An Algorithm for Promoter Mapping of ALGT (ECF Sub Family of Sigma Factor)

Dr. MOHAMMAD ALAMGEER
Bioinformatician & Software Engineer
Assistant Professor
Department of Information Systems
King Khalid University, Kingdom of Saudi Arabia (KSA).

ABSTRACT:- Cystic fibrosis (CF) is the most common lethal inheritable disease in Caucasians [1]. The primary contributors to the high morbidity and mortality in CF are the chronic respiratory infections caused by bacterial pathogens [2]. The predominant CF pathogen is Pseudomonas aeruginosa, and over 90% of CF patients eventually become colonized with this organism [3]. A classical feature of P. aeruginosa strains infecting CFpatients is that they mutate into the mucoid, exopolysaccharide alginate-overproducing form, in a process referred to as conversion to the mucoid phenotype [4]. The conversion of Pseudomonas aeruginosa to the mucoid phenotype coincides with the establishment of chronic respiratory infections in cystic fibrosis (CF). A major pathway of conversion to mucoidy in clinical strains of P. aeruginosa is dependent upon ctivation of the alternative sigma factor AlgU (P. aeruginosa _E) [5]. At the genetic level, the conversion to mucoidy in P. aeruginosaoccurs via mutations in a cluster of genes encoding the alternative sigma factor AlgU [6], also known as AlgT [7, 8], and an array of AlgU regulators: MucA, MucB, MucC, and MucD [9, 10, 11, 12].
Extracytoplasmic function (ECF) sigma factors constitute a diverse family of proteins, within the class of the sigma 70 subunit of RNA polymerase.

KEYWORDS:- Promotor Mapping, AlgT, Sigma Factors, ECF, Escherichia coli

I. BACKGROUND

Pseudomonas aeruginosa is an opportunistic pathogen that causes chronic infections in Cystic Fibrosis patients. Frequent nosocomial infections are caused by this pathogen. Many clinical isolates in particular from Cystic fibrosis patients exhibit a mucoid phenotype. This is due to copious production of the polysaccharide alginate. Alginate is an important virulence determinant for *Pseudomonas aeruginosa*. It is believed that it inhibits phagocytosis and potentially limits antibiotic efficacy due to limited antibiotic penetration. We have been interested in understanding the regulation and production of alginate in *P. aeruginosa.* AlgT is a member of the ECF subfamily of sigma factors. It has been shown to control expression from the 18 kb biosynthetic operon for alginate. Members of the ECF family of sigma factors exist in a diverse group of organisms where they respond to various forms of extra cytoplasmic stimuli.

II. PROMOTER MAPPING OF AlgT (ECF SUB FAMILY OF SIGMA FACTOR)

The extracytoplasmic function (ECF) sigma factors are found in a diverse range of bacteria and many are activated to transcribe their regulons in response to a change in environmental conditions [13]. For example, ECF sigma factors regulate iron uptake and heat-shock responses in Escherichia coli [14], alginate biosynthesis and exotoxin secretion in Pseudomonas aeruginosa [15], carotenoid biosynthesis in Myxococcus Xanthus [16] and expression of the thioredoxin system in response to oxidative stress in Streptomyces coelicolor [17]. Here I present a proteomic analysis to examine the regulon for AlgT. We have identified and demonstrated that *dsbA* is under transcriptional control of AlgT. I present here characterization of this gene and additional potential components of the AlgT regulons.

There is a recursive program to find the AlgT promoter region in nucleotide sequence. This algorithm compute the sub sequence(s) – $s(S_1, S_2, \ldots, S_n)$ present in sequence (S) with pattern in which sequence start with pattern of AlgT promoter and end with Stop codon (TAA / TAG / TGA). The pattern of promoter is – XXXXXX……XXX[XXXX]X……XXX……Stop. In this pattern X represents nucleotide base pair - A, C, T, and G. In this pattern the first gap should be 16 or 17 base pairs long, second gap should be 4 to 10 base pairs long and the third gap is up to any number of base pairs with multiplication of 3 (triplet codon for protein synthesis) followed by Stop codon. The minimum length of initial pattern for mapping is 37.

Let S be a sequences in which we want to find out the presence of promoter region. To compute the same, we must first need to check the length of sequence S larger than minimum length of pattern. If found then proceed the process to find pattern of promoter

region. First we need to find the presence of first, second, and third sub strings of patter in entire length of given nucleotide sequence and then if found then need to find the position of the presence of respective string. Let the position be p_1, p_2, and p_3 respectively. Now we need to find the gap length between p_1 & p_2 and p_2 & p_3. Let the respective gap be $g1$ and g_2. If the first gap length ($_{g1}$) is equal to 16 or 17 and the second gap length (g_2) is between 4 to 10, then we need to find Stop codon after p_3. If Stop codon present then the substring from the initial position of p_1 to stop codon is the sub sequence s(S_1) which promote the sigma factor for ECF sub family. Now recursively we find the next sub sequences $s(S_2, S_3. . . . . , S_n)$ in new string left after stop codon position.

Algorithm *Promoter Mapping of AlgT - ECF Subfamily*

Input: nucleotide sequence (*S*)
Output: Position and Subsequence of AlgT - P(*S*)

```
pro_seq1 ← xxxxxx
pro_seq2 ← xxx[xxxx]x
pro_seq3 ← xxx
pro_seq ← pro_seq1, pro_seq2, pro_seq3
min_length ← length of AlgT Sequence pattern
new_seq ← S
l ← length of S
for k ← l to min_length do
        if length(new_seq) > min_length then
                new_seq ← promap (new_seq)
promap (new_seq)
for j ← 1 to 3 do
        indices[j] ← push (indices[j], position(new_seq, pro_seq[j]) – len(pro_seq[j]))
for i ← 0 to indices[0] do
        for j ← 0 to indices[1] do
                p1 ← indices[0][i] + 1
                p2 ← indices[1][j] + 1
                if p1 < p2 and flag1 = false then g1 ← p2 - p1 – len(pro_seq[0]
                        if g1 = 16 or g1 = 17 then flag1 ← true
if flag1 = true then
        for k ← 0 to indices[2] do
                p3 ← indices[2][k] + 1
                if p2 < p3 and flag2 = false then g2 ← p3 - p2 - len(pro_seq[2]
                if g2 >= 4 and g2 <= 10 then flag2 ← true
if flag2 = true then
        for j ← (p3+2) to length(new_seq) do step 3
                if subsequence(new_seq, j, 3) = Stop Codon then p4 ← j
        result_seq ← subsequence (new_seq, p1 - 1, p4 - p1 + 4)
return subsequence(new_seq, p4)
```

Figure 1: An algorithm for promoter mapping of AlgT (ECF sub family of sigma factor)

Time complexity:

The time complexity of this programming operation depends on the length of genomic sequences. To see this, first we need to match all three pattern of subsequence part of promoter and make a separate list of positions of their presence in entered nucleotide sequence. This is the dominant term in the time complexity. Now we need to match pairs of such positions in which gap g_1 and g_2 of required length present. After getting the gaps, we need to find stop codon just after third subsequence at interval of

tinplate codon. If found then we can say that promoter region present otherwise not present in nucleotide sequence. After getting the pattern of promoter region, we start to search next promoter region with making a new nucleotide sequence (part of original nucleotide sequence after stop codon of previous pattern).

Let we are entering the nucleotide sequence of length l base pairs log. Total list of positions found for all three subsequences of pattern is $p_1[n_1]$, $p_2[n_2]$, and $p_3[n_3]$ respectively. The sum of all searches is computed as –

$$l \text{ x } 6 + l \text{ x } 5 + l \text{ x } 3 \ldots\ldots\ldots \quad \text{(i)}$$

Out of these positions only one from each subsequence is responsible to make a single pattern for promoter mapping. Now we need to perform number of all possible matches to get proper required gaps g_1 and g_2 between two successive sub sequences of promoter region by getting positions from n_1, n_2, and n_3. The sum of all possible comparisons can be computed as –

$$n_1 \text{ x } n_2 \text{ x } n_3 \ldots\ldots\ldots\text{(ii)}$$

The above all possible search can be computed in closed form as follows –

$$[(l \text{ x } 6 + l \text{ x } 5 + l \text{ x } 3) + (n_1 \text{ x } n_2 \text{ x } n_3)] \ldots\ldots \quad \text{(iii)}$$

If gap g1 of proper length found, then we find gap g2 of proper length. If both g1 and g2 present then we can say that promoter region present. After that to find stop codon, start search just after third subsequence at interval of multiplication of three base pairs. If the pattern of promoter mapping is found, then the total number of iterations required to execute the program is –

$$j = k, k = p_4$$

$$\sum \ [((l\text{-}k) \text{ x } 6 + (l\text{-}k) \text{ x } 5 + (l\text{-}k) \text{ x } 3) + (n_1 \text{ x } n_2 \text{ x } n_3)] + [(l\text{-}k) / 3]\ldots\ldots\text{(iv)}$$

$$j = 0, k = 0$$

If the pattern of promoter mapping is not found, then the total number of iterations required to execute the program is –

$$[(l \text{ x } 6 + l \text{ x } 5 + l \text{ x } 3) + (n_1 \text{ x } n_2 \text{ x } n_3)] \ldots\ldots\text{(vi)}$$

III. RESULTS AND PROOF

This algorithm used to compute the presence of promoter region of AlgT ECF sub family in nucleotide sequence. This accepts nucleotide sequence of length l (longer than length of promoter region). After computing, it returns subsequence(s) of pattern matched region commonly present inside the nucleotide sequence.

To compute the AlgT algorithm, let we input a nucleotide sequence *S of length l*, where $l = 292$.

ACGATAGATACAGATAGATCCTGAACTGATA
GACAGATAGATACACTGATACAGAACAAGAT
AGGAACTTAGCATAGATATAGCAGTTCTAAA
GCATGCAATGACGATAGATACAGATAGATCT
GATAGACAGATAGATACACGATAGATACAG
ATAGATCCTGAACTGATAGACAGATAGATAC
ACTGATACAGATAAGACAAGATAGGAACTTA
GCATAGATATAGCAGTTCTGAAGCATGCAAT
GACGATAGATACAGATAGATCCTGAACTGAT
AGACAGATAGATAC

First it finds the list of positions of first, second and third subsequence pattern of promoter from first position of nucleotide sequence S, where value of k is 0. Let pattern of subsequences are – GAACTT, TCT[ACTG]A, and ATG respectively. The patter for Stop codon is TAA/TAG/TGA. After computing, the values for positions of subsequences are - $p_1(65, 210)$, $p_2(88, 122, 233)$, and $p_3(96, 101, 241, 246)$.

Now it finds possible gaps g_1 and g_2 between $p_1[n_1]$ & $p_2[n_2]$ and $p_2[n_2]$ & $p_3[n_3]$. To compute the same, we do following comparisons -

$p_1[1] < p_2[1]$
So the gap $g_1 = p_2[1] - p_1[1] - 6 = 88 - 65 - 6 = 17$
$g_1 = 16$ or 17 (between required gap length). So we precede the process.

$p_2[1] < p_3[1]$
So the gap $g_2 = p_3[1] - p_2[1] - 5 = 96 - 88 - 5 = 3$
g_2 is not between 4 to 10 17 (between required gap length). So we need to compute the g_2 value with next position of p_3 list.

$p_2[1] < p_3[2]$
So the gap $g_2 = p_3[2] - p_2[1] - 5 = 101 - 88 - 5 = 8$
$g_2 >= 4$ and $g_2 <= 10$. So we precede the process.
Now we need to find the position p_4 for presence of Stop codon. The number of base pairs (gap g_4) between p_4 and third subsequence must be in multiplication of three.
Position of Stop codon (p_4) after third subsequence $p_3[2] + 5 = 146$

$g_3 = 146 - p_3[2] - 3 = 146 - 101 - 3 = 42$
42 is in multiplication of 3.
That means the pattern for promoter mapping found and the pattern is the consequence stretch of subsequence of positions from p1 to p4 and is –

GAACTTAGCATAGATATAGCAGTTCTAAAGCATGCAATGACGATAGATACAGATAGATCTGATAG
ACAGATAGATACACGATAG

By recursion process, the length of subsequences decreases. If the length of new nucleotide sequence is sufficient, compute the algorithm recursively. Now to find the other promoter region, we need to repeat the same process with new sequence of nucleotide starts from stop codon.

Now for next execution, the sequence is –

ATACAGATAGATCCTGAACTGATAGACAGAT
AGATACACTGATACAGATAAGACAAGATAG

GAACTTAGCATAGATATAGCAGTTCTGAAGC
ATGCAATGACGATAGATACAGATAGATCCTG
AACTGATAGACAGATAGATAC

The length of nucleotide sequence (l = 144) is sufficient to execute the algorithm. Again first we compute the positions of first, second and third subsequence of promoter pattern from first position of S, where value of $k = p_4 = 145$. After computing, the values for positions of subsequences are - $p_1(62)$, $p_2(85)$, and $p_3(93, 98)$.

Now it finds possible gaps g_1 and g_2 between $p_1[n_1]$ & $p_2[n_2]$ and $p_2[n_2]$ & $p_3[n_3]$ with following comparisons -

$p_1[1] < p_2[1]$
So the gap $g_1 = p_2[1] - p_1[1] - 6 = 85 - 62 - 6 = 17$
g_1 = 16 or 17 (between required gap length). So we precede the process.

$p_2[1] < p_3[1]$
So the gap $g_2 = p_3[1] - p_2[1] - 5 = 93 - 85 - 5 = 3$
g_2 is not between 4 to 10 17 (between required gap length). So we need to compute the g_2 value with next position of p_3 list.

$p_2[1] < p_3[2]$
So the gap $g_2 = p_3[2] - p_2[1] - 5 = 98 - 85 - 5 = 8$
g_2 >= 4 and g_2 <= 10. So we precede the process.

Now we need to find the position p_4 for presence of Stop codon. The number of base pairs (gap g_4) between p_4 and third subsequence must be in multiplication of three.
Position of Stop codon (p_4) after third subsequence $p_3[2] + 5 = 121$
$g_3 = 122 - p_3[2] + 3 = 122 - 98 - 3 = 21$
21 is in multiplication of 3.
That means the pattern for promoter mapping found and the pattern is the consequence stretch of subsequence of positions from p1 to p4 and is –

GAACTTAGCATAGATATAGCAGTTCTGAAGCATGCAATGACGATAGATACAGATAGATCCTG
A

Now we go to repeat first step to find next region with new sequence of nucleotide starts from stop codon. Now for next execution, the new sequence is –

ACTGATAGACAGATAGATAC

The length of nucleotide sequence (l = 20) is not sufficient to execute the algorithm. So algorithm exits from further iteration.

IV. AVAILABILITY

Through bioinformatics approach, GenSolution is a website, accessible at the URL http://www.gensolution.org (now partner site of AZ Group of Education & Technology: http://www.az-group.org). This website provides biological databases and modules to solve biological problems by computational method. The author also implements the algorithm for promoter mapping of AlgT (ECF sub family of signa factor). This program is accessible through drop down menu option of SeqComparison present at website home page of GenSolution web portal. It accepts nucleotide sequence in text area (Figure 2). If promoter for AlgT present in entered sequence then it displays desire result (Figure 3) otherwise it displays a message - “Sorry! Promoter region for AlgT does not exist”.

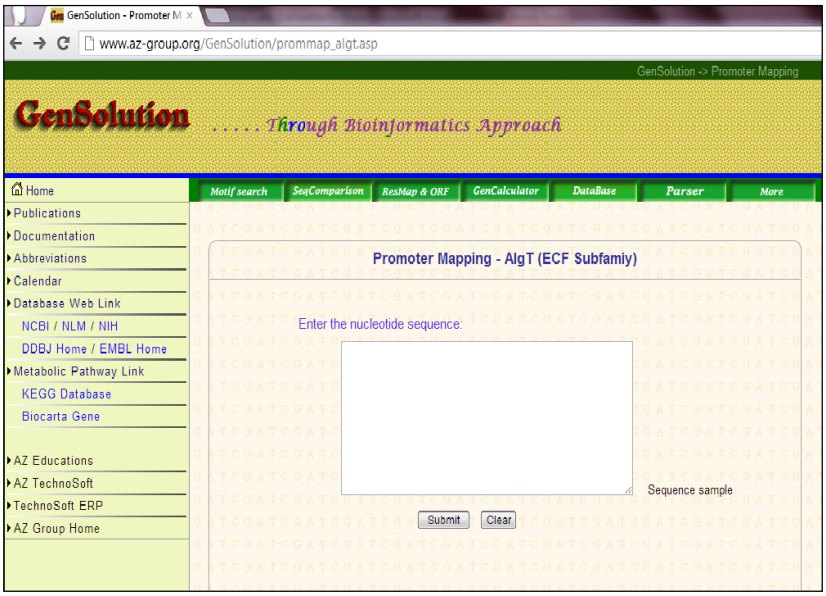

Figure 2: Form to accept nucleotide sequence for promoter mapping of AlgT(ECF sub family of sigma factor)

If user entering listed nucleotide sequence then after submitting it will displays result (Figure 3) with showing all possible fragment of promoter for AlgT with specific position indication.

Entered Sequence:

acgatagata cagatagatc ctgaactgat agacagatag atacactgat acagaacaag ataggaactt agcatagata tagcagttct aaagcatgca atgacgatag atacgatag atctgataga cagatagata cacgatagat acagatagat cctgaactga tagacagata gatacactga tacagataag acaagataag aacttagcat agatatagca gttctgaagc atgcaatgac gatagataca gatagataat gaactgatag acagatagat ac

Result page:

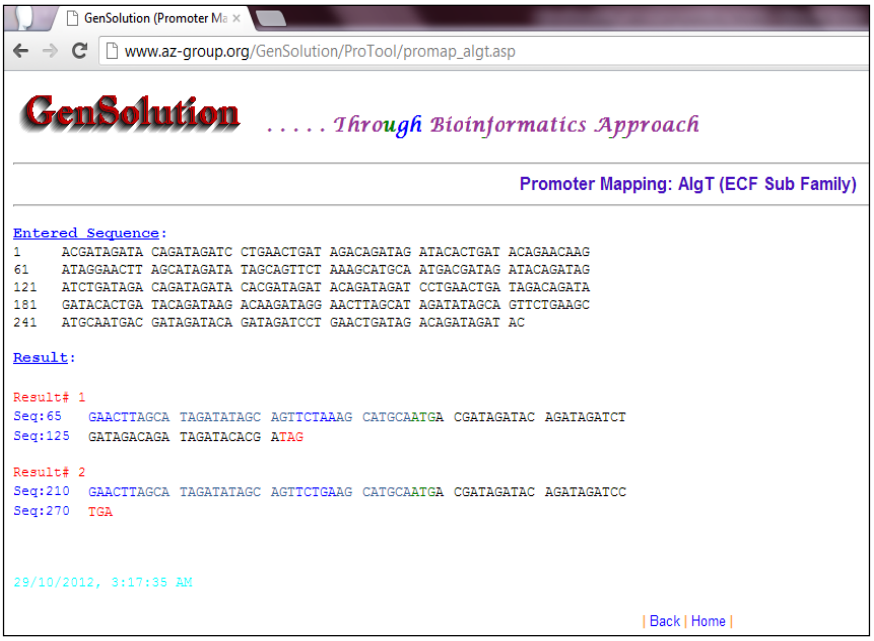

Figure 3: Result page of promoter mapping of AlgT(ECF sub family of sigma factor)

V. ACKNOWLEDGEMENTS

The author thanks Dr. Sonal Malhotra for discussions and encouragements. He gratefully acknowledges the partial support of Dr. Moinudin Khan and Dr. Abdul Ilah. The author also thanks Dr. Kulvinder Singh Saini and Dr. V.C. Kalia for giving me opportunity to realize such types of biological problem during the project training at Ranbaxy and IGIB research center respectively. I am thankful to my wife Mrs. Shameema Rahman, family members, and parents for their support, freedom and, motivation. Above all I thank "Almighty Allah".

REFERENCES

[1] Welsh, M. J., L.-C. Tsui, T. F. Boat, and A. L. Beaudet. 1995. Cystic fibrosis, p. 3799–3876. In C. R. Scriver, A. L. Beaudet, W. S. Sly, and D. Valle (ed.), The metabolic and molecular basis of inherited disease, vol. III. McGraw-Hill, Inc., New York, N.Y.

[2] Tatterson, L. E., J. F. Poschet, A. Firoved, J. Skidmore, and V. Deretic. 2001. CFTR and Pseudomonas infections in cystic fibrosis. Front. Biosci. 6:D890–D897.

[3] FitzSimmons, S. C. 1993. The changing epidemiology of cystic fibrosis. J. Pediatr.122:1–9.

[4] Govan, J. R. W., and V. Deretic. 1996. Microbial pathogenesis in cystic fibrosis: mucoid Pseudomonas aeruginosa and Burkholderia cepacia. Microbiol. Rev. 60:539–574.

[5] Aaron M. Firoved, J. Cliff Boucher. and Vojo Deretic. 2002. Global Genomic Analysis of AlgU (_E)-Dependent Promoters (Sigmulon) in Pseudomonas aeruginosa and Implications for Inflammatory Processes in Cystic Fibrosis. JOURNAL OF BACTERIOLOGY, Feb. 2002, p. 1057–1064

[6] Martin, D. W., B. W. Holloway, and V. Deretic. 1993. Characterization of a locus determining the mucoid status of Pseudomonas aeruginosa: AlgU shows sequence similarities with a Bacillus sigma factor. J. Bacteriol. 175: 1153–1164.

[7] Brightbill, H. D., D. H. Libraty, S. R. Krutzik, R. B. Yang, J. T. Belisle, J. R. Bleharski, M. Maitland, M. V. Norgard, S. E. Plevy, S. T. Smale, P. J. Brennan, B. R. Bloom, P. J. Godowski, and R. L. Modlin. 1999. Host defense mechanisms triggered by microbial lipoproteins through toll-like receptors. Science 285:732–736.

[8] Flynn, J. L., and D. E. Ohman. 1988. Cloning of genes from mucoid Pseudomonas aeruginosa which control spontaneous conversion to the alginate production phenotype. J. Bacteriol. 170:1452–1460.

[9] Boucher, J. C., J. Martinez-Salazar, M. J. Schurr, M. H. Mudd, H. Yu, and V. Deretic. 1996. Two distinct loci affecting conversion to mucoidy in Pseudomonas aeruginosa in cystic fibrosis encode homologs of the serine protease HtrA. J. Bacteriol. 178:511–523.

[10] Boucher, J. C., M. J. Schurr, H. Yu, D. W. Rowen, and V. Deretic. 1997. Pseudomonas aeruginosa in cystic fibrosis: role of mucC in the regulation of alginate production and stress sensitivity. Microbiology 143:3473–3480.

[11] Martin, D. W., M. J. Schurr, M. H. Mudd, and V. Deretic. 1993. Differentiation of Pseudomonas aeruginosa into the alginate-producing form: inactivation of mucB causes conversion to mucoidy. Mol. Microbiol. 9:497–506.

[12] Martin, D. W., M. J. Schurr, M. H. Mudd, J. R. W. Govan, B. W. Holloway, and V. Deretic. 1993. Mechanism of conversion to mucoidy in Pseudomonas aeruginosa infecting cystic fibrosis patients. Proc. Natl. Acad. Sci. USA 90: 8377–8381.

[13] Helmann, 2002; Lonetto et al., 1994; Raivio & Silhavy, 2001.

[14] Braun, 1997; De Las Penas et al., 1997.

[15] Hershberger et al., 1995; Ochsner et al., 1996.

[16] Gorham et al., 1996.

[17] Kang et al., 1999; Paget et al., 1998.

