行政院國家科學委員會專題研究計畫　成果報告

基因編碼使用與基因表達模式的關聯--水稻、阿拉伯芥、玉米之比較研究

計畫類別：個別型計畫
計畫編號：NSC92-2313-B-216-002-
執行期間：92 年 08 月 01 日至 93 年 07 月 31 日
執行單位：中華大學生物資訊學系

計畫主持人：趙雅婷

計畫參與人員：邢禹依

報告類型：精簡報告

處理方式：本計畫可公開查詢

中　華　民　國 93 年 11 月 2 日

一、中文摘要

在許多物種的基因組中都可以觀察到同義編碼的使用並非是隨機的，而會偏向使用某些特定的編碼。先前的研究顯示高等植物的核基因在同義編碼的使用上偏向使用第三位置核酸為 C 或 G 的編碼。過去由於缺乏廣域的基因表達的數量化資料，無法進行基因表達與基因編碼偏差的統計相關分析。本研究使用水稻、玉米與阿拉伯芥的 SAGE 資料，分析基因表達量與同義編碼使用之關連。除了現有的編碼偏差測量方法，我們並使用編碼第三位置的 GC 含量作為評量指數。對於水稻與阿拉伯芥而言，所有的 SAGE 基因庫皆得到基因表達量與編碼偏差之間具有顯著的線性相關的結果，而第三位置編碼偏差亦與基因表達量呈顯著的直線相關。不論是水稻、玉米或 阿拉伯芥，基因編碼偏差與編碼區域長度皆有顯著的線性相關關係，顯示編碼區域較短的基因在基因編碼所受到選擇壓力較小。
關鍵詞：同義基因編碼偏差、基因表達系列分析、水稻、玉米、阿拉伯芥、生物資訊、編碼適應指數、有效編碼數、第三位置 GC 含量

二、英文摘要

Nonrandom usage of synonymous codons has been observed in genomes of many species, where synonymous codon usage appears an overall ‘bias’ towards a specific subset of codons. In many cases this codon bias reflects the genome composition bias and is probably resulted either from mutational biases, or from natural selection, or both. Early studies have reported the trends in synonumous codon usage in higher plant genes. However, lack of the global quantitative expression data, the association of gene expression patterns and codon usage bias remain obscure. The application of serial analysis of gene expression (SAGE) technique has allowed for the quantitative analysis of a large number of transcripts. As a consequence, we were able to study the pattern of codon usage from the perspective of transcription. We use SAGE data sets to investigate the gene expression levels of large sets of genes available for rice, Arabidopsis and maize. Spearman’s rank correlation coefficients were calculated for the relationship of gene expression level to the measures of synonymous codon usage bias such as the codon adaptation index (CAI), effective number of codons (Nc), and the GC content of the third codon position (gc3). In both rice and Arabiodpsis, we observed a significant correlation between codon usage and gene expression levels. Besides, a negative correlation for expression levels vs. the coding sequence length and a negative correlation between coding sequence length and the codon usage bias are observed.

Keywords: synonymous codon usage bias, serial analysis of gene expression (SAGE), rice, maize, Arabidopsis, bioinformatics, codon adaptation index, effective number of codons, gc3 content.

三、報告內容

(一)前言與文獻探討

包括高等植物在內的許多物種，其基因編碼之中同義編碼的出現並非是完全隨機的，各種同義編碼的使用頻率會隨基因而異，例如阿拉伯芥的 photosynthetic genes 與 housekeeping genes 的編碼比較偏好 GC，而在一些特定的組織與逆境(如脫水、低溫)之下表達的基因則有微弱的 AT 編碼偏好 (Chiapello et al., 1998) 。Fennoy and Bailey-Serres(1993)分析玉米 101 個基因發現，基因編碼的偏差是由第三位置核酸之 GC 出現頻率的差異而來的：light-regulated chloroplast proteins, ABA-induced proteins, histones, and anthocyanin biosynthetic enzymes 等等的基因編碼偏好使用 GC，而 stirage proteins 與 regulatory proteins 如 transposases, kinases, phosphatases 和 transcription factors 等的基因編碼則較為隨機。基因編碼使用的偏差反映了整個基因組的核酸組成的偏差，其可能是由突變的偏差、自然選擇、或是這些因子所共同造成的現象。先前對於高等植物之基因編碼的相關研究僅限於探討各種功能基因的編碼偏差的趨勢(如 Fennoy and Bailey-Serres, 1993; Morton, 1998; Chiapello et al., 1998 等等)，少有對基因表達量與基因編碼偏差進行統計的相關分析。Duret 與 Mouchiroud (1999)利用 EST 序列資料進行分析，在線蟲和果蠅都很明顯的觀察到基因編碼偏差較大的基因其表達量較高的現象，而在雙子葉植物的 Arabidopsis 這種相關卻不明顯。另一方面，單子葉植物如水稻、玉米其基因表達與基因編碼偏差的關聯仍有待解明。近年來新發展的基因表達系列分析(serial analysis of gene expression, SAGE；Velculescu et al., 1995)技術，可以在單一實驗中同時調查細胞之中數千種 mRNA 的濃度，並且得到量化的結果，為基因表達模式的研究提供極為有用的資料。應用水稻、玉米與阿拉伯芥的 SAGE 資料使我們得以全面探討基因表達量與基因編碼偏差之關聯，並比較單子葉植物與雙子葉植物之間基因編碼偏差的趨勢，而我們的研究結果亦可望應用於其他缺乏全轉錄組資料的高等植物。

(二) 材料與研究方法

(1) 基因表達資料：我們一共分析了五個水稻 SAGE 基因庫，兩個阿拉伯芥 SAGE 基因庫與一個玉米的 SAGE 基因庫。水稻的 SAGE 基因庫中有兩個源自於種子(代號 5DAP, 10DAP)，分別具有 6029 及 9032 個獨特短標籤(unique tags), tag 總數分別為 11488，26627。根的基因庫具有 16566 個獨特短標籤，tag 總數為 47882；另外還有兩個葉子的基因庫(代號 HL, IL)，各有 5588，8722 獨特短標籤，總 tag 數分別為 11061 及 21047。阿拉伯芥與玉米的 SAGE 基因庫皆下載自 NCBI 之 Gene Expression Omnibus(GEO) (http://www.ncbi.nlm.nih.gov/geo/)。其中 GSM30396 源自於 *Arabidopsis thaliana* (Col-4) aerial tissue，具有 21426 個獨特序列，序列總數為 79754。GSM8646 源自於 *Arabidopsis thaliana* seedling roots，此基因庫當中出現至少兩次的獨特短序列

有 4399 個，序列總數為 32104。GSM23443 源自於 maize seedlings with primary roots 12-20 mm in length，此基因庫當中出現至少兩次的獨特短序列有 14850 個，序列總數為 135571。

(2) 註解 SAGE tags：由於 tag 的長度僅為 10 個鹼基，使用 Blast 比對時需要放大 E-value，並需要根據 alignment 的位置一一判斷其可能性，非常不易判讀。因此本研究不直接使用 Blast 做 tags 之註解，而以製作 SAGE map 的方式來得到 tag 與基因的對應關係。此方法利用 GeneIndex、mRNA 以及全長 CDNA 等序列，在決定序列方向之後由最靠近 3’ 端的 NlaIII sites (CATG) 擷取與其 3’ 端相臨的 10 個鹼基，連同其所由來的 Gene Index 或 CDNA 序列編號、註解等等構成 SAGE map，也就是 SAGE tags 的索引表。其中，我們以註解的全長 CDNA 序列計算編碼區域之長度。我們由全長 CDNA 資料擷取 Tag 時，具有相同 3’ 端 Tags 之序列經 BLAST 序列比對確定為同一 Cluster 者，採用最長之序列。我們的方法不僅用於水稻 SAGE 基因庫，亦使用於阿拉伯芥與玉米的基因庫。其中阿拉伯芥與玉米的註解與 NCBI 之阿拉伯芥與玉米的 SAGE map 相符合，可知以我們的程序所註解的水稻基因庫結果應是極為可靠的。

(3) 有效編碼數(Effective number of codons, Nc)：將氨基酸依其同義編碼數目區分為 5 型，對於第 t 型的氨基酸而言，以 p_i 表示第 i 種同義編碼的使用頻率(i =1, 2, …, k，當 t=1,2,3,4,5; k 分別為 1,2,3,4,6)，n 為該種氨基酸的總使用次數，其同質性之計算為： $\hat{F}_t = (n\sum_{i=1}^{k} p_i^2 - 1)/(n-1)$

對每一型的每一種氨基酸皆以上式計算其 $\hat{F}$ 值，以 $\overline{\hat{F}_t}$ 表示第 t 型氨基酸的同質性的估計值的平均，則有效編碼數(Nc)之估計為：$\hat{N}_c = 2 + 9/\overline{\hat{F}_2} + 1/\overline{\hat{F}_3} + 5/\overline{\hat{F}_4} + 3/\overline{\hat{F}_6}$

(4)編碼適應指數(Codon Adaptation Index, CAI)

第 a 個氨基酸的第 i 種同義編碼的相對同義編碼使用值(Relative Synonymous Codon Usage, RSCU)以下列公式計算：

$RSCU_{ai} = \dfrac{x_{ai}}{\frac{1}{n_a}\sum_{i=1}^{n_a} x_{ai}}$，其中 x_{ai} 為第 a 個氨基酸的第 i 種同義編碼的使用次數觀測值，n_a 為第 a 個氨基酸的同義編碼數，式子中，分母的部分就是在所有的同義編碼被使用的機會相等的假設之下，第 i 種同義編碼的使用次數的期望值。

第 a 個氨基酸的同義編碼中最常被使用的編碼的 RSCU 值以 $RSCU_{a\max}$ 表示，定義第 a 個氨基酸的第 i 種同義編碼的相對適應性(w_{ai})為： $w_{ai} = RSCU_{ai} / RSCU_{a\max}$

RSCU 通常由高表達量的基因計算而得。對任一基因而言，其編碼適應指數(Codon Adaptation Index, CAI)即為其所有編碼的 RSCU 值的幾何平均除以同樣氨基酸組成之下的 CAI 的最大值，亦即 $CAI = CAI_{obs} / CAI_{\max}$。

其中，$CAI_{obs} = (\prod_{l=1}^{L} RSCU_l)^{1/L}$ ，$CAI_{\max} = (\prod_{l=1}^{L} RSCU_{l\max})^{1/L}$，$L$ 為該基因的氨基酸個

數。

(5) 編碼第三位置 GC 含量：本研究除了計算上述常用編碼偏差評量之外，並計算編碼三位置的 GC 使用率，作為編碼偏差之評估方法。

(三)結果與討論

考慮到 SAGE tag 可能存在的定序誤差，計算時首先將各基因庫當中僅出現一次的獨特短序列去除。除去各基因庫當中無法定位的 tag 之後，各基因庫當中編碼序列長度、有效編碼數(Nc)、編碼適應指數(CAI)與編碼第三位置 GC 含量(gc3)之分佈的統計量如 table 1 與 table 2 所示。Nc 的值越小，CAI 的值越高，或是 gc3 的值越高皆表示編碼偏差的情形較為明顯。一般而言水稻的 Nc 值較阿拉伯芥與玉米的小，gc3 的值較阿拉伯芥的 gc3 還要高。在此我們使用 Spearman 等級相關分析編碼偏差、編碼區域之長度、編碼第三位置 GC 含量與基因表達量的關聯。雖然我們去除了僅出現一次的獨特短序列以及無法定位的獨特短序列資料，這樣作並不會影響各基因表達量、編碼區域長度和編碼偏差評量指數之相對排序結果。分析水稻的 SAGE 基因庫資料顯示，編碼區域長度與基因表達量為負相關，其中有四個基因庫的 r_s 皆為-0.13 (p<0.01)，僅有 5DAP 基因庫的 r_s 皆為-0.08(p<0.01)。分析阿拉伯芥兩個基因庫資料亦得到類似的結果：r_s=-0.22, p<0.01 for GSM8646；r_s=-0.15, p<0.01 for GSM30396。然而玉米的分析結果卻顯示編碼區域長度與基因表達量之間存在著正相關：r_s=0.19, p<0.01。

分析水稻、阿拉伯芥、玉米之基因編碼偏差、編碼第三位置 GC 含量與基因表達量的關聯如 Table 3 與 Table 5 所示。對於水稻與阿拉伯芥而言，儘管在各基因庫當中表現的基因不全然相同，所有的基因庫皆得到基因表達量與編碼偏差之間具有顯著的線性相關(Table 3 與 Table 5)的結果。此外第三位置編碼偏差亦與基因表達量呈顯著的直線相關。唯一的玉米基因庫則未觀察到基因編碼偏差與基因表達量的線性相關。

不論是水稻、玉米或 阿拉伯芥，基因編碼偏差與編碼區域長度都有顯著的線性相關關係(Table 4 與 Table 6)，也就是說具有較長的編碼區域的基因其編碼偏差的情形較不明顯，此結果顯示編碼區域較短的基因在基因編碼所受到選擇壓力似乎較小。

水稻與阿拉伯芥皆具有全長 cDNA 序列資料庫以及完整的基因體序列，對於 SAGE tag 的定位以及編碼偏差的計算皆可得到較準確的結果。玉米雖然具有豐富的 EST 序列，但是卻缺乏全長 cDNA 資料庫和基因體序列，影響了可以正確定位的 tag 數目，同時也連帶影響編碼偏差的計算，這可能是造成本研究中無法觀察到編碼偏差與基因表達的相關趨勢的原因。經由本研究分析得知，高等植物的單子葉植物如水稻，雙子葉植物如阿拉伯芥皆有基因編碼偏差較大的基因其表達量較高的傾向。此外本研究所選用的編碼第三位置的 GC 含量亦與基因表達量有顯著的負相關。編碼第三位置的 GC 含量之計算較 Nc 與 CAI 的計算來得簡易許多，因此可以快速運用於缺乏全轉錄體資料的其他高等植物。

Table 1. Statistics of the codon bias, proteins length for five rice gene expression data sets

	Rice SAGE Libraries				
Variables	5DAP	10DAP	root	HL	IL
Nc	45.88(10.34)[a]	46.29(10.24)	45.86(10.45)	44.68(10.78)	45.13(10.63)
CAI	0.72(0.11)	0.72(0.11)	0.72(0.10)	0.73(0.11)	0.73(0.11)
gc3	0.64(0.19)	0.64(0.20)	0.66(0.20)	0.67(0.20)	0.67(0.20)
protein length	294.08(202.97)	308.24(208.92)	316(193.99)	298.16(191.23)	296.69(195.30)

[a] : the values in parentheses show standard deviations.

Table 2. Statistics of the codon bias, proteins length for A. thaliana and maize SAGE data sets

	SAGE Libraries		
Variables	GSM30396 (A. thaliana)	GSM8646 (A. thaliana)	GSM23443 (maize)
Nc	52.33(4.44) [a]	51.83(4.96)	52.55(6.35)
CAI	0.76(0.03)	0.77(0.03)	0.56(0.07)
gc3	0.44(0.07)	0.45(0.07)	0.80(0.10)
protein length	380.77(214.05)	352.11(211.95)	915.55(695.76)

[a] : the values in parentheses show standard deviations.

Table 3. Spearman rank correlation coefficients (r_s) between the gene expression levels and three different codon bias estimators calculated for five rice gene expression data sets

codon bias	Rice SAGE Libraries				
estimators	5DAP	10DAP	root	HL	IL
Nc	-0.20**	-0.21**	-0.20**	-0.28**	-0.27**
CAI	0.23**	0.23**	0.25**	0.30**	0.28**
gc3	0.19**	0.20**	0.19**	0.28**	0.27**

**: significant at the 1% level

Table 4. Spearman rank correlation coefficients (r_s) between the protein length and three different codon bias estimators for five rice gene expression data sets

codon bias estimators	Rice SAGE Libraries				
	5DAP	10DAP	root	HL	IL
Nc	0.29**	0.31**	0.21**	0.29**	0.27**
CAI	-0.17**	-0.22**	-0.14**	-0.14**	-0.14**
gc3	-0.19**	-0.24**	-0.15**	-0.15**	-0.16**

**: significant at the 1% level

Table 5. Spearman rank correlation coefficients (r_s) between the gene expression levels and three different codon bias estimators calculated for A. thaliana and maize SAGE data sets

codon bias estimators	SAGE Libraries		
	GSM30396 (A. thaliana)	GSM8646 (A. thaliana)	GSM23443 (maize)
Nc	-0.15**	-0.23**	-0.05
CAI	0.11**	0.25**	-0.019
gc3	0.26**	0.28**	0.05

**: significant at the 1% level

Table 6. Spearman rank correlation coefficients (r_s) between the protein length and three different codon bias estimators for A. thaliana and maize SAGE data sets

codon bias estimators	SAGE Libraries		
	GSM30396 (A. thaliana)	GSM8646 (A. thaliana)	GSM23443 (maize)
Nc	0.11**	0.16**	0.13**
CAI	-0.07**	-0.10**	-0.26**
gc3	-0.27**	-0.33**	-0.31**

**: significant at the 1% level

四、計畫成果自評

本研究的 SAGE 定位方法應用在阿拉伯芥與玉米的註解，其結果與 NCBI 的 SAGE map 資料相符合，可知本研究所註解的水稻基因庫結果應是極為可靠的。藉由 SAGE 的全轉錄體數量化基因表達資料，我們以水稻、玉米和阿拉伯芥為對象進行基因表達與編碼偏差關聯的研究分析，得以對高等植物的基因編碼偏差與基因表達、編碼序列長度的關連

有全面的了解。目前研究成果已整理投稿當中。

五、參考文獻

Chiapello, H., Lisacek, F., Caboche, M. & Henaut, A. (1998) Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. Gene 209, GC1-GC38.

Duret, L. & Mouchiroud, D. (1999) Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis. Proc. Natl. Acad. Sci. U S A 96, 4482-4487.

Fennoy, S.L. & Bailey-Serres, J. (1993) Synonymous codon usage in *Zea mays* L. nuclear genes is varied by levels of C and G-ending codons. Nucleic Acids Res. 21, 5294-5300.

Morton, B.R. (1998) Selection on the Codon Bias of Chloroplast and Cyanelle Genes in Different Plant and Algal Lineages. J. Mol. Evol. 46, 449–459.

Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial analysis of gene expression. *Science* 270: 484-487.

Wright, F. (1990) The 'effective number of codons' used in a gene. Gene 87, 23-29.