ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
Volume 469

BIOCHEMICAL ENGINEERING IV

Edited by Henry C. Lim and K. Venkatasubramanian

The New York Academy of Sciences
New York, New York

Cover: The cover illustration of the soft-bound edition shows *Spodoptera frugiperda* cells 48 hours postinfection with *Autographa californica* nuclear polyhedrosis virus. (For details, see page 281.)

Library of Congress Cataloging-in-Publication Data

Biochemical engineering IV.

(Annals of the New York Academy of Sciences, ISSN 0077-8923; v. 469)
Based on papers from the Fourth International Conference on Biochemical Engineering held by the Engineering Foundation, Oct. 1–5, 1984 in Galway, Ireland.
Bibliography: p.
Includes index.
1. Biochemical engineering—Congresses. I. Lim, Henry C., 1935– . II. Venkatasubramanian, K. III. International Conference on Biochemical Engineering (4th : 1984 : Galway, Galway) IV. Engineering Foundation (U.S.) V. Series. VI. Title: Biochemical engineering. 4. VII. Title: Biochemical engineering four.
Q11.N5 vol. 469 500 s 86-12535
[TP248.3] [660'.63]
ISBN 0-89766-330-0
ISBN 0-89766-334-9 (pbk.)

SP
Printed in the United States of America
ISBN 0-89766-333-0 (cloth)
ISBN 0-89766-334-9 (paper)
ISSN 0077-8923

Preface

This volume is a collection of papers on recent developments in biochemical engineering presented at the Fourth International Conference on Biochemical Engineering held in Galway, Ireland, on October 1–5, 1984. The conference sessions explored the following topics: Recent Advances in Genetic Engineering; Design, Operating Strategies, and Separation for Bioprocesses; Mass Transfer and Transport in Bioreactors; Biokinetics and Bioreactor Modeling; Tissue Cultures; Biological Production of Chemicals and Energy; Estimation, Optimization, and Control of Bioreactors; and Microbial Production of Fuels.

This conference was organized once again by the Engineering Foundation. We gratefully acknowledge major financial support from the National Science Foundation and the New York Academy of Sciences. We are also deeply appreciative of financial contributions made by the following organizations and companies: Novo Industri, the Japanese Society of Enzyme Engineering, Kyowa Hakko Kogyo Co. Ltd., Abbott Laboratories, Dorr-Oliver, Inc., the Monsanto Company, Ajinomoto Company, Inc., Celanese Research Co., Ciba-Geigy, E. I. du Pont de Nemours & Company, Fujisawa Pharmaceutical Company, W. R. Grace & Company, the H. J. Heinz Company, Hoffmann-LaRoche, Inc., Imperial Chemical Industries, Miles Laboratories, Inc., Rhone-Poulenc, Inc., the Standard Oil Company (Indiana), Merck & Company, Inc., Gist-Brocades, N.V., and Biotechnical Resources, Inc. We are particularly grateful to Professor S. Suzuki of the Saitoma Institute of Technology for his help in securing financial support in Japan.

With deepest gratitude we thank Mr. Harold Comerer, Director, Dr. Sandford Cole, and the staff of the Engineering Foundation for their help in organizing the conference. With great pleasure we acknowledge the impeccable editorial work of Mr. Thomas Cohn and the Editorial Department of the New York Academy of Sciences. We would also like to express our deep appreciation to Professor Patrick Fottrell and his staff at the National University of Ireland for their superb handling of the local arrangements in Galway. We thank Ms. Janet Taylor and Ms. Connie Marsh for their secretarial assistance. It is also a pleasure to acknowledge gratefully the invaluable assistance of Mr. Yong Keun Chang of Purdue University.

Henry C. Lim
K. Venkatasubramanian

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

Volume 469
May 22, 1986

BIOCHEMICAL ENGINEERING IV[a]

Editors and Conference Organizers
HENRY C. LIM and K. VENKATASUBRAMANIAN

CONTENTS

[a]This volume is the result of the Fourth International Conference on Biochemical Engineering, held by the Engineering Foundation with the support of the New York Academy of Sciences and the National Science Foundation, on October 1–5, 1984, in Galway, Ireland.

Part III. Mass Transfer and Transport in Bioreactors

Part IV. Biokinetics and Bioreactor Modeling

Part V. Tissue Cultures

Part VI. Biological Production of Chemicals and Energy

Part VII. Estimation, Optimization, and Control of Bioreactors

Part VIII. Microbial Production of Fuels

Genetic Engineering of Extracellular Enzyme Systems of *Bacilli*

D. J. McCONNELL,[a] B. A. CANTWELL,[a,b] K. M. DEVINE,[a]
A. J. FORAGE,[b] B. M. LAOIDE,[a] C. O'KANE,[a]
J. F. OLLINGTON,[a,c] AND P. M. SHARP[a]

[a]*Trinity College*
Dublin, Ireland

[b]*Arthur Guinness Son & Co. Ltd.*
Dublin, Ireland

[c]*Biocon Ltd.*
Carrigaline
County Cork, Ireland

INTRODUCTION

The *Bacilli* produce a wide variety of extracellular enzymes some of which are manufactured and used on an industrial scale.[1-3] The enzymes have many applications, for example, in the production of food and drink. There is a growing interest in the use of genetic engineering technology to increase the understanding of the genetics and biochemistry controlling the production and excretion of these enzymes. With the knowledge so gained, it should be easier to improve the level of production of these enzymes and other gene products in *Bacilli*.

The *Bacilli* have many advantages for this kind of development. The genus includes *B. subtilis*, which has the best-characterized genetics of gram-positive bacteria.[4] The molecular biology of the *Bacilli* has been extensively studied,[5] and methods for molecular cloning have been developed.[6] A review of the literature indicates that 20 genes coding for extracellular enzymes have been cloned from nine different *Bacillus* species (TABLE 1). In this paper we present an analysis of some of the elements related to the expression of these genes—promoters, terminators, ribosome binding sites, codon usage, and signal sequences.

CLONING AND EXPRESSION OF *BACILLUS* EXTRACELLULAR ENZYMES

Carbohydrases

The amylolytic enzymes have been most extensively investigated and genes for the α- and β-amylases that hydrolyze the α-1,4-glucosidic bonds in starch and related polysaccharides have been isolated and characterized (TABLE 1). Both the liquefying type of α-amylase as produced by *B. licheniformis*, *B. coagulans*, and *B. amyloliquefaciens* and the saccharifying enzyme produced by *B. subtilis* have been cloned. These two types of α-amylase are distinguished on the basis of their ability to degrade starch. The liquefying α-amylase converts starch to maltodextrins while the saccharifying enzyme converts it to maltose and glucose. The α-amylases of strains of *B. licheniformis* are of particular interest as they are active at the high temperatures used in the

TABLE 1. Molecular Cloning of *Bacillus* Extracellular Enzymes[a]

Enzyme	Donor Organism	Host Organism	DNA Sequence	Mutant	Reference
Alkaline phosphatase	*B. licheniformis*	*E. coli*	—		26
α-amylase	*B. amyloliquefaciens*	*B. subtilis*	Yes	Yes	75, 76, 77, 78
	B. coagulans	*B. subtilis*	—		9, 79
		E. coli			
	B. licheniformis	*B. subtilis*	Partial		7, 34
	B. megaterium	*E. coli*	—		79
	B. stearothermophilus	*B. subtilis*	—		80, 81, 82
		B. stearothermophilus	—		
		E. coli			
	B. subtilis	*B. subtilis*	Yes		83, 84, 85, 86
		E. coli			
β-amylase	*B. cereus*	*E. coli*	—		79
CM-cellulases	*Bacillus* sp.	*E. coli*	—	No	16
β-glucanase	*B. subtilis*	*E. coli*	Yes	No	17, 18, 19, 20
		S. cerevisiae			
Hemolysin	*B. cereus*	*B. subtilis*	—	Non-producer	25
		E. coli			
β-lactamase	*B. cereus*	*B. subtilis*	Yes	Non-producer	87
		E. coli			
Levansucrase	*B. subtilis*	*B. subtilis*	—	Yes	88, 89
		E. coli			
Neutral protease	*B. stearothermophilus*	*B. subtilis*	—	Yes	90
		B. stearothermophilus			
Penicillinase	*B. licheniformis*	*B. licheniformis*	Yes	Non-producer	21, 23, 91, 92, 93
		B. subtilis			94, 95, 96
		B. stearothermophilus			
		E. coli			
		P. aeruginosa			
Subtilisin	*B. amyloliquefaciens*	*B. subtilis*	Yes	Yes	31
	B. subtilis	*B. subtilis*	Yes		32
Xylanase	*B. pumilus*	*E. coli*	—	Yes	97, 98, 99
	B. subtilis	*E. coli*	—		

[a]The information under Mutant refers to whether there are *B. subtilis* mutants deficient in the enzymatic activity. In some cases (labeled Nonproducer) *B. subtilis* strains are not known to produce the activity.

industrial degradation of starch (85°–115°C). Genes for α-amylases from two strains of *B. licheniformis* have been cloned in our laboratories,[7,8] the enzymes having temperature optima of 93°C and at least 100°C, respectively. A third α-amylase gene has been cloned from an organism identified as *B. coagulans*, and it is probably related to the *B. licheniformis* gene.[8,9] Genes for α-amylases have been cloned from four other species of *Bacilli* (TABLE 1), and it will be of considerable interest to compare their characteristics. It should be possible to gain a greater understanding of the structural basis of the enzymatic activities and the differences in temperature optima that extend from about 50°C to at least 100°C. This knowledge will be useful in designing enzymes with novel desirable properties and in constructing genes to encode them.

A series of genetic studies has shown that the α-amylase gene of *B. subtilis* is under a complex genetic control system that is poorly understood at the molecular level. Mutations have been found that together have enhanced the level of production by up to 2,000-fold,[10,11] and some of these increase the production of several other extracellular enzymes including levansucrase, protease,[11,12] and β-glucanase (Gormley, unpublished observations). As more knowledge is obtained on the molecular genetic basis of these systems, it should become easier to place other genes under their control. There are some indications that the genetic control systems of different species of *Bacilli* may be related, for example, the *B. licheniformis* FD02 α-amylase gene is catabolite repressed in its natural host and also after molecular cloning into *B. subtilis*[13] (also Laoide, unpublished observations).

The β-glucanases, including cellulases, of *Bacilli* have been characterized less than the amylolytic enzymes despite the fact that β-glucanases are widely distributed throughout the genus. Enzymes capable of hydrolyzing (1 → 3)-β-, (1 → 6)-β-, and (1 → 4)-β-D-glucans have been described,[3,14,15] but so far only genes coding for (1 → 4)-β-D-glucan endohydrolases have been cloned. Two genes coding for CM-cellulases have been isolated from an alkalophilic *Bacillus* species,[16] while in our own laboratory a gene coding for a (1 → 3), (1 → 4)-β-D-glucan endohydrolase has been isolated from *B. subtilis*. The latter enzyme hydrolyzes (1 → 4)-β-glucosyl linkages only where the glucosyl residue is substituted at the C(0)3 position in the mixed (1 → 3), (1 → 4)-β-D-glucans found in barley and the lichen *Cetraria islandica*. The donor strain of *B. subtilis* used in this cloning experiment also exhibits a CM-cellulase and a weak cellobiohydrolase activity (Cantwell, unpublished observations), but the genes encoding these enzymes have yet to be isolated. The (1 → 3), (1 → 4)-β-D-glucan endohydrolase is expressed in *E. coli*[17,18] and in *S. cerevisiae*[19] (Cantwell and Brazil, unpublished observations). Experiments on the localization of the enzyme in *E. coli* have shown that up to 20% of the enzyme produced is excreted into the extracellular medium (Gormley, unpublished observations). Low levels of expression of the enzyme in *S. cerevisiae* were first achieved in this laboratory by placing the gene under control of the yeast *iso*-1-cytochrome *c* (CYC-1) promoter. Subsequently, the yeast alcohol dehydrogenase (ADH-1) promoter was used to direct the synthesis of higher levels of expression of β-glucanase in yeast. Attempts to express the gene in yeast using *Bacillus* control sequences alone did not yield detectable levels of enzyme. The complete sequence of this β-glucanase gene has been published.[20]

Other carbohydrases of *Bacillus* that have been cloned and expressed in *E. coli* include the genes for xylan degradation from *B. subtilis* and *B. pumilus* and the levansucrase gene from *B. subtilis* (TABLE 1).

Proteases

The proteases represent another major group of *Bacillus* extracellular enzymes. There are two main types, the metal and serine proteases. The metal proteases require

Ca^{2+} for stability and Zn^{2+} for activity. They have pH optima at or near neutrality, and for this reason are sometimes referred to as neutral proteases. The second type, the serine proteases, also known as the subtilisins, have alkaline pH optima and a serine residue at or near the active site. Three different proteolytic enzymes have been cloned—two subtilisins, from *B. subtilis* and *B. amyloliquefaciens* and a neutral protease from the thermophile *B. stearothermophilus* (TABLE 1).

β-Lactamases

Many *Bacillus* species are known to produce β-lactamases. Only two of these enzymes have been investigated in detail. These are the penicillinases of *B. cereus* and *B. licheniformis*. Expression of both enzymes has been achieved in *B. subtilis* and in *E. coli* (TABLE 1) while expression of the penicillinase of *B. licheniformis* has also been studied in the thermophile *B. stearothermophilus* where the cloned gene was expressed even at 60°C.[21] The control system of the *B. licheniformis* enzyme is believed to involve a repressor coded by a gene linked to the structural gene of the enzyme.[22-24]

Others

The hemolysin gene from *B. cereus*[25] and the alkaline phosphatase gene from *B. licheniformis*[26] have also been isolated by molecular cloning.

TRANSCRIPTIONAL CONTROL SYSTEMS

Transcriptional control systems in bacteria involve several different kinds of interactions between proteins and nucleic acid sequences at the beginning and end of structural genes. RNA polymerase interacts at promoters, attenuators, and terminators, while accessory proteins such as repressors or positive control factors interact at operators and positive control sites, respectively. These proteins and sequences are essential components of the mechanism determining the maximal rate of transcription, and the regulation of that rate by external factors.

Knowledge of the transcriptional control systems of genes coding for extracellular enzymes will be valuable in manipulating the expression of these genes and in adapting these systems to control the expression of heterologous genes.

In *B. subtilis*, RNA synthesis is catalyzed by a set of RNA polymerases that differ in one subunit, the σ factor, which determines the promoter-specificity of the enzyme.[27-30] Five such factors are known—σ^{43}, σ^{37}, σ^{32}, σ^{29}, and σ^{28}—and corresponding promoter sequences have been identified. Consensus sequences have been proposed based on nine sequences for σ^{43} promoters, two for σ^{37}, two for σ^{32}, four for σ^{29} and two for σ^{28}.[28] The σ^{29} RNA polymerase is known to transcribe genes during sporulation and may be specifically related to sporulation.

It is of interest to define the promoter sequences of the genes coding for extracellular enzymes since these will provide information related to the level and mode of transcription. The sequences to the 5′ side of eight different genes of *Bacillus* extracellular enzymes have been reported, and the authors of the original papers have suggested possible promoters (TABLE 2). In the case of five genes, the evidence is minimal, based only on location and on agreement with the σ^{43} promoter consensus. For three genes the evidence is strengthened by experimental determination of the site of initiation of RNA synthesis. It is likely that the two subtilisin genes are transcribed

TABLE 2. Promoter Sequences[a]

Gene	Organism	σ^{43} Promoters −35 −10 TTGACA....17–18 bases....TATAAT *B. subtilis* σ^{43} Consensus	Distance to ATG	RNA Polymerase Subunit Type	Promoter Determination	References
α-Amylase	*B. amyloliquefaciens*	TTGTTAATTTTACTGATATGTAAAATATAAT	30	σ^{43}	C	77
α-Amylase	*B. licheniformis*	TTGTTA−−AAAATTCGGAATATTTATACAAT	34	σ^{43}	C	34
α-Amylase	*B. subtilis*	TTGAAA−GGAGGAAGCGGAAGAATGAAGTAA	56	σ^{43}	C	84, 100
	B. subtilis	TTGAAA−−−GGAGGAAGCGGAAGAATGAAGT	58	σ^{43}	C	
	B. subtilis	TTGATA−−−GAGTGATTGTGATAATTTAAAT	123	σ^{43}	C	
	B. subtilis	TTGTTT−TGATAGAGTGATTGTGATAATTTT	127	σ^{43}	C	
β-Glucanase	*B. subtilis*	TTGACC−GATGTTCCCTTTTGAAAGAATCAT	46	σ^{43}	C	20
β-Lactamase	*B. cereus*	TTGTTA−−ATAATTCTCATACTAGCTATAAA	82	σ^{43}	C	101
Penicillinase	*B. licheniformis*	TTGCAT−−TTAAATCTTACATATGTAATACT	165	σ^{43}	C, SEQ	94, 95
		Other Promoters				
Subtilisin	*B. amyloliquefaciens*	GGTCTA−−−CTAAAATATTATTCCATACTATACAATT	16	σ^{37}	C, BAL	31
Subtilisin	*B. subtilis*	P_2 AGTCTTTTAAGTAAGTCTACTCTGAATTTTTTTA	19	σ^{37}	S1	32, 102
		P_1 AAATTCA−CAGAATAGTCTTTTAAGTAAGTCTA	34	Unknown	S1	

[a]The *B. subtilis* σ^{43} consensus sequence is taken from Stephens *et al.*[43] The putative −35 and −10 sequences are underlined. The distance to the ATG is the number of base pairs between the end of the −10 sequence and the ATG thought to be the initiator codon of the gene. Promoters have been determined by consensus (C), by sequencing the RNA (SEQ), by Bal31 digestion (BAL) or by S1 nuclease mapping (S1). In the cases of the "other promoters," the sequences underlined are those related to the −35 and −10 regions of other σ^{37} promoters.[43] The distance to the ATG is the distance from the end of the sequence shown.

by the σ^{37} RNA polymerase,[31,32] and the evidence is strong that the *B. licheniformis* penicillinase gene is transcribed by the σ^{43} RNA polymerase. In this case the distance between the −10 sequence and the ATG at the beginning of the gene is remarkably long—165 bases. The data on other promoters are limited with definitive evidence on the initiation of transcription still outstanding.

Terminator sequences in *Bacilli* have not been studied extensively.[27] By analogy with terminators for *E. coli* RNA polymerase, which is closely related to the σ^{43} RNA polymerase, it is expected that terminators of *Bacilli* should be comprised of a region of dyad symmetry followed by a run of T's. Sequences of this form are found just following the ends of the genes for the α-amylases of *B. amyloliquefaciens* and *B. subtilis,* and after the penicillinase gene of *B. licheniformis* and the subtilisin gene of *B. amyloliquefaciens* (TABLE 3). There is direct evidence that transcription does terminate in the region of the designated sequence at the end of the penicillinase gene.[33] The end of the β-glucanase gene has a complex set of inverted repeat sequences. The β-glucanase gene and the α-amylase gene of *B. licheniformis* are preceded by sequences that resemble the form of typical *E. coli* terminators.[20,34]

TRANSLATIONAL CONTROL SEQUENCES

The rate of translation of a mRNA of *E. coli* is controlled by a sequence at the 5′ end of the mRNA molecule called the Shine-Dalgarno sequence or ribosome binding site, and that is partially complementary to a sequence at the 3′ end of the *E. coli* 16S rRNA with which it interacts. The same has been suggested for other prokaryotic systems.[35] In gram-positive genes potentially strong ribosome binding site sequences have been identified in cloned DNA sequences just to the 5′ side of putative initiation codons.[36] These sequences were initially identified as partially complementary to the sequence at the 3′ end of the *B. subtilis* 16S rRNA and there is direct experimental evidence that changes in such sequences alter the rate of gene expression.[37] *Bacillus* mRNA ribosome binding sites show, in general, rather extensive complementarity to this 16S rRNA sequence.

The sequences of potential ribosome binding sites for eight genes coding for *Bacillus* extracellular enzymes are shown in TABLE 4. These have usually been identified by the authors of the cited papers on the basis of location (just to the 5′ side of the specified initiation codon) and complementarity to the 3′ end of the *B. subtilis* 16S rRNA. The application of this second criterion to mRNA sequences from *Bacillus* species other than *B. subtilis* is justified by the fact that 16S rRNA 3′ end sequences are strongly conserved. A 15-base sequence at the 3′ end of *B. subtilis* 16S rRNA is identical to the sequence in the same position in *B. stearothermophilus* 16S rRNA, and furthermore is identical to the *E. coli* sequence over the 12 bases most usually postulated to be directly involved in complex formation.

The sequences underlined in TABLE 4 give the minimum free-energy value for the interaction with the 3′ end of 16S rRNA calculated according to Tinoco *et al.*[38] In some cases the designated sequences or the free-energy values differ slightly from those of the cited authors. The free-energy values (−11.4 to −22.0 kcalories) are all in the range found by McLaughlin *et al.*,[36] (−11.6 to −21.0) supporting their hypothesis that the interactions between 16S rRNA and ribosome binding sites in gram-positive bacteria are usually stronger than in *E. coli* where the ΔG values range from −4 to −22 kcalories.[35] The number of bases between the last base of the designated ribosome binding site and the first base of the putative initiation codon in the sequences in TABLE 4 varies from 3 to 10. This range is not significantly different from the range (5 to 11) cited by McLaughlin *et al.*[36]

TABLE 3. Terminator Sequences[a]

Gene	Organism	Stop Codon	Inverted Repeats	References
α-Amylase	*B. amyloliquefaciens*	TAA	GGTAATAAAAAAACACCTCCAAGCTGAGTGCGGGTATCAGCTTGGAGGTGCGTTTATTTTTT	Takkinen *et al.*[77]
α-Amylase	*B. subtilis*	TGA	GGGCAAGGCTAGACGGGACTTACCGAAAGAAACCATCAATGATGGTTTCTTTTTTG.... (38 bases) ....GTTTGTGAAGTGTTGCACAATATAAATGTGAAATACTTCACAAACAAAAA	Yang *et al.*[100]
β-Glucanase	*B. subtilis*	TAA	TGCCAAATGTGAAAGAACCTGCTGCAATATAGCAGGCTCTTATGATTGTAATGAGAAGTTGT	Murphy *et al.*[20]
Penicillinase	*B. licheniformis*	TAA	AGTCACTTGGTGATCAAGCTCATATCATTGTCCGGCAATGGTGTGGGCTTTTTTTG	Neugebauer *et al.*[95]
Subtilisin	*B. amyloliquefaciens*	TAA	ACATAAAAAACCGGCCTTGGCCCCGCCGGTTTTTTATTATTTTTCTTCCTCC	Wells *et al.*[31]
Subtilisin	*B. subtilis*	TAA	TAGTAAAAAGAAGCAGGTTCCTCCATACCTGCTTCTTTTTATTTGTCAGCATCCT	Stahl & Ferrari[102]

[a]The putative stop codon at the end of each gene and the sequences that form inverted repeats are underlined.

TABLE 4. Ribosome Binding Sites[a]

Gene	Organism		ΔG	Spacing	References
α-Amylase	*B. amyloliquefaciens*	5′ AAAUGAGAGGGAGAGGAAACAUG 3′	−15.8	4	77
α-Amylase	*B. licheniformis*	AUUGAAAGGGGAGGAGAAUCAUG	−19.6	3	34
α-Amylase	*B. subtilis*	UCAAAUAAGGAGUGUCAAGAAUG	−14.8	6	84, 100
β-Glucanase	*B. subtilis*	AAUGAAAGGGGAAUGCCAAUAUG	−18.8	9	20
β-Lactamase	*B. cereus*	CUAAAAAUUUGGAAGGAAUGAUG	−11.4	4	87, 101
Penicillinase	*B. licheniformis*	AACGGAGGGAGACGAUUUUGAUG	−18.6	10 (7)	94, 95
Subtilisin	*B. amyloliquefaciens*	AAAAAAGGAGAGGAUAAAGAGUG	−22.0	5	31
Subtilisin	*B. subtilis*	UUAAAAGGAGAGGGUAAAGAGUG	−18.0	5	102, 32
		3′ UCUUUCCUCCACUAG 5′			
		B. subtilis 16S rRNA 3′ end			

[a]The sequences that give the lowest free-energy value ΔG when made complementary with the 3′ sequence of 16S rRNA have been underlined. The ΔG values were calculated according to Tinoco *et al.*[38] The spacing is the number of bases between the last underlined base and the putative initiator codon. In one case the β-lactamase of *B. cereus*, there are two AUG codons in phase beside one another, but one is only one base from the putative ribosome binding site. In the case of the penicillinase of *B. licheniformis* there is a UUG in phase beside the AUG. The codon UUG is known to be used sometimes to initiate protein synthesis.[40]

Atkins[39] noted that the stop codons UAA or UGA (but not UAG) are often found within the region of 15 nucleotides to the 5′ side of prokaryotic initiation codons, and he has inquired whether these are part of the recognition signal for ribosomes. This has been questioned as more sequences have been analyzed.[35,40] In the eight *Bacillus* sequences in TABLE 4, UAA or UGA occurs in seven cases, while UAG is never found in the region of 20 nucleotides to the 5′ side of the initiation codon. It remains to be shown whether this is significant, but the data are in support of Atkins's suggestion.

The suggested initiation codons for the eight genes in TABLE 4 include 6 AUG and 2 GUG. The GUG codons occur in the two homologous subtilisin genes that also have very similar sequences to the 5′ side of the GUG with a run of 17 out of 18 bases identical. One gene, the penicillinase of *B. licheniformis,* has a UUG triplet in frame and in a position suggesting that it might act as an initiation codon. This triplet has been shown to initiate protein synthesis in one *S. aureus* gene[36] in four cases in *E. coli* (reviewed in Kozak[40]), and probably does so in some other *Bacillus* genes including *spoϕ*H of *B. licheniformis,*[41] *cat*-86 of *B. pumilus,*[42] and the 0.3-kb gene of *B. subtilis.*[43]

CODON USAGE IN *BACILLI*

It has been suggested that codon usage may be related to gene expression with different subsets of synonymous codons being preferred in genes of high and low expression.[44–46] If heterologous genes are to be highly expressed in a microorganism and if codon usage is important in controlling expressivity, then codon usage in a heterologus gene should match as closely as possible the codon usage of well-expressed genes of the host. With this in mind, we have computed codon usage in 17 *Bacillus* genes including six coding for extracellular enzymes.

The codon usages are compiled in TABLE 5. Two important extracellular enzyme genes were not included in the *Bacillus* data primarily because of homology to sequences already in the list. These were the *B. licheniformis* α-amylase (which is only partially sequenced), homologous to the *B. amyloliquefaciens* α-amylase, and the *B. amyloliquefaciens* subtilisin, homologous to the *B. subtilis* enzyme. Thirteen of the seventeen *Bacillus* genes chosen are from *B. subtilis, B. amyloliquefaciens,* and *B. licheniformis,* three species with DNA of about 45% GC content. The DNAs of the other species have lower reported GC contents with values for *B. cereus* at 32–36% being clearly less than the major group (CRC Handbook of Biochemistry and Molecular Biology, 2nd Edition).

In the *Bacillus* genes codon usage is highly nonrandom, both across the whole code and within synonymous groups of codons. Significantly, this pattern of nonrandom synonymous codon usage is quite different from that seen in compilations of data for *E. coli* (52 genes, see Alff-Steinberger),[47] bacteriophage T7 (50 genes),[48] and yeast (17 genes, compiled from the EMBL Nucleotide Sequence Data Library). If codon usage is linked to gene expressivity,[49–51] these differences imply that codon usage may be a factor affecting variation in heterologous gene expression between *Bacillus, E. coli,* and yeast.

Several factors appear to contribute to nonrandom synonymous codon usage in other organisms, including tRNA abundance,[46] third base pyrimidine bias yielding codon-anticodon interactions of intermediate strength[45] and a global excess of RNY (purine–anybase–pyrimidine) codons.[52] There are insufficient data available on *Bacillus* tRNA abundance, but examination of TABLE 5 shows that neither of the two latter hypotheses apply to *Bacillus.* It has been reported that the influence of third base

TABLE 5. Codon Usage[a]

		N	B	E	Y			N	B	E	Y			N	B	E	Y			N	B	E	Y
UUU	Phe	145	(1.43	0.74	0.78)	UCU	Ser	83	(1.23	1.40	2.61)	UAU	Tyr	138	(1.38	0.80	0.67)	UGU	Cys	19	(0.88	0.86	1.72)
UUC		58	(0.57	1.26	1.22)	UCC		56	(0.83	1.60	1.44)	UAC		62	(0.62	1.20	1.33)	UGC		24	(1.12	1.14	0.28)
UUA	Leu	115	(1.59	0.40	1.13)	UCA		82	(1.21	0.42	0.67)	UAA	term	9	(1.80	2.25	2.29)	UGA	term	5	(1.00	0.52	0.00)
UUG		65	(0.90	0.52	3.28)	UCG		48	(0.71	0.69	0.25)	UAG		1	(0.20	0.23	0.71)	UGG	Trp	76	(1.00	1.00	1.00)
CUU	Leu	108	(1.49	0.45	0.45)	CCU	Pro	64	(1.28	0.47	0.97)	CAU	His	90	(1.45	0.75	0.91)	CGU	Arg	49	(1.32	3.39	1.24)
CUC		33	(0.46	0.45	0.10)	CCC		27	(0.54	0.28	0.28)	CAC		34	(0.55	1.25	1.09)	CGC		25	(0.67	2.15	0.16)
CUA		36	(0.50	0.11	0.68)	CCA		38	(0.76	0.65	2.68)	CAA	Gln	117	(1.18	0.47	1.75)	CGA		30	(0.81	0.19	0.03)
CUG		77	(1.06	4.06	0.36)	CCG		71	(1.42	2.60	0.07)	CAG		82	(0.82	1.53	0.25)	CGG		25	(0.67	0.17	0.03)
AUU	Ile	174	(1.54	1.07	1.52)	ACU	Thr	61	(0.72	0.99	1.70)	AAU	Asn	167	(1.19	0.53	0.64)	AGU	Ser	53	(0.78	0.35	0.71)
AUC		111	(0.99	1.83	1.32)	ACC		43	(0.51	2.06	1.47)	AAC		114	(0.81	1.47	1.36)	AGC		84	(1.24	1.53	0.32)
AUA		53	(0.47	0.09	0.17)	ACA		147	(1.73	0.28	0.57)	AAA	Lys	256	(1.50	1.52	0.70)	AGA	Arg	64	(1.72	0.06	4.16)
AUG	Met	126	(1.00	1.00	1.00)	ACG		89	(1.05	0.67	0.26)	AAG		85	(0.50	0.48	1.30)	AGG		30	(0.81	0.04	0.39)
GUU	Val	95	(1.18	1.45	1.94)	GCU	Ala	127	(1.19	1.02	2.19)	GAU	Asp	182	(1.30	0.91	1.11)	GGU	Gly	87	(0.89	1.92	3.22)
GUC		62	(0.77	0.58	1.33)	GCC		81	(0.76	0.85	1.19)	GAC		98	(0.70	1.09	0.89)	GGC		105	(1.08	1.58	0.41)
GUA		81	(1.00	0.88	0.30)	GCA		127	(1.19	0.88	0.52)	GAA	Glu	227	(1.32	1.45	1.67)	GGA		139	(1.43	0.19	0.18)
GUG		85	(1.05	1.09	0.44)	GCG		91	(0.85	1.24	0.09)	GAG		118	(0.68	0.55	0.33)	GGG		59	(0.61	0.31	0.18)

[a]Codon usage (N) is shown for 17 *Bacillus* genes and also the relative usage of synonymous codons in *Bacillus* (B), *E. coli* (E), and *S. cerevisiae* (Y). The *Bacillus* genes used were: *B. amyloliquefaciens* α-amylase,[77] *B. cereus* β-lactamase,[101] *B. licheniformis ermD*,[103] penicillinase, *spo* φH,[41] *B. megaterium* protein C,[104] *B. pumilus cat*-86,[42] *B. subtilis* 0.3-kb gene,[43] α-amylase,[100] β-glucanase,[20] *purF*,[105] *spo* φF, subtilisin,[102] *trpC*, *trpD*, *trpE*,[106] *B. thuringiensis* crystal protein.[107] The yeast genes used were: *S. cerevisiae* actin, Adh I, Adh II, β-tubulin, URF at centromere, *iso*-1-cytochrome *c*, *iso*-2-cytochrome *c*, glyceraldehyde-3-phosphate dehydrogenase, histone 2A1, histone 2B1, HIS 1, HIS 4, invertase, MES 1, acid phosphatase, pyruvate kinase, and tryptophan synthetase. All DNA sequences were from the EMBL Nucleotide Sequence Data Library (Release 3) unless otherwise referenced.

pyrimidine bias is more evident in highly expressed genes in *E. coli*.[44] The only *Bacillus* genes showing significant bias are the *B. amyloliquefaciens* subtilisin gene and the *B. megaterium* protein C gene, although it might be expected that some of the extracellular enzymes would be "highly expressed." Though there are differences between the *Bacillus* genes in exact pattern of nonrandom codon usage, as measured by heterogeneity chi-squares (data not shown), the heterogeneity is mainly between species in the current sample. The 12 genes in the subset of *B. subtilis* and *B. licheniformis* are comparatively homogeneous with the exception of the subtilisin gene. The differences for the genes from the other species may reflect the nature of these particular genes rather than a difference between species.

SIGNAL SEQUENCES OF EXTRACELLULAR ENZYMES IN *BACILLI*

The signal sequence hypothesis for protein export first proposed by Blobel and Dobberstein[53] and subsequently modified[54] has been supported by evidence from a variety of prokaryotic and eukaryotic systems. A key component of this hypothesis has been that proteins destined for export from the cell contain a sequence at the NH_2 terminus that functions to initiate transport through the membrane. It is proposed that this sequence is recognized by receptors in the membrane that anchor the nascent protein and ribosome and allow cotranslational tunneling of the nascent peptide through the membrane. In many cases the signal peptide sequence is subsequently cleaved by the enzyme signal peptidase.

Extracellular enzymes of *Bacilli* provide a good system for the study of protein export from cells. Attention has recently been focused on the signal sequences of these enzymes and their role in enzyme export from the cell. Constructions in which the promoter and signal sequence of α-amylase from *B. amyloliquefaciens* have been fused with the structural genes for β-lactamase (from *E. coli*) and human interferon have resulted in export of these proteins from *B. subtilis*.[55,56] Ohmura *et al.*[57,58] achieved export of β-lactamase of *E. coli* from *B. subtilis* by fusing the promoter and signal sequence of an α-amylase gene from a *Bacillus* strain to the β-lactamase structural gene. Similar constructions have achieved the secretion of human insulin.[59] Ohmura *et al.*[57] further analyzed the regions of the signal sequence that are necessary to achieve protein export by constructing a series of plasmids containing deletions in the region coding for the signal sequence. Although the signal sequence for α-amylase contains 41 amino acids, it was found that signal sequences containing 33 and 31 amino acids were sufficient to achieve protein export while no export was found with a signal sequence of only 28 amino acids. It appears, then, that information necessary for the export of extracellular enzymes from *Bacillus* is contained within a leader sequence, similar to exported proteins in other prokaryotes and eukaryotes. It is thus appropriate to compare the features of *Bacilli* signal sequences with those of the better characterized *E. coli* system.

The advent of cloning has allowed a compilation of signal sequences for a variety of *Bacillus* extracellular enzymes (TABLE 6). These sequences can be compared with those of *E. coli* under four headings—hydrophilic region, hydrophobic region, the charged amino acids that separate these regions, and the point at which the signal sequence is cleaved to yield mature protein. The sequences listed in TABLE 6 are somewhat arbitrarily aligned by the charged amino acids that separate the hydrophilic region from the hydrophobic region. Lys-Arg or Lys-Lys is found in all eight sequences listed. These amino acid doublets are found in many other prokaryotic signal sequences where they seem to occur closer to the initiating methionine residue.[60] The conservation