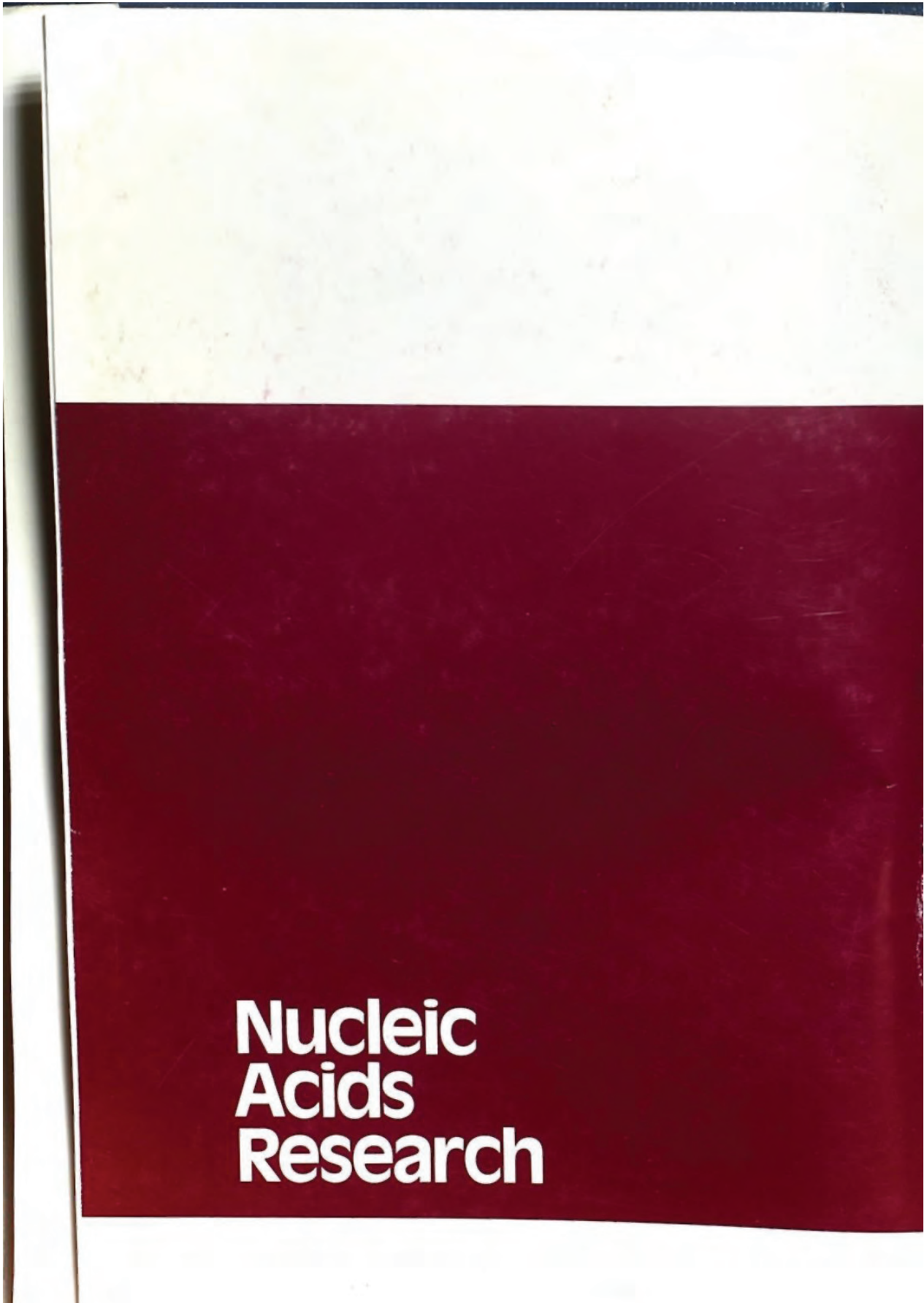




Nucleic Acids Research

Nucleic Acids Research

Volume 17 number 2 January 26, 1989



FEB 0 9 89

IRL PRESS
—at—
OXFORD UNIVERSITY PRESS
Oxford New York Tokyo

ISSN 0305 1048
Codon NARHAD

...t announcement
s...

Nucleic Acids Research

submit your paper for *Nucleic Acids Research* on floppy disc. It will
set form. No special coding of your text and no change in the
citation are involved. Our conversion system will automatically
disc into high-quality typesetting.
al originated in

and version 2 Macintosh

ompatibles), NEC or Apple Macintosh

papers for direct camera-ready
uously increasing the flexibility of the
as possible can benefit from improved

an information pack, which gives full
ion procedure.

Department:
n, Oxford OX8 1JJ, UK

n VA 22101-0850, USA

Contents

Nucleic Acids Research
Volume 17 number 2 January 25, 1989

SURVEY AND SUMMARY

- Codon usage in plant genes 477
E.E.Murray, J.Lotzer and M.Eberle

MOLECULAR BIOLOGY

- Sequence requirement for specific interaction of an enhancer binding protein
(EBP1) with DNA 499
L.Clark and R.T.Hay

- The PreA₆₉₅ precursor protein of Alzheimer's disease A4 amyloid is encoded
by 16 exons 517
H.G.Lemaire, J.M.Salbaum, G.Multhaup, J.Kang, R.M.Bayney, A.Unterbeck,
K.Beyreuther and B.Müller-Hill

- DNA bending and binding factors of the human β -actin promoter 523
T.Kawamoto, K.Makino, S.Orita, A.Nakata and T.Kakunaga

- Detection of specific DNA sequences using antibodies recognizing UV-labelled
DNA 539
V.Lund, B.H.Lindqvist and G.Eggset

- Replication timing of 10 developmentally regulated genes in *Physarum*
polycephalum 553
G.Pierron, M.Benard, E.Puvion, R.Flanagan, H.W.Sauer and D.Pallotta

- Multiple DNA repair activities for 3'-deoxyribose fragments in *Escherichia*
coli 587
C.Bernelot-Moens and B.Demple

- Regulation of the human β -actin promoter by upstream and intron domains 601
S.-Y.Ng, P.Gunning, S.-H.Liu, J.Leavitt and L.Kedes

- CMER, an RNA encoded by human cytomegalovirus is most likely
transcribed by RNA polymerase III 631
R.Marschalek, E.Amon-Böhm, J.Stoerker, S.Klages, B.Fleckenstein and
T.Dingermann

- Thymidylate synthase gene expression is stimulated by some (but not all)
introns 645
T.Deng, Y.Li and L.F.Johnson

- Excess information at bacteriophage T7 genomic promoters detected by a
random cloning technique 659
T.D.Schneider and G.D.Stormo

- Compensatory mutations demonstrate that P8 and P6 are RNA secondary
structure elements important for processing of a group I intron 675
C.L.Williamson, N.M.Desai and J.M.Burke

- Characterization of single base substitutions in edited apolipoprotein B
transcripts 691
G.E.Tennyson, C.A.Sabatos, T.L.Eggerman and H.B.Brewer Jr

Contents (continued)

Identification of a conserved sequence in the non-coding regions of many human genes L.A. Donehower, B.L. Slagle, M. Wilde, G. Darlington and J.S. Butel	699
Construction of mutant and chimeric genes using the polymerase chain reaction F. Vallette, E. Mege, A. Reiss and M. Adesnik	723
Pleiotropic derepression of developmentally regulated cellular and viral genes by c-myc proto-oncogene products in undifferentiated embryonal carcinoma cells R. Onclercq, A. Lavenue and C. Cremisi	735
The modulatory element TNNCT affects transcription of a <i>Drosophila</i> tRNA ^{Ala} gene without affecting transcription complex stability F.G. Sajjadi and G.B. Spiegelman	755
Transcriptional activation of <i>Xenopus</i> class III genes in chromatin isolated from sperm and somatic nuclei A.P. Wolffe	767
Species variation in transcription factor IIIA C.J. Gaskins, R.M. Fiser-Littell, A.L. Duke and J.S. Hanas	781
ENZYMOLOGY	
Rat mammary gland fatty acid synthase: localization of the constituent domains and two functional polyadenylation/termination signals in the cDNA M. Schweizer, K. Takabayashi, T. Laux, K.-F. Beck and R. Schreglmann	567
CHEMISTRY	
Complex structural behavior of oligopurine-oligopyrimidine sequence cloned within the supercoiled plasmid P. Parniewski, G. Galazka, A. Wilk and J. Klysik	617
A novel technique for the rapid preparation of mutant RNAs G.F. Joyce and T. Inoue	711
FOR THE RECORD	
Sequence papers	
Nucleotide sequence of two mouse histone H4 genes V.S. Meier, R. Böhni and D. Schümperli	795
Nucleotide sequence of a 25S rRNA gene from tomato T. Kiss, M. Kis and F. Solymosy	796
Complete nucleotide sequence of the hsp70 gene of <i>T. cruzi</i> J.M. Requena, M.C. Lopez, A. Jimenez-Ruiz, G. Morales and C. Alonso	797
Nucleotide sequence of rye chloroplast DNA fragment, comprising <i>psbD</i> , <i>psbC</i> and <i>trnS</i> genes A.A. Bukharov, V.L. Kolosov, O.N. Klezovich and A.S. Zolotarev	798
The complete nucleotide sequence of bovine liver cytochrome b ₅ mRNA R.J. Cristiano and A.W. Steggle	799
Nucleotide sequence of the <i>Escherichia coli</i> porin thermoregulatory gene <i>envY</i> M.D. Lundrigan, M.J. Friedrich and R.J. Kadner	800

Contents (continued)

Nucleotide sequence of the <i>deoC</i> gene of <i>Mycoplasma pneumoniae</i> S. Loechel, J.M. Inamine and P. Hu	801
Sequence of a cDNA for mouse thrombomodulin and comparison of the predicted mouse and human amino acid sequences W.A. Dittman and P.W. Majerus	802
The tRNA ^{Leu} (anticodon TTU) gene and its upstream sequence coding for a homolog of the <i>E. coli</i> large ribosome-subunit protein L14 in the <i>Tetrahymena</i> mitochondrial genome Y. Suyama and F. Jenney	803
Nucleotide sequence of cytochrome P450 L1A1 (lanosterol 14 α -demethylase) from <i>Candida albicans</i> M.H. Lai and D.R. Kirscht	804
The <i>SSA1</i> and <i>SSA2</i> genes of the yeast <i>Saccharomyces cerevisiae</i> M.R. Slater and E.A. Craig	805
Method papers	
A new vector and RNase H method for the subtractive hybridization K. Kuze, A. Shimizu and T. Honjo	807
Improved control of partial DNA restriction enzyme digest in agarose using limiting concentrations of MG ⁺⁺ H.M. Albertsen, D. Le Paslier, H. Abderrahim, J. Dausset, H. Cann and D. Cohen	808
A sequence homologous to kappa-deleting element is located 5' to the human J κ locus M.A. Kennedy, C.M. Morris and P.H. Fitzgerald	809
Preparation of single stranded insert DNA free of vector sequences J. Biernat, U.B. Göbel and H. Köster	810
Rapid carrier and prenatal diagnosis of Duchenne and Becker muscular dystrophy R.G. Roberts, C.G. Cole, K.A. Hart, M. Bobrow and D.R. Bentley	811
Cloning of synthetic oligodeoxynucleotides may result in high frequency promoter mutations in <i>E. coli</i> B.F. Bauer and W.M. Holmes	812
Characterization of the <i>ung1</i> mutation of <i>Escherichia coli</i> U. Varshney and J.H. van de Sande	813
Rapid method for preparation and cleavage of bacterial DNA for pulsed-field gel electrophoresis J.L. Flanagan, L. Ventra and A.S. Weiss	814
Direct recovery of <i>in vitro</i> transcripts in a protected form suitable for prolonged storage and shipment at ambient temperatures I. Jupin, D.E. Slat, P.A.C. Watkins and T.M.A. Wilson	815
Oncogene detection by enzymatic amplification on flow sorted chromosomes A. Schmitz, S. Olschwang, B. Chaput, G. Thomas and G. Frelat	816
Insertion of rare cutting sites nearby genes allows their rapid physical mapping: localization of the <i>E. coli</i> map locus C.L. Smith, G. Condemine, S.-Y. Chang, E. McGary and S. Chang	817

Contents (continued)

Construction of representative genomic DNA libraries using one microgram of DNA M.Lu	818
Increased efficiency of the Taq polymerase catalyzed polymerase chain reaction S.A.Krawetz, R.T.Pon and G.H.Dixon	819
RFLP REPORTS	
Four RFLPs of the human insulin receptor gene: PstI, KpnI, RsaI (2 RFLPs) N.J.Cox, R.S.Spielman, C.R.Kahn, D.Muller-Wieland, K.M.Kriauciunas and R.Taub	820
A new EcoRI polymorphism for the insulin receptor gene D.Accili, S.Elbein, C.McKeon and S.I.Taylor	821
An allelic polymorphism of the angiotensinogen gene in mice W.M.Clouston and R.J.Richards	822
Isolation and mapping of a polymorphic DNA sequence (pB3.811) on chromosome 12 [D12S33] R.Rimokh, J.P.Rouault, K.Wahbi, M.Gadoux, M.Lafage, P.Cornillet, J.Samarut, D.Germain and J.P.Magaud	823
Two polymorphic TaqI sites at the human NGFR locus (17q12 - 17q22) E.C.Wright, P.R.Fain, D.F.Barker and M.V.Chao	824
A moderately frequent HindIII polymorphism at the human NGFR locus (17q12 - 17q22) E.C.Wright, P.R.Fain, D.F.Barker and M.V.Chao	825
A rare PvuII RFLP at the CRYB1 locus (17q11.2 - 17q12) D.F.Barker, P.R.Fain, E.C.Wright, K.Nguyen and L.-C.Tsui	826
MspI RFLP at CRYB1 locus (17q11.2 - 17q12) D.F.Barker, P.R.Fain, E.C.Wright, K.Nguyen and L.-C.Tsui	827
A two allele DNA polymorphism of the human phenylethanolamine N-methyltransferase (hPNMT) gene identified by HG1A I M.R.Hoehe, W.H.Berrettini and E.E.Baetge	828
Author index	829
Corrigenda	830

Subscriptions

An annual subscription to *Nucleic Acids Research* comprises 24 issues plus one supplement, with an Annual Author and Subject Index. Subscriptions are entered on a calendar year basis only. Price includes handling and delivery, by air beyond Europe.

Annual subscription rate (Volume 17, 1989): US\$690. Not applicable to the UK and Japan—please write for separate details. Back volume prices and subscription rates for individuals are also available on request.

Orders. Orders and payments from, or on behalf of, subscribers in the various geographical areas shown below should be sent to the IRL Press office indicated.

The Americas: IRL Press, PO Box Q, McLean, VA 22101-0850, USA.

Japan: payment in Yen to sole agents (details on request), USACO Corporation, Tsutsumi Building, 13-12 Shimbashi 1-chome, Minato-ku, Tokyo 105, Japan.

Rest of the World: IRL Press, Pinkhill House, Southfield Road, Eynsham, Oxford OX8 1JJ, UK. Tel: (0865) 882283; telex: 83147 (IRL); fax: (0865) 882890.

Advertising. To advertise in *Nucleic Acids Research* contact IRL Press (US office) in the Americas or IRL Press (UK office) in the Rest of the World (see addresses above).

© IRL Press at Oxford University Press 1989. All rights reserved by the publisher.

Back volumes of this journal are available in hardcover from the publisher and in 16 mm microfilm, 35 mm microfilm and 105 microfiche from University Microfilms International, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, USA. Copies of articles published are also available from UMI.

Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, is granted by IRL Press for libraries and other users registered with the Copyright Clearance Center (CCC) Transactional Reporting Service, provided that the base fee of \$3.00 per copy is paid directly to CCC, 21 Congress St., Salem, MA 01970, USA. 0305-1048/89 \$3.00 + 0. Special requests, such as copying for general distribution, or for advertising or promotional purposes, should be addressed to the Production Editor, *Nucleic Acids Research*, IRL Press, Pinkhill House, Southfield Road, Eynsham, Oxford OX8 1JJ, UK.

Typeset by Infotype and printed by Information Printing Ltd, Oxford, UK on Permanent Paper.

New techniques to complement existing manuals

DNA Cloning Volume III a practical approach



Published
in the
Practical
Approach
series

Edited by D M Glover, Imperial College of Science and Technology, London

The techniques relied upon by molecular biologists continue to develop and find wider applications. *DNA Cloning Volume III*, like volumes I and II, will extend and enhance existing cloning manuals. Its contents are diverse. Presented together for the first time, they form a repertoire of the latest techniques for cloning and expression of DNA molecules. Each chapter is written by experts in the field. The authors describe the protocols they use – providing the minute details which make the difference between success and failure in the laboratory.

Contents

Application of plasmids containing promoters specific for phage-encoded RNA polymerases *P F R Little and I J Jackson* • Choice and use of cosmid vectors *P F R Little* • Genetic approaches to the cloning, modification and characterization of cosmid clones and clone libraries *A Poustka and H Lehrach* • The purification of eukaryotic polypeptides expressed in *E coli* *F A O Marston* • Production and purification of polyclonal antibodies to the foreign segment of β -galactosidase fusion proteins *S B Carroll and A Laughon* • Production of monoclonal antibodies against fusion proteins produced in *E coli* *S A Mole and D P Lane* • Expression and secretion of foreign genes in yeast *B L A Carter, M Irani, V L MacKay, R L Seale, A V Sledziewski and R A Smith* • The use of vectors based on gene amplification for the expression of cloned genes in mammalian cells *C R Bebbington and C C G Hentschel* • Retroviral vectors *A M C Brown and M R D Scott* • The production of transgenic mice by the microinjection of cloned DNA into fertilized one-cell eggs *D Murphy and J Hanson*

October 1987;
272pp, hardbound:
1 85221 049 4;
US\$54.00/£27.00;
softbound:
1 85221 048 6;
US\$36.00/£18.00

Save 15%

Volumes I, II and III:
Soft 1 85221 069 9;
US\$92.00/£46.00

US\$ prices valid in the Americas and Japan; £ prices definitive elsewhere. Prices subject to change without notice. Please send payment with order.

Prepaid orders to:

IRL Press Ltd PO Box 1, Eynsham, Oxford OX8 1JJ, UK
PO Box Q, McLean, VA 22101-0850, USA

IRL PRESS
at
OXFORD UNIVERSITY PRESS

Volume 17 Number 2 1989

Nucleic Acids Research

Codon usage in plant genes

Elizabeth E Murray^{1,2*}, Jeff Lotzer¹ and Mary Eberle¹

¹Agrigenetics Advanced Sciences Company, 5649 East Buckeye Road, Madison, WI 53713 and
²McArdle Laboratory for Cancer Research, University of Wisconsin, Madison, WI 53706, USA

Received October 6, 1988; Accepted December 14, 1988

Abstract:

We have examined codon bias in 207 plant gene sequences collected from Genbank and the literature. When this sample was further divided into 53 monocot and 154 dicot genes, the pattern of relative use of synonymous codons was shown to differ between these taxonomic groups, primarily in the use of G+C in the degenerate third base. Maize and soybean codon bias were examined separately and followed the monocot and dicot codon usage patterns respectively. Codon preference in ribulose 1,5 bisphosphate and chlorophyll a/b binding protein, two of the most abundant proteins in leaves was investigated. These highly expressed are more restricted in their codon usage than plant genes in general.

Introduction:

With the exception of Met and Trp, all amino acids are encoded by two to six synonymous codons. In the majority of species studied to date, an organism's use of synonymous codons is not random (1-5). However, detailed characterization of specific patterns of codon usage have been reported primarily for unicellular organisms, including *E. coli* (6-8), *Bacillus* (9,10), *Agrobacterium* (11) and yeast (12-17). The pattern of codon usage in higher eukaryotes has been examined in only a limited number of species, including *Drosophila* (19) and man (13, 18-21).

In the last three years, a large number of DNA sequences of higher plant genes have been reported, enabling us to extend the initial analyses of plant codon usage previously reported (13, 18, 22,23). We have used an expanded sample of 207 plant gene sequences to examine some general observations about codon usage.

In general, genes within a taxonomic group exhibit similarities in codon choice, regardless of the function of these genes. Thus an estimate of the overall use of the genetic code by an taxonomic group can be obtained by summing codon frequencies of all its sequenced genes. This species-specific codon choice has been called a "codon dialect" by Ikemura (13). Here we report on the "codon dialect" of 207 plant genes. We have broken this sample down into monocotyledonous and dicotyledonous plants to determine whether these broader taxonomic groups are characterized by different patterns of synonymous codon preference. Finally, we report the codon dialect of maize and soybean, since over 25 genes have been sequenced in each of these agronomically important species.

Bias in codon choice within genes in a single species appears related to the level of expression of the protein encoded by that gene (6, 7, 12-18). Codon bias is most extreme in highly expressed proteins of *E. coli* and yeast. In these organisms, a strong positive

correlation has been reported between the abundance of an isoaccepting tRNA species and the favored synonymous codon. In one group of highly expressed proteins in yeast, over 96% of the amino acids are encoded by only 25 of the 61 available codons (15). These 25 codons are preferred in all sequenced yeast genes, but the degree of preference varies with the level of expression of the genes. Recently, Hoekema and colleagues (24) report that replacement of these 25 preferred codons by minor codons in the 5' end of the highly expressed yeast gene PGK1 results in a decreased level of both protein and mRNA. They conclude that biased codon choice in highly expressed genes enhances translation and is required for maintaining mRNA stability in yeast. The degree of codon bias may be a factor to consider when engineering high expression of heterologous genes in yeast and other systems.

In plants, the ribulose 1,5 biphosphate small subunit (RuBPC SSU) and chlorophyll a/b binding protein (CAB) gene families encode two of the most abundant proteins in leaves. These genes have been sequenced in a number of different plants, and we have examined their codon usage in detail to determine whether they are more biased than other plant genes.

Data and Methods:

The 207 plant genes included in the sample are detailed in Tables 1 and 2. Sources for the data were Genbank (release 55) or original publications, referenced in Tables 1 and 2 and listed in the Appendix. Genes for which only genomic sequence was available were included, and introns were removed before codon usage data was generated. Partial sequences available for some genes were included when reading frame could be determined.

Homologous genes in different species, or multigene gene families within a species have been sequenced, including those for zein, glycinin, vicilin, CAB and RuBPC SSU. Multiple sequences of these genes were included in the sample if they differed by a minimum of 10% in the base composition of their protein coding regions. As a result, this sample may contain some bias towards codon usage in highly expressed genes.

Genbank sequences were extracted using the GENBANK program (25). Codon usage tables were compiled using the program CODONFREQUENCY from the program library of the University of Wisconsin Genetics Computer group (26).

Results:

Plant genes coding for proteins with a wide variety of functions have now been sequenced. We have tabulated the sequences of 207 plant genes (Tables 1 and 2) from 6 monocot and 36 dicot species. These proteins are present in a wide range of plant tissues at varying levels of expression. However, to date, only a few plant genes encoding rare proteins and/or mRNAs have been sequenced.

We have calculated the codon usage profile of the pooled plant sample and separate codon usage profiles for the monocotyledonous and dicotyledonous groups of plants (Table 3). This division reveals that the relative use of synonymous codons differs between the monocots and the dicots. Since the monocot sample is one third smaller than the dicot sample, we were concerned that the relative abundance of storage protein genes could skew its codon usage profile. Accordingly, we calculated codon usage in the monocots without these genes (Table 3).

Table 1
Descriptions and sources of 53 monocot genes included in the analysis.

GENUS/SPECIES	GENBANK	PROTEIN	REF
<i>Avena sativa</i>	ASTAP3R	Phytochrome 3	1
<i>Hordeum vulgare</i>	BLVALR	Aleurain	2
	BLYAMY1	α amylase 1	3
	BLYAMY2	α amylase 2	4
	BLYCHORD1	Hordein C	5
	BLYGLUCB	β glucanase	6
	BLYHORB	B1 hordein	7
	BLYPAPI	Amylase/protease inhibitor	8
	BLYTHIAR	Toxin α hordothionin	9
	BLYUBIQR	Ubiquitin	10
		Histone 3	11
		Leaf specific thionin 1	11
		Leaf specific thionin 2	12
		Plastocyanin	13
<i>Oryza sativa</i>	RICGLUTG	Glutein	14
		Glutein	15
<i>Triticum aestivum</i>	WHTAMYA	α amylase	16
	WHTCAB	CAB	17
	WHTEMR	Em protein	18
	WHTGIR	gibberellin responsive protein	19
	WHTGLGB	γ gliadin	20
	WHTGLIABA	α/β gliadin Class AII	21
	WHTGLUTI	High MW glutenin	22
	WHTH3	Histone 3	23
	WHTH4091	Histone 4	24
	WHTRBCB	RuBPC small subunit	25
<i>Secale cereale</i>	RYESECGR	γ secalin	26
<i>Zea mays</i>	MZEA1G	40.1 kD A1 protein (NADPH-dependent reductase)	27
	MZEA2CTIG	Actin	28
	MZEA2H11F	Alcohol dehydrogenase 1	28
	MZEA2H2NR	Alcohol dehydrogenase 2	29
	MZEA1D	Aldolase	30
	MZEA2T	ATP/ADP translocator	31
	MZEEG2R	Glutelin 2	32
	MZEEG2T3B	Glutathione S transferase	33
	MZEH3C2	Histone 3	34
	MZEH4C14	Histone 4	35
	MZEHSP701	70 kD Heat shock protein, exon 1	35
	MZEHSP702	70 kD Heat shock protein, exon 2	36
	MZELHCP	CAB	37
	MZEMPL3	Lipid body surface protein L3	38
	MZEPEPCR	Phosphoenolpyruvate carboxylase	39
	MZERBCS	RuBPC small subunit	40
	MZESUSYSG	Sucrose synthetase	41
	MZETP12	Triosephosphate isomerase 1	42
	MZEZE20M	19 kD zein	43
	MZEZE20M	19 kD zein	43
	MZEZE15A3	15 kD zein	44
	MZEZE16	16 kD zein	45
	MZEZE19A	19 kD zein	46
	MZEZE22A	22 kD zein	46
	MZEZE22B	22 kD zein	47
		Catalase 2	48
		Regulatory C1 locus	48

Data was obtained from GenBank (release 55) or, when no Genbank file name is specified, directly from the published source.

In general, the most important factor in discriminating between monocot and dicot patterns of codon usage is the percentage G+C content of the degenerate third base. In monocots, 16 of 18 amino acids favor G+C in this position, while dicots only favor G+C in 7 of 18 amino acids.

Table 2
Descriptions and sources of 154 genes dicot included in the analysis.

GENUS/ SPECIES	GENBANK	PROTEIN	REF
<i>Anemone majus</i>	AMACHS	Chalcone synthetase	49
<i>Arabidopsis thaliana</i>	ATHADH	Alcohol dehydrogenase	50
	ATHH3GA	Histone 3 gene 1	51
	ATHH3GB	Histone 3 gene 2	51
	ATHH4GA	Histone 4 gene 1	51
	ATHLHCP1	CAB	52
	ATHTUBA	α tubulin	53
		5-enolpyruvyl-4-hydroxy-3-phosphate 54 synthetase	
<i>Bertholletia excelsa</i>		High methionine storage protein	55
<i>Brassica campestris</i>		Acyl carrier protein	56
<i>Brassica napus</i>	BNANAP	Napin	57
<i>Brassica oleracea</i>	BOLSLSGR	S-locus specific glycoprotein	58
<i>Carica papaya</i>	CENCONA	Concanavalin A	59
<i>Chlamydomonas reinhardtii</i>	CPAPAP	Papain	60
	CREC552	Preprocytochrome	61
	CRERBCS1	RuBPC small subunit gene 1	62
	CRERBCS2	RuBPC small subunit gene 2	62
<i>Cucurbita pepo</i>	CUCPHI	Phytochrome	63
<i>Cucumis sativus</i>	CUSGMS	Glyoxosomal malate synthetase	64
	CUSLHCPA	CAB	65
	CUSSSU	RuBPC small subunit	65
<i>Daucus carota</i>	DAREXT	Extensin	66
	DAREXTR	33 kD extensin related protein	67
<i>Dolichos biflorus</i>	DBILECS	seed lectin	68
<i>Flavaria trinervia</i>	FTRBCR	RuBPC small subunit	69
<i>Glycine max</i>	SOY7SAA	7S storage protein	70
	SOYACTIG	Actin 1	27
	SOYCHPI	CPI protease inhibitor	71
	SOYGLY1A1A	Glycinin A1a Bx subunits	72
	SOYGLY1AAB	Glycinin A5A4B3 subunits	73
	SOYGLY1AB	Glycinin A3/14 subunits	74
	SOYGLY1R	Glycinin A2B14 subunits	75
	SOYHSP175	Low M W heat shock proteins	76
	SOYLGBI	Leghemoglobin	77
	SOYLEA	Lectin	78
	SOYLOX	Lipoxygenase 1	79
	SOYNOD20G	20 kDa nodulin	80
	SOYNOD23G	23 kDa nodulin	81
	SOYNOD24H	24 kDa nodulin	82
	SOYNOD26B	26 kDa nodulin	83
	SOYNOD26R	26 kDa nodulin	84
	SOYNOD27R	27 kDa nodulin	85
	SOYNOD35M	35 kDa nodulin	86
	SOYNOD75	75 kDa nodulin	86
	SOYNODR1	Nodulin C31	87
	SOYNODR2	Nodulin E27	87
	SOYPRP1	Proline rich protein	88
	SOYRUBP	RuBPC small subunit	89
	SOYURA	Urease	90
	SOYHSP26A	Heat shock protein 36A	91
		Nuclear-encoded chloroplast heat shock protein	92
		22 kDa nodulin	80
		β 1 tubulin	93
		β 2 tubulin	93
		Seed α globulin (vicilin)	94
		Seed β globulin (vicilin)	94
		RuBPC small subunit	95
		2S albumin seed storage protein	96
		Wound-induced catalase	97
		CAB	98
		RuBPC small subunit	99
		leghemoglobin I	100

GENUS/ SPECIES	GENBANK	PROTEIN	REF
<i>Lycopersicon esculentum</i>	TOMBIOBR	Biotin binding protein	101
	TOMETHYBR	Ethylene biosynthesis protein	102
	TOMPG2AR	Polygalacturonase-2a	103
	TOMPSI	Tomaio photosystem I protein	104
	TOMRBCSA	RuBPC small subunit	105
	TOMRBCSB	RuBPC small subunit	105
	TOMRBCSC	RuBPC small subunit	106
	TOMRBCSD	RuBPC small subunit	107
	TOMRRD	RuBPC small subunit	108
	TOMWIPG	Wound induced proteinase inhibitor I	109
	TOMWIPH	Wound induced proteinase inhibitor II	110
		CAB 1A	110
		CAB 1B	110
		CAB 3C	110
		CAB 4	111
		CAB 5	111
		Leghemoglobin III	112
<i>Medicago sativa</i>	ALFLB3R	RuBPC small subunit	113
<i>Mesembryanthemum crystallinum</i>		RuBPC small subunit	114
<i>Nicotiana glauca</i>	TOBATP21	Mitochondrial ATP synthase β subunit	115
		Nitrate reductase	116
		Glutamine synthetase	117
		Endochitinase	118
		A subunit of chloroplast G3PD	118
		B subunit of chloroplast G3PD	118
		C subunit of chloroplast G3PD	119
		Pathogenesis related protein 1a	119
		Pathogenesis-related protein 1c	120
		Pathogenesis related protein 1b	121
		Peroxisome	122
		RuBPC small subunit	123
		TMV-induced protein homologous to thaumatin	124
		Cellulase	125
<i>Persea americana</i>	AVOCEL	Chalcone synthase	126
<i>Petroselinum hortense</i>	PHOCHL	CAB 13	126
<i>Petunia sp.</i>	PETCAB13	CAB 22L	126
	PETCAB22R	CAB 22R	126
	PETCAB25	CAB 25	127
	PETCAB37	CAB 37	127
	PETCAB91R	CAB 91R	128
	PETCHSR	Chalcone synthase	129
	PETGCR1	Glycine-rich protein	130
	PETRBCS08	RuBPC small subunit	130
	PETRBCS11	RuBPC small subunit	131
		70 kDa heat shock protein	132
		Chitinase	133
		Phytohemagglutinin E	133
		Phytohemagglutinin L	134
		Glutamine synthetase 1	134
		Glutamine synthetase 2	135
		Leghemoglobin	136
		Lectin	137
		Phenylalanine ammonia lyase	138
		α phaseolin	139
		β phaseolin	139
		Arcein seed protein	140
		Chalcone synthase	141
		Seed albumin	142
		CAB	143
		Glutamine synthetase (nodule)	143

GENUS / SPECIES	GENBANK	PROTEIN	REF
	PEALECA	Lectin	144
	PEALEGA	Legumin	145
	PEARUBPS	RuBPC small subunit	146
	PEAVIC2	Vicilin	147
	PEAVIC4	Vicilin	147
	PEAVIC7	Vicilin	147
		Alcohol dehydrogenase 1	148
		Glutamine synthetase (leaf)	143
		Glutamine synthetase (root)	143
		Histone 1	149
		Nuclear encoded chloroplast heat shock protein	92
<i>Raphanus sativus</i>		RuBPC small subunit	150
<i>Ricinus communis</i>	RCCAGG	Agglutinin	151
	RCCRIN	Ricin	152
	RCCICL4	Isochorate lyase	153
<i>Silene pratensis</i>	SIPFDX	Ferrodoxin precursor	154
	SIPPCY	Plastocyanin precursor	155
<i>Senapis alba</i>	SALGAPDH	Nuclear gene for G3PD	156
<i>Solanum tuberosum</i>	POTPAT	Patatin	157
	POTINHVI	Wound-induced proteinase inhibitor	158
	POTLSIG	Light-inducible tissue specific ST-LS1 gene	159
	POTPI2G	Wound-induced proteinase inhibitor II	160
	POTRBCS	RuBPC small subunit	161
		Sucrose synthetase	162
<i>Spinacia oleracea</i>	SPIACP1	Acyl carrier protein I	163
	SPIOEC16	16 kDa photosynthetic oxygen-evolving protein	164
	SPIOEC23	23 kDa photosynthetic oxygen-evolving protein	165
	SPIPCG	Plastocyanin	165
	SPIPS33	33 kDa photosynthetic water oxidation complex precursor	166
		Glycolate oxidase	167
<i>Vicia faba</i>	VFALBA	Leghemoglobin	168
	VFALEB4	Legumin B	169
		Vicilin	170

Data was obtained from GenBank (release 55) or, when no Genbank file name is specified, directly from the published source.

Table 3

Codon usage in pooled sequences of higher plant genes.

AmAcid	Codon	Plants n = 207		Dicots n = 154		Monocots n = 53		Monocots No Storage Proteins n = 39	
		No.	%	No.	%	No.	%	No.	%
Gly	GGG	731	15	449	12	282	21	267	22
Gly	GGA	1629	32	1399	38	230	17	193	16
Gly	GGT	1477	29	1231	34	246	18	207	17
Gly	GGC	1179	24	596	16	583	44	543	45
Glu	GAG	2102	57	1498	51	604	75	568	79
Glu	GAA	1616	43	1419	49	197	25	154	21
Asp	GAT	1458	50	1269	58	189	27	162	24
Asp	GAC	1441	50	927	42	514	73	503	76
Val	GTG	1354	31	956	29	398	36	338	37
Val	GTA	491	11	402	12	89	8	52	6
Val	GTT	1478	34	1270	39	208	19	154	17
Val	GTC	1045	24	642	20	403	37	362	40

AmAcid	Codon	Plants		Dicots		Monocots		Monocots No Storage Proteins	
		No.	%	No.	%	No.	%	No.	%
Ala	GCG	546	11	211	6	335	22	284	24
Ala	GCA	1156	22	916	25	240	16	137	12
Ala	GCT	1901	37	1530	42	371	24	254	21
Ala	GCC	1548	30	960	27	588	38	510	43
Arg	AGG	742	26	540	25	202	26	163	25
Arg	AGA	707	24	633	30	74	9	50	8
Ser	AGT	581	13	493	14	88	8	51	6
Ser	AGC	887	20	605	18	282	26	225	27
Lys	AAG	2241	66	1600	61	641	86	609	86
Lys	AAA	1139	34	1034	39	105	14	98	14
Asn	AAT	1137	41	982	45	155	25	106	23
Asn	AAC	1646	59	1188	55	458	75	356	77
Met	ATG	1356	100	982	100	374	100	311	100
Ile	ATA	505	16	419	18	86	11	48	8
Ile	ATT	1241	40	1051	45	190	24	128	21
Ile	ATC	1374	44	873	37	501	65	433	71
Thr	ACG	343	11	184	8	159	21	146	22
Thr	ACA	745	24	636	27	109	14	73	11
Thr	ACT	990	31	842	35	148	19	116	18
Thr	ACC	1082	34	721	30	361	46	319	49
Trp	TGG	790	100	605	100	185	100	171	100
End	TGA	68	33	50	33	18	34	15	37
Cys	TGT	432	40	338	44	94	30	69	27
Cys	TGC	647	60	423	56	224	70	185	73
End	TAG	48	24	29	19	19	36	11	27
End	TAA	88	43	72	48	16	30	14	35
Tyr	TAT	743	37	630	43	113	21	68	16
Tyr	TAC	1267	63	838	57	429	79	354	84
Leu	TTG	1185	22	965	26	220	14	108	9
Leu	TTA	412	8	363	10	49	3	19	2
Phe	TTT	1047	40	887	45	160	25	95	20
Phe	TTC	1597	60	1106	55	491	75	392	80
Ser	TCG	343	8	192	6	151	14	139	17
Ser	TCA	768	17	649	19	119	11	67	8
Ser	TCT	1009	22	844	25	165	15	112	13
Ser	TCC	896	20	621	18	275	26	237	29
Arg	CGG	198	7	95	5	103	13	94	14
Arg	CGA	214	7	181	8	33	4	23	3
Arg	CGT	534	18	441	21	93	12	67	10
Arg	CGC	520	18	241	11	279	36	268	40
Gln	CAG	1465	43	787	41	678	46	457	60
Gln	CAA	1912	57	1125	59	787	54	305	40
His	CAT	575	48	465	54	110	33	85	30
His	CAC	625	52	398	46	227	67	202	70
Leu	CTG	792	15	347	9	445	28	371	33
Leu	CTA	434	8	281	8	153	9	59	5
Leu	CTT	1273	24	1032	28	241	15	151	13
Leu	CTC	1189	22	691	19	498	31	434	38
Pro	CCG	492	13	236	9	256	23	224	30
Pro	CCA	1507	39	1126	42	381	34	202	27
Pro	CCT	1063	28	874	32	189	17	118	15
Pro	CCC	755	20	469	17	286	26	212	28

n = the number of DNA sequences in the sample. No. is the number occurrences of a given codon in the sample. % is the percent occurrence for each codon within a given amino acid in the sample. (See text for description of the samples).

Table 4
Codon usage in pooled sequences of higher plant genes.

AmAcid	Codon	Soybean n = 29		Maize n = 26		CAB n = 17		RuBP SSU n = 20	
		No.	%	No.	%	No.	%	No.	%
Gly	GGG	90	16	95	16	42	8	16	9
Gly	GGA	189	33	78	13	167	32	95	51
Gly	GGT	193	33	129	21	196	37	32	17
Gly	GGC	102	18	302	50	118	23	43	23
Glu	GAG	310	51	368	81	178	71	139	74
Glu	GAA	301	49	84	19	73	29	49	26
Asp	GAT	244	62	87	24	53	29	39	33
Asp	GAC	150	38	277	76	128	71	79	67
Val	GTG	219	37	227	40	62	21	93	36
Val	GTA	77	13	36	6	24	8	7	3
Val	GTT	227	38	99	17	118	39	87	33
Val	GTC	75	12	209	37	96	32	73	28
Ala	GCG	42	8	211	24	26	5	16	5
Ala	GCA	170	30	115	13	61	12	42	14
Ala	GCT	208	37	237	27	225	45	110	38
Ala	GCC	139	25	324	36	192	38	125	43
Arg	AGG	88	22	109	26	21	15	17	12
Arg	AGA	119	30	28	7	33	24	31	21
Ser	AGT	117	18	29	5	15	5	21	8
Ser	AGC	129	20	150	28	84	27	56	22
Lys	AAG	278	58	367	90	186	85	176	85
Lys	AAA	204	42	41	10	34	15	30	15
Asn	AAT	168	40	56	19	52	30	35	26
Asn	AAC	248	60	246	81	119	70	102	74
Met	ATG	184	100	210	100	111	100	115	100
Ile	ATA	109	24	35	8	10	6	1	1
Ile	ATT	219	49	100	24	61	40	63	43
Ile	ATC	118	27	284	68	83	54	83	56
Thr	ACG	29	7	114	26	10	6	5	3
Thr	ACA	128	29	48	11	35	22	21	13
Thr	ACT	151	35	72	16	61	38	59	36
Thr	ACC	124	29	212	47	54	34	79	48
Trp	TGG	82	100	84	100	99	100	86	100
End	TGA	5	18	7	26	15	88	2	11
Cys	TGT	63	40	29	21	16	39	7	9
Cys	TGC	95	60	110	79	25	61	72	91
End	TAG	9	32	14	52	0	0	1	5
End	TAA	14	50	6	22	2	12	16	84
Tyr	TAT	135	49	38	14	23	19	17	10
Tyr	TAC	139	51	240	86	99	81	151	90
Leu	TTG	175	24	116	13	118	30	79	36
Leu	TTA	79	11	28	3	15	4	6	3
Phe	TTT	166	46	69	20	106	40	32	20
Phe	TTC	193	54	278	80	160	60	125	80
Ser	TCG	39	6	89	16	17	5	10	4
Ser	TCA	125	19	56	10	46	15	48	19
Ser	TCT	140	22	75	14	83	26	33	13
Ser	TCC	94	15	145	27	69	22	89	34
Arg	CGG	17	4	54	13	7	5	1	1
Arg	CGA	41	10	13	3	6	4	3	2
Arg	CGT	70	18	45	11	50	36	48	33
Arg	CGC	64	16	165	40	20	15	44	31
Gln	CAG	181	41	311	59	36	37	75	51
Gln	CAA	261	59	219	41	60	62	73	49
His	CAT	124	63	49	29	16	32	4	18
His	CAC	73	37	122	71	34	68	18	82

AmAcid	Codon	Soybean		Maize		CAB		RuBP SSU	
		No.	%	No.	%	No.	%	No.	%
Leu	CTG	75	10	289	31	29	7	27	12
Leu	CTA	60	8	78	9	6	2	9	4
Leu	CTT	184	26	147	16	134	34	56	25
Leu	CTC	148	21	261	28	88	23	43	20
Pro	CCG	55	8	149	27	29	10	13	6
Pro	CCA	346	47	126	23	137	47	72	34
Pro	CCT	236	32	109	20	73	25	60	29
Pro	CCC	95	13	164	30	54	18	66	31

n = the number of DNA sequences in the sample. No. is the number occurrences of a given codon in the sample. % is the percent occurrence for each codon within a given amino acid in the sample. (See text for description of the samples).

The G ending codons for Thr, Pro, Ala and Ser are avoided in both monocots and dicots because they contain C in codon position II. The CG dinucleotide is strongly avoided in plants (23) and other eukaryotes (27), possibly due to regulation involving methylation. In dicots, XCG is always the least favored codon, while in monocots this is not the case. The doublet TA is also avoided in codon positions II and III in most eukaryotes (27), and this is true of both monocots and dicots.

Grantham and colleagues (18) have developed two codon choice indices to quantify CG and TA doublet avoidance in codon positions II and III. XCG/XCC is the ratio of codons ending to T-ending triplets with T as the second base. These indices have been calculated for the plant data in this paper (Table 5) and support the conclusion that monocot and dicot species differ in their use of these dinucleotides. This pattern of synonymous codon usage is not dependent on the inclusion of storage protein genes in the monocot sample, since the pooled codon usage data for monocots without storage proteins is even less like the dicot pattern (Table 3). Not surprisingly, the pooled plant sample resembles the dicot pattern more than that of monocots, since almost three times as many dicot sequences as monocot sequences were available.

For two species, soybean and maize, larger sequence samples were available (29 and 28 genes respectively), so that species-specific codon usage profiles could be calculated (Table 4). Not surprisingly, the maize codon usage pattern resembles that of monocots in general. Since these sequences represent over half of the monocot sequences available. The codon profile of the maize subsample is even more strikingly biased in its preference for G+C in

Table 5
Avoidance of CG and UA doublets in codons position II-III.

Group	Plants	Dicots	Monocots	Monocots no storage protein	Maize	Soybean	RuBP SSU	CAB
XCG/XCC	40	30	61	62	67	37	18	22
XTA/XTT	37	35	47	34	43	41	9	13

XCG/XCC and XTA/XAA values are multiplied by 100.

Table 6
Codon frequencies in 17 plant CAB genes.

		MZW LHC P	WHIT CAB	ATH LHC PI	CUS LHC PA	LGI AB 19	LYS ¹ LUT 1A	LYS ² LUT 1B	LYS ³ LUT 3C	LYS ⁴ LUT 4	LYS ⁵ LUT 5	PET CAB 13	PET CAB 22L	PET CAB 22R	PET CAB 25	PET CAB 37	PET 91R	PEA CAB 80
Gly	GGG	5	5	2	2	12	1	3	1	2	2	0	1	2	2	0	1	1
Gly	GGA	1	0	18	9	0	12	11	8	12	11	12	11	12	12	15	10	12
Gly	GGT	1	11	8	13	0	13	16	4	13	14	16	13	14	15	15	16	13
Gly	GGC	25	15	5	7	20	5	2	3	7	4	3	6	2	3	4	2	5
Glu	GAG	15	13	12	12	15	9	12	6	7	8	11	10	11	10	8	11	8
Glu	GAA	0	0	3	3	0	7	3	2	10	8	3	5	5	4	8	5	7
Asp	GAT	0	1	3	3	0	3	2	1	3	4	5	5	5	5	4	3	6
Asp	GAC	11	11	8	8	13	7	9	4	8	7	6	5	5	6	7	8	5
Val	GTG	7	7	2	6	5	3	2	1	3	2	3	4	3	4	4	3	3
Val	GTA	0	0	1	0	0	2	2	1	1	3	4	2	2	2	1	2	2
Val	GTT	0	4	8	7	1	6	9	7	13	9	9	3	8	9	10	9	6
Val	GTC	11	6	7	4	11	7	5	4	1	3	6	9	5	5	1	6	5
Ala	GCG	10	6	1	0	5	0	1	0	0	0	1	0	0	1	0	1	0
Ala	GCA	0	2	3	2	0	6	1	7	2	4	4	2	4	4	7	6	7
Ala	GCT	2	5	13	16	0	18	19	11	17	14	17	18	20	14	13	14	14
Ala	GCC	22	18	16	9	26	9	9	8	6	4	0	10	7	13	9	10	7
Arg	AGG	0	0	4	2	0	1	1	1	1	2	1	1	0	1	4	0	2
Arg	AGA	0	0	2	1	1	3	3	2	3	1	3	3	3	2	1	3	2
Ser	AGT	0	0	1	0	0	1	0	0	0	1	1	2	3	2	1	2	1
Ser	AGC	8	7	7	6	4	6	6	5	3	3	4	4	5	4	3	4	5
Lys	AAG	13	11	12	11	12	10	10	7	9	7	11	13	12	12	11	14	11
Lys	AAA	0	2	0	2	0	2	4	1	4	3	4	2	3	2	1	1	1
Asn	AAT	0	2	1	6	0	3	5	2	2	3	6	5	4	5	4	3	1
Asn	AAC	9	8	10	5	9	7	5	6	10	6	5	5	6	6	6	7	9
Met	ATG	7	7	7	5	6	7	7	6	6	5	7	7	7	7	6	7	7
Ile	ATA	0	1	1	0	0	0	0	0	0	3	2	0	0	1	0	1	1
Ile	ATT	1	5	3	3	0	3	5	2	11	6	1	2	5	4	7	1	2
Ile	ATC	7	4	3	5	12	6	3	4	3	3	5	5	4	4	6	5	4
Thr	ACG	5	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0
Thr	ACA	0	0	5	1	0	2	3	1	3	4	2	1	3	2	4	3	1
Thr	ACT	0	2	2	4	0	2	4	1	5	4	5	7	5	6	6	4	4
Thr	ACC	4	6	3	4	4	3	5	1	1	1	4	4	3	2	1	3	5

Trp	TGG	6	6	6	6	6	6	3	0	6	6	6	6	6	6	6	6	6
End	TGA	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Cys	TGT	0	2	0	1	1	1	2	0	2	0	2	3	1	3	1	2	0
Cys	TGC	3	1	1	1	1	2	1	1	2	0	2	3	1	3	1	2	0
End	TAG	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
End	TAA	0	1	0	0	0	1	1	0	0	3	2	2	4	2	3	1	1
Tyr	TAT	0	1	0	2	0	1	0	0	0	5	5	6	4	5	5	7	7
Tyr	TAC	8	6	7	6	6	6	6	2	8	5	5	6	4	5	5	7	7
Leu	TTG	1	4	13	13	0	8	9	4	2	5	7	10	8	7	6	9	12
Leu	TTA	0	0	0	1	0	3	4	1	3	3	1	0	0	2	0	0	3
Phe	TTT	0	5	4	4	0	7	5	6	12	10	8	7	7	9	10	8	4
Phe	TTC	14	9	12	11	18	9	11	3	6	6	9	8	9	8	8	8	11
Ser	TCG	3	6	0	1	2	0	0	2	1	0	0	1	1	0	0	0	0
Ser	TCA	0	0	5	2	0	4	2	3	2	3	1	3	2	2	6	1	7
Ser	TCT	0	4	2	3	0	6	8	6	4	3	7	6	8	7	3	8	8
Ser	TCC	7	4	5	5	8	4	5	1	2	2	2	5	4	3	1	5	6
Arg	CGG	1	1	0	0	3	0	0	0	0	0	2	0	0	0	1	0	0
Arg	CGA	0	2	1	0	0	0	0	0	0	4	3	4	4	3	4	4	3
Arg	CGT	0	2	2	4	0	4	4	0	5	4	3	4	4	3	4	4	3
Arg	CGC	7	4	0	0	6	0	0	1	1	0	0	0	0	0	0	0	1
Gln	CAG	4	4	2	4	10	2	0	2	2	1	0	0	1	0	2	1	1
Gln	CAA	0	2	2	2	0	3	4	2	8	5	4	4	3	5	8	3	5
His	CAT	0	0	1	1	0	1	3	1	2	2	0	0	0	1	2	0	2
His	CAC	3	3	2	2	3	2	0	1	1	1	3	3	3	2	1	3	1
Leu	CTG	9	1	0	1	13	0	1	0	0	1	1	0	1	0	0	0	1
Leu	CTA	0	0	3	0	0	0	0	0	0	0	1	0	1	1	0	0	0
Leu	CTT	2	7	6	7	0	11	9	7	12	8	7	11	9	7	12	12	7
Leu	CTC	14	16	4	2	11	1	4	2	2	2	5	1	4	4	3	2	5
Pro	CCG	12	3	1	1	6	0	1	0	2	1	2	0	0	0	0	0	0
Pro	CCA	2	3	8	10	0	10	11	6	8	10	10	9	8	10	9	9	14
Pro	CCT	0	3	5	5	0	7	6	3	4	2	6	6	7	6	5	6	2
Pro	CCC	5	8	4	4	10	2	2	2	3	1	5	1	2	2	1	4	2

Absolute frequencies of each codon are calculated from sequences in GenBank release 55 or the literature (see Tables 1 and 2). Sequences not in GenBank release 55 are CAB genes referenced in Table 2 as follows:

- ¹ *Lycopersicon esculentum* CAB 1A
- ² *Lycopersicon esculentum* CAB 1B
- ³ *Lycopersicon esculentum* CAB 3C
- ⁴ *Lycopersicon esculentum* CAB 4
- ⁵ *Lycopersicon esculentum* CAB 5

Table 7
Codon frequencies in 20 plant RuBPC SSU genes.

		WHI	MZR	CRE	CRE	CUS	FTR	SOY	HNN	LGI	TOM	TOM	TOM	TOM	MES	RAP	TOB	PET	PET	FEA	POT
		RBC	BCS	RBC	RBC	SSU	BCR	RUB	RUB	RLI	RBC	RBC	RBC	RBC	CRY	SAT	RBP	RBC	RBC	RUB	RBC
		B	SI	S2				P	ISC	SBPC	SA	SB	SC	SD			CO	SOB	SII	PS	S
Gly	GGG	3	1	0	0	0	0	0	0	3	1	1	1	1	1	1	0	1	1	1	1
Gly	GGA	1	1	0	0	7	7	4	5	0	0	6	0	7	6	8	6	5	4	7	8
Gly	GGT	0	0	0	0	1	2	3	4	0	2	2	2	2	4	0	0	3	2	3	2
Gly	GGC	3	9	4	4	3	2	7	1	3	0	1	2	0	0	1	4	2	4	2	0
Glu	GAG	9	5	6	6	5	9	6	6	11	8	8	8	8	6	7	8	7	5	8	8
Glu	GAA	0	0	0	0	4	2	3	4	0	3	3	3	3	2	4	5	2	3	5	3
Asp	GAT	0	0	1	1	2	0	4	1	0	3	3	3	3	1	2	3	3	3	4	2
Asp	GAC	5	8	7	7	6	4	3	5	6	3	3	3	3	4	6	2	3	3	2	4
Val	GTG	5	5	10	10	4	6	4	3	5	3	4	4	3	5	3	6	5	4	5	4
Val	GTA	0	0	0	0	1	0	1	0	0	0	0	1	0	2	1	0	0	0	1	0
Val	GTT	0	0	0	0	6	3	3	5	1	8	7	5	8	4	5	7	4	5	7	9
Val	GTC	4	6	9	9	3	6	2	3	3	3	3	3	3	6	4	1	4	2	1	2
Ala	GCG	2	1	2	2	0	0	0	1	1	2	0	0	2	0	0	1	0	0	1	2
Ala	GCA	0	0	0	0	2	2	1	0	0	1	5	6	3	3	2	4	4	4	1	2
Ala	GCT	1	1	5	3	9	4	7	7	0	8	6	8	8	6	7	5	8	6	6	6
Ala	GCC	4	10	22	23	3	8	5	7	6	4	3	1	4	7	4	6	3	7	3	5
Arg	AGG	0	1	0	0	4	0	3	1	1	0	1	1	0	0	0	1	1	3	1	0
Arg	AGA	0	1	0	0	0	2	1	3	1	2	2	2	2	1	2	2	3	2	4	2
Ser	AGT	0	0	0	0	4	2	0	0	0	1	2	2	1	1	2	2	3	1	1	1
Ser	AGC	3	6	2	3	3	1	2	2	2	5	4	3	5	2	4	3	3	3	2	4
Lys	AAG	8	9	10	10	8	12	9	10	9	7	8	9	7	10	11	10	11	10	10	7
Lys	AAA	0	1	0	0	2	2	2	3	0	3	3	1	3	1	2	1	0	1	3	3
Asn	AAT	1	0	0	0	2	0	2	1	0	3	4	3	1	3	0	2	4	3	1	3
Asn	AAC	2	7	8	8	6	7	5	5	4	7	4	4	7	6	6	7	3	3	1	7
Met	ATG	3	6	9	9	6	6	7	7	3	6	6	7	6	7	5	5	5	6	6	6
Ile	ATA	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Ile	ATT	1	5	6	5	7	4	1	5	1	0	5	6	4	5	1	5	4	5	4	5
Ile	ATC	5	5	6	5	7	4	5	5	6	3	3	4	3	5	4	4	5	3	3	3
Thr	ACG	1	3	0	0	0	2	1	0	1	0	0	0	0	0	0	0	0	0	0	0
Thr	ACA	2	0	0	0	0	0	0	1	0	2	4	0	0	2	1	2	0	0	1	5
Thr	ACT	0	0	1	0	4	3	3	0	2	5	4	2	2	4	5	7	5	5	4	1
Thr	ACC	1	6	5	4	3	5	6	7	3	5	2	5	5	7	6	3	2	4	2	1
Irp	TGG	3	4	4	5	4	5	4	5	4	5	5	5	5	4	4	5	5	5	1	5
End	TGA	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0

Cys	TGT	0	0	0	0	1	1	0	2	0	0	1	0	0	1	1	4	0	0	0	0
Cys	TGC	4	5	4	4	5	1	4	1	1	4	3	4	4	4	4	4	4	4	5	4
End	TAG	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
End	TAA	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1
Tyr	TAT	0	0	0	0	0	2	1	1	1	1	1	1	1	1	1	1	1	2	0	0
Tyr	TAC	7	10	7	7	6	7	8	8	8	9	10	8	9	5	9	9	11	6	10	7
Leu	TTG	0	0	0	0	1	0	0	1	0	1	0	1	0	1	0	1	0	0	0	0
Leu	TTA	0	0	0	0	2	2	2	0	0	3	3	1	3	0	3	1	4	0	4	2
Phe	TTT	0	0	0	0	4	3	5	6	7	8	6	8	0	1	6	6	1	7	5	5
Phe	TTC	7	8	9	8	9	7	8	6	5	6	5	6	1	1	0	0	0	0	0	0
Ser	TCG	0	4	1	2	0	1	0	1	1	1	0	1	1	1	1	0	0	0	1	4
Ser	TCA	0	0	0	0	5	2	2	3	0	3	4	1	3	1	3	1	3	2	4	1
Ser	TCT	1	0	1	2	4	1	1	0	1	1	3	2	1	1	3	1	4	5	7	5
Ser	TCC	2	6	7	7	5	2	6	6	3	5	3	4	5	4	6	3	4	5	7	5
Arg	CGG	0	1	0	0	0	0	0	1	0	0	1	0	1	1	1	0	0	0	0	0
Arg	CGA	0	0	0	0	0	0	0	0	0	1	0	2	3	3	4	3	2	2	4	5
Arg	CGT	1	0	0	0	3	3	3	3	1	2	2	1	1	2	1	1	1	0	0	0
Arg	CGC	4	6	11	11	1	1	2	1	2	2	1	1	2	1	2	1	1	0	0	0
Gln	CAG	4	7	10	10	4	3	3	3	3	3	2	5	3	4	3	4	2	3	2	1
Gln	CAA	0	0	0	0	4	5	4	4	1	5	7	4	5	3	3	6	7	6	5	4
His	CAT	0	0	0	0	0	1	0	0	0	0	0	1	1	0	2	1	1	2	1	0
His	CAC	1	1	0	0	1	0	2	1	0	0	1	1	0	2	1	1	2	1	1	2
Leu	CTG	3	8	9	9	1	0	1	0	2	0	0	1	0	3	0	1	0	0	0	0
Leu	CTA	0	0	0	0	1	1	1	1	0	0	0	1	0	2	4	6	4	3	4	3
Leu	CTT	1	0	0	0	3	5	3	4	0	4	4	2	4	2	3	1	1	5	2	2
Leu	CTC	3	4	0	0	5	2	2	5	1	2	2	2	1	2	3	1	1	5	2	2
Pro	CCG	2	7	3	3	0	1	0	1	2	0	0	0	0	1	0	0	0	0	0	0
Pro	CCA	2	0	0	0	4	4	6	5	0	4	4	5	3	4	4	5	4	4	7	5
Pro	CCT	3	0	2	0	3	4	1	3	0	4	5	3	4	4	3	5	5	3	1	3
Pro	CCC	1	5	9	11	4	4	3	4	6	2	1	3	2	5	2	0	2	4	1	2

Absolute frequencies of each codon are calculated from sequences in GenBank release 55 or the literature (see Table 1 and 2). Sequences not in GenBank release 55 are RuBPC SSU genes referenced in Table 2 as follows:

¹ *Mesembryanthemum crystallinum* RuBPC SSU

² *Raphanus sativus* RuBPC SSU

codon position III. However, the soybean codon usage pattern is almost identical to the general dicot pattern, even though it represents a much smaller portion of the entire dicot sample.

In order to determine whether the coding strategy of highly expressed genes such as RuBPC SSU and CAB is more biased than that of plant genes in general, we calculated codon usage profiles for subsets of these genes (20 and 17 sequences respectively) (Table 4). The RuBPC SSU and CAB pooled samples are characterized by stronger avoidance of the codons XCG and XTA than in the larger monocot and dicot samples (Tables 4 and 5). Although most of the genes in these subsamples are dicot in origin (17/20 and 15/17), their codon profile resembles that of the monocots in that G+C is preferred in the degenerate base III.

The use of pooled data for highly expressed genes may obscure identification of species-specific patterns in codon choice. Therefore, we have tabulated the codon choices of individual genes for RuBPC SSU (Table 6) and CAB (Table 7). The preferred codons of the maize and wheat genes for RuBPC SSU and CAB are more restricted in general than are those of the dicot species. Matsuoka et al. (28) noted the extreme codon bias of the maize RUBPC SSU gene as well as two other highly expressed genes in maize leaves, CAB and phosphoenolpyruvate carboxylase. These genes almost completely avoid the use of A+T in codon position III, although this codon bias was not as pronounced in non-leaf proteins such as ADH, zein 22 kDa subunit, sucrose synthetase and ATP/ADP translocator. Since the wheat SSU and CAB genes have a similar pattern of codon preference, this may reflect a common monocot pattern for these highly expressed genes in leaves. The CAB gene for Lemna and the RuBPC SSU genes for Chlamydomonas share a similar extreme preference for G+C in codon position III.

In dicot CAB genes, however, A+T degenerate bases are preferred by some synonymous codons (e.g. GCT for Ala, CTT for Leu, GGA and GGT for Gly). In general the G+C preference in position III is less pronounced for both RuBPC SSU and CAB genes in dicots than in monocots.

Discussion:

Because of the degenerate nature of the genetic code, only part of the variation contained in a gene is expressed in its protein. It is clear that variation between degenerate base frequencies is not a neutral phenomenon since systematic codon preferences have been reported for bacterial, yeast and mammalian genes. Analysis of a large group of plant gene sequences indicates that synonymous codons are used differently by monocots and dicots. These patterns are also distinct from those reported for *E. coli*, yeast and man (13,18).

In general, the plant codon usage pattern more closely resembles that of man and other higher eukaryotes than unicellular organisms, due to the overall preference for G+C content in codon position III (13,18). Monocots in this sample share the most commonly used codon for 13 of 18 amino acids as that reported for a sample of human genes (18), although dicots favor the most commonly used human codon in only 7 of 18 amino acids.

Several earlier discussions of plant codon usage have focussed on the differences between codon choice in plant nuclear genes and in chloroplasts (18,23). Chloroplasts differ from the nuclear genome of higher plants in that they encode only 30 tRNA species (29, 30). Since

chloroplasts have restricted their tRNA genes, the use of preferred codons by chloroplast-encoded proteins appears more extreme. However, a positive correlation has been reported between the level of isoaccepting tRNA for a given amino acid and the frequency with which this codon is used in the chloroplast genome (31).

Our analysis of the expanded plant sample confirms earlier reports that the nuclear and chloroplast genomes in plants have distinct coding strategies. The codon usage of monocots in this sample is distinct from chloroplast usage, sharing the most commonly used codon for only 1 of 18 amino acids. Dicots in this sample share the most commonly used codon of chloroplasts in only 4 of 18 amino acids. In general, the chloroplast codon profile more closely resembles that of unicellular organisms, with a strong bias towards the use of A+T in the degenerate third base.

In unicellular organisms, highly expressed genes use a smaller subset of codons than do weakly expressed genes, although yeast and *E. coli* use distinct preferred codons in some cases. Sharp and Li (12) report that codon usage in 165 *E. coli* genes reveals a positive correlation between high expression and increased codon bias. Bennetzen and Hall (15) and others (12-14, 17, 18) have described a similar trend in codon selection in yeast. Codon usage in these highly expressed genes correlates with the abundance of isoaccepting tRNAs in both yeast and *E. coli*. If, as Ikemura (16) has proposed, the good fit of abundant yeast and *E. coli* mRNA codon usage to isoacceptor tRNA abundance promotes high translation levels and high steady state levels of these proteins, then the potential for high levels of expression of plant genes in yeast or *E. coli* could be limited by their codon usage. Hoekema et al. (24) report that replacement of the 25 most favored yeast codons with rare codons in the 5' end of the highly expressed gene PGK1 leads to a decrease in both mRNA and protein. These results indicate that codon bias should be considered when engineering high expression of foreign genes in yeast and other systems.

A number of researchers have attempted to express plant genes in yeast (32-34) and *E. coli* (35-37). In the case of wheat α -gliadin (32), α -amylase (33) genes, and maize zein genes (34), low levels of expression have been reported in yeast. Neill et al. (32) have suggested that the low levels of expression of α -gliadin in yeast may be due in part to codon usage bias, since α -gliadin codons for Phe, Leu, Ser, Gly, Tyr and especially Glu do not correlate well with the abundant yeast isoacceptor tRNAs. In *E. coli*, however, soybean glycinin A2 (35) and wheat RuBPC SSU (36, 37) are expressed adequately.

Not much is known about the makeup of tRNA populations in plants. Viotti et al. (38) report that maize endosperm actively synthesizing zein, a storage protein rich in glutamine, leucine, and alanine, is characterized by higher levels of accepting activity for these three amino acids than are maize embryo tRNAs. This may indicate that the tRNA population of specific plant tissues may be adapted for optimum translation of highly expressed proteins such as zein. To our knowledge, no one has experimentally altered codon bias in highly expressed plant genes to determine possible effects of the protein translation in plants to check the effects on the level of expression. Our data indicate that the highly biased RuBPC SSU and CAB genes would be good candidates for such an experiment.

Acknowledgements:

We would like to thank Sharon Locken for bibliographic assistance and Lynette Bausch, Lynda Brereton and Sarah Weske for assistance in data entry.

*To whom correspondence should be addressed

References

1. Grantham, R. (1980) Trends Biochem. Sci. 5, 327-331.
2. Grantham, R., Gautier, C. and Gouy, M. (1980) Nucl. Acids Res. 8, 1893-1912.
3. Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pavé, A. (1980) Nucl. Acids Res. 8, r49-r62.
4. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. and Mercier, R. (1981) Nucl. Acids Res. 9, r43-r74.
5. Aota, S.-I., Gojobori, T., Ishibashi, F., Maruyama, T. and Ikemura, T. (1988) Nucl. Acids Res. 16, r315-r402.
6. Ikemura, T. (1981) J. Mol. Biol. 146, 1-21.
7. Gouy, M. and Gautier, C. (1982) Nucl. Acids Res. 10, 7055-7074.
8. Sharp, P.M. and Li, W.-H. (1986) J. Mol. Evol. 24, 28-38.
9. Ogasawara, N. (1985) Gene 40, 145-150.
10. McConnell, D.J., Catwell, B.A., Devine, K.M., Forage, A.J., Laiode, B.M., O'Kane, C., Ollington, J.F. and Sharp, P.M. (1986) Ann. NY Acad. Sci. 469, 1-17.
11. Winans, S.C., Allenza, P., Stachel, S.E., McBride, K.E. and Nester, E.W. (1987) Nucl. Acids Res. 15, 825-837.
12. Sharp, P.M. and Li, W.-H. (1986) Nucl. Acids Res. 14, 7734-7749.
13. Ikemura, T. (1985) Mol. Biol. Evol. 2, 13-34.
14. Boer, H.A., de and Kastelein, R.A. (1986) In: From Gene to Protein: Steps Dictating the Maximal Level of Gene Expression (eds. J. Davis, B. Reznikoff and L. Gold). Butterworths, New York.
15. Bennetzen, J.L. and Hall, B.D. (1982) J. Biol. Chem. 257, 3026-3031.
16. Ikemura, T. (1982) J. Mol. Biol. 158, 573-597.
17. Sharp, P.M., Touhy, T.M.F. and Mosurski, K.R. (1986) Nucl. Acids Res. 14, 5125-5143.
18. Grantham, R., Perrin, P. and Mouchiroud, D. (1986) Oxford Surveys in Evol. Biol. 3, 48-81.
19. Russell, D.W., Smith, M., Williamson, V.M. and Young, E.T. (1983) J. Biol. Chem. 258, 2674-2682.
20. Hastings, K.E.M. and Emerson, C.P. (1983) J. Mol. Evol. 19, 214-218.
21. Bernardi, G. and Bernardi, G. (1985) J. Mol. Evol. 22, 363-365.
22. Lycett, G.W., Delauney, A.J. and Croy, R.R.D. (1983) FEBS Letters 153, 43-46.
23. Boudraa, M. (1987) Genet. Sel. Evol. 19, 143-154.
24. Hoekma, A., Kastelein, R.A., Vasser, M.A. and Boer, H.A. de (1987) Mol. Cell. Biol. 7, 2914-2924.
25. Martinez, H.M., Katzung, B. and Farrah, T. (1984) Sequence Analysis Programs Manual, University of California, San Francisco.
26. Devereux, J. and Haerberli, P. (1983) Program Library of the University of Wisconsin Genetics Computer Group, Madison, Wisconsin.
27. Grantham, R., Greenland, T., Louail, S., Mouchiroud, D., Prato, J.L., Gouy, M. and Gautier, C. (1985) Bull. Inst. Pasteur. 83, 95-148.
28. Matsuka, M., Kano-Murakami, Y., Tanaka, Y., Ozeki, Y. and Yamamoto, N. (1987) J. Biochem. 102, 673-676.
29. Wakasugi, T., Ohme, M., Shinozaki, K. and Sugiura, M. (1986) Plant Mol. Biol. 7, 385-392.
30. Shinozaki, K., Ohme, M., Tanaka, M., Wakasugi, T., Hayashida, N., Matsubayashi, T., Zaita, N., Chonwongse, J., Obokata, J., Yamaguchi-Shinohara, K., Ohto, C., Torazawa, K., Meng, B.Y., Sugita, M., Deno, H., Kamogashira, T., Yamada, K., Kusuda, J., Takaiwa, F., Kato, A., Tohdoh, N., Shimada, H. and Sugiura, M. (1986) EMBO J. 5, 2043-2049.
31. Pfizinger, H., Guillemaut, P., Weil, J.-H. and Pillay, D.T.N. (1987) Nucl. Acids Res. 15, 1377-1386.
32. Neill, J.D., Litts, J.C., Anderson, O.D., Greene, F.C. and Stiles, J.I. (1987) Gene 55, 303-317.
33. Rothstein, S.J., Lahners, K.N., Lazarus, C.M., Baulcombe, D.C. and Gatenby, A.A. (1987) Gene 55, 353-356.
34. Coraggio, I., Compagno, C., Martegani, E., Ranzi, B.M., Sala, E., Alberghina, L. and Viotti, A. (1986) EMBO J. 5, 459-465.
35. Fuzakawa, C., Udaka, K., Murayama, A., Higuchi, W. and Totsuka, A. (1987) FEBS Lett. 224, 125-127.
36. Vies, S.M. van der, Bradley, D. and Gatenby, A.A. (1986) EMBO J. 5, 2439-2444.
37. Gatenby, A.A., Vies, S.M. van der, and Rothstein, S.J. (1987) Eur. J. Biochem. 168, 227-231.
38. Viotti, A., Balducci, C. and Weil, J.H. (1978) Biochim. Biophys. Acta 517, 125-132.

Appendix I: References for Tables 1 and 2

1. Hershey, H.P., Barker, R.F., Idler, K.B., Lissemore, J.L. and Quail, P.H. (1985) Nucl. Acids Res. 13, 8543-8559.
2. Rogers, J.C. Dean, D. and Heck, G.R. (1985) Proc. Natl. Acad. Sci. U.S.A. 82, 6512-6516.
3. Knox, C.A.P., Sonthayanon, B., Chandra, G.R. and Muthukrishnan, S. (1987) Plant Mol. Biol. 9, 3-17.
4. Forde, B.G., Kreis, M., Williamson, M.S., Fry, R.P., Pywell, J., Shewry, P.R., Bunce, N. and Mifflin, B.J. (1985) EMBO J. 4, 9-15.
5. Fincher, G.B., Lock, P.A., Morgan, M.M., Lingelbach, K., Wettenthal, R.E.H., Mercer, J.F.B., Brandt, A. and Thomsen, K.K. (1986) Proc. Nat. Acad. Sci. U.S.A. 83, 2081-2085.
6. Forde, B.G., Heyworth, A., Pywell, J. and Kreis, M. (1985) Nucl. Acids Res. 13, 7327-7339.
7. Mundy, J. and Rogers, J.C. (1986) Planta 169, 51-63.
8. Ponz, F., Paz-Ares, J., Hernandez-Lucas, C., Garcia-Olmedo, F. and Carbonero, P. (1986) Eur. J. Biochem. 156, 131-135.
9. Gausing, K. and Barkardottir, R. (1986) Eur. J. Biochem. 158, 57-62.
10. Chojceki, J. (1986) Carlsberg Res. Commun. 51, 211-217.
11. Bohlmann, H. and Apel, K. (1987) Mol. Gen. Genet. 207, 446-454.
12. Nielsen, P.S. and Gausing, K. (1987) FEBS Lett. 225, 159-162.
13. Takaiwa, F., Ebinuma, H., Kikuchi, S. and Oono, K. (1987) FEBS Lett. 221, 43-47.
14. Higuchi, W. and Fukazawa, C. (1987) Gene 55, 245-253.
15. Baulcombe, D.C., Huttly, A.K., Martienssen, Barker, R.F. and Jarvis, M.G. (1987) Mol. Gen. Genet. 209, 33-40.
16. Lamppa, G.K., Morelli, G. and Chua, N.-H. (1985) Mol. Cell Biol. 5, 1370-1378.
17. Litts, J.C., Colwell, G.W., Chakerian, R.L. and Quatrano, R.S. (1987) Nucl. Acids Res. 15, 3607-3618.
18. Baulcombe, D.C., Barker, R.F. and Jarvis, M.G. (1987) J. Biol. Chem. 262, 13726-13735.
19. Rafalski, J.A. (1986) Gene 43, 221-229.
20. Okita, T.W., Cheesbrough, V. and Reeves, C.D. (1985) J. Biol. Chem. 260 8203-8213.
21. Sugiyama, T., Rafalski, A., Peterson, D. and Sö ll, D. (1985) Nucl. Acids Res. 13, 8729-8737.
22. Tabata, T., Fukasawa, M. and Iwabuchi, M. (1984) Mol. Gen. Genet. 196, 397-400.
23. Tabata, T. and Iwabuchi, M. (1984) Gene 31, 285-289.
24. Smith, S.M., Bedbrook, J. and Speirs, J. (1983) Nucl. Acids Res. 11, 8719-8734.
25. Kreis, M., Forde, B.G., Rahman, S., Mifflin, B.J. and Shewry, P.R. (1985) J. Mol. Biol. 183, 499-502.

26. Schwarz-Sommer, Z., Shepherd, N., Tacke, E., Gierl, A., Rohde, W., Leclercq, L., Mattes, M., Berndtgen, R., Peterson, P.A. and Saedler, H. (1987) *EMBO J.* 6, 287-294.
27. Shah, D.M., Hightower, R.C. and Meagher, R.B. (1983) *J. Mol. Appl. Genet.* 2, 11-126.
28. Dennis, E.S., Sachs, M.M., Gerlach, W.L., Finnegan, E.J. and Peacock, W.J. (1985) *Nucl. Acids Res.* 13, 727-743.
29. Kelley, P.M. and Tolan, D.R. (1986) *Plant Physiol.* 82, 1076-1080.
30. Baker, A. and Leaver, C.J. (1985) *Nucl. Acids Res.* 13, 5857-5867.
31. Prat, S., Cortadas, J., Puigdomènech, P. and Palau, J. (1985) *Nucl. Acids Res.* 13, 1493-1504.
32. Moore, R.E., Davies, M.S., O'Connell, K.M., Harding, E.L., Wiegand, R.C. and Tiemeier, D.C. (1986) *Nucl. Acids Res.* 14, 7227-7235.
33. Chaubet, N., Philipps, G., Chaboute, M.-E., Ehling, M. and Gigot, C. (1986) *Plant Mol. Biol.* 6, 253-263.
34. Philipps, G., Chaubet, N., Chaboute, M.-E., Ehling, M. and Gigot, C. (1986) *Gene* 42, 225-229.
35. Rochester, D.E., Winer, J.A. and Shah, D.M. (1986) *EMBO J.* 5, 451-458.
36. Matsuoka, M., Kano-Murakami, Y. and Yamamoto, N. (1987) *Nucl. Acids Res.* 15, 6302.
37. Vance, V.B. and Huang, A.H.C. (1987) *J. Biol. Chem.* 262, 11275-11279.
38. Izui, K., Ishijima, S., Yamaguchi, Y., Katagiri, F., Murata, T., Shigesada, K., Sugiyama, T. and Katsuki, H. (1986) *Nucl. Acids Res.* 14, 1615-1628.
39. Lebrun, M., Waksman, G. and Freyssinet, G. (1987) *Nucl. Acids Res.* 15, 4360.
40. Werr, W., Frommer, W.-B., Maas, C. and Starlinger, P. (1985) *EMBO J.* 4, 1373-1380.
41. Marchionni, M. and Gilbert, W. (1986) *Cell* 46, 133-141.
42. Geraghty, D.E., Messing, J. and Rubenstein, I. (1982) *EMBO J.* 1, 1329-1335.
43. Marks, M.D., Lindell, J.S. and Larkins, B.A. (1985) *J. Biol. Chem.* 260, 16451-16459.
44. Prat, S., Pérez-Grau, L. and Puigdomènech, P. (1987) *Gene* 52, 41-49.
45. Pedersen, K., Devereux, J., Wilson, D.R., Sheldon, E. and Larkins, B.A. (1982) *Cell* 29, 1015-1026.
46. Marks, M.D. and Larkins, B.A. (1982) *J. Biol. Chem.* 257, 9976-9983.
47. Bethards, L.A., Skadsen, R.W. and Scandalios, J.G. (1987) *Proc. Natl. Acad. Sci. U.S.A.* 84, 6830-6834.
48. Paz-Ares, J., Ghosal, D., Wienand, U., Peterson, P.A. and Saedler, H. (1987) *EMBO J.* 6, 3553-3558.
49. Sommer, H. and Saedler, H. (1986) *Mol. Gen. Genet.* 202, 429-434.
50. Chang, C. and Meyerowitz, E.M. (1986) *Proc. Natl. Acad. Sci. USA* 83, 1408-1412.
51. Chaboute, M.-E., Chaubet, N., Philipps, G., Ehling, M. and Gigot, C. (1987) *Plant Mol. Biol.* 8, 179-191.
52. Leutwiler, L.S., Meyerowitz, E.M. and Tobin, E.M. (1986) *Nucl. Acids Res.* 14, 4051-4064.
53. Ludwig, S.R., Oppenheimer, D.G., Silflow, C.D. and Snustad, D.P. (1987) *Proc. Natl. Acad. Sci. U.S.A.* 84, 5833-5837.
54. Klee, H.J., Muskopf, Y.M. and Gasser, C.S. (1987) *Mol. Gen. Genet.* 210, 437-442.
55. Altenbach, S.B., Pearson, K.W., Leung, F.W. and Sun, S.S.M. (1987) *Plant Mol. Biol.* 8, 239-250.
56. Rose, R.E., DeJesus, C.E., Moylan, S.L., Ridge, N.P., Scherer, D.E. and Knauf, V.C. (1987) *Nucl. Acids Res.* 15, 7197.
57. Ericson, M.L., Roedin, J., Lenman, M., Glimelius, K., Josefsson, L.-G. and Rask, L. (1986) *J. Biol. Chem.* 261, 14576-14581.
58. Nasrallah, J.B., Kao, T.-H., Goldberg, M.L. and Nasrallah, M.E. (1985) *Nature* 318, 263-267.

59. Carrington, D.M., Auffret, A. and Hanke, D.E. (1985) *Nature* 313, 64-67.
60. Cohen, L.W., Coghlan, V.M. and Dihel, L.C. (1986) *Gene* 48, 219-227.
61. Merchant, S. and Bogorad, L. (1987) *J. Biol. Chem.* 262, 9062-9067.
62. Goldschmidt-Clermont, M., and Rahire, M. (1986) *J. Mol. Biol.* 191, 421-437.
63. Sharrock, R.A., Lissimore, J.L. and Quail, P.H. (1986) *Gene* 47, 287-295.
64. Smith, S.M. and Leaver, C.J. (1986) *Plant Physiol.* 81, 762-767.
65. Greenland, A.J., Thomas, M.V. and Walden, R.M. (1987) *Planta* 170, 99-110.
66. Chen, J. and Varner, J.E. (1985) *EMBO J.* 4, 2145-2151.
67. Chen, J. and Varner, J.E. (1985) *Proc. Natl. Acad. Sci. USA* 82, 4399-4403.
68. Schnell, D.J. and Etzler, M.E. (1987) *J. Biol. Chem.* 262, 7220-7225.
69. Adams, C.A., Babcock, M., Leung, F. and Sun, S.M. (1987) *Nucl. Acids Res.* 15, 1875.
70. Schuler, M.A., Ladin, B.F., Pollaco, J.C., Freyer, G. and Beachy, R.N. (1982) *Nucl. Acids Res.* 10, 8245-8261.
71. Hammond, R.W., Foard, D.E. and Larkins, B.A. (1984) *J. Biol. Chem.* 259, 9883-9890.
72. Negoro, T., Momma, T. and Fukazawa, C. (1985) *Nucl. Acids Res.* 13, 6719-6731.
73. Momma, T., Negoro, T., Hirano, H., Matsumoto, A., Udaka, K. and Fukazawa, C. (1985) *Eur. J. Biochem.* 149, 491-496.
74. Fukazawa, C., Momma, T., Hirano, H., Harada, K. and Udaka, K. (1985) *J. Biol. Chem.* 260, 6234-6239.
75. Momma, T., Negoro, T., Udaka, K. and Fukazawa, C. (1985) *FEBS Lett.* 188, 117-122.
76. Nagao, R.T., Czarnecka, E., Gurley, W.B., Schöffl, F. and Key, J.L. (1985) *Mol. Cell. Biol.* 5, 3417-3428.
77. Brisson, N. and Verma, D.P.S. (1982) *Proc. Natl. Acad. Sci. USA* 79, 4055-4059.
78. Vodkin, L.O., Rhodes, P.R. and Goldberg, R.B. (1983) *Cell* 34, 1023-1031.
79. Shibata, D., Steczko, J., Dixon, J.E., Hermodson, M., Yazdanparast, R. and Axelrod, B. (1987) *J. Biol. Chem.* 262, 10080-10085.
80. Sandal, N.N., Bojsen, K. and Marcker, K.A. (1987) *Nucl. Acids Res.* 15, 1507-1519.
81. Mauro, V.P., Nguyen, T., Katinakis, P. and Verma, D.P.S. (1985) *Nucl. Acids Res.* 13, 239-249.
82. Katinakis, P. and Verma, D.P.S. (1985) *Proc. Natl. Acad. Sci. USA* 82, 4157-4161.
83. Jacobs, F.A., Zhang, M., Fortin, M.G. and Verma, D.P.S. (1987) *Nucl. Acids Res.* 15, 1271-1280.
84. Fortin, M.G., Morrison, N.A. and Verma, D.P.S. (1987) *Nucl. Acids Res.* 15, 813-824.
85. Nguyen, T., Zelechowska, M., Foster, V., Bergmann, H. and Verma, D.P.S. (1985) *Proc. Natl. Acad. Sci. USA* 82, 5040-5044.
86. Franssen, H.J., Nap, J.-P., Gludemans, T., Stiekema, W., van Dam, H., Govers, F., Louwerse, J., van Kammen, A., and Bisseling, T. (1987) *Proc. Natl. Acad. Sci. USA* 84, 4495-4499.
87. Sengupta-Gopalan, C., Pitas, J.W., Thompson, D.V. and Hoffman, L.M. (1986) *Mol. Gen. Genet.* 203, 410-420.
88. Hong, J.C., Nagao, R.T. and Key, J.L. (1987) *J. Biol. Chem.* 262, 8367-8376.
89. Berry-Lowe, S.L., McKnight, T.D., Shah, D.M. and Meagher, R.B. (1982) *J. Mol. Appl. Genet.* 1, 483-498.
90. Krueger, R.W., Holland, M.A., Chisholm, D. and Polacco, J.C. (1987) *Gene* 54, 41-50.
91. Czarnecka, E., Nagao, R.T., Key, J.L. and Gurley, W.B. (1988) *Mol. Cell Biol.* 8, 1113-1122.
92. Vierling, E., Nagao, R.T., DeRocher, A.E. and Harris, L.M. (1988) *EMBO J.* 7, 575-581.
93. Guiltinan, M.J., Ma, D.-P., Barker, R.F., Bustos, M.M., Cyr, R.J., Yadegari, R. and Fosket, D.E. (1988) *Plant Mol. Biol.* 10, 171-184.
94. Chlan, C.A., Borroto, K., Kamalay, J.A. and Dure, L. (1987) *Plant Mol. Biol.* 9, 533-546.

95. Waksman, G. and Freyssinet, G. (1987) *Nucl. Acids Res.* 15, 1328-1328.
96. Allen, R.D., Cohen, E.A., Vonder Haar, R.A., Adams, C.A., Ma, D.P., Nessler, C.L. and Thomas, T.L. (1987) *Mol. Gen. Genet.* 210, 211-218.
97. Sakajo, S., Nakamura, K. and Asahi, T. (1987) *Eur. J. Biochem.* 165, 437-442.
98. Karlin-Neumann, G.A., Kohorn, B.D., Thornber, J.P. and Tobin, E.M. (1985) *J. Mol. Appl. Genet.* 3, 45-61.
99. Stiekema, W.J., Wimpee, C.F. and Tobin, E.M. (1983) *Nucl. Acids Res.* 11, 8051-8061.
100. Konieczny, A. (1987) *Nucl. Acids Res.* 15, 6742.
101. Hoffman, N.E., Pichersky, E., and Cashmore, A.R. (1987) *Nucl. Acids Res.* 15, 3928.
102. Holdsworth, M.J., Bird, C.R., Ray, J., Schuch, W. and Grierson, D. (1987) *Nucl. Acids Res.* 15, 731-739.
103. Grierson, D., Tucker, G.A., Keen, J., Ray, J., Bird, C.R. and Schuch, W. (1986) *Nucl. Acids Res.* 14, 8595-8603.
104. Hoffman, N.E., Pichersky, E., Malik, V.S., Castresana, C., Ko, K., Darr, S.C. and Cashmore, A.R. (1987) *Proc. Natl. Acad. Sci. U.S.A.* 84, 8844-8848.
105. Pichersky, E., Bernatzky, R., Tanksley, S.D. and Cashmore, A.R. (1986) *Proc. Natl. Acad. Sci. USA* 83, 3880-3884.
106. McKnight, T.D., Alexander, D.C., Babcock, M.S. and Simpson, R.B. (1986) *Gene* 48, 23-32.
107. Lee, J.S., Brown, W.E., Graham, J.S., Pearce, G., Fox, E.A., Dreher, T.W., Ahern, K.G., Pearson, G.D. and Ryan, C.A. (1986) *Proc. Natl. Acad. Sci. USA* 83, 7277-7281.
108. Graham, J.S., Pearce, G., Merryweather, J., Titani, K., Ericsson, L.H. and Ryan, C.A. (1985) *J. Biol. Chem.* 260, 6561-6564.
109. Pichersky, E., Bernatzky, R., Tanksley, S.D., Breidenbach, R.B., Kausch, A.P. and Cashmore, A.R. (1985) *Gene* 40, 247-258.
110. Pichersky, E., Hoffman, N.E., Malik, V.S., Bernatzky, R., Tanksley, S.D., Szabo, L. and Cashmore, A.R. (1987) *Plant Mol. Biol.* 9, 109-120.
111. Ray, J., Bird, C., Maunders, M., Grierson, D. and Schuch, W. (1987) *Nucl. Acids Res.* 15, 10587.
112. Kiss, G.B., Végh, Z. and Vincze, E. (1987) *Nucl. Acids Res.* 15, 3620.
113. DeRoche, E.J., Ramage, R.T., Michalowski, C.B. and Bohnert, H.J. (1987) *Nucl. Acids Res.* 15, 6301.
114. Boutry, M. and Chua, N.-H. (1985) *EMBO J.* 4, 2159-2165.
115. Calza, R., Huttner, E., Vincentz, M., Rouzé, P., Galangau, F., Vaucheret, H., Chérel, I., Meyer, C., Kronenberger, J., and Caboche, M. (1987) *Mol. Gen. Genet.* 209, 552-562.
116. Tingey, S.V. and Coruzzi, G.M. (1987) *Plant Phys.* 84, 366-373.
117. Shinshi, H., Mohnen, D. and Meins, F. Jr. (1987) *Proc. Natl. Acad. Sci. USA* (1987) 84, 89-93.
118. Shih, M.-C., Lazar, G. and Goodman, H.M. (1986) *Cell* 47, 73-80.
119. Pfizner, U.M. and Goodman, H.M. (1987) *Nucl. Acids Res.* 15, 4449-4465.
120. Cornelissen, B.J.C., Hooft van Huijsduijnen, R.A.M., Van Loon, L.C. and Bol, J.F. (1986) *EMBO J.* 5, 37-40.
121. Lagrimini, L.M., Burkhardt, W., Moyer, M. and Rothstein, S. (1987) *Proc. Natl. Acad. Sci. USA* 84, 7542-7546.
122. Mazur, B.J. and Chui, C.-F. (1985) *Nucl. Acids Res.* 13, 2373-2386.
123. Cornelissen, B.J.C., Hooft van Huijsduijnen, R.A.M. and Bol, J.F. (1986) *Nature* 321, 531-532.
124. Tucker, M.L., Durbin, M.L., Clegg, M.T. and Lewis, L.N. (1987) *Plant Mol. Biol.* 9, 197-203.
125. Reimold, U., Kröger, M., Kreuzaler, F. and Hahlbrock, K. (1983) *EMBO J.* 2, 1801-1805.
126. Dunsmuir, P. (1985) *Nucl. Acids Res.* 13, 2503-2518.
127. Stavton, M.M., Black, M., Bedbrook, J. and Dunsmuir, P. (1986) *Nucl. Acids Res.* 14, 9781-9796.
128. Koes, R.E., Spelt, C.E., Reif, H.J., van den Elzen, P.J.M., Veitkamp, E. and Mol, J.N.M. (1986) *Nucl. Acids Res.* 14, 5229-5239.
129. Condit, C.M. and Meagher, R.B. (1986) *Nature* 323, 178-181.
130. Turner, N.E., Clark, W.G., Tabor, G.J., Hironaka, C.M., Fraley, R.T. and Shah, D.M. (1986) *Nucl. Acids Res.* 14, 3325-3342.
131. Winter, J., Wright, R., Duck, N., Gasser, C., Fraley, R. and Shah, D. (1988) *Mol. Gen. Genet.* 211, 315-319.
132. Broglie, K.E., Gaynor, J.J. and Broglie, R.M. (1986) *Proc. Natl. Acad. Sci. USA* 83, 6820-6824.
133. Hoffman, L.M. and Donaldson, D.D. (1985) *EMBO J.* 4, 883-889.
134. Gebhardt, C., Oliver, J.E., Forde, B.G., Saarelainen, R. and Mifflin, B.J. (1986) *EMBO J.* 5, 1429-1435.
135. Lee, J.S. and Verma, D.P.S. (1984) *EMBO J.* 3, 2745-2752.
136. Hoffman, L.M. (1984) *J. Mol. Appl. Genet.* 2, 447-453.
137. Edwards, K., Cramer, C.L., Bolwell, G.P., Dixon, R.A., Schuch, W. and Lamb, C.J. (1985) *Proc. Natl. Acad. Sci. USA* 82, 6731-6735.
138. Slightom, J.L., Drong, R.F., Klassy, R.C. and Hoffman, L.M. (1985) *Nucl. Acids Res.* 13, 6483-6498.
139. Osborn, T.C., Alexander, D.C., Sun, S.S.M., Cardona, C. and Bliss, F.A. (1988) *Science* 240, 207-210.
140. Ryder, T.B., Hedrick, S.A., Bell, J.N., Liang, X., Clouse, S.D. and Lamb, C.J. (1987) *Mol. Gen. Genet.* 210, 219-233.
141. Higgins, T.J.V., Beach, L.R., Spencer, D., Chandler, P.M., Randall, P.J., Blagrove, R.J., Kortt, A.A., and Guthrie, R.E. (1987) *Plant Mol. Biol.* 8, 37-45.
142. Cashmore, A.R. (1984) *Proc. Natl. Acad. Sci. USA* 81, 2960-2964.
143. Tingey, S.V., Walker, E.L. and Coruzzi, G.M. (1987) *EMBO J.* 6, 1-9.
144. Gatehouse, J.A., Brown, D., Evans, I.M., Gatehouse, L.N., Jones, D., Preston, P. and Croy, R.R.D. (1987) *Nucl. Acids Res.* 15, 7642.
145. Lycett, G.W., Croy, R.R.D., Shirsat, A.H. and Boulter, D. (1984) *Nucl. Acids Res.* 12, 4493-4506.
146. Coruzzi, G., Broglie, R., Edwards, C. and Chua, N.-H. (1984) *EMBO J.* 3, 1671-1679.
147. Lycett, G.W., Delauney, A.J., Gatehouse, J.A., Gilroy, J., Croy, R.R.D. and Boulter, D. (1983) *Nucl. Acids Res.* 11, 2367-2380.
148. Llewellyn, D.J., Finnegan, E.J., Ellis, J.G., Dennis, E.S. and Peacock, W.J. (1987) *J. Mol. Biol.* 195, 115-123.
149. Gantt, J.S. and Key, J.L. (1987) *Eur. J. Biochem.* 166, 119-125.
150. Guidet, F. and Fourcroy, P. (1988) *Nucl. Acids Res.* 16, 2336.
151. Roberts, L.M., Lamb, F.L., Pappin, D.J.C. and Lord, J.M. (1985) *J. Biol. Chem.* 260, 15682-15686.
152. Halling, K.C., Halling, A.C., Murray, E.E., Ladin, B.F., Houston, L.L. and Weaver, R.F. (1985) *Nucl. Acids Res.* 13, 8019-8033.
153. Beeching, J.R. and Northcote, D.H. (1987) *Plant Mol. Biol.* 8, 471-475.
154. Smeeckens, S., van Binsbergen, J. and Weisbeek, P. (1985) *Nucl. Acids Res.* 13, 3179-3194.
155. Smeeckens, S., de Groot, M., van Binsbergen, J. and Weisbeek, P. (1985) *Nature* 317, 456-458.
156. Martin, W. and Creff, R. (1986) *Eur. J. Biochem.* 159, 323-331.
157. Mignery, G.A., Pikaard, C.S., Hannapel, D.J. and Park, W.D. (1984) *Nucl. Acids Res.* 12, 7987-8000.
158. Cleveland, T.E., Thornburg, R.W. and Ryan, C.A. (1987) *Plant Mol. Biol.* 8, 199-207.

159. Eckes, P., Rosahl, S., Schell, J. and Willmitzer, L. (1986) *Mol. Gen. Genet.* 205, 14-22.
160. Keil, M., Sanchez-Serrano, J., Schell, J. and Willmitzer, L. (1986) *Nucl. Acids Res.* 14, 5641-5650.
161. Wolter, F.P., Fritz, C.C., Willmitzer, L., Schell, J. and Schreier, P.H. (1988) *Proc. Natl. Acad. Sci. USA* 85, 846-850.
162. Salanoubat, M. and Belliard, G. (1987) *Gene* 60, 47-56.
163. Scherer, D.E. and Knauf, V.C. (1987) *Plant Mol. Biol.* 9, 127-134.
164. Jansen, T., Rother, C., Steppuhn, J., Reinke, H., Beyreuther, K., Jansson, C., Andersson, B. and Herrmann, R.G. (1987) *FEBS Lett.* 216, 234-240.
165. Rother, C., Jansen, T., Tyagi, A., Tittgen, J. and Herrmann, R.G. (1986) *Curr. Gen.* 11, 171-176.
166. Tyagi, A., Hermans, J., Steppuhn, J., Jansson, Ch., Vater F. and Herrmann, R.G. (1987) *Mol. Gen. Genet.* 207, 288-293.
167. Volokita, M. and Somerville, C.R. (1987) *J. Biol. Chem.* 262, 15825-15828.
168. Kuhse, J. and Pühler, A. (1987) *Plant Sci.* 49, 137-143.
169. Baumlein, H., Wobus, U., Pustell, J. and Kafatos, F.C. (1986) *Nucl. Acids Res.* 14, 2707-2720.
170. Bassüner, R., Hai, N.V., Jung, R., Saalbach, G. and Müntz, K. (1987) *Nucl. Acids Res.* 15, 9609.

Sequence requirement for specific interaction of an enhancer binding protein (EBP1) with DNA

Lilian Clark and Ronald T.Hay

Department of Biochemistry and Microbiology, University of St Andrews, St Andrews,
 Fife KY16 9AL, UK

Received November 25, 1988; Accepted December 12, 1988

ABSTRACT

Short DNA sequence motifs have been identified in viral and cellular enhancers which represent the binding sites for a variety of *trans*-acting factors. One such HeLa cell factor, EBP1, has been purified and shown to bind to sequences in the SV40 enhancer. The PRDII element in the human β -interferon gene regulatory element (IRE) shows strong sequence similarity to the EBP1 binding site in the SV40 enhancer. We demonstrate here that EBP1 binds to its sites in the SV40 enhancer and IRE in a similar manner, making base specific contacts over one complete turn of the DNA double helix. Mutational analysis of the EBP1 sites in the IRE and SV40 enhancer has identified the DNA sequence requirements necessary for specific EBP1/DNA complex formation. In addition, 34 DNA sequences related to the EBP1 binding site were analysed for their ability to bind EBP1. Sequences constituting high affinity binding sites possess the sequence 5'-GG(N)₆CC-3'. Single base pair changes in the region between the conserved Gs and Cs can generally be tolerated although it is clear that these intervening bases contribute to binding affinity. Mutations in the recognition site which could lead to gross structural changes in the DNA abolish EBP1 binding.

INTRODUCTION

Transcriptional control elements are generally grouped into two distinct classes of *cis*-acting regulatory sequences, promoters and enhancers, which can overlap both physically and functionally (1). Within eukaryotic promoters, DNA sequences responsible for defining the RNA start site and the rate of initiation of transcription have been identified (2; 3; 4) with efficient transcription being dependent on the correct spatial arrangement of the various promoter elements (5; 6; 7). Considerably more flexibility is exhibited with enhancer sequences in that they are able to activate transcription of linked genes essentially independent of position or orientation with respect to the responding promoter. This enhancing effect was first identified in nucleotide sequences present in the small DNA tumour virus Simian Virus 40 (SV40) (8; 9) and since then the SV40 enhancer, which serves as the prototype enhancer element, has been the subject of intense investigation.

Mutational analysis has defined short DNA segments, or motifs, within the SV40 enhancer which, though having little enhancing effect in isolation, when multimerised, or combined with other motifs, can act in concert to generate enhancer activity (10; 11; 12; 13). It appears that each of these DNA sequence motifs represents the binding site for at least one *trans*-acting factor (14) and it is likely to be the interaction between various *trans*-acting factors with the transcriptional machinery that is responsible for enhancer function.

Although the SV40 enhancer has been shown to be active in a variety of cell types, it also demonstrates distinct cell type specificities (15). Activation of transcription by individual DNA sequence motifs is cell type specific *in vivo* (16; 17; 13) and this is paralleled by cell type specificity in the binding of cellular factors to these motifs *in vivo* (18; 19).