvature, free energy and SIDD.

Precision, recall, and f-score were defined as follows,

$$precision = \frac{TP}{TP + FP} \qquad (1)$$

$$recall = \frac{TP}{TP + FN} \qquad (2)$$

$$f\text{-}score = \frac{2\,\mathrm{x} precision \mathrm{x} recall}{precision + recall} \qquad (3)$$

where *TP*, *TN*, *FP*, and *FN* are the numbers of true positives, true negatives, false positives and false negatives, respectively.

A comparison of free energy, curvature, and SIDD structural profiles is shown in the following figures. To create the structural data, each sequence value in *E. coli K12* was converted to its corresponding numeric structural property value. Next, the average value at each location was computed for all promoters (for the 500 bp region from -400 to +100, with respect to transcription start sites at +1).

Signatures of structural properties

Figure 1 is DNA free energy. High free energy values indicate low stability, and indicate regions where strand separation is more likely to occur. Figure 1 shows a low stability region from -100 to +50, with respect to the TSS. A distinctive peak appears near -10. So, the -10 region may be the least stable.

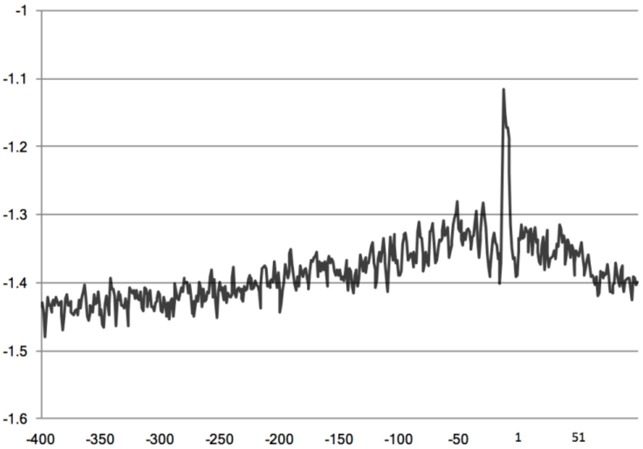

Figure 1: Average free energy values for the promoter regions

Similar changes in promoter regions can be seen in Figure 2 for SIDD, represented as G(x). G(x) corresponds to the incremental free energy needed for the base pair at position x to always remain open. It begins a noticeable decrease until its lowest points near -35 and -10, then begins an increase.

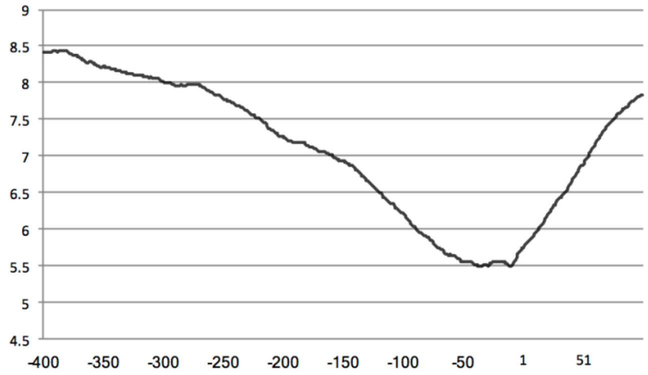

Figure 2: Average SIDD G(x) values for the promoter regions

Curvature increases from -400 to its highest at -53, before beginning to decrease. All three properties show noticeable increases or decreases in promoter regions and distinctive spikes near some known promoter indicators, such as -10, and -35. Thus, structural properties appear to be good candidates for identifying promoter regions.

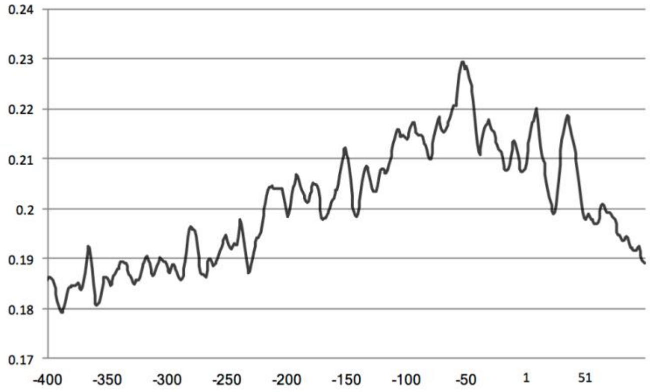

Figure 3: Average DNA curvature values for the promoter regions

Evaluation

Weka's J48 decision tree was used to perform the classifications of promoters and non-promoters. The construction of the training and testing sets is described in the methods sections. The f-score was computed for curvature, free energy and SIDD. For free energy, the resulting f-score was 67.1% (promoter 50.9%, non-promoter 74.9%); SIDD 70.9% (promoter 56.4%, non-promoter 77.8%); and curvature 61.5% (promoter 42%, non-promoter 71.8%). All methods performed better at identifying non-promoters than promoters. SIDD performed best overall, followed closely by free energy, and then curvature with the lowest f-score.

One of the major challenges in biology is the correct identification of promoter regions. Computational methods based on motif searching have been the traditional approach taken. This study has shown that DNA structural properties, such as free energy, curvature, and stress-induced duplex destabilization (SIDD) are useful in promoter classification, as well.

Future research will involve combining multiple structural-based predictors with sequence-based methods. For example, in [5] it was shown that SIDD was not directly related to primary sequences or unique motifs, and not positively correlated with DNA curvature. Thus, using SIDD with other predictive sequence and structural properties, particularly those not strongly correlated, may be fruitful. In addition, it may be useful to determine whether a classifier trained on one genome predicts well on others. Also, combining multiple classifiers as part of a voting system, such as an ensemble, may prove beneficial.

References

1 Gerald Z. Hertz, Gary D. Stormo. "*Escherichia coli* promoter sequences: analysis and prediction"; Methods in Enzymology, Volume 273, Pages 30-42, 1996.

2 Araceli M. Huerta, Julio Collado-Vides. "Sigma 70 promoters in *Escherichia coli*: specific transcription in dense regions of overlapping promoter-like signals". Journal of Molecular Biology, Volume 333, Issue 2, Pages 261-278, October 2003.

3 Czuee Morey, Sushmita Mookherjee, Ganesan Rajasekaran, Manju Bansal. “DNA Free Energy-Based Promoter Prediction and Comparative Analysis of Arabidopsis and Rice Genomes”. Plant Physiology, Volume 156, Issue 3, Pages 1300-1315, April 2011.

4 Charles Bland, Abigail S. Newsome, Aleksandra Markovets. “Promoter prediction in *E. coli* based on SIDD profiles and Artificial Neural Networks”. BMC Bioinformatics, Volume 11, Supplement 6, October 2010.

5 Huiqan Wang, Craig J. Benham. “Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress". BMC Bioinformatics, 7: 248, May 2006.

6 Aditi Kanhere, Manju Bansal. “A novel method for prokaryotic promoter prediction based on DNA stability”. BMC Bioinformatics, 6:1, January 2005.

7 Aditi Kanhere, Manju Bansal. “Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes”. Nucleic Acids Research, Volume 33, Issue 10, Pages 3165-3175, June 2005.

8 Limor Kozobay-Avraham, Sergey Hosid, Alexander Bolshoy. “Involvement of DNA curvature in intergenic regions of prokaryotes”. Nucleic Acids Research, Volume 34, Issue 8, Pages 2316–2327, May 2006.

9 Huiqan Wang, Michiel Noordewier, Craig J. Benham. “Stress-induced DNA duplex destabilization (SIDD) in the E. coli genome: SIDD sites are closely associated with promoters”. Genome Research, Volume 14, Issue 8, Pages 1575-1584, August 2004.

10 ReqlonDB [http://regulondb.ccg.unam.mx]

11 John SantaLucia Jr. “A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics”. Proceedings of the National Academy of Sciences USA, Volume 95, Number 4, Pages 1460-1465, February 1998.

12 E.S. Shpigelman, E.N. Trifonov, A. Bolshoy. “Curvature: software for the analysis of curved DNA”. Computer Applications in the Biosciences, Volume 9, Issue 4, Pages 435-440, August 1993.

13 Curvature [http://www.lfd.uci.edu/~gohlke/dnacurve/]

14 Weka [http://www.cs.waikato.ac.nz]

СРАВНЕНИЕ МЕТОДОВ КЛАССИФИКАЦИИ ПРОМОУТЕРНЫХ РЕГИОНОВ В *E. COLI* НА ОСНОВЕ СТРУКТУРНЫХ СВОЙСТВ ДНК

Кармен Райт[1]***, Дж. Кармен***[2]***, А.С.Ньюсом***[3]***, Ч. Блэнд***[4]
[1] Ph.D, и.о. ассоц. профессора
[2] магистрант,
[3] Ph.D, директор биоинформатики и и.о.ассоц. профессора
[4] Ph.D, доцент,
[1]Государственный университет Джексон, Джексон, Миссисипи, США
[2,3,4] Государственный университет долины Миссисипи
Итта Бена, Миссисипи, США, asnewsome@mvsu.edu, (662) 254-3395

Одной из основных проблем в биологии является правильная идентификация регионов - промоутеров. Традиционно применялись вычислительные методы, основанные на поиске мотивов. Исследования показали, что структурные свойства ДНК, такие как свободная энергия, кривизна и дестабилизация дуплекса, вызванные стрессом (SIDD), также полезны в классификации промоуторов. В этой статье эти свойства сравнивались для их эффективности при правильной классификации промоуторов. При использовании дерева решений для классификации промоутора, основанного на структурных свойствах ДНК, SIDD продемонстрировал небольшое улучшение по сравнению со свободной энергией и кривизной, при этом значения f-score составляли 70,9%, 67,1% и 61,5% соответственно.

Ключевые слова: *классификация промоуторов, кривизна ДНК, SIDD, свободная энергия*

ДНҚ-ның ҚҰРЫЛЫМДЫҚ ҚАСИЕТІНІҢ НЕГІЗІНДЕ E.COLI ПРОМОУТЕРЛІ АЙМАҒЫНЫҢ КЛАССИФИКАЦИЯ ӘДІСТЕРІН САЛЫСТЫРУ

Кармен Райт[1], Дж. Кармен[2], А.С. Ньюсом[3], Ч. Блэнд[4]
[1] Ph.D, қаум.профессор,
[2] магистрант,
[3] Ph.D, Биоинформатика директоры және қаум.профессор,
[4] Ph.D, доцент,
[1]Джексон мемлекеттік университеті, Джексон, Миссисипи, АҚШ
[2,3,4] Миссисипи алқабының мемлекеттік университеті,
Итта Бена, Миссисипи, АҚШ, asnewsome@mvsu.edu, (662) 254-3395

Биологиядағы маңызды мәселелердің бірі – промоутер аймақтарын дұрыс анықтау. Есептеу әдісі мотивті іздеу негізінде дәстүрлі тәсіл ретінде қолданылған. Зерттеу көркесеткендей, бос энергия, ДНҚ қисықтығы стресспен туындайтын дуплекс дестабилизациясы (SIDD) секілді ДНҚ-ның құрылымдық қасиеті, промоторлар классификациясына қажет. Бұл мақалада промоторлар классификациясының тиімділігі үшін осы құрылымдарға салыстыру жүргізілді. ДНҚ – ның құрылымдық қасиетіне негізделген, промоторлар классификациясына шешім ағашын қолдану барсында, SIDD бос энергия мен ДНҚ қисықтығына қарағанда шамалы жақсарғанын көрсетті, сондай-ақ сәйкесінше f-score 70,9%,67,1% және 61,5% болды.

Түйін сөздер: *промоуторлар классификациясы, ДНҚ қисықтығы, SIDD, бос энергия*