International Journal of Computational Bioinformatics and In Silico Modeling
Vol. 2, No. 2 (2013): 89-93
Research Article
Open Access

ISSN: 2320-0634

Modified Pattern Matching (MPM) Algorithm for Detecting the Ribosome Binding Sites: Trends for Bioinformatics Analysis of *Escherichia coli* K-12 (MG 1655) Genome

Junaid Bin Ahsan[1], Adnan Mannan[2], Rasel Das[3], Muhammad Ibrahim Khan[1], Md. Arifuzzaman*[4]

[1] *Department of Computer Science & Engineering, Chittagong University of Engineering & Technology, Chittagong, Bangladesh.*
[2] *Department of Genetic Engineering and Biotechnology, University of Chittagong, Chittagong-4331, Bangladesh.*
[3] *Nanotechnology and Catalysis Research Centre, University of Malaya, 50603 Kuala Lumpur, Malaysia.*
[4] *Department of Biochemistry and Biotechnology, University of Science & Technology Chittagong, Chittagong, Bangladesh.*

*Corresponding author: Md. Arifuzzaman, e-mail: larif67@yahoo.com

Received: 22 January 2013 Accepted: 12 February 2013 Online: 01 March 2013

ABSTRACT

Shine-Dalgarno (SD) sequence is found just 5' to the translation initiation codon, part or all of a polypurine domain UAAGGAGGU is found in the prokaryotic mRNA ribosome binding site (RBS). The importance of the SD sequence for identification of the translation initiation site on the mRNA by the ribosome is very much clear, and thereby, translational efficiency is strongly affected by the spacing between the SD and the initiation codon. Although, whether there is a unique optimal spacing is not as clear. The definitions of the spacing as well as secondary structures have been complicated and obscured matters. A systematic study installed by the development of a novel modified algorithmic approach to detect the most probable RBS which is located at upstream of any gene of *E. coli* genome were undertaken by us, and in addition we have also developed an algorithm which picks up the absolute spacers between each RBS and AUG start codon. Moreover, in order to determine their expression rate in *E. coli* K-12 (MG1655) genomes all genes on the basis of their spacer lengths between RBS and AUG start codons of that gene were classified. A new insight to disclose the actual happening and ambiguity of *E. coli* K-12 (Mg1655) genome in a more efficient manner might be provided by this global gene analysis.

Keywords: Shine dalgarno sequence, upstream, codon, ribosome binding site, highly expressed gene

INTRODUCTION

The functional regulatory sites such as gene promoters and regulatory regions are frequently localized by pattern recognition [1]. In prokaryotes the 16S ribosomal RNA drives the ribosome to bind with specific mRNA ribosome binding site (RBS), called SD (Shine-Dalgarno) sequence [2-4]. This binding site is generally extended almost 20-25 nucleotides upstream of AUG start codon and consisting of polypurine nucleotide sequences (UAAGGAGGU) in prokaryotes [2, 3 and 5]. The space between the SD and AUG varies from gene to gene in a genome, with the average being seven nucleotides [3 and 4]. But, some scientists have proposed this range may vary from 5-13 nucleotides [6]. So, there are some uncertainties about absolute location and spacer lengths between SD and AUG sequences in bacteria. But, these long and short spacers between the SD and AUG sequences have considerable effects on the rate of translation of any gene [7 and 8]. For this variability, it is difficult to pick up absolute RBS from an mRNA of a gene. Therefore, pattern recognition might play an important role to solve this problem to deal with every gene in an efficient manner. Such recognition is not only important to detect regulatory

sites but also it can detect the signal for identification of any gene. Every gene might have an RBS in any organism genome. Most of the conventional methods existed to detect the SD sequence are often failed because of SD's degenerative and flexibility characters. For this, the present study has installed by the development of a novel modified algorithmic approach to detect the most probable RBS which is located at upstream of any gene of *E. coli* genome. In addition, we have also developed an algorithm which picks up the absolute spacers between each RBS and AUG start codon. Further, we classified highly expressed genes on the basis of their spacer lengths between RBS and AUG start codons in order to determine their absolute expression rate in *E. coli* K-12 (MG1655) genomes. This global gene analysis might give a new insight to disclose the actual happening and ambiguity of *E. coli* K-12 (Mg1655) genome efficiently.

MATERIALS AND METHODS

For pattern recognition of SD sequence, we have retrieved the upstream region of each gene of *E. coli* K-12 (Mg1655) genome (4185 genes) from the database 'ecogene' (http://ecogene.org/) [9].

Implementation

Detection of SD by Modified Pattern Matching Algorithm "MPM"

We have proposed here, a modified pattern matching (MPM) algorithm according to the maximum percentage matching to separate the spacers by pointing out the SD sequence "AGGAGGU" [2]. We call this algorithm "MPM" as a nutshell. The algorithm works as follows: first it takes a string, "S" containing a portion of upstream of a gene 25 nucleotides. Then it checks for SD sequence with the "AGGAGGT". It returns the substring which has maximum match with "AGGAGGT" and which has maximum "G". Once the SD is pointed out, we extract the spacers and save it into a file. The MPM algorithm is shown below.

Procedure MPM(S)

```
{
len:=strlen(S);
count1:=(len-6);
count2:=0,count3=0,max=0,mG=0;
for i:=0 to (count1-1)
{
temp:=substring of S(length:=i to (i+6));
count2:=perc1("AGGAGGT",temp);
count3:=perc2("AGGAGGT",temp);

if(mG<count3)
{
if(max<count2)
{
max:=count2;
mG:=count3;
SD:=S;
}}}}
```

procedure perc1 (real_SD,substring)

```
{
total=0;
for i:=0 to 6
{
if(substring[i]==real_SD[i])
{
total+:=1;
}
}
total=total*100/7;
return total;
}
```

procedure perc2(real_SD,substring)

```
{
G=0;
for i=0 to 6
{
if(real_SD[i]==substring[i])
{
if(substring[i]=='G')
{
G+=1;} } }
return G;
}
```

The two sub algorithm perc1 and perc2 returns maximum match with "AGGAGGT" and maximum "G" in that substring.

Finding out the consensus sequence of each group

We retrieved all the highly expressed genes of *E. coli* from the "List of Highly Expressed Genes" at the (http://genomes.urv.es/HEG-DB/consulta/cons_heg.php?nom[]=ecoli) [10]. According to their spacer lengths we grouped them on the basis of conserved consensus spacer sequence by using a novel conserved algorithm. This algorithm takes the strings of equal length. Then it returns the character which is optimum in each column of string array. If the column of string array contains more than one character of equal intensity, then it shows too. The conserved algorithm is shown below together with two sub maximum finding algorithm.

procedure conserve (S,n,m) //S is the array of strings of equal length,n is the total number of //strings,m is the string length

```
{
str[1:n][1:m];//two dimensional array of strings
ary[1:4];
A=0,C=0,G=0,T=0;
for j=0 to (m-1)
{
for i=0 to (n-1)
{
if(S[i][j]=='A')
{
A+=1;
}
elseif(S[i][j]=='C')
{
```

```
C+=1;
}
elseif(S[i][j]=='G')
{
G+=1;
}
elseif(S[i][j]=='T')
{
T+=1;
}
}
ary[0]=A;
ary[1]=C;
ary[2]=G;
ary[3]=T;
m1=max1(ary);
m2=max2(ary);
l=0;
for k=0 to 3
{
if(m2==ary[k])
flag[j][l]=k;
else
flag[j][l]=-1;
l++;
}
}
s[4]={'A','C','G','T'};
for l=0 to (m-1)
{
for k=0 to 3
{
if (flag[l][k]!=-1)
  print s[flag[l][k]];
 } } }
```

procedure max1(x)//x is the array of length 4

```
{
r=0,m=0;
for i=0 to 3
{
if(m<x[i])
{
m=x[i]
r=i;
} }
return r;
}
```

procedure max2(x) //x is the array of length 4

```
{
m=0;
for i=0 to 3
{
if(m<x[i])
{
m=x[i];
}
}
return m;
```

RESULTS AND DISCUSSION

Development of New Modified Algorithm to Detect the SD Sequence

Although there are several types of pattern recognition algorithms exists, but this novel algorithmic approach is more efficient to detect ribosome binding site in mRNA. This algorithm is able to pick up the highest similar sequences to standard (UAAGGAGGU) within the 25 nucleotides of the upstream of any gene in genome of *E. coli.* Traditional "pattern matching algorithm" [11] is unable to deal with this task. For handling the global genes of *E. coli* genome and decreasing the error rate, we installed modified "pattern matching algorithm" depicted as flow chart below (Fig. 1). This algorithm is able to pick up the highest similar SD sequence along with spacer between SD and AUG start codon of any gene in *E. coli* genome.

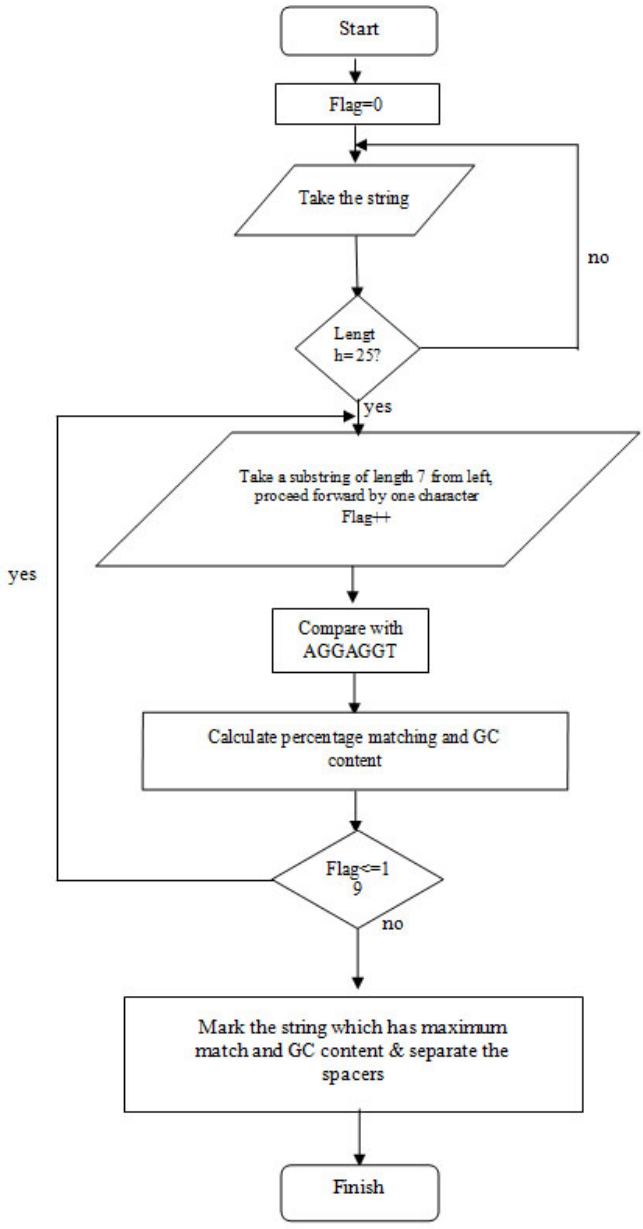

Figure 1. Development of novel modified algorithm for detecting the RBS or SD sequence. Here mg means maximum of G and mp means maximum percentage.

Construction of an Algorithm for Conserved Spacers Analysis

We have sorted all the highly expressed gene spacers by using "Bubble sort algorithm" [12]. Further we developed a new algorithm for detecting the conserved spacer sequences among them. The flowchart is given in Fig. 2

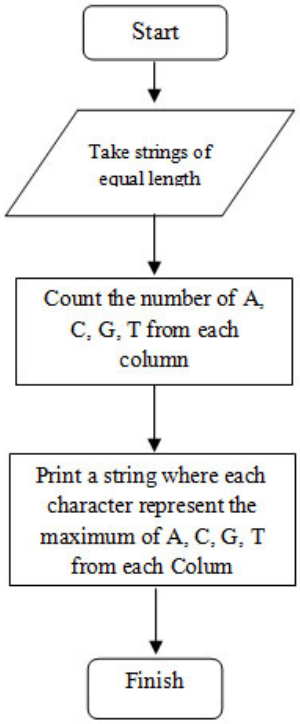

Figure 2. Flowchart of conserved spacer sequence algorithm

We have picked up and compiled all the ribosome binding sides of all genes in *E. coli* K-12 (MG 1655) genome by using our new algorithm MPM (Modified Pattern Matching) algorithm (Supplementary file 1). Although traditional pattern matching algorithm exist [11]; but this new algorithm is efficiently able to predict the SD sequence efficiently. For further bioinformatics analysis of highly expressed genes in *E. coli,* we have classified them into 16 groups on the basis of conserved spacer sequences length (from 3 to 18) and want to see whether the spacer length has an effect on the gene expression (Supplementary file 2). Interestingly, we have observed that the effects of spacer's length on the highly expressed gene expression. Since, it has been shown that, the number of highly expressed genes decreases with increasing conserved spacer length (Fig. 3). In addition, there are 16 groups which contain 16 spacers (from 3 to 18) and amazingly we have observed that, all spacers contained maximum Adenine (A) in their sequence composition (Supplementary file 2).

So, the spacer length can affect the translation of any genes in *E. coli.* This finding is fully consistent with published hypothesis [6], that is, SD-AUG spacing plays a significant role in the process of translation initiation and aligned spacing is the most appropriate measure of spacing [6].

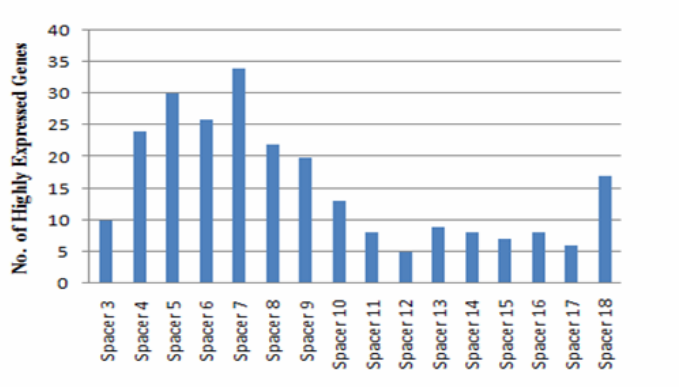

Figure 3. Statistical Analysis of Highly Expressed Genes of *E. coli* on the Basis of Spacers Length

CONCLUSION AND FUTURE STUDY

Ribosome binding sites in mRNA are degenerated strongly in their sequences. So, the traditional pattern recognition algorithm is not suitable. It causes a paradigm shift in the development of a new algorithm MPM which is able to minimize ambiguities of RBS sensing. But in future we may also use the Neural Network paradigms, Self Organizing Maps (SOM) or Adaptive Resonance Theory (ART) to optimize the final results of this study which might be able to evaluate all true results of RBS detection. It would also be nice to apply Decision Trees and Regression pattern analysis to the RBS recognition problems. Another finding was that, the effect of conserved spacer lengths on gene expression. Furthermore, future studies are needed to give answer of a question why high percentage Adenine (A) nucleotides are present in spacers of highly expressed genes in *E. coli* genome.

REFERENCES

1. Oliveira MFDS, Mendes DQ, Ferrari LI and Vasconcelos ATR (2004). Ribosome binding site recognition using neural networks. J of Gen. Mol Bio 27 (4): 644-650.
2. Shine J and Dalgarno L (1974). The 3'-Terminal Sequence of Escherichia coli 16S Ribosomal RNA: Complementarity to Nonsense Triplets and Ribosome Binding Sites. Proc. Natl Acad. Sci. USA. 71: 1342-1346.
3. Gold L (1988). Posttranscriptional regulatory mechanisms in Escherichia coli. Ann. Rev. Biochem. 57: 199-233.
4. Kozak M (1983). Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. Microbiol Rev. Mar. 47(1):1–45.
5. Steitz JA (1969). Polypeptide chain initiation nucleotide sequences of the three ribosomal binding sites in bacteriophage R17 RNA. Nature 224: 957-967.
6. Chen H, Bjerknes M, Kumar R, Jay K (1994). Determination of the optimal aligned spacing between the Shine- Dalgarno sequence and the translation initiation codon of Escherichia coli mRNAs. Nuc Acid. Res, 22 (23): 4953-4957.
7. Roberts TM, Bikel I, Yocum RR, Livingston DM, Ptashne M (1979). Synthesis of simian virus 40 t antigen in Escherichia coli. Proc. Natl. Acad. Sci. USA. 76(11): 5596–5600.
8. Singer BS, Gold L, Shinedling ST, Colkitt M, Hunter LR, Pribnow D, Nelson MA (1981). Analysis in vivo of translational mutants of the rIIB cistron of bacteriophage T4. J. Mol. Biol. 149(3):405–432.
9. Rudd KE (2000). EcoGene: a genome sequence database for Escherichia coli K-12. Nucleic Acids Res. 28: 60-64.
10. Puigbo P, Guzman E, Romeu A and Garcia-Vallve S (2007) OPTIMIZER: A web server for optimizing the codon usage of

DNA sequences. Nucleic Acids Res. Vol. 35: No. suppl_2 W126-W131.
11. Cormen TH, Leiserson CE, Rivest RL, Stein C (2001) [1990]. Introduction to Algorithms (2nd ed.). MIT Press and McGraw-Hill. ISBN 0-262-03293-7. Pp. 909.
12. Lipschutz S (2006). Schaum's Outline of Theory and Problems of Data Structures (Schaum's Outlines); Mcgraw-Hill (December 1986), ISBN: 978-0070380011

This manuscript contains supplementary data.

Excel Data:
http://bioinfo.aizeonpublishers.net/content/2013/2/suppl_89-93.xls

Access data from:
http://bioinfo.aizeonpublishers.net/content/2013/2/suppl_89-93.pdf

*****