行政院國家科學委員會專題研究計畫　成果報告

搜尋非人類靈長動物基因啟動子條控序列

計畫類別：個別型計畫
計畫編號：NSC93-2213-E-216-040-
執行期間：93 年 08 月 01 日至 94 年 07 月 31 日
執行單位：中華大學生物資訊學系

計畫主持人：張慧玫

報告類型：精簡報告

處理方式：本計畫涉及專利或其他智慧財產權，2 年後可公開查詢

中　華　民　國 94 年 10 月 28 日

行政院國家科學委員會專題研究計畫成果報告

搜尋非人類靈長動物基因啟動子調控序列

Mining putative promoter regulatory elements in nonhuman primates

計畫編號：NSC93-2213-E-216-040

執行期限：93 年 8 月 1 日至 94 年 7 月 31 日

主持人：張慧玫 助理教授　　中華大學生物資訊系

E-mail: wmchang@chu.edu.tw

計畫參與人員：蘇美蘭　　中華大學資訊系碩士生

中文摘要

在這篇論文中，我們從 NCBI 的 GeneBank 裡取得 222 條非人類靈長物種啟動子序列，為 59 個基因共有 33 個物種當做我們研究的資料。接著對於這些序列執行基本的序列分析，作多重序列比對，並使用 BLAST 作 homology-hit 分析。為探究啟動子上的推測性調控因子，我們使用兩種方法:以 PCMC 方法找出靈長類等區位高出現率短序列，和以多重序列比對次相似區結合 TRANSFAC 資料庫，來分析人類和靈長物種啟動子推測性調控序列，得到的結果運用我們發展出的圖形化介面和 sequence logos 呈現出來。

PCMC 的統計高出現率區的結果，與以多重序列比對經我們發展出的圖形化介面顯現相似區的結果，大致相合。以多重序列比對次相似區結合 TRANSFAC 資料庫方法，找出推測性調控序列中，CREB, SRF, IRF-2, c-Myb 靈長序列偏好，得到從 JASPAR 脊椎動物轉錄序列資料庫類似序列偏好之印證。而本論文也找出其他資料庫未有的新的推測性靈長調控序列，其序列偏好以統計式的 sequence logos，供未來分子生物實驗作功能測試。

關鍵詞: 靈長物種啟動子、多重序列比對、圖形化介面、序列統計代表圖示

Abstract

We retrieved the 222 nonhuman primate promoter sequences of 59 genes from the GeneBank of NCBI with 33 species as the starting dataset for our study. We performed general analysis of these promoter sequences: multiple sequence alignment (MSA) and homology-hit method using BLAST (basic local alignment search tool). Then, to determine the putative regulatory elements, we analyze primate promoter sequences useing two different methods by PCMC program for primate-specific, position corresponding, highly-representative short sequences and by the less-conserved regions first obtained through MSA, then extracting putative elements from TRANSFAC database. Further these results were presented visualization tool developed by us by and by statistical sequence logos.

PCMC program for primate-specific, position corresponding, highly-representative

short sequences is helpful for analyzing putative primate regulatory elements from highly similar promoters. The results of PCMC program and those of our visualizing tool showing MSA trends are consistent. The results of the less-conserved regions through MSA plus TRANSFAC screening show that the primate consensus sequence preference by us for CREB, SRF, IRF-2, and c-Myb elements is consistent with those found in vertebrate transcriptional JASPAR database. De novel primate consensus sequence preference for putative regulatory elements uniquely found by the present study are also obtained and listed in sequence logos for future molecular functional analysis.

Keywords: primate prommoter, multiple sequence alignment, graphic interface, statistic sequence logos

Background:

Transcription regulatory elements (also named as cis-acting elements or promoter elements) are sequences within promoter regions of a gene that bind to transcription factors so that the gene can be turned on under the regulatory condition of the transcription factors and elements. Promoters are generally accepted as a TATAAA motif located at 30 nucleotides upstream (called position -30) relative to the transcription start site and a G-C enriched region downstream of the transcription start site. Computational methods for detecting transcription start sites include TATA box motif detection, hidden Markov models, and neural networks. Correlation of experimental measurements and theoretical prediction of promoters for human genome is unsatisfactory due to poor sensitivity, high false positives, and poor positional accuracy. TRANSFAC contains eukaryotic *cis*-acting regulatory DNA elements and *trans*-acting factors from yeast to human. It consists of six cross-linked tables: SITE, CELL, FACTOR, CLASS, MATRIX, and GENE. Therefore, such TRANSFAC-associated prediction programs as MatInspector, which uses a library of matrices selected from the TRANSFAC MATRIX table. Other software as AliBaba2, Match 1.0, and TFBLAST also have been developed from TRANSFAC. These public prediction programs *in silico* for transcription elements binding sites. In the present study, first we collected 222 nonhuman primate promoter sequences of 59 genes from NCBI by aligning with those of human and rodents. The sequence analysis is preformed by using multiple sequence alignments and bioinformatics tools.

Experimental Results

General layout of BLAST hit

Our starting dataset was 222 nonhuman primate promoter sequences of 59 genes from NCBI in 30 species from 19 genuses. To get a general evolutionary trend, we used BLAST program of default settings to find the similarity hits of promoter sequences against non-repeated nucleotide database of NCBI. As shown in Figure 1, the X-axis is a sliding scale of stringency or –log E-values. Most genes but not VHL gene were shown higher hit numbers in the lower –log E-values, indicating that the higher possibility of the potential orthologs are detected. Great numbers of orthologs are

found in medium stringency with –log E-values of 40 through 100 for major histocompatibility complex (MHC) class I genes (dotted lines).

Results of Multiple Sequence Alignments

To investigate the phylogenic relationship, we performed multiple sequence alignment (MSA) using ClustalW on workbench website. Results from MSA were districted into two different regions: less-conserved (of less than 100% similarity) and conserved region (of 100% similarity). The score matrix of pairwise distance using ClustalW can also show similarity between two sequences (by *p*-distance). The results of score matrix revealed that 39 out of 47 genes have higher similarity range from 80% to 100%. Only 8 genes have lower score from 3% to 74%, including APP, BMP2, CCR5, GPHA, IL-4, MID1, TNFA, and VHL, indicating these genes might be fast-evolving genes.

Detection of Putative Transcription Regulatory Elements

To detect the putative transcription regulatory elements, we use two different methods.

Results of Promoter Cluster Motif Classification (PCMC)

The highly-presented short sequences were searched using extraction tool—PCMC to retrieve the common sequence fragments of 5 to 20 bps in length shared among different sequence entries with each overall input length limited within 30 kb. To show changes of position of a shared characteristic sequence among entries, we observed degrees of position difference in elements of related promoter sequences by a graphic user interface as shown in Figure 5. This is particularly useful for promoter genes with lower-similarity (28-55%) such as APP, GPHA, TNFA, and VHL, containing less shared characteristic sequences reveal exiguously non-aligned graphs.

Results of Aligned TRANSFAC Database

The second analysis method is to screen for known transcription factor database (TRANSFAC) in Signal Scan website. Results from cross-species multiple alignment using MSA were searched for the putative binding sites in both conserved and less-conserved regions and the numbers of sites were counted. As expected, most genes have putative binding sites found within the two regions.

Gene-specific

To show changes of position, a visualization tool was developed by us. In Figure 2, similar allocations of the related putative regulatory elements were showed in AFP, nerve growth factor, and APP promoters. Great changes in sequence and position were observed in denoted promoters of GPHA, TNFA genes. Species differences were found in the Brain-2 / N-Oct 3 and VHL promoter sequences. In Brain-2 / N-Oct 3 gene, the distances between GC-rich sequences for SP1 or CP1 are farther away in Pongo organism than in two other Great Apes

(Pan and Gorilla). In the VHL promoter sequences, it is found that higher conservation both in sequence and distance between NF-1 and SP1 elements is in Gorilla and Papio than in Pan and Macaca. This is consistent with general evolution analysis by MSA of VHL promoter sequences. It is found that putative CP1 and CTF binding sites were found frequently only in conserved regions, while putative GATA-1 and Pit-1 binding sites were only in less-conserved regions. Other putative binding sites are sporadically frequent in both conserved and less-conserved regions.

Analysis of Consensus Transcription Regulatory Elements

Related regulatory elements were collected from the less-conserved regions to observe the variations among elements (Figure 3). Graphical view of sequence logos according to position frequency was used for the putative regulatory elements as sequence representations. The resultant non-human primate sequence logos were compared to those of human and rodents retrieved from JASPAR database as shown in Figure 4. High frequency for particular nucleotide distinct out at each position was observed for CREB, SRF, IRF-2, c-Myb elements and consensus binding sequences were extracted. In contrast, no particular nucleotide distinct out at each position was observed in AP-2, Sp1, and GATA-1 elements, suggesting that these elements could contain multiple sites widely bound on relatively less-strained promoter sequence.

Species-specific

The consensus binding sequences of TATA and CAAT box of promoters from experiment-supported, primate-specific from NCBI, vertebrate-specific by JASPER, and vertebrate-specific ones collected by us were compared and run through sequence logo analysis. TATA box consensus sequences (TTCTAAA) were obtained shown in Figure 4. It is worth of knowing that few sequences were left from primate promoters after trimmed with human est (expressed sequence tag) clones. This suggests that the usages of transcription starts between human and other primates might be dramatically different.

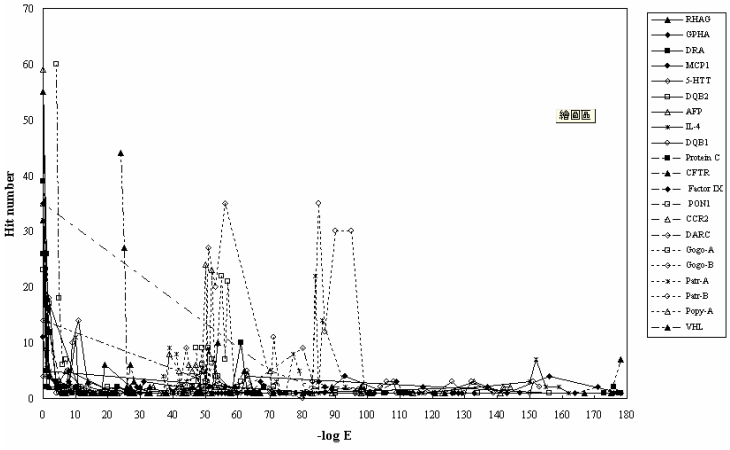


Figure 1 The relation of hit numbers and E-values (-log E-value scale) used BLAST program.

Organism	CAAT 序列	TATA 序列
Gorilla gorilla	GACCAAT	TTCTAAA
	GCCAAT	TCTAAA
		CATAAA
Pan troglodytes	GCCAAT	TTCTAAA
		CATAAA

Pongo pygmaeus	GCCAAT	TTCTAAA
Macaca mulatta	CCAAT	TATATAA

Table 1 The species preference of identifiable CAAT box and TATA box annotated by NCBI of primate

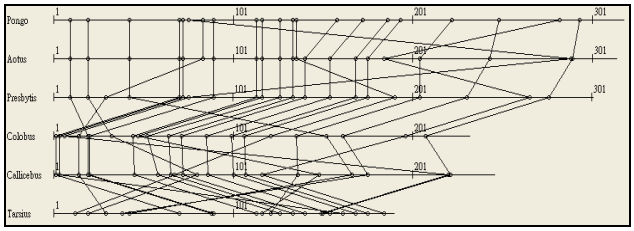


Figure 2 Diagram format by a visualization tool for putative conserved regulatory

Clustal Distance Matrix			
(1)	(2)	(3)	(4)
(1) 0.000			
(2) 0.000			
(3) 0.003	0.003		
(4) 0.011	0.011	0.017	
(5) 0.273	0.273	0.268	0.254

(1) Pongo pygmaeus

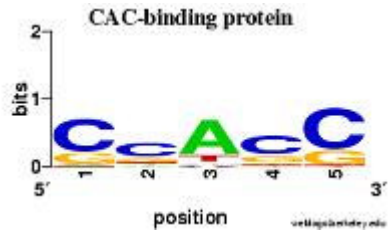
(2) Pan troglodytes

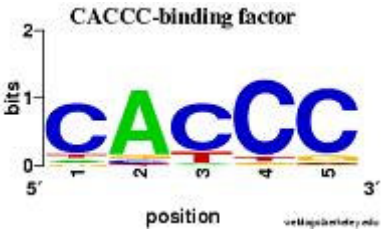
(3) Gorilla gorilla

(4) Homo sapiens

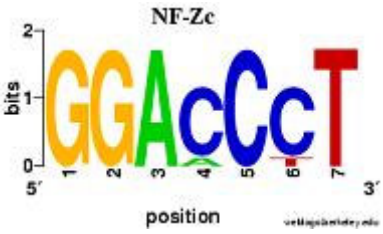
(5) Mus musculus

Nucleotide proportional change





Inverted sequence variation



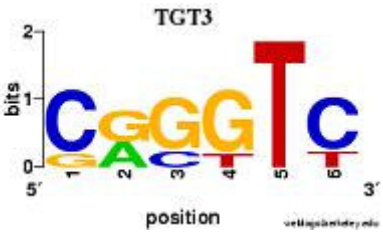
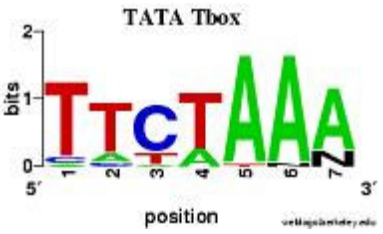
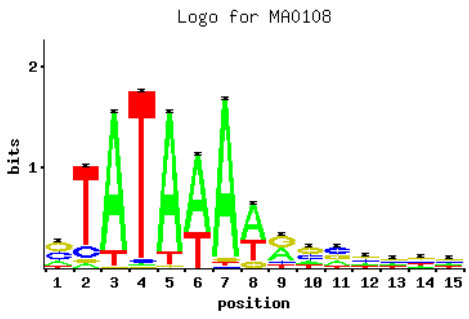


Figure 3 Sequence representation and position frequency of consensus promoter elements in non-human





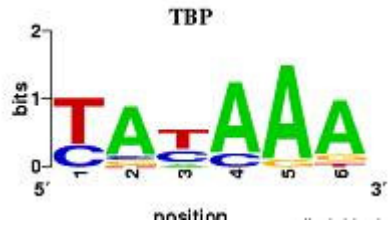


Figure 4 Comparison of TATA box from NCBI, JASPAR database , and our study.

Conclusion

It is worth of knowing that few sequences were left from primate promoters after trimmed with human est (expressed sequence tag) clones. This suggests that the usages of transcription starts between human and other primates might be dramatically different. An average of a human gene encoding 450 amino acids would span a genomic region of 27 kb and 43% of the human genome is repetitive elements suggesting that human genes are not spaced efficiently. We do understand the possibility of our primate dataset might contain some primate-specific repetitive elements undetected by our analysis. Our detection power is partially diluted since we use TRANSFAC vertebrate standards for cross-examining the highly representative short sequences after the PCMC analysis. Many highly representative short sequences after the PCMC analysis might be functional only in primates. Promoter reporter analysis aimed at putative regulatory elements and primate expression microarrays looking for the co-regulation promoter elements might all be helpful.

計畫成果自評：

感謝國科會支持本研究，使得國內生物資訊研究開展起步漸而邁向將來的豐富成熟，雖然本研究距離成果拓展至實際應用仍嫌太早，但對於生物資訊探勘研究，本年度計劃協助了一位碩士班研究生完成了一篇論文，其部分整理後，以英文版已發表在國際蛋白體研討會，並期累積更多結果整理成國際期刊。

Reference:

Audic, S. and Claverie, J. M. (1997) Detection of eukaryotic promoters using Markov transition matrices. *Comput. Chem.*, 21: 223-227.

Crooks GE, Hon G, Chandonia JM, Brenner SE. (2004) WebLogo: A sequence logo generator. *Genome Res.*, 14: 1188-1190.

Dermitzakis, ET; Clark, AG. (2002) Evolution of Transcription Factor Binding Sites in Mammalian Gene Regulatory Regions: Conservation and Turnover. *Mol. Biol. Evol.*, 9: 1114-1121.

Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, 262: 208-214.

Prestridge, D.S. (1991) SIGNAL SCAN: A computer program that scans DNA sequences for eukaryotic transcriptional elements. *Comput. Appl. Biosci.*, 7(2): 203-206.

Quandt, K., Frech, K., Karas, H., Wingender, E., Werner, T. (1995) MatInd and MatInspector: new fast and versatile

tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.* 23: 4878-4884.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, WW and Lenhard, B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 32(1): D91-4.

Schneider TD, Stephens RM. (1990) Sequence Logos: A New Way to Display Consensus Sequences. *Nucleic Acids Res.*, 18: 6097-6100.

Stormo GD. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, 16:16-23.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, 22: 4673-4680.