Specific binding sites and DNA microarray analysis for gene dependency

Tomoyuki Ueda (210006)

School of Information Science,
Japan Advanced Institute of Science and Technology

February 13, 2003

Keywords: gene dependency , specific sequences , statistical analysis , similarity.

The word 'genome' comes from 'gene' and 'chromosome'. and suggests the information of genes on a chromosome. Now, The genome science is shifting the focus form structual information of a gene to functional information of a gene[1]. Various factors affect the production process of protein[2]. The protein is called 'transcription regulatory factor' , when the protein that be producted by a geneA affects transcription of a geneB. Therefore, a geneB is depend on a geneA. In fact, a geneB is regulated by a geneA, when the protein that isproducted by a geneA binds to the transcription regulatory region of a geneB, and regulate the transcription of a geneB. There is a gene dependency. In this bout, a geneA is called 'regulatory gene'. It is known that there are similar sequences among transcription regulatory region of those genes that are regulated by a common transcription regulatory factor. There are known researches about promoter sequence[3]. Their are mathematical models and algorithmical methods for trying to identify promoter sequences. The methods concern both searching in a genome for a previously defined consensus and extracting a consensus for a set of sequences. Such methods were often tailored for either eukaryotes or prokaryotes although this does not preclude use of the same method for both types of organisms. There are great hopes that these methods are a good way of discovering unknown genes, and

unknown members of group that be regulated by a common gene[10]. In other research, it is shown that if genes have a common specific binding sequence, they have a common regulation[4]. This is significance for gene dependency in statistical. So, Analysis of transcription regulatory regions lights on gene dependency estimation. Other existing researches for gene dependency are Blloean Network[5], Baysian Network[6], S-system[7] and so on. Their researches are proposing some graphical network models for gene dependency.But,their researches used only DNA microarray data. DNA microarray technologies have made it possible to measure the expression levels of thousands of genes simultaneously under different conditions. But, this data have many problems[5]. For example, Their problems are low reliability, some solutions, assortment explosionand so on. Then, in our study, gene dependency is deduced by analysis of specific binding sites and DNA microarray data. Therefore, our study will be high reliability compared to known researches that used only DNA microarray data. On the occasion of our study, we use three datas of Bacillus subtilis[9]. One of datas is DNA microarray datas that are offered by Kyusyuu Univ. Other datas are DNA sequence that is published on WEB site (NBCI: http//www.nbci.nlm.nih.gov/) and known gene dependencies that are published on WEB site (DBTBS: http://dbtbs.hgc.jp) The first of our study is that putative specific sequences are abstracted from transcription regulatory regions by statistical analysis. Then, covering these sequences, max score of similarity between local regions on comparing genes is intergenal similar score. The second , it is estimated that a gene cluster have a common regulation, when a gene cluster is contained with two groups. One of groups has high similar score. Another group has high expression intensity more than threshold. Here is our procedure. The first step, we have transcription regulatory regions of each gene. Their lengths are no fewer than 0[ℓp] , nor more than 529[ℓp] . Their start positions are start positions of each gene. The next step, we study frequency of appearance of strings that have several lengths on transcription regulatory regions of each gene. Several lengths are no fewer than 3 , nor more than 12 . As a result of this research, frequency of appearance of strings is predictably in strings of long length. we study about short string ,because long string is expressed in short string. O_s is frequency of appearance for each string.

E_s is expectation. Degree of specificity defined as $(O_s - E_s)/E_s$. Then, we observe transcription regulatory regions that are expressed in degree of specificity. We discovered that known specific sequences are contained with minus regions. Moreover, we study this regions again. As a result, we take a threshold -0.1 in degree of specificity. This value and under cover 90[%] of known specific sequecnces. In our study, high sepecific sequences are object of similarity among genes. Each transcription regulatory regions have consecutive region that be 30[bp] long. we called this consecutive region window. Degree of smilarity is calculated specific sequences in these windows. and similar score is greatest it . With the use of our method, we studied some genes that are known for gene dependency. So, those genes will be contained with specific areas that have high similar score among their sequences and high expression intensity on graph. In a gene phoP, similar score of gene cluster that be affected by phoP is more higher than average of all gene score. But, those windows with high score are binding sites of sigma factor. In a group ccpA, similarity score look like rundom values. But, sum of these score is higher than average of all gene score. In this case, a score of a member of group is not ramdom in compairing of a member of this group with other member. Consequently, we must consider similarity with each similar score, each score position, method for evaluation of total score, best parameters. Also,in a general way, transcription is effected by many factors. So, transcription has immense complexity. Our study offer one of way for gene dependency up to connect with gene dependency and gene structure and gene expression. We try again our study with their modifications.

References

[1] K.Matsubara compilation "Genome function",Nakayama book store,2000

[2] M.Horikoshi compilation "Gene expression",chuugai igakusya,2001

[3] Ann Vanet,Laurent Marsan,Mrie-France Sagot, "Promoter sequences and algorithmical methods for identifying them",1999 Editions scientifiques et medicales Elsevier SAS, Microbiol.150(1999),779-799

[4] T.Ogura "Analysis of Transcription regulatory region and DNA microarray data for gene dependency",JAIST master thesis,2003

[5] H.Kitano "system biology"、syuujunsya、2001

[6] N.Friedman, M.Linial,I.Nachman,D.Pe'er "Using bayesian Network to Analyze Expression Data",Journal of Computational Biology, vol.95,no.25,pp.14863-14868,1998

[7] T.Takagi M.Tomita compilation "Bioinformatics and Information biology",Nakayama book store,2000

[8] Yishai M.Fraenkel,Yael Mandel,Devorah Friedberg and Hanah Mrgalit "Identification of common motifs in unaligned DNA sequence : application to Escherichia coli Lrp regulon",CABIOS vol.11 no.4(1995) 379-387

[9] F.kunst, N.Ogasawara, and other researchers "The complete genome sequence of the Gram-positive bacterium Bacillus sutilis"、Nature vol.390 、November 1997

[10] Jaak Vilo, Alvis Brazma, Inge Jonassen, Alan Robinson, Esko Ukkonen 'Mining for putative regulatory elements in the yeast genome using gene expression data"、American Association for Artifical Intelligence、2000