NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

ISBN 0-315-53213-0

The Molecular Isolation and Characterization of Endo-ß-1,4-glucanase Genes from *Bacillus polymyxa* and *Bacillus circulans*

by

Stephen Baird, B.I.S.

A thesis submitted to the Faculty of Graduate Studies
in partial fulfillment of the requirements for the degree of
Master of Science

Department of Biology
University of Ottawa
Ottawa, Ontario

UNIVERSITÉ D'OTTAWA
UNIVERSITY OF OTTAWA

Abstract

Cellulose the major structural component of plants is degraded by a variety of fungi and bacteria which produce extracellular enzymes called cellulases. An endocellulase gene has been isolated from the *Bacillus polymyxa* NRCC 2822 and *Bacillus circulans* NRCC 9024 in *E. coli* using the phage M13mp9 as the primary cloning vector. A 1.6 kbp PstI DNA fragment from *B. polymyxa* and a 3.4 kbp HindIII DNA fragment from *B. circulans* which contained these endo-ß-1,4-glucanase genes were selected by the detection of the enzymatic activity from the gene product using either the Congo Red assay or Ostazin Brilliant Red H3B dyed cellulose. Both genes appear to be the only endo-ß-1,4-glucanase genes contained in their respective parental hosts. The endo-ß-1,4-glucanase gene from *B. circulans* appears to be very homologous to the endo-ß-1,4-glucanase gene of *B. subtilis* PAP115 using restriction endonuclease site mapping, DNA-DNA hybridization, the 'S1 nuclease protection assay', and examination of the protein products from the genes using SDS-PAGE. The *B. polymyxa* gene appeared to be quite different from the *B. circulans* and *B. subtilis* genes using these methods.

Expression of the *B. polymyxa* endoglucanase gene in *E. coli* was due to an in-frame fusion of the 5′end of the coding region and the 5′end of the *lac*Z α-peptide gene at the PstI site of either M13mp9 or pUC9. A 2.35 kbp EcoRI DNA fragment from *B. polymyxa* which overlapped the 1.6 kbp PstI fragment was isolated containing the remaining 5′portion of the endoglucanase gene. When the gene was reconstituted by fusing the 5′and 3′ portions, it did not express in *E. coli*. The 1.6 kbp PstI DNA fragment and the 1.5 kbp non-overlapping region of the 2.35 kbp EcoRI fragment were sequenced. No promoter sequences which could allow transcription in *E. coli* were found, but an open reading frame, ORF-2, coding for a putative protein of 187 amino acids was found 374 base pairs 5′ to the endoglucanase translational start codon and appeared from the sequence data to be transcribed divergently from the endoglucanase gene. The ORF-2 protein sequence had a 28 % amino acid similarity with the NH_2-terminal half of the *E. coli lac* repressor protein. The endoglucanase gene contained a 398 codon open reading frame which would produce an enzyme of theoretically 44 kDa in size. Deletions of only 12 codons from the 3′end of the endoglucanase gene stopped any detectable endoglucanase activity from being produced in *E. coli*.

All known published cellulase sequences to date and the putative amino acid sequence of the *B. polymyxa* endoglucanase were compared to each other in pairwise comparisons. The *B. polymyxa* endoglucanase had a 34 % amino acid similarity to the *Clostridium thermocellum cel*B endoglucanase sequence but very little similarity to any endoglucanase from other *Bacillus* species. The eight endoglucanase amino acid sequences of *T. reesei* EG3, *Schizophyllum commune* EG2, *Erwinia chrysanthemi celZ*, *B. subtilis* PAP115, *Bacillus* sp. strain no. 1139, *Bacillus* sp. strain N-4 EG N2, *B. polymyxa* and *C. thermocellum cel*B were all aligned together. A consensus sequence of Asn-Glu-Pro has been postulated as being part of the active site. Hydropathy plots of the region surrounding this consensus sequence also show some similarity to each other. The sequence similarities of these eight endoglucanases suggest that they are part of a subclass of endo-ß-1,4-glucanases with a distinct enzymatic mechanism or function.

RÉSUMÉ

La cellulose, composant structural majeur des végétaux, est dégradée par de nombreux champignons et bactéries, grâce à leurs cellulases extracellulaires. Deux gènes codant pour une endocellulase, provenant l'un de *Bacillus polymyxa* NRCC 2822 et l'autre de *Bacillus circulans* NRCC 9024, ont été clonés dans *Escherichia coli* à l'aide du phage M13mp9 utilisé comme vecteur primaire. Un fragment d'ADN de *B. polymyxa* coupé par PstI de 1,6 kbp et un fragment d'ADN de *B. circulans* coupé par HindIII de 3,4 kbp, contenant tous deux le gène de l'endo-ß-1,4-glucanase, ont été selectionnés par détection de l'activité enzymatique de leur produit à l'aide du test au rouge Congo ou au rouge brillant Ostazin H3B, colorant spécifique de la cellulose. Ces deux gènes semblent être les seuls gènes d'endo-ß-1,4-glucanase contenus dans leur génome respectif. Comme le montrent les cartes des sites de restriction, les hybridations ADN-ADN, le test "protection par la nuclease S1" et l'étude des protéines produites par ces gènes avec la technique SDS-PAGE, le gène de l'endo-ß-1,4-glucanase de *B. circulans* parait avoir beaucoup d'homologie avec celui de *B. subtilis*. L'utilisation de ces méthodes montre que le gène de *B. polymyxa* semble assez différent de ceux de *B. circulans* et *B. subtilis*.

L'expression du gène de l'endoglucanase de *B. polymyxa* dans *E. coli* a été obtenue par fusion de l'extrémité 5′ de la région codante à l'extrémité 5′ du gène *lacZ* α-peptide au site PstI de M13mp9 ou pUC9. Nous avons isolé un fragment d'ADN de *B. polymyxa* coupé par EcoRI de 2.35 kbp qui chevauche le fragment PstI de 1.6 kbp et qui contient la partie 5′ du gène de l'endoglucanase. Mais le gène reconstitué en fusionnant les parties 3′ et 5′ ne s'exprime pas dans *E. coli*. Le fragment PstI et la region non chevauchante de 1.5 kbp du fragment EcoRI ont été séquencés. Aucun promoteur, permettant une transcription dans *E. coli* n'à été détecté mais une zone de lecture, ORF-2, codant pour une hypothétique protéine de 187 acides aminés a été trouvée 347 paires de bases en amont du codon de départ de la transcription du gène de l'endoglucanase et semble d'après les données du séquençage être transcrite de facon divergente à celle du gène de l'endoglucanase. La séquence de la protéine ORF-2 est similaire à 28% avec la moitié amine terminale de la protéine *lac* represseur de *E. coli*. Le gène de l'endoglucanase contient une zone de 398 codons de lecture ouverte qui produirait théoriquement une enzyme de 44 kDa. Une délètion de seulement 12 codons à l'extrémité 3′ du gène supprime toute expression d'activité endoglucanasique dans *E. coli*.

La possible séquence d'amino-acides de l'endoglucanase de *B. polymyxa* a été comparée aux autres séquences de cellulases publiées. La séquence de l'endoglucanase de *B. polymyxa* présente une similarité de 34% avec celle de *Clostridium thermocellum celB* alors qu'elle est assez différente de celle de toutes les autres espèces de *Bacillus*. La comparaison des séquences des amino acides de huit endoglucanases (*T. reesei* EG2, *Schizophyllum commune* EG2, *Erwinia chrysanthemi celZ*, *B. subtilis* PAP115, *Bacillus* sp. strain no.1139, *Bacillus* sp. strain N-4 EG N2, *B.polymyxa* and *C. thermocellum celB*) présentent en commun la séquence Asn-Glu-Pro qui est proposée comme faisant partie du site actif. La région entourant cette séquence présente également une certaine similarité mise en évidence par la courbe d'hydropathie. La similarité des séquences de ces huit endoglucanases suggèrent qu'elles appartiennent à une même sous-classe d'enzymes qui correspondrait à une fonction ou un mécanisme enzymatique identiques.

To Christine

Acknowledgements

Before starting research in this laboratory I had very limited experience in molecular biology and a very limited knowledge of biology in general. I believe this thesis demonstrates that I have picked up a few concepts and techniques for which I have many people to thank. Foremost, I would like to thank Dr. Vern Seligy for allowing me the privilege of doing research at the National Research Council of Canada in his lab and for his caring attitude which he has exhibited to me over the years in words and actions. Thanks to Dr. Doug Johnson for his financial and personal support as well as always having the time to talk. I would also like to thank one of my first teachers of molecular biology, Mr. Michael Dove, whose ability for hard work cannot serve as a model because it is at an unattainably high level. Dr. Shahragim Tajbakhsh has helped me through the years that I have been at NRC in every way a friend can. I am also thankful for the help and discussion of other present and ex-lab-mates, Bob, Prakash, Mary, Ivan, Ron, Chusa, Gilles and everyone in Dr. G. Willicks lab. Many people at NRC have helped me do my work whose time I am also grateful for. Lise Bramall was responsible for the output of the hydropathy plots and the maintenance of the COMPMOLGEN programs. I thank Paul Berti for taking the time to run the SEQDP program for me. Dr. M. Yaguchi was kind enough to provide the unpublished amino acid sequence of *Schizophyllum commune* EG2. Dr. W. Sung synthesized the unique sequencing primers. I'd like to thank Dr. Peter Biely for the gift of OBR-cellulose. My appreciation for the help over the years goes to the Library staff and the people in Stores (i.e. Terry). I would also like thank my friends and family who have been concerned and supported me during my graduate work.

Table of Contents

List of Figures

List of Tables

Abbreviations

aa	amino acid(s)
ATCC	American Type Culture Collection
BRL	Bethesda Research Laboratories
Bluo-gal	halogenated indoyl-ß-D-galactoside
CBH	cellobiohydrolase
CMC	carboxymethylcellulose
CMCase	enzyme capable of endohydrolytic cleavage of CMC
CTAB	cetyl trimethyl ammonium bromide
dATP	2'-deoxyadenosine-5'-triphosphate
[α-^{32}P]dATP	2'-deoxyadenosine-5'-(α-[^{32}P])triphosphate
[α-^{35}S]dATP	2'-deoxyadenosine-5'-(α-[^{35}S]thio)triphosphate
dCTP	2'-deoxycytidine-5'-triphosphate
DEAE	diethylaminoethyl
dGMP	2'-deoxyguanosine-5'-phosphate
dGTP	2'-deoxyguanosine-5'-triphosphate
DNase I	deoxyribonuclease I
DNS	dinitrosalicylic acid
DP	degree of polymerization
dsDNA	double stranded deoxyribonucleic acid
dTTP	2'-deoxythymidine-5'-triphosphate
EDTA	ethylene diamine tetra acetic acid
EG	endoglucanase
EtBr	ethidium bromide
ETOH	ethanol
Eσ	RNA polymerase holoenzyme
F'episome	extrachromosomal DNA in *E. coli* required for pilus formation and infection by the phage M13
FPA	filter paper activity
g	gram(s)
HEWL	hen egg-white lysozyme
hr	hour(s)
IPTG	isopropyl-ß-D-thiogalactopyranoside
IU	International Units
kbp	kilobasepair
kDa	kilodalton
l	litre
L	Luria Bertani media
mCi	milli Curie
min.	minute(s)
ml	millilitre
M	molar
mM	milli molar
mm	millimetre
mMol	milli mole(s)
M13	single stranded DNA *E. coli* bacteriophage
NBRF	National Biomedical Research Foundation
ng	nano gram(s)
NRCC	National Research Council of Canada
nt	nucleotide(s)
°C	degrees centigrade
OBR	Ostazin Brilliant Red H3B dye

ORF	open reading frame, string of codons with no stop codons
ORF-2	558 nucleotide open reading frame upstream of *B. polymyxa* endoglucanase gene
pfu	plaque forming unit(s)
REN	restriction endonuclease(s)
RNA	ribonucleic acid
RNase A	ribonuclease A from bovine pancreas
rpm	revolutions per minute
SDS	sodium dodecylsulfate
SDS-PAGE	sodium dodecylsulfate - polyacrylamide gel electrophoresis; method for electrophoresing proteins under denaturing condition
SSC	saline-sodium citrate solution; 0.15 M NaCl, 0.015 M Na citrate
ssDNA	single stranded DNA
TBE	Tris-Borate-EDTA buffer
TE	Tris-EDTA buffer; 10 mM Tris, pH 8.0, 1 mM EDTA
Tris	Tris(hydroxymethyl)aminomethane
µg	micro gram
µl	micro litre
USDA	United States Department of Agriculture
UTR	untranslated region (of mRNA or gene)
w/v	weight per volume
2YT	Two times Yeast extract - Tryptone media
σ	RNA polymerase sigma factor

I. INTRODUCTION

The first observations of cellulolytic activity by an enzyme *in vitro* were reported at the beginning of this century (Seilliere, 1906; see review Goksøyr and Eriksen, 1980). Studies on the function of various cellulase enzymes was significantly escalated by the second world war when the United States military called for long range basic research on fungi responsible for deteriorating army equipment and supplies made of cotton material at outposts in the South Pacific (see review by Reese and Mandels, 1984). The resulting work, in particular by Elwyn Reese, Mary Mandels, Ralph Siu and K. Selby, laid the foundation of our present knowledge about cellulolytic fungi and cellulases, the various enzymes that degrade cellulose. The oil crisis of 1973 increased the research of cellulolytic organisms. The main objective was to apply these organisms in the conversion of lignocellulosic matter into alternate forms of energy to alleviate dependency on imported oil. Economic and environmental considerations should provide the impetus to continue research in this area given the massive quantities of cellulose waste now originating from household garbage, industry, forestry and agriculture. These concerns have inspired my research into the biology and some of the molecular characteristics of cellulase enzymes.

The first section of this introduction raises some questions about the desirability of our continued dependence on the oil industry as a source of transportation fuel. It also examines the need to produce an alternative liquid transportation fuel, such as ethanol, from cellulosic material. Cellulases are responsible for a key step in producing ethanol from cellulose, but they also have other beneficial uses which will also be explained. In subsequent sections, the sources and structure of cellulose will be described as well as a description of the sources of cellulases, the specific enzymes which are termed cellulases, and the proposed mechanism(s) of attack used by these enzymes to hydrolyse cellulose into simple sugars. The work in this thesis is mainly concerned with the molecular manipulation of cellulase genes which was done to gain more information about the enzymes for eventual application of cellulases in biotechnology. Three aspects of cloning

cellulase genes for biotechnological application will be discussed: (a) the number of different cellulase enzymes needed by an organism to fully digest cellulose; (b) the delineation of a catalytic site for possible protein engineering; and (c) the present status of cellulase gene expression in heterologous hosts. The final two sections of the introduction will describe cellulases from *Bacillus* sp., their importance and the objective of this thesis.

A. The Canadian Energy Situation

In Canada, approximately 25% of the oil used is imported (Statistics Canada, 1987). Foreign oil is used chiefly by the eastern provinces which do not have ready access to the western provinces' oil supplies. Therefore in spite of our rich oil resources we are dependent on foreign oil supplies. Canada exports a good proportion of its indigenous crude oil which still makes it a net producer. However, of the total revenue from the production of oil in Canada in 1986, 52% was controlled by foreign owned subsidiaries. This control by foreign interests rises to 60% if the downstream sector of the industry (i.e. refining, petrochemicals, distribution, etc.) is included (Petroleum Monitoring Agency, 1987). The net income of the foreign controlled companies in 1986 was over a billion dollars while the Canadian controlled enterprises lost over 3 billion dollars. Part of the reason for this large separation of net income between foreign and Canadian controlled oil producers was due to efforts to increase Canadian ownership. This move appeared to be effective since in 1982 about 75% of the industry had been under foreign control. Unfortunately, most of the financial burden of increased Canadian control of the industry was borne by Canadian tax payers (MacDonald, 1982). Much of the money Canadian consumers use for oil products will go to foreign interests because of the regional disparity of oil production and the controlling foreign interests in the Canadian oil field.

The transport sector is the single largest oil consumer. Road vehicles account for 80 % of the consumption in this sector. (Statistics Canada, 1987). One solution to the problems associated with Canadian oil-dependency is to find an alternative transportation fuel which is indigenously produced and controlled in all regions of this country. In Canada

several substitutes have already been examined which include: alcohols (ethanol and methanol), liquified natural gas, propane, and hydrogen (Special Committee on Alternative Energy and Oil Substitution, 1981). Each of these fuels burns much more efficiently and pollutes less than the present fuels used. Of all these substitutes, the alcohols are the only ones that can be produced in all regions of the country from either coal, natural gas or biomass.

While the production costs of methanol from coal or natural gas allow this fuel to be competitive with gasoline, the most optimistic studies indicate that the production of alcohol from lignocellulosic matter is only marginally economical. Several pilot plants with processes that enzymatically degrade cellulose have been built to produce ethanol and other products from lignocellulosic substrates. The wholesale cost (Can. dollars, 1981) of a litre of ethanol from these and other proposed plants varies from 23.4 cents to 88.6 cents (Stone et al., 1982; Emert et al., 1980; Clausen and Gaddy, 1983; Parker et al., 1981; Huff, 1981). The 1987 net price of a litre of gasoline from Canadian refiners averages 29.0 cents (Petroleum Monitoring Agency, 1988). However, these comparisons are inconclusive as they do not take into account government subsidies of the industry or the volatile price of crude oil. Present day crude oil prices have dropped ~45% since 1985 and are now lower than the average price of crude oil in 1981. These large price swings are not the result of free market forces but more the result of oil cartels and world-political forces.

There is also a question about whether the net energetics of producing ethanol from cellulose are positive (Jawetz, 1981) or negative (Lewis, 1977). This must also be considered.

B. Commercial Uses of Cellulases

The cellulose hydrolysis process can be commercially beneficial for the manufacturing of products other than alcohol. Many types of flora are important producers of drugs and chemical compounds in the form of naturally synthesized secondary metabolites. Many drugs today are chemically synthesized derivatives of plant compounds.

The liberation of proteins or compounds produced by plants may become more important as large scale plant cell cultures are developed for production of these products (Fowler, 1983). The reuse of the plant cell biomass from cell cultures would also require the breakdown of the cellulose in the cell walls (Smith et al., 1985). An example of the extraction of compounds from another type of plant material is described by Szakacs-Dobozi et al. (1988) where mustard seeds are pretreated with cellulases to increase the yield of mustard oil which contains the commercially important allyl isothiocyanate.

The most recent list of applications of cellulases and cellulolytic organisms was made by Mandels (1985). It included the production of soluble sugars as substrates for the growth of organisms used in the fermentation industry or for growing yeasts as a 'single cell protein' for animal feeds. Enzymatic removal of cellulose from wood can produce high quality lignin for use as a chemical raw material in the manufacturing of adhesives and resins (Muller and Glasser, 1983). In the food industry cellulases and cellulolytic organisms could be used in the production of foodstuffs and more recently they have been examined as way of converting apple waste to microbial biomass utilizable in animal feed (Friedrich et al., 1987). Preliminary experiments have also been done with the aim of converting waste cellulose into edible oil products using the cellulolytic fungi *Trichoderma reesei* (Brown and Hasan, 1988).

Cellulases have been produced commercially for the food industry in Japan for at least 20 years (Toyama, 1969; Yamada, 1977) and elsewhere for the scientific market. The technology for using cellulases exists and only depends upon the economics of the application. A list of pilot plants designed for cellulose hydrolysis and fermentation can be found in the review by Coughlan (1985).

C. Cellulose

1. Sources

As has been pointed out in many papers discussing world carbon sources, cellulose is the most abundant organic carbon molecule on the planet. It is produced mainly by

plants. Since cellulose is the major structural component of plant cell walls it is also very resistant to degradation. It is encased by other more soluble sugar polymers called hemicellulose and the second most important structural component of plant cell walls, the complex polymer containing phenol-propane units termed lignin. Woody tissue like stems and roots have the most cellulose and leaves have the least. The purest form of natural cellulose is found in the seed hair of plants of the genus *Gossypium* (cotton plants). Plant cellulose is obviously very significant for the pulp and textile industries. Its preservation is thus important for these and other applications.

Other sources of cellulose include some species of lower eukaryotes and prokaryotes (Haigler, 1985). The exoskeletons of invertebrates consist of a modified cellulose, chitin which is a linear polymer of N-acetyl-D-glucosamine connected by ß-1,4 linkages. However, some marine organisms of the *Tunicata* class have a form of cellulose in their outer mantles called Tunicin (Nevell and Zeronian, 1985; Warwicker, 1985). Fibril formation of cellulose produced by bacteria (i.e. *Acetobacter* sp.) has been studied to gain a greater understanding of native cellulose structure (Haigler, 1985). These bacteria have even been considered as industrial producers of cellulose since their product would be free of lignin (Nevell and Zeronian, 1985).

2. Structure

Cellulose is a polymer of glucopyranose rings in the chair configuration attached by ß-1,4-linkages. A survey on the measured degree of polymerization (DP) of various sources of cellulose reported a range from 15 to 14,000 glucose units (Cowling and Brown, 1969). In plants a number of cellulose strands or fibrils are found grouped together in a crystalline form called microfibrils. These microfibrils are oriented in different directions to the axis of the plant cell dependant upon which layer of the cell wall they occur. In the inner and outer layers of a plant cell wall the cellulose microfibrils are covered with lignin blocking any attack from cellulolytic enzymes.

The lack of solubility of cellulose is another major problem in the efficient use of

cellulolytic enzymes. ß-1,4 linked glucose polymers with only a DP of 8 are insoluble (Matheson and McCleary, 1985). The pyranose rings are rotated 180° relative to the next attached glucose unit. This produces a ribbon like structure which can strongly hydrogen bond to neighbouring cellulose chains in an ordered crystalline fashion. In contrast both the α-1,4 linkage of glucose monomers in amylose and the ß-1,3 linkage of glucose in laminarin produce a helical conformation which increases water solubility (Matheson and McCleary, 1985). Amylose, which is considered a carbohydrate storage molecule for plants, can be broken down more easily than cellulose because of its increased solubility.

The strong intermolecular bonding of neighbouring cellulose strands and ordered structure that it creates makes the interstrand space too small for proteins to diffuse into the microfibril. The cellulose strands in a microfibril are somewhere between 5-10Å apart. There is a variety of evidence which suggests that the cellulose surface must be accessible to solutes of 40 to 60 Å in order to allow for enzyme attack (Ladish et al.,1983). This corresponds to a protein around 60 kDa. in size. If a typical cellulase protein has a size range of 35-160Å in diameter (Ladish et al.,1983), the only cellulose accessible to enzymes would be on the outside of the microfibril.

Presently there is still controversy as to whether the cellulose chains are in parallel or anti-parallel orientation to the neighbouring chains on the same plane (Preston, 1986). Cellulose strands are in an anti-parallel configuration when the pyranose rings of one strand are in 180° orientation relative to the pyranose rings of the neighbouring strands. It is accepted that dissolved cellulose recrystallizes in anti-parallel strands, but may not necessarily exist as such in native cellulose (Preston, 1986). Knowledge of this structural aspect is important for the understanding of degradation by cellulases of native cellulose as it is believed that one reason for the multiplicity of cellulase enzymes produced by many organisms is to overcome the different stereospecific structures generated by these different chain orientations (Wood, 1985; Wood et al., 1988)

D. Cellulases

Several reviews have been written on enzymatic hydrolysis of cellulose (Bisaria and Ghose, 1981; Chahal, 1982; Eriksson, 1982; Ladish et al., 1983; Finch and Roberts, 1985; Wood, 1985). There have also been several reviews about cellulases (Goksøyr and Eriksen, 1980; T.-M. Enari, 1983; Coughlan, 1985; Coughlan and Ljungdahl, 1988), cellulolytic organisms (Ljungdahl and Eriksson, 1985), thermophilic cellulolytic organisms (Margaritis, 1986), the regulation of cellulolytic activity (Stutzenburger, 1985) and the cloning of cellulases (Béguin et al., 1987). The 14 reviews cited here are only a partial list.

1. Sources

Cellulases are produced by a variety of organisms including bacteria, actinomycetes, fungi, algae and plants (Finch and Roberts, 1985; Coughlan, 1985; Wood, 1985; Wong et al., 1977). Examples of cellulases produced by eukaryotes, other than fungi, like protozoa and metazoon such as nematodes, anthropods, molluscs, echinoderms, and annelids, are rare or questionable. The problem has been in determining whether the cellulases are produced by the organism per se or some symbiotic microbe located in the host's gut. The presence of cellulolytic bacteria and fungi in the intestinal flora of ruminants and other herbivores as well as in the lower intestine of humans appears to be beneficial (Ljungdahl and Eriksson, 1985; Wolin and Miller, 1983; Wedkind et al., 1988). It has long been accepted that the protozoan from the gut of termites is cellulolytic but studies on pure cultures of these organism are still termed "putatively axenic cultures" (Odelson and Breznak, 1985). Shipworms of the Teredinidae previously thought to be cellulolytic have been found to contain a symbiotic nitrogen-fixing cellulolytic bacterium (Coughlan, 1985). However, several other organisms, the crustacea of the *Limnoria* genus, certain nematodes, and silverfish are believed to produce cellulase activity in absence of other microorganisms (Finch and Roberts, 1985). The molecular isolation and characterization of genes responsible for the cellulase activity from organisms should definitively establish the source of these enzymes without the problems of attaining live

axenic cultures.

There are three general groups of fungi which hydrolyze wood in nature. The first is the white-rot fungi which degrade both the cellulose and lignin in wood at an equal rate. Some white-rot fungi will degrade preferentially the lignin. The decay pattern caused by these fungi is now termed white pocket rot (Otjen and Blanchette, 1984). The degradation takes place near the hyphae of the fungi, particularly the growing tips. The second and third groups; soft-rot and brown-rot fungi, degrade only the carbohydrate portions of the wood, the cellulose and hemicellulose. The hyphae of these fungi grow into neighbouring cells through the lumen of the wood cells and holes which naturally occur in wood cell walls called pits and capillaries. The secondary layer of the wood cell which does not contain much lignin is degraded and eventually the cell collapses. The brown-rot fungi in contrast to the other two groups, seem to produce a diffusible substance which contributes to the degradation of cellulose at a distance from the hyphae and also do not seem to produce an exoglucanase (Ljungdahl and Eriksson, 1985).

2. Mechanisms of Attack

As already mentioned there are several classes of enzymes which make up the cellulase enzyme system. The first is generally referred to as an endoglucanase or more correctly endo-1,4-ß-D-glucan 4 glucanohydrolase (EC 3.2.1.4). This enzyme attacks randomly at regions having amorphous cellulose structure or soluble cellulose derivatives like carboxymethylcellulose (CMC) and is sometimes called CMCase. The random cutting by a true endoglucanase can be measured by its ability to lower the viscosity of a solution containing a soluble cellulose derivative like CMC or hydroxyethylcellulose. The exoglucanase or 1,4-ß-D-glucan cellobiohydrolase (EC 3.2.1.91) in contrast will not cut internally in the CMC chain and can only cleave off cellobiose units available at the non-reducing end. Modified glucose moieties in a cellulose strand with groups attached cannot be used as a substrate for either of the glucanases (Reese et al., 1950) which means very little hydrolysis of CMC can be done by an exoglucanase before the enzyme is stopped.

CMC used for the analysis of cellulases usually have a substituted group every 7th glucose unit or less often, this allows enough space for an endoglucanase to bind and hydrolyze the cellulose. Exoglucanases have the ability to produce reducing sugars from crystalline substrates like the commercially prepared Avicel which is susceptible to only minimal attack by endoglucanases. The ability to attack one of the two substrates is used most often in defining an isolated cellulase, but this may become inappropriate as more information is presented about enzymes that cannot be properly categorized by these definitions (i.e. CBH I of *T. reesei*, see Coughlan and Ljungdahl, 1988; Hayashida and Mo, 1986). One other enzyme which is not often found but can be termed a cellulase is the glucohydrolase or 1,4-ß-D-glucan glucohydrolase (EC 3.2.1.94). This enzyme is very similar to cellobiohydrolase (CBH) in that it attacks the cellulose strand in an "exo" fashion, but instead of releasing cellobiose units it releases glucose units.

There are several other enzymes which also play some role in cellulose hydrolysis by removing oligosaccharide end-products which might inhibit the further action of cellulases. ß-glucosidases (ß-D-glucoside glucohydrolase, EC 3.2.1.21) will hydrolyze cellobiose or cleave off glucose units from oligosaccharides. This enzyme is but one way to change cellobiose into a utilizable form for an organism. There are several other enzymes which do not hydrolyze cellobiose but are steps in alternative pathways for utilizing the resulting soluble oligosaccharide products of cellulose hydrolysis. The enzymes cellobiose phosphorylase and cellobiose epimerase are produced by some cellulolytic organisms instead of ß-glucosidase. In order to help in the simultaneous degradation of lignin, white-rot fungi transfer electrons from cellobiose to lignin or lignin degradation products during its oxidation by cellobiose dehydrogenase.

The first major theory on the mode of attack by cellulases was put forward by Reese et al. (1950). This theory designated one component, C_1, of the cellulase complex as nonhydrolytic. The C_1 enzyme would first attack the cellulose and make it more susceptible to hydrolysis by the subsequent component C_X. This theory developed from

the observation that some culture filtrates would hydrolyze native cellulose while others were able to only hydrolyze soluble cellulose derivatives like carboxymethylcellulose (CMC).

$$\text{CELLULOSE} \xrightarrow{C_1} \text{REACTIVE CELLULOSE} \xrightarrow{C_X} \text{CELLOBIOSE} \xrightarrow{\beta\text{-glucosidase}} \text{GLUCOSE}$$

This theory is not altogether accepted as the mode of cellulase action today. The C_1 protein was suggested to have been a "hydrogen bondase" that would disrupt hydrogen bonds between parallel chains of cellulose so the C_X enzyme would be free to hydrolyze the separated cellulose strand (Siu, 1951). What is thought to be the C_1 enzyme has been isolated and is now described as a ß-1,4-exoglucanase protein (Wood and McCrae, 1977). Sometimes the C_1 labelling is still used to describe this category of protein even if it is accepted that it is not the initial step.

The next general theory describing the mode of action of cellulases was put forward by Eriksson (see discussion of Leatherwood, 1969). This theory was modelled on what was known at the time about the cellulase systems of the white-rot fungi and the soft-rot fungi *T. reesei*. The general mode of attack has the endoglucanase cleaving randomly in amorphous sections of the cellulose microfibrils leaving substrate for the exoglucanase to remove small oligosaccharides which would be in turn hydrolyzed to glucose by the ß-glucosidase. In this theory the endo and exoglucanase work in a synergistic manner, that is both enzymes together hydrolyze more cellulose than if both enzymes worked alone. Neither of these two initial theories seem to encompass the data that now exists about the many different cellulolytic organisms and newly isolated components involved in cellulose hydrolysis.

The most current mechanism put forward on the hydrolysis of cellulose by cellulase systems incorporates the work of many groups and is described in two excellent reviews (Coughlan, 1985; Coughlan and Ljungdahl, 1988). The first step, echoing the theory of Reese, is the 'amorphogenesis' of the crystalline cellulose. Several non-hydrolytic factors

have been isolated which are believed to render the cellulose more susceptible to enzymatic attack. These factors include: the iron-containing microfibril generating factor found in *T. reesei* (Griffin et al., 1984); H_2O_2/Fe produced by brown-rot fungi (Koenigs, 1975) and decaying plants (Miller, 1986); the Yellow Affinity Factor (YAS) and S_L component of *Clostridium thermocellum* which bind to the cellulose and promote a greater adsorption of enzymes to the substrate (Ljungdahl et al., 1983; Wu and Demain, 1988). Binding factors may be common to other cellulolytic organisms (Lamed et al., 1987). The carboxy portion of an endoglucanase from *Humicola grisea* has been shown by electron microscopy to bind strongly to Avicel and split off layers of strands from the microfibril without producing anymore reducing ends (Hayashida and Mo, 1986). Similarly acting protein domains have also been reported in the cellobiohydrolase of *Aspergillus ficum* (Hayashida et al., 1988) and possibly CBH I of *T. reesei* (Tilbeurgh et al., 1985) and the cellulases of *Cellulomonas fimi* (Miller et al., 1988). The function of these domains are not easily testable and therefore may or may not be commonly found in cellulase systems.

The next stage of cellulose degradation is done by the endoglucanases and exoglucanases with some synergistic effect. More than one type or subclass of endoglucanase and exoglucanase may be needed for the different stereospecific forms of cellulose that present themselves to the enzymes, this is summed up by Wood et al.(1988). The resulting oligosaccharides are then cleaved to glucose units, phosphorylated or oxidized by the enzymes already mentioned.

Each different organism may contain one or another of the components mentioned in their cellulase system. It is obvious that all of the different mechanisms used are not yet fully understood as well defined exoglucanases have not been isolated for brown-rot fungi and cellulolytic bacteria. The exoglucanase isolated from the bacteria *Cellulomonas fimi* is active against the substrate CMC which is used to select for endoglucanases (Miller et al., 1988).

3. Multiplicity of Enzymes

Almost all cellulolytic organisms studied to date produce a multiplicity of different cellulase enzymes. This has been especially noted with endoglucanases where species of *Trichoderma* produced at least 6 different extracellular forms (Beldman et al., 1985; Sheir-Neiss and Montenecourt, 1984; Labudova and Farkas, 1983). It has been argued that some of the different enzymes could be coded by one gene. So far only two endoglucanase genes have been isolated from *T. reesei*. Some of the proteins detected could be isoforms produced by a different amount of post translational processing like glycosylation and proteolytic cleavage or poor purification of enzymes which have formed a complex. This has been shown to be the case by Willick and Seligy (1985), and by Paice et al. (1984) with endoglucanases from *Schizophyllum commune* as well as with enzymes from *T. reesei* (Wood et al., 1988). With the increase in molecular isolation of many cellulase genes, it appears many fungal and bacterial organisms have more than one genetically distinct endoglucanase and/or exoglucanases. Some examples of these organisms are shown in Table 1.

Table 1. List of organisms with multiple cellulase genes.

Organism	Distinct genes isolated	Reference
T. reesei	2 endoglucanases	Penttilä, et al., 1986
	2 cellobiohydrolases	Niku-Paavola, et al., 1985
C. thermocellum	6 endoglucanases	Béguin et al., 1988
Clostridium cellulolyticum	2 endoglucanases	Faure et al., 1988
Bacteroides succinogenes	6 endoglucanases	Crosby et al., 1984.
Bacillus sp.N-4	3 endoglucanases	Kim et al., 1987.
		Sashihara et al.,1984

As already mentioned, it has been postulated that the different forms of similarly acting enzymes are necessary to accommodate the different steric positioning of the ß-1,4 glycosyl linkages in native cellulose (Wood, 1985; Wood et al., 1988). The catalytic mechanism might remain the same, but the binding of the enzyme would have to be

different.

Genetic multiplicity may prove to be more the rule than the exception as more isolation of genes takes place. *Bacillus subtilis* and related species appear to be the exceptions as some of the few microbes that have only one cellulase gene. Five groups independently have cloned only one endoglucanase from several strains of *B. subtilis*. Four of the genes have a very high DNA sequence homology (see Lee and Pack, 1987 and Results Section G2) and no other cellulases have been yet detected by any of the groups (Seligy et al.,1987; Koide et al., 1986; Robson and Chambliss, 1986; Lee and Pack, 1987). The fifth group, Sharma et al.(1987) have cloned a gene whose restriction endonuclease map is very similar to the genes already cloned by the other four laboratories.

4. Active Sites

It has been proposed that the catalytic mechanisms used for enzymatic cleavage of ß-linkages of glycoside polymers might all be similar (Vernon and Bank, 1963). For this reason, the well studied lysozyme enzyme has been used as a model for cellulases, as well as for amylases and xylanases (Paice and Jurasek, 1979; Yaguchi et al., 1983). Lysozyme cleaves the ß-1,4-linkage in polymers of alternating units of N-acetylmuramic acid and N-acetylglucosamine, the major polysaccharide of many bacterial cell walls. In the general model for the action of lysozyme, the enzyme introduces strain on the conformation of the glucopyranose rings which in turn lowers the activation energy for hydrolysis of the connecting linkage. A glutamic acid residue (Glu 35 in hen egg-white lysozyme, HEWL) donates a proton to the linking oxygen while a negatively charged aspartic residue (Asp 52, HEWL) stabilizes the C-1 atom when the electrons shared with the oxygen are pulled away. The reaction is completed with the addition of an -OH from water to the C-1 atom (Zubay, 1983). Yaguchi et al. (1983) showed some similarity between hen egg-white lysozyme and the N-terminal amino acid sequence of an endoglucanase from *S. commune*. Later kinetic data (Clark and Yaguchi, 1985) demonstrated that carboxyl groups took part in the cleavage of ß-1,4-linkages but their exact location in the protein was not determined.

Comparison of amino acid sequences of CBH I from *T. reesei* and lysozyme from bacteriophage T4 suggested that Glu-65 and Asp-74 of CBH I might be the analogous catalytic residues (Paice et al., 1984). However, the subsequent sequence comparison of EG I of *T. reesei* to CBH I, while showing over 50% homology between the two amino acid sequences, did not show a conservation of these two sites (Pentillä et al., 1986). The exact location of the active site is not yet known for any cellulase.

E. Heterologous Expression

Most of the cellulase genes already isolated have been expressed to some degree in a heterologous host. The heterologous hosts are different for bacterial and fungal genes. No fungal genes have been expressed in *E. coli*, only in yeasts. All of the bacterial cellulase genes so far cloned have been expressed in *E. coli*. Host genetic compatibility has eliminated the need for extensive genetic engineering with bacterial cellulase genes. Two reasons behind the development of heterologous expression of cellulases are: 1) increased production rates of the enzymes and 2) broadening the substrate range and application of the heterologous host.

Even though cellulases are considered to be secondary catabolic enzymes, their production by fungi probably represents the highest titers of extracellular proteins by living organisms in culture. Levels of over 20 g/l of proteins are commonly measured from cultures of hyperproducing cellulase mutants of *T. reesei* (Saddler, 1986; Allen, 1983). The choice high-level producer of heterologously expressed higher eukaryotic gene products are the insect cell lines infected with Baculovirus derivatives which produce up to 1 g/l of a specific protein inside the cells (Lucknow and Summers, 1988). From bacteria, mutants of *B. subtilis* have produced titers up to 2.5 g/l of a specific extracellular protein, a-amylase (Yoneda, 1982). The rate of cellulase production by fungi facilitated by host mutation may have been brought to its highest level. One limiting factor that still exists is the slow growth rate of the fungi. Although bacteria divide much faster than fungi, cellulolytic bacteria do not seem to secrete a full cellulase system into the culture supernatants. *C. thermocellum*

will produce a culture filtrate which can degrade crystalline cellulose but only during the stationary phase of growth (Ljungdahl et al., 1983). Such a result may be a sign of cell lysis and not of the secretion of all the cellulases. This is probably one of the reasons why very little work has been done on the production of cellulases from bacteria. Therefore, there is a lack of data to compare with the extensive bioengineering work done with *T. reesei*. A survey of production rates of hyperproducing mutant strains of *T. reesei* lists levels of 2.6 to 125 IU of filter paper activity (FPA) per litre per hour (Ghose and Ghosh, 1979; Saddler, 1986). Some similar data does exist for the production of endoglucanases from bacteria. In *T. reesei* strain RL-P37 rates of 2910 IU $l^{-1}hr^{-1}$ of CMCase activity and 108 IU $l^{-1}hr^{-1}$ FPA were reported by Sheir-Neiss and Montenecourt (1984). These rates compare with the highest rates of production by fungi mentioned already. In contrast the production rate of the CMCase of the cellulolytic bacterium *Sporocytophaga myxococcoides* was measured at 6000 IU $l^{-1}hr^{-1}$ (Vance et al., 1981). This bacterium was not tested for FPA and may not be representative of cellulase production by bacteria as a *B. subtilis* endoglucanase gene cloned into *Bacillus megaterium* only produced 170 IU $l^{-1}hr^{-1}$ of CMCase activity (Lee and Pack, 1987). If an accurate comparison of rates of production is to be made, the specific amounts of endoglucanases being produced and their specific activities must be known. The problem of making comparisons like this is often compounded because the units of activity used from different laboratories is not equivalent. The International Energy Association (IEA) is presently trying to find consistent and standard assay procedures to be used internationally for measuring amounts of the protein and enzymatic activity.

The rate of production is not important if a heterologous host uses the foreign gene to allow its use of a new substrate (i.e. cellulose). The host organism could immediately remove the end-products of cellulose degradation that cause end-product inhibition of cellulase activity (Woodward and Arnold, 1981). This should result in less cellulase needed to be produced for the same amount of hydrolysis. As stated earlier in section A and B, previous objectives of cellulase research were in using the end-product sugars for alcohol

fermentation. Fermentative organisms with cellulases should be able to use the simple sugar end-products from cellulose hydrolysis as they are produced. The best studied organisms for fermenting sugars into alcohol are yeasts from the genus *Saccharomyces*. The industrially important "common baker's yeast" *Saccharomyces cerevisiae* has been used to express several ß-glucanases and they are listed in Table 2.

Table 2. Cellulases expressed by transformed *S. cerevisiae* strains.

Enzyme	Genetic origin	Reference
ß-1,3-1,4-endoglucanase	*B. subtilis*	Hinchliffe and Box, 1984
ß-1,4-endoglucanase A	*C. thermocellum*	Sacco et al., 1984
ß-1,4-endoglucanase cenA	*Cellulomonas fimi*	Skipper et al., 1985
ß-1,4-endoglucanase	*B. subtilis*	Seligy et al., 1987
ß-1,4-endoglucanase EGI	*T. reesei*	Pentillä et al., 1986
ß-1,4-endoglucanase EGI	*T. reesei*	Van Ardsell et al., 1987
cellobiohydrolases, CBHI CBHII	*T. reesei*	Pentillä et al., 1988
ß-1,4-endoglucanase cenA ß-1,4-exoglucanase cexA	*C. fimi*	Wong et al., 1988

The studies on heterologous-expression of cellulase genes in this yeast provide important information on the secretory capabilities of the yeast as well as its biotechnological potential. Only the transformed *S. cerevisiae* produced by Wong et al. (1988) has been shown to hydrolyze cellulose from pretreated wood but this new organism did not hydrolyze the cellulose substrate as well as a normal cellulolytic microbe even though it could also ferment the end-products. It should be stressed that *S. cerevisiae* is only a moderate producer of extracellular enzymes to begin with and that it is one of many organisms which can ferment sugars into ethanol. There are also many other organisms that are able to carry out commercially important fermentations resulting in end-products such as citric acid, butandiol, acetic acid, etc, which are not produced appreciably by or at all *S. cerevisiae*.

Other organisms used to express cloned cellulase genes are given in the following examples. The yeast *Schwanniomyces alluvius* which already has both a cellobiase and an

amylase system was transformed with an *B. subtilis* endoglucanase gene (Seligy et al., 1987). An endo-ß-1,4 glucanase gene from *B. subtilis* has also been transferred into *Zymomonas anaerobia* in order to increase its substrate range (Yoon et al., 1988). The gene apparently was not transcribed well and was not secreted by *Z. anaerobia* in contrast to the 120 fold higher expression in *E. coli*. Due to problems isolating *Thermomonospora fusca* cellulase genes in *E. coli*, Ghangas and Wilson (1987) examined *Streptomyces lividans* as a possible cloning host. It was thought that some of the *T. fusca* sequences were lethal to *E. coli*. A previously isolated *T. fusca* endoglucanase gene (Collmer and Wilson, 1983) was transferred to *S. lividans* and successfully expressed (Ghangas and Wilson, 1987). Although *S. lividans* previously could grow on cellobiose, the recombinant organism could not grow on cellulose. Two other reports of heterologous expression of cellulase genes are: *C. thermocellum* genes being expressed in *Streptomyces albus* (Romaniec et al, 1988); and an endoglucanase gene of *T. reesei* in an *Aspergillus* species (Barnett and Shoemaker, 1988).

F. Bacilli and Cellulase

There has been no well-documented case of a *Bacillus* sp. that produces a cellulase enzyme complex capable of degrading native cellulose, for this reason bacilli are not considered to be truly cellulolytic. However, a number of bacilli have been shown to have endoglucanase activity. Including the five separate genetic isolations of endoglucanases from *B. subtilis* already mentioned in section D.3, there are a total of 12 endoglucanase genes that have been isolated from *Bacillus* species (see review by Cantwell et al., 1988). Two of these will be described in greater detail later in this study. Other than *B. subtilis* and the newly isolated unclassified *Bacillus* species, there have been reports of 6 other bacilli that have cellulolytic activity. The species are: *B. licheniformis* (Dhillon et al., 1985; Knösel, 1971), *B. polymyxa* (Greaves, 1971; Fogarty and Griffin, 1973a), *B. cereus* (Thayer, 1978) and *B. circulans* (Buchanan and Gibbons, 1974), *B. brevis*, *B. firmus*, and *B. pumilus* (Knösel, 1971). In each case, the activity was reported to be solely endo-cellulolytic.

Although bacilli do not produce a complete cellulolytic enzyme complex, there are

at least two advantages for considering the use of *Bacillus* cellulases for industrial applications. The specific activity of an endoglucanase from *Bacillus subtilis* DLG was compared to that of the extensively studied endoglucanase activity of *T. reesei* QM9414 and found to be superior on a noncrystalline cellulosic substrate, trinitrophenyl-carboxymethyl cellulose (Robson and Chambliss, 1984). Lo and coworkers (1988) have made similar findings with the *B. subtilis* PAP115 endoglucanase when compared to the fungal *S. commune* endoglucanase, EG1. Secondly, none of the *Bacillus* enzymes studied appear to be glycosylated, which does not appear to be the case for enzymes from the cellulolytic bacteria *C. fimi* (Langsford, 1984). This makes the heterologous expression of these enzymes adaptable to many systems. In contrast, all fungal cellulases studied to date have been found to be glycosylated (Coughlan, 1985). The pattern of glycosylation has not been faithfully repeated in heterologous hosts. Considerable work is still required to understand the need and function of glycosylation of these enzymes.

G. Research Objectives

Previous to the study described in this thesis a ß-1,4 endoglucanase gene from *B. subtilis* PAP115 had been isolated by researchers at the Pulp and Paper Research Institute of Canada (Seligy et al., 1987) and characterized at the National Research Council of Canada (MacKay et al., 1986; Seligy et al., 1987; Lo et al., 1988). These results along with the studies on the two alkalophilic *Bacillus* sp. no.1139 and N-4 (Horikoshi and Fukumori, 1988) suggested that all *Bacillus* endoglucanase genes were related. To study this aspect further I looked at 15 other *Bacillus* species for endoglucanase activity. DNA-DNA hybridization was also used to see if any of the *Bacillus* sp. had genes with DNA homology to the PAP115 endoglucanase gene. As a result of this work, new endoglucanase genes from *B. circulans* and *B. polymyxa* were cloned using *E. coli* as the expression host. Characterization of the DNA of these genes showed that the *B. circulans* and *B. subtilis* genes were very similar, but the *B. polymyxa* gene was clearly different.

The major objective of the characterization work presented here was the

sequencing of the endoglucanase gene from *B. polymyxa* since it had shown no sequence homology by DNA-DNA hybridization to the *B. subtilis* gene and therefore possibly represented a unique endoglucanase enzyme in the *Bacillus* genus. The DNA sequence of the *B. polymyxa* endoglucanase gene was examined for regulatory sequences and the putative amino acid sequence was compared for similarity with all available (published) amino acid sequences for cellulase enzymes to date. At this time no catalytic residues of a cellulase have been determined. An alignment was made with eight of the cellulase amino acid sequences to find residues that may have been conserved. The consensus sequences so obtained, when examined with the data on the catalytic mechanism modelled after that of lysozyme, suggest that some of the residues have possibly been conserved for function.

II. MATERIAL AND METHODS

A. Media

Bacillus strains were grown in Luria Bertani (L) broth containing 10 g Bacto-Tryptone, 5 g Yeast extract, and 10 g NaCl per litre (Maniatis et al., 1982). *E.coli* was grown in liquid cultures of L or 2YT (16 g Bacto-Tryptone, 10 g Yeast Extract, 2 g NaCl per litre) (Messing, 1983) and on plates of B media containing 10 g Bacto-Tryptone, 8 g NaCl, and 1 ml of 1% vitamin B_1 solution per litre (Messing, 1983).

Plates were made by adding 20 g/l agar to liquid medium. Top agar for plating M13 phage was made from liquid medium supplemented with 5 g agarose/l. Substrates of CMC and Ostazin Brilliant Red H3B dyed (OBR) cellulose were added to plates or top agar at a concentration of 0.1% w/v. OBR cellulose was added to the top agar only. Ampicillin was added to media at a concentration of 35 µg/ml for the selection of cells containing pUC derived plasmids.

Experiments requiring minimal media used, in g/l : KCl 1.0; $MgCl_2 \cdot 2H_2O$, 0.2; NaH_2PO_4, 5.4; $Na_2HPO_4 \cdot 2H_2O$, 7.0; $CaCl_2 \cdot 2H_2O$, 0.25; $FeSO_4 \cdot 7H_2O$, 0.0005; $MnSO_4 \cdot 4H_2O$, 0.001; and Bacto-peptone, 10.0 (Fogarty and Griffin, 1973b) or, where stated, 1% (w/v) Yeast Nitrogen Base supplemented with 1% (w/v) of various carbon sources.

B. Bacterial Strains and Vectors

The *Bacillus* strains used in this thesis are listed in Table 3 in the Results, section A.

The following *Escherichia coli* strains were used for cloning or expression: DH5αF′ [F′,*end*A1, *hsd*R17(r_k^-,m_k^+), *sup*E44, *thi*-1, lambda$^-$, *rec*A1, *gyr*A96, *rel*A1, Δ(*lac*ZYA-*arg*F), ϕ80*dlac*ZΔM15] (see Jeese, 1986 for DH5α); HB101 [F$^-$, *hsd*R514 (r_b^-,m_b^-), *rec*A13, *ara*-14, *pro*A2, *lac*Y1, *gal*K2, *rps*L20 (Sm^r), *xyl-5*, *mtl*-1, *sup*E44, *lambda*$^-$] (Maniatis et al., 1982); JM101 [Δ*lacpro*, *sup*E, *thi*, F′ *tra*D36, *pro*AB, *lac*I^q ZΔM15] (Messing, 1983); and JM103 [Δ*lacpro*, *sup*E, *thi*, *str*A, *sbc*B15, *end*A, *hsp*R4, F′ *tra*D36, *pro*AB, *lac*I^q ZΔM15]

(Messing, 1983).

E. coli cell lines were stored as described by Hanahan (1985). Cells carrying the F′ episome were grown on M9 plates before preparation for storage at -80°C. M9 media is made up by dissolving 6 g of Na_2HPO_4, 3 g of KH_2PO_4, 0.5 g of NaCl, and 1 g of NH_4Cl in 978 ml H_2O, 20 g of agar is added and the mixture is autoclaved. To this is added the following sterilized solutions: 1 ml of 1 M $MgSO_4 7 H_2O$, 10 ml of 20% glucose, 1 ml of 1% vitamin B_1, and 10 ml of 0.01 M $CaCl_2$ (Messing, 1983).

The vectors used from the M13 series were: mp9 (Vieira and Messing, 1982), mp18, and mp19 (Yanish-Perron et al., 1984). From the pUC series of plasmids, pUC 8 and pUC 9 were used (Vieira and Messing, 1982).

C. DNA Isolation and Purification of DNA from Various Sources

1. Genomic DNA from *Bacillus*

Isolation of genomic DNA was performed using the phenol extraction method of Saito and Miura (1963). RNA was removed from the DNA preparations with ~3 µg RNase A/µg DNA; RNase A was prepared according to Maniatis et al. (1982). DNA was recovered by precipitation with two volumes of ethanol.

2. Plasmids and Replicative Form (RF) DNA of M13

Plasmids and the replicative forms of M13 were purified using an alkaline lysis method described by Maniatis et al. (1982). Cells were grown 10 to 20 hr using 1 to 10 ml cultures. Proteins were precipitated after the lysis by using 5 M K acetate, 3 M acetic acid.

3. Single stranded DNA from M13 phage

Single stranded DNA from M13 phage was prepared by two methods depending on the end use of the preparation. Except for sequencing, ssDNA was prepared from cultures of 1.5 to 10 ml grown from 8 to 20 hr as described by Carlson and Messing (1984). Phage were precipitated from culture supernatants for 20-30 min. at 4° C. with one tenth volume of 20% PEG and either 2.5 M Na Acetate or 3.5 M NH_4 Acetate. The resuspended phage were extracted once each with an equal volume of phenol, phenol/chloroform/isoamyl

alcohol (25:24:1), and chloroform/isoamyl alcohol (24:1) before precipitating the DNA with 2 volumes of ethanol.

Single stranded DNA to be used only for sequencing was prepared by an alternative method without phenol using cetyl trimethyl ammonium bromide (CTAB) in a protocol modified from Del Sal and Schneider (1986). To precipitate phage, 200 µl of a solution of 20% PEG and 3.5 M NH_4 Acetate were added to 1 ml of culture supernatant and left for 20 min. at 4°C in 1.5 ml eppendorf tubes. The phage was pelleted by centrifuging on a benchtop eppendorf centrifuge for 10 minutes. This precipitation step replaces the use of DNase I for removing residual *E. coli* genomic DNA (Del Sal and Schneider, 1986). PEG remaining on the side of the tube was wiped off using tissue paper without disturbing the phage pellet. The pellet was resuspended in 1 ml of H_2O and then mixed with 100 µl of 10X template preparation solution (2% CTAB, 500 mM Tris, and 100 mM EDTA). After 5 min. at room temperature (RT) the precipitated DNA was collected by centrifugation for 10 min. and resuspended with vigorous vortexing in 500 µl of H_2O followed by mixing with 2 volumes of ETOH. The DNA was collected by centrifuging again and was washed with 70% ETOH before resuspending in 15 µl of TE buffer (10 mM Tris, pH 8.0, and 1mM EDTA). Yields were approximately 2 µg of DNA.

4. Specific Fragments

Fragments of dsDNA were prepared for cloning or radioactive labelling by one of the following methods.

(a) DNA transfer to DEAE paper - DNA was isolated by electrophoresing onto diethylaminoethyl (DEAE) cellulose paper (NA45, Schleicher and Schuell) as described by Dretzen et al. (1981) with some refinements as made by Lizardi et al. (1984). The DNA was eluted off the paper by placing it in a 1.5 ml eppendorf with a solution of 50 to 150 µl of 2.5 M Na Acetate pH 7.5, 20 mM Tris pH 7.6, 2 mM EDTA. The tube was heated to 65°C for 30 min. then vortexed and centrifuged to keep the paper at the bottom of tube. The elution buffer was removed and an equivalent amount was added again. After eluting,

the second aliquot of buffer was added to the first and extracted with 3 volumes of water saturated butanol to remove any ethidium bromide. The DNA was precipitated with 2 volumes of ETOH followed by a 70 % ETOH wash and resuspended in TE at a concentration of ~2 to 500 ng DNA/µl.

(b) DNA electro-elution - Fragments of DNA were also purified from preparative agarose gels using the IBI Electro-Eluter as described by the manufacturers instructions. A slice of agarose containing the desired DNA was placed in the Electro-Eluter and the DNA was eluted into a 7.5M NH_4Acetate cushion. The DNA was precipitated from the ammonium acetate by mixing with 2 volumes of ETOH. The precipitated DNA was washed with 70% ETOH and resuspended in TE buffer.

(c) Double stranded DNA fragments generated from single stranded M13 clones - Double stranded DNA was generated from ssDNA of M13 clones by methods which relied on having the desired ssDNA in both orientations in M13. Clones of opposite orientation were hybridized in restriction endonuclease (REN) buffer for 20-30 min. at 65°C. This created dsDNA only at the insert site and reconstituted the RE recognition site at each end of the insert. If the opposite orientation clones were in sister vectors (i.e. M13mp18 and M13mp19 or M13mp8 and M13mp9) the whole multiple cloning site was reconstituted as dsDNA. Following the hybridization, the duplex DNA was digested with the appropriate REN of choice. The DNA was used directly for cloning experiments or purified further by preparative agarose gel electrophoresis and one of the procedures described in (a) or (b).

The hybridized ssDNA clones were also recovered as dsDNA by digesting the single stranded portions with the ssDNA nucleases, S1 (Seligy et al., 1987) or mung bean nuclease. This procedure is termed the 'S1 nuclease protection assay'. As above, the DNA was hybridized in the enzyme buffer (S1 and mung bean nuclease buffer made as described in Maniatis et al., 1982) at 65°C. for 20-30 minutes. Approximately one unit of enzyme was added per µg of DNA for a 20-30 minute digestion at 37°C. The reaction was stopped by adding EDTA to a final concentration of 25 mM. The resulting fragments were purified by

consecutively passing through two 1 ml spun columns containing Sephadex G50 beads equilibrated with TE (Maniatis et al., 1982) or purified as already described in (a) or (b), starting with electrophoresis in an agarose gel.

D. Cloning Methodology

1. Ligations

All ligations of DNA used the same buffer which contained 50 mM Tris (pH 7.4), 10 mM $MgCl_2$, and 10 mM dithiothreitol (Maniatis et al., 1982). ATP was added separately to a final concentration of 1 mM. Approximately 1 unit of T4 DNA ligase was added per µg of DNA in the reaction. Reactions were either left for 3-4 hr at RT or at 4°C overnight. The ligated DNA in the various reactions were either used to transform *E. coli* immediately or stored frozen at -20°C.

DNA fragments and vector DNA were prepared for ligations by one of three ways:

(a) Unmodified DNA - The vector DNA was simply cut with one or two restriction enzymes and added to the insert DNA at an approximate molar concentration of 1:3, vector:insert.

(b) 5'PO_4 removed - The vector DNA was treated with Calf Intestinal Phosphatase (CIP) (Boerhinger Mannheim) according to the manufacturer at a concentration of 1 unit CIP/µg vector for the last 30 min. of its digestion with a given restriction enzyme. Digestion of the DNA was then stopped by adding EDTA to 25 mM after which the solution was extracted once with an equal volume of a phenol/chloroform/isoamyl alcohol mixture and once with chloroform/isoamyl alcohol. The DNA was precipitated with 2 vol. ETOH. The vector DNA was added to the insert in the same ratios as described in (a).

(c) 3'ends partially filled in - A cloning procedure preventing the recircularization of vector DNA during the ligation of vector DNA to insert DNA was also used (Zabarovsky and Allikmets, 1986). M13mp19 was cut with Xba I and partially filled in with Klenow dCTP and dTTP as described (Zabarovsky and Allikmets, 1986) with the exception that

only 2 units instead of 10 units of Klenow were used per µg of DNA. Genomic DNA of *B. polymyxa* was cut with Hind III and filled in with dGMP and dAMP. The vector and genomic DNA were mixed together with ligation reaction components and incubated for 24 hr at 4°C.

2. Transformation of *E. coli*

Transformation of *E.coli* with plasmids and with M13 DNA was performed according to Hanahan (1983) with modifications for the transformation of M13 DNA. After incubating the competent *E. coli* cells with M13 DNA on ice, the cells were heat shocked at 42°C for 90 sec. in 17mm X 100mm plastic tubes. An equivalent volume of late log cells was added to the tubes with 3-4 ml soft agar, the mixture was briefly mixed and then plated directly (Messing, 1983). No incubation at 37°C was allowed after the heat shock took place, as is customary with cells transformed with plasmids.

E. Analysis of Proteins

1. Enzymatic Assays

(a) Plate

M13 phage and plasmids containing inserts of DNA responsible for expressing active ß-1,4-endoglucanase were detected using two different colorimetric assays. The main assay used, included carboxymethylcellulose (Hercules, DS: 7, medium viscosity) in the bottom and the top agar at a concentration of 0.1% and employed Congo Red stain as described by Teather and Wood (1982). In the alternative assay, used only for the detection of M13 clones, Ostazin Brilliant red H3B dyed cellulose (Biely et al., 1985) was added to the top agar (0.1% w/v).

(b) Liquid

The levels of endoglucanase activity produced by various clones and host organisms were quantified using liquid assays of culture supernatants. Endoglucanase enzyme activity was measured and calculated from 24 hr culture supernatants using the viscometric assay (Hulme, 1971). A Fisher F887 viscometer was used at an incubation temperature of 40°C

to measure the change in the viscosity of a 0.5 % CMC solution in 0.01 M Na acetate, pH 6.0 buffer due to endocleavage by the cellulase enzymes. Units are expressed in the number of ß-1,4 bonds broken min^{-1} ml^{-1}.

Enzymatic activity was also measured by the detection of reducing ends produced by the cleavage of CMC using dinitrosalicylic acid (Miller, 1959). One half of a ml of diluted supernatant was added to 0.5 ml of 0.5 % CMC, in 0.01 M Na acetate pH 6.0, and incubated at 40°C for 30 min. One ml of DNS reagents was added to stop the reaction and then the solution was boiled for 10 min. Units are expressed in glucose equivalents produced min^{-1} ml^{-1}. This assay was routinely used to quantify enzymatic activity, but no results from the use of this assay are presented in this thesis.

Total protein in culture supernatants was assayed using the BioRad protein assay system following the procedure of Bradford (1976). Bovine serum albumin (Sigma) was used as the protein standard.

2. E'ɔctrophoresis

Intracellular and extracellular proteins from recombinant *E. coli* were electrophoresed in denaturing gels according to Laemmli (1970) using a BioRad Mini-Protean II Dual Slab Gel electrophoresis apparatus. The typical gel consisted of a 4% polyacrylamide stacking gel and a 10 % polyacrylamide base. Gels were run according to the manufacturers recommended voltage.

For intracellular proteins, cells were pelleted and washed 2 times in 1/10 the original cell volume then resuspended in the same volume of TE. The cell mixture was added to an equal volume of 2X Laemmli loading buffer. Cell lysis was completed by heating the samples for 5 min. in a bath of boiling water before loading onto the gel.

Extracellular proteins were prepared and concentrated by two methods:

(a) Ethanol precipitation - Two and one half volumes of ETOH were added to the culture supernatant and left for three hr at RT (Petre et al., 1981). The protein was centrifuged down at 10 K rpm for 10 min. in a Sorvall SA-600 rotor. The remaining ETOH in the

pellet was evaporated with N_2 and the pellet was resuspended in 1/100th of the original supernatant volume of TE.

(b) Lyophilization - The culture supernatant was first lyophilized then resuspended in 1/50th of the original cell culture volume of TE. Aliquots of 100 µl were then passed through two spun columns of Bio-Gel P6 polyacrylamide beads (BioRad) equilibrated with 0.01 M Na Acetate, pH 6.0. The solution was lyophilized again and then resuspended in 15 µl of TE.

For either procedure the proteins were added to 2X Laemmli loading buffer, boiled for 3 min. and loaded onto the gel. After the gels were run, they were stainesd with 0.25% Coomassie brilliant blue R-250, 40% methanol, and 10% Acetic Acid for 20 to 30 minutes. Destaining took place with 40% MEOH, and 10% acetic acid for 3 hr with constant agitation and several changes of destaining solution.

F. Analysis of DNA

1. Electrophoresis

DNA was electrophoresed in agarose gels of usually 0.7% agarose according to Maniatis et al. (1982) using Tris-EDTA-Borate (TEB) buffer. After electrophoresis, DNA in the gels was detected by soaking the gels in an aqueous solution of 1 µg/ml ethidium bromide (EtBr) for 5 min., followed by a destaining of the gel in H_2O for 20 minutes. Stained DNA was made visible to the naked eye by a hand held long-wavelength emitter UV lamp or photographed through a red, B+W S.6 filter while illuminating the gel with short-wavelength UV light.

2. Transfer to Membrane

(a) Agarose gels

DNA was transferred to nylon membranes (Biotrans, New England Nuclear) using an alkaline blotting procedure (Reed and Mann, 1985). A tray containing a 3 cm layer of paper covered with 3 layers of Whatman 3MM was filled to about 2 cm with transfer

solution consisting of 0.3 M NaOH and 1.5 M NaCl. To aid in the transfer of large DNA fragments the DNA in the gel had been partially depurinated by soaking the gel in 0.25 M HCl for 7 to 10 min. at RT (Wahl et al., 1979). The gel was then placed in the alkali transfer solution and left for 20 min. to cleave the DNA at the depurinated sites. Following this, the gel was laid on the Whatmann paper in the tray, onto which the membrane was placed followed by another 2 layers of Whatman 3MM, an adequate layer of absorbant paper, and a weight of approximately 500 g. The average transfer lasted 4 hr before neutralizing the membrane in a solution of Tris pH 7.5, 1.5 M NaCl. The filter was then air dried and baked for 30 min. at 80°C before hybridizing under standard conditions.

(b) Plates

DNA from plaques of M13 phage were transferred from agar plates to a nylon derivative membrane (HyBond, Amersham) using the very simple methodology of Mayaux et al. (1987). After growing the *E. coli* cells overnight the plates were chilled for 1 hr. at 4°C. The pre-cut membranes were placed in contact with the surface of the plates for 5 minutes. Filters were then gently removed from the plates and washed 4 times for 3-5 min. in 10 X SSC, air dried, and baked for 30 min. at 80°C. Standard hybridization conditions were then used.

3. Hybridization

All hybridizations of radioactive probes to DNA immobilized onto membranes were carried out in heat sealable plastic bags at 65°C for 15 to 20 hr. Membranes were pre-hybridized for at least 1 hr in hybridization buffer similar to that described in the original GeneScreenPlus protocol (NEN, 1983), i.e. 1 M NaCl, 1% SDS, and 100 µg/ml Salmon Sperm DNA. After pre-hybridization, the buffer was removed and 10 to 12 ml of fresh hybridization buffer was added with the denatured probe. Hybridizations were carried out at 65°C for 15 to 20 hr. Membranes were washed 2 times in 100 ml 2X SSC for 5 min. at RT with constant agitation, 2 times in 250 ml of 2X SSC, 1.0% SDS at 65°C for 20-30 min., and 2 times in 250 ml of 0.1X SSC for 15 min. at RT with constant agitation. The

membranes were then air dried and exposed to X-ray film for 20 min. to 5 days at RT or -80°C.

4. Preparation of Radioactive DNA Probes

Radioactive DNA probes were labelled with [α-^{32}P]dATP (800 mCi./mMol) and were prepared from purified DNA fragments by nick translation as described by Maniatis et al. (1982). Probes were also made using a random hexamer oligonucleotide priming kit following the manufacturer's instructions (Boerhinger Mannheim). This method was developed by Feinberg and Vogelstein (1983). Unincorporated radioactive nucleotides were removed from all probes by passing through two, 1 ml spun columns containing Sephadex G50 (Pharmacia) dextran beads (Maniatis et al., 1982). Probes were denatured with 1/3 of a volume of 1M NaOH and neutralized just prior to their addition to the hybridization bag with an equal volume of 1 M Tris pH 7.5. Probes were denatured for all subsequent uses by placing in boiling water for 5 minutes.

G. Sequencing of DNA Fragments Coding for *B. polymyxa* Endoglucanase Gene

The DNA fragments were sequenced by first creating a family of ordered sequential deletions in a method described by Dale and coworkers (1985). Single stranded DNA of the full length clones in M13mp9 or mp19 was hybridized with the oligomer RD20 (R. Brousseau, NRCC), which created double stranded DNA region covering the Eco RI site. This allowed the restriction enzyme Eco RI to cleave the ssDNA template, 3′ to the site of the inserted heterologous DNA. T4 DNA polymerase (IBI), which has a 3′->5′ but not a 5′->3′exonuclease activity, was used to resect the ssDNA at intervals for a controlled time period in the absence of dNTPs. The ssDNA was then tailed with dGMP using terminal deoxynucleotidyl transferase in reaction conditions optimized for an addition of 4-6 nucleotides. The oligomer RD20 was again hybridized to the DNA template circularizing the linear strands by complementing the G-tail at the 3′end and the 15 nucleotide sequence at the 5′ end. T4 DNA ligase was then used to covalently close the circularized molecule. The ligated DNA was transformed as already described.

The next day the deletion clones were picked with toothpicks and each grown in 1.5 ml of 2YT medium with 40 µl of late log JM101 or DH5αF′ *E. coli* cells. The clones were roughly sized by electrophoresing lysed supernatants in 0.7% agarose gels of 1X TBE buffer without EtBr for 14 hr at 75 volts. Single stranded DNA was prepared from clones that seemed to represent the complete DNA sequence with approximately 200 base pair gaps between each. Insert size of each deletion clone was more precisely determined prior to sequencing using the S1 nuclease protection assay as described in Material and Methods, section 4.C(c).

Sequencing was performed by either of two dideoxy-chain termination methods. The first method was as described by Messing (1983) using the large fragment of DNA Polymerase I (Klenow) and [α-^{35}S]dATP (Biggin et al., 1983; Ornstein and Kashdan, 1985) instead of [α-^{32}P]dATP. The second method used modified T7 DNA Polymerase (supplied as Sequenase) as described by the manufacturer, United States Biochemical Corporation (USB). The latter procedure was used throughout most of the work because it yielded superior sequencing results. Approximately 500 ng of template DNA were used per reaction instead of the 1 to 2 µg. suggested by USB.

The majority of the sequencing reactions were electrophoresed on 43 cm gels (BRL S2 sequencing units) using wedge spacers (BRL) with a 1:4, top to bottom thickness ratio. Buffer and acrylamide solutions were used in the concentrations given in Carlson and Messing (1984). All 43 cm gels were run at 55 to 60 watts constant power for approximately 1.75 hr. The gels were fixed in 5% methanol and 5% acetic acid either for 10 min. (for 4 mm thick gels) or 30 min. (for wedge gels) before drying (Orstein and Kashdan, 1985). Wedge gels needed to be separated from the glass plates during fixing and removal of the urea otherwise the bottoms surfaces of the gels would not dry tack free. Gels were exposed for 20 to 72 hr directly on Kodak XRP film at RT.

H. Sequence Analysis of DNA and Protein

1. DNA

DNA sequences were compiled and partially analyzed using the IBI Sequence Analysis Programs written by J. Pustell and Kafatos (1982, 1984). Data were also analyzed by the Sequence Analysis package written by Stephens (1985).

2. Protein

Published amino acid sequences and putative translation products of published DNA sequences were compared in pairs using the MATCH program. This program was created by the Biomathematics group of NRCC (now called Biomolecular Modelling) and is based on the algorithm of Needleman and Wunsch (1970). Percentage of similarity of amino acid sequences is based on the number of direct matches of residues given by the MATCH program. The number calculated is relative to the number of residues in the smaller of the two sequences in a single comparison.

Pairwise amino acid alignments were also produced by the MATCH program. Multiple alignments were produced by overlaying several pairwise alignments and re-aligned manually on a word processor.

The relative evolutionary distance was calculated for a few comparisons using the SEQDP program written by Kanehisa (1982). In the program, distances were calculated using the weighting of the Dayhoff Matrix (Dayhoff et al., 1978). The Dayhoff Matrix provides a different weight for each amino acid. Conservation of rare amino acids from one sequence to another is given a larger weight than more common amino acids. Sequence alignments were also made by a method based on the algorithm of Needleman and Wunsch (1970). The distance from the comparison of two sequences is related to 20 comparisons made with randomized sequences produced from the original sequences. The differences between the distances of the comparison of the original sequences and the mean of the comparisons with the randomized sequences are expressed in standard deviations from the mean of a Gaussian distribution. The standard deviation is generated

from the 20 comparisons of the random sequences. For sequences that were not close enough to show any evolutionary significance (i.e., less than a standard deviation from the mean), the probability of the similarity being only random was also calculated.

The ORF-2 putative protein amino acid sequence was used to search for similar sequences to it in the Protein Identification Resource/National Biomedical Research Foundation (PIR/NBRF) Protein Sequence Data Bank (release 16.0, May '88) and the SWISS-PROT sequence database (release 7.0, May '88) which is managed by the European Molecular Biology Laboratory. The search was done by the XFASTP program of Bionet based on the algorithm of Lipman and Pearson (1985).

Hydropathy plots were generated from the HYDRO program in the COMPMOLOGEN DNA/Protein analysis package developed at NRCC by the Biomathematics group. Values for each residue were calculated relative to 6 flanking residues according to Kyte and Doolittle (1982).

III. RESULTS

A. Primary screening of *Bacillus* species for endoglucanase activity

In the interest of knowing which bacilli produced cellulase enzymes, a sample strain of each *Bacillus* species in the NRCC culture collection was assayed for endocellulolytic activity. A list of the *Bacillus* species and strains screened for potential endoglucanase production is presented in Table 3. Each strain was grown on plates made up with L medium and 0.1% CMC at the temperature described in Table 3. After colonies had grown to a minimum of 2 mm in size, they were examined for endoglucanase activity by the Congo Red (CR) CMC plate assay of Teather and Wood (1982). No assays were carried out on cell extracts of the strains that were CR^- because endoglucanases, by virtue of the large substrate molecules that they attack, must be transported outside of the cell membrane and cell wall. As indicated in Table 3, activity was only exhibited by the *B. circulans* and *B. polymyxa* strains and by all the *B. subtilis* strains with the exception of NRCC 9044.

DNA was purified from each of these same strains with the exception of some of the *B. subtilis* strains. Genomic DNA was only purified from *B. subtilis* strains 3052, 9044, 9045, PAP 115. After digestion with PstI, and electrophoresis on agarose gels, DNA from the gels was transferred onto a nitrocellulose filter and subjected to DNA-DNA hybridization according to Southern (1975). As a cellulase gene probe, the 3.1 kbp PstI DNA fragment containing the *B. subtilis* PAP115 endoglucanase gene (MacKay et al., 1986) was labelled with radioactive dAMP and hybridized to the bound DNA at low stringency. In Fig. 1B a radioactive band appears in all lanes at approximately 3.5 to 4 kbp in size. This is not indicative of a positive DNA/DNA hybridization because lane d of Fig. 1A, shows a failed extraction of genomic DNA from *B. stearothermophilus* which also produces a radioactive band in Fig. 1B. These bands might be protein from the DNA samples. In Fig. 1B positive hybridization occurred with only the *B. subtilis* (lane j) and the

Table 3. *Bacillus* species and strains screened for endoglucanase activity.

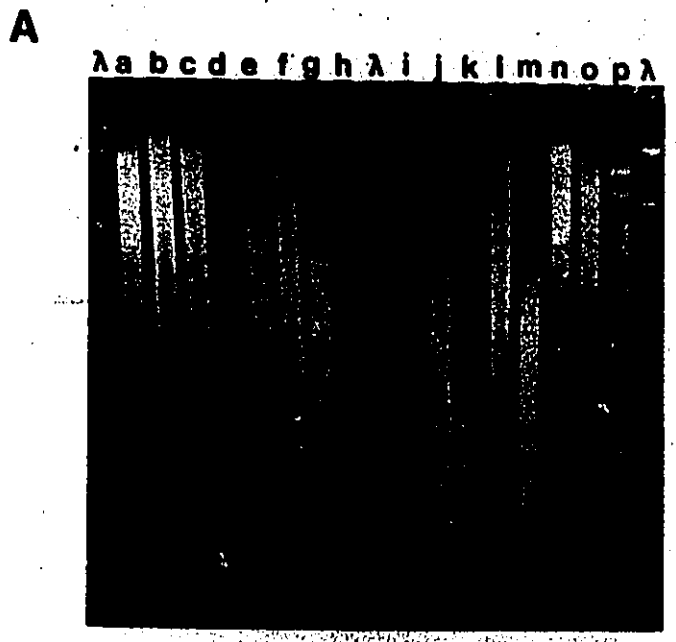
Species	NRCC strain no.	Temp.C.	Other strain no./Remarks	Activity +/-
B.alvei	9062	30		-
B.amyloliquifaciens	2147	37		-
*B.brevis**	9084	37	USDA 616	-
*B.cereus**	9003	30		-
*B.circulans**	9024	30	USDA 729	+
B.coagulans	2851	37		-
*B.licheniformis**	9012	37		-
B.macerans	9074	37		-
B.megaterium	2852	30		-
B.pantothenticus	9026	30	+ 4% NaCl	-
*B.polymyxa**	2822	30		+
*B.pumilus**	9066	37		-
B.sphaericus	9027	30		-
B.stearothermophilus	9001	55		-
*B.subtilis**	3052	37	ATCC6051	+
	9041		USDA231	+
	9044			-
	9045			+
	9048			+
	9049			+
	9051		USDA703	+
	9052		USDA1003	+
	9053			+
	9057			+
	9058		U. of Alberta 231	+
			PAP 115	+
B.thuringiensis subsp. *kurstaki*	2278	37		-

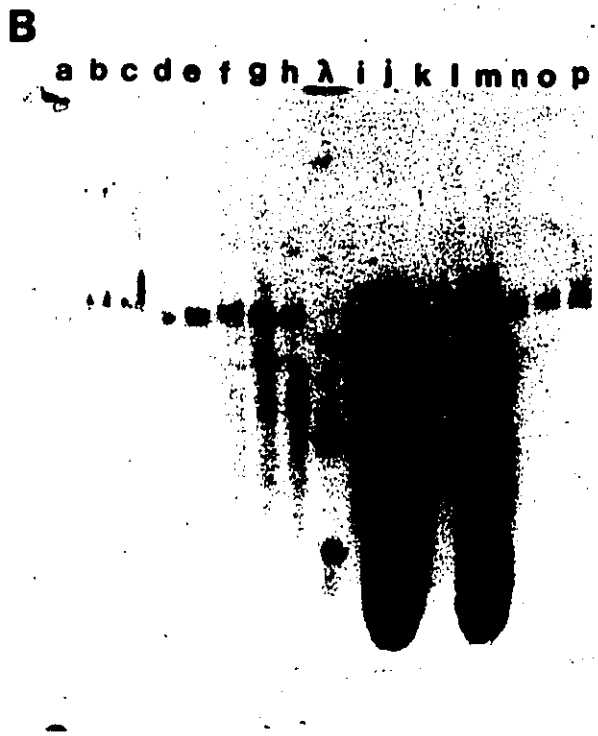
Species marked with * have been reported as having strains which produce cellulase (See Introduction, section F).

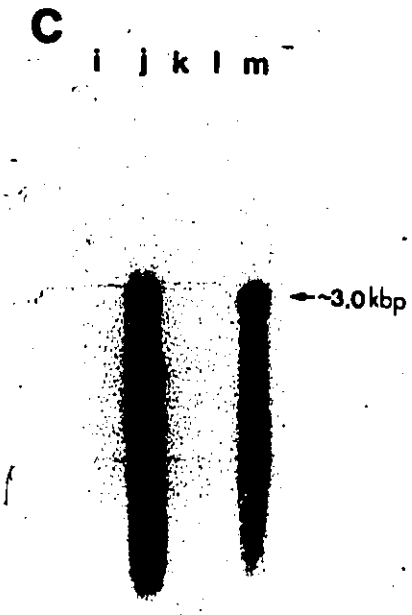
FIGURE 1. DNA-DNA HYBRIDIZATION OF 15 DIFFERENT BACILLI GENOMES PROBED WITH THE *B. subtilis* PAP115 ENDOGLUCANASE GENE.

The 3.1 kbp PstI fragment of *B. subtilis* PAP115 encoding the endoglucanase gene was labelled with radioactive [^{32}P] dAMP and hybridized to genomic *Bacillus* DNA of the stained gel in 1(A) according to method of Southern (1975). The results of this hybridization are shown in 1(B). A shorter exposure of the hybridized filter to x-ray film is shown for selected lanes in 1(C). All the genomic DNA was cut with PstI (listed below).

λ = lambda phage DNA cut with Hind III
a = *B. thuringiensis*
b = *B. amyloliquifaciens*
c = *B. brevis*
d = *B. stearothermophilus*
e = *B. pantothenticus*
f = *B. cereus*
g = *B. alvei*
h = *B. coagulans*
i = *B. licheniformis*
j = *B. subtilis*
k = *B. macerans*
l = *B. polymyxa*
m = *B. circulans*
n = *B. sphaericus*
o = *B. pumilus*
p = *B. megaterium*


A
λa b c d e f g h λ i j k l m n o p λ


B
a b c d e f g h λ i j k l m n o p


C
i j k l m
~3.0kbp

B. circulans (lane m) chromosomal DNA. A shorter exposure of the same hybridization filter shown in Fig.1C revealed more discreet DNA fragments of similar size for both of these types of DNA.

B. Cloning of the *B. circulans* and *B. polymyxa* endo-ß-1,4-glucanase genes

A library of genomic DNA of both *B. circulans* and *B. polymyxa* was constructed by ligating partially PstI digested genomic DNA with the *E. coli* bacteriophage M13mp9. Each ligation yielded approximately 10^5 individual clones per µg of genomic DNA after transfection of *E. coli* JM103 with the ligated material. Since all cloned bacterial endocellulase genes studied to date are expressed to some extent in *E. coli*, the Congo Red assay was used to screen for positive endoglucanase producing clones. Results in Fig. 2A and 2B show examples of the two types of enzymatic plate assays used for selecting an actively expressed endoglucanase gene in *E. coli*. A mixture of active and inactive clones of the endo-ß-1,4-glucanase gene from *B. polymyxa* were plated on the two different substrates used. The visual contrast of the Congo Red assay is greater between positive and negative activity than seen with OBR-cellulose. For this reason, the Congo Red assay was used for primary screening in all the original cloning experiments. Both the *B. subtilis* PAP115 and *B. circulans* clones gave haloes 2 to 3 times larger than those exhibited by the *B. polymyxa* clones. However after subcloning of the *Bacillus* DNA using the pUC plasmids no species difference in apparent levels of endocellulase expression were observed (data not presented).

Two CR^+ clones (Bp-P1, Bp-P2) were isolated from the *B. polymyxa* bank of approximately 10,000 clones. Clone Bp-P1 contained a 1.6 kbp single PstI DNA fragment while the clone Bp-P2 contained the same 1.6 kbp fragment and a 0.2 kbp PstI DNA fragment. The removal of the 0.2 kbp DNA from Bp-P2 did not change the activity exhibited by this clone. The orientation of the fragments relative to the transcription of the ß-galactosidase α-peptide coding gene in M13mp9 was the same in both of these clones. Three clones (Bc-P1, Bc-P2 and Bc-P3) were isolated from the 10,000 clones making up the

FIGURE 2. EXAMPLES OF PLATE ASSAYS FOR ENDOGLUCANASE ACTIVITY OF E. COLI CELLS INFECTED WITH RECOMBINANT M13.

A mixture of M13mp9 phage with and without the *B. polymyxa* endoglucanase gene were used to infect *E. coli* JM101 cells and plated with top agar containing either 0.1% CMC for (A), the Congo red plate assay or (B), 0.1% Ostazin Brilliant Red H3B (OBR) dyed cellulose for the detection of endoglucanase activity. White arrows point to plaques with cells expressing endoglucanase and black arrows point to plaques with cells which do not..


A


B

B. circulans bank. Each of the three clones contained the same 3.1 kbp PstI DNA fragment and all were in the same orientation in the vector relative to the α-peptide gene.

Since M13 vectors have been suspected of not being able to propagate large recombinant fragments, similar banks of *B. polymyxa* and *B. circulans* DNA were also made using the pUC 8 plasmid. The *B. polymyxa* genomic DNA partially digested with PstI was used in this library construct. However for the *B. circulans* DNA library construct, partially HindIII digested DNA was used. No endoglucanase expressing clones were found in the *B. polymyxa* bank, but 2 positive clones (pCH.7 and pCH.9) were found in the *B. circulans* bank. Both plasmids contained a 3.4 kbp HindIII fragment that when digested with PstI, produced a 3.1 kbp piece of DNA. Since the restriction endonuclease (REN) sites of all *B. circulans* clones recovered were identical, it was assumed that the same genomic section of DNA was cloned in each case.

C. DNA characterization of clones

The isolated *B. polymyxa* PstI and the *B. circulans* HindIII DNA fragments were subcloned into M13mp19 in the opposite orientation to which they were originally found (*B. polymyxa* - Bp-op19, *B. circulans* - Bc-H4). Since the *B. circulans* gene is expressed in both orientations, this implies that this gene uses its own *E. coli*-like promoter sequence(s). However this is not the case for the *B. polymyxa* gene since expression of this gene occurred in only one orientation, suggesting that expression may be dependent on the promoter of the ß-galactosidase α-peptide gene.

In order to show that the clones from each genome were unique, the clones were hybridized with the genomic DNA of *B. subtilis* PAP115, *B. circulans* and *B. polymyxa*. Each genomic DNA was digested with EcoRI, HindIII and PstI, followed by electrophoresis in a 0.7% agarose gel and transfer onto a nylon membrane. As described in the legend of Fig. 3, the immobilized DNA was probed with either a 3.4 kbp HindIII DNA fragment containing the endoglucanase gene from *B. circulans* or a 1.6 kbp PstI segment of DNA containing part of the endoglucanase gene from *B. polymyxa*. The data in

FIGURE 3. *B. subtilis*, *B. circulans* AND *B. polymyxa* GENOMIC DNA HYBRIDIZED WITH EITHER THE *B. circulans* OR *B. polymyxa* ENDOGLUCANASE GENES.

The DNA of gels (A) and (B) were transferred onto nylon membranes as described in section F.2(a) of the Material and Methods. Gel A was hybridized with the *B. circulans* 3.4 kbp Hind III fragment encoding the endoglucanase gene. The results of this hybridization are shown in (C). Gel B was hybridized with the 1.6 kbp Pst I fragment encoding the *B. polymyxa* endoglucanase gene and the results are shown in (D).

Bs = *B. subtilis* PAP115
Bc = *B. circulans*
Bp = *B. polymyxa*
λ = lambda phage DNA cut with HindIII
M = DNA size markers, see Appendix D

H = Hind III
E = Eco RI
P = Pst I

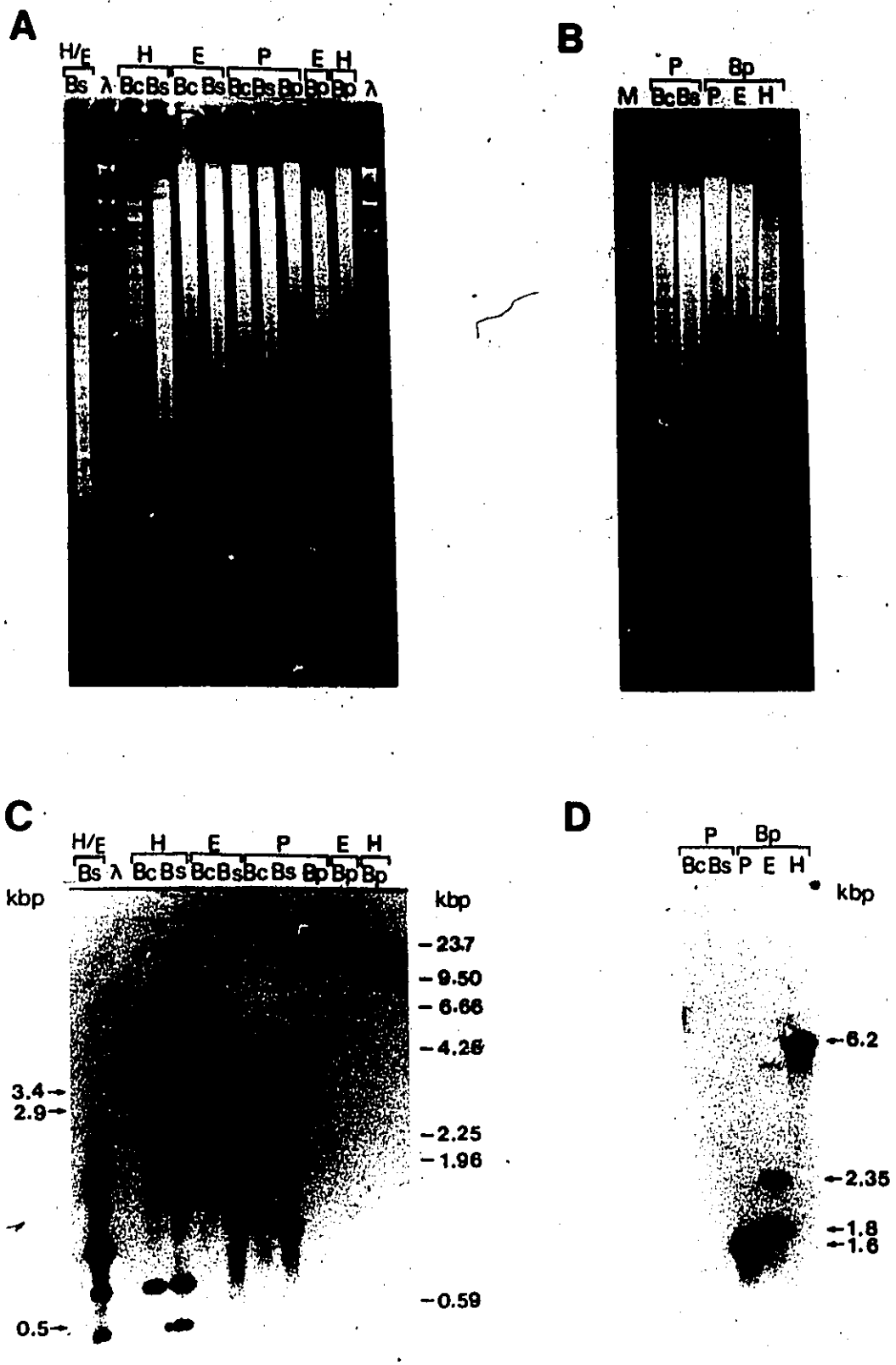

A
H/E
H
E
P
E
H
Bs
λ
Bc
Bs
Bc
Bs
Bc
Bs
Bp
Bp
Bp
λ
B
P
Bp
M
Bc
Bs
P
E
H
C
H/E
H
E
P
E
H
Bs
λ
Bc
Bs
Bc
Bs
Bc
Bs
Bp
Bp
Bp
kbp
kbp
23.7
9.50
6.66
4.26
3.4
2.9
2.25
1.96
0.59
0.5
D
P
Bp
Bc
Bs
P
E
H
kbp
6.2
2.35
1.8
1.6

Fig. 3 show that the *B. polymyxa* gene is quite different from the genes of either *B. circulans* or *B. subtilis*. Since the data only shows the *B. polymyxa* gene probe hybridizing to a single fragment in both the lanes of HindIII and PstI digested *B. polymyxa* genomic DNA, it suggests that this is a single copy gene.

B. circulans and *B. subtilis* as shown originally in Fig.1, have homologous fragments that are very similar and only appear different in Fig. 3B by the additional HindIII site in the *B. subtilis* genome. This homology is reinforced by the S1 nuclease protection assay using one of the clones of *B. circulans* (Bc-H4) hybridized to the opposite M13 *B. subtilis* clone, Bs-P3.2, (opposite orientation *B. subtilis* clone is Bs-P3.3) containing the 3.1 kbp PstI fragment (see Mackay et al., 1986). In Fig 5 the results show a single band at approximately 3 kbp. Although the sensitivity of this assay is unknown in terms of the minimum number of non-identical nucleotides (nt) required before S1 is able to degrade the heteroduplex, the result of Fig 5 suggests that there are no large nonhomologous regions between the *B. circulans* and *B. subtilis* clones. In addition to this homology, both *B. subtilis* and *B. circulans* share some genomic DNA sequence outside of the probe segment as indicated in Fig. 3B. These fragments seen in Fig. 3B as the 1.5 kbp EcoRI, 1.1 kbp PstI and the 0.7 kbp HindIII size fragments of DNA are not detected on x-ray film exposed to the filter for only a short time.

The REN maps of the three endoglucanase genes are presented in Fig 4. These maps were constructed from the analysis of genomic blots as well as sequence information of the *B. polymyxa* gene presented later (see Results, sections F.1 and F.2). The map of the *B. subtilis* gene agrees with that published by Mackay et al. (1986). The different HindIII site was assigned to the 3′ end of the *B. circulans* gene. This was found by first hybridizing together *B. circulans* clones of opposite orientations that are in M13, and then cutting with EcoRI and digesting the non-hybridizing single stranded DNA with S1 nuclease. The resultant fragments are a 1.8 kbp which is similar in size expected for the 5′ end of the *B. subtilis* clone (MacKay et al., 1986), and a 1.6 kbp fragment which was larger than the

FIGURE 4. RESTRICTION ENDONUCLEASE MAPS OF *B. subtilis*, *B. circulans*, AND *B. polymyxa* ENDOGLUCANASE GENES AND SURROUNDING GENOMIC DNA.

Maps were constructed from the data provided in Fig. 3, other REN data not shown and the sequence data from Fig. 11 and Fig. 12. Arrows denote the putative region and direction of transcription of genes. Location and direction of transcription of the *B. subtilis* gene was shown by MacKay et al. (1986). ORF-2 on the *B. polymyxa* clone E.1.2 (see section E of Results) refers to a 2nd open reading frame found in the sequence data of Fig. 11 which is possibly transcribed.

E = EcoRI
H = HindIII
P = PstI

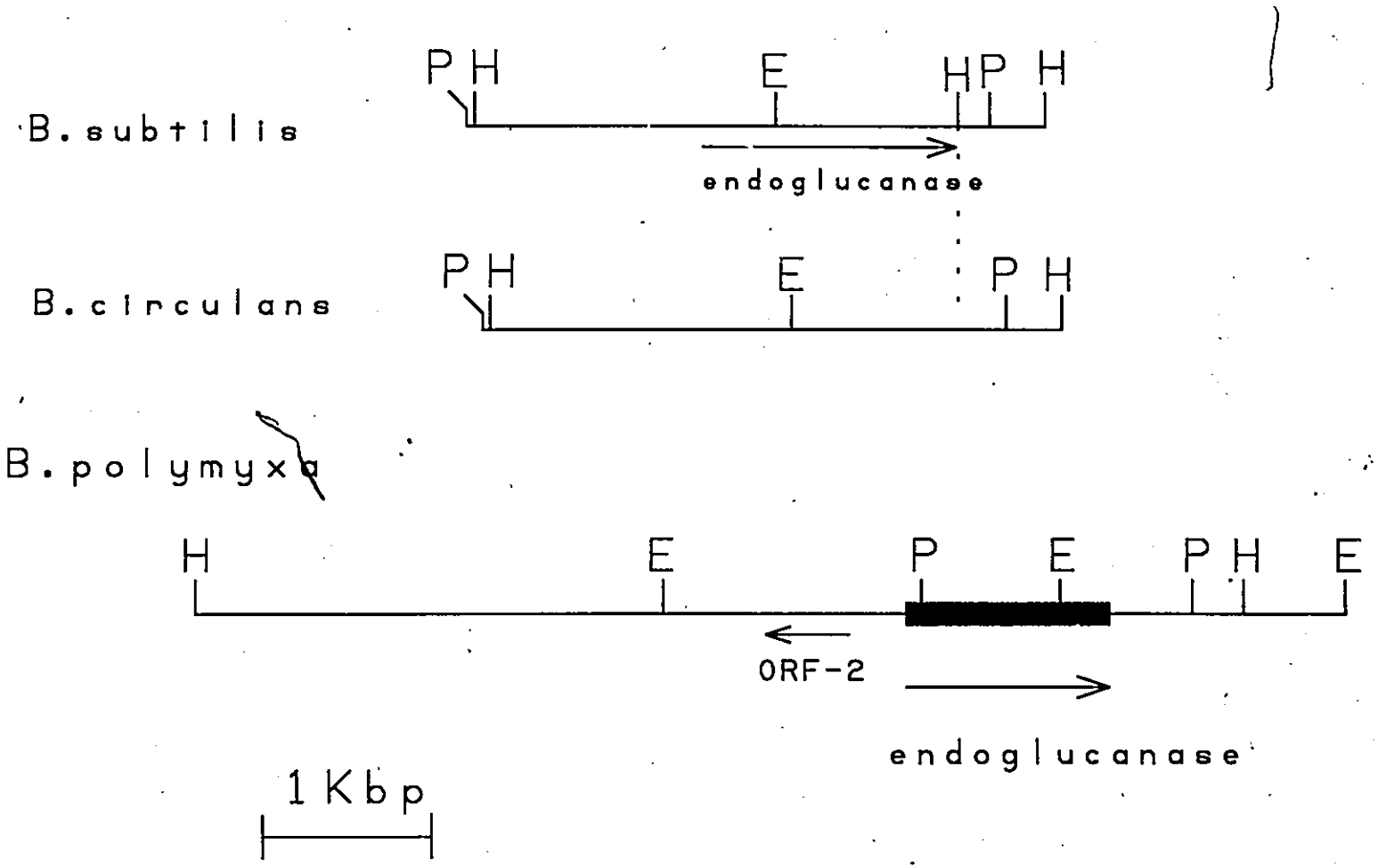
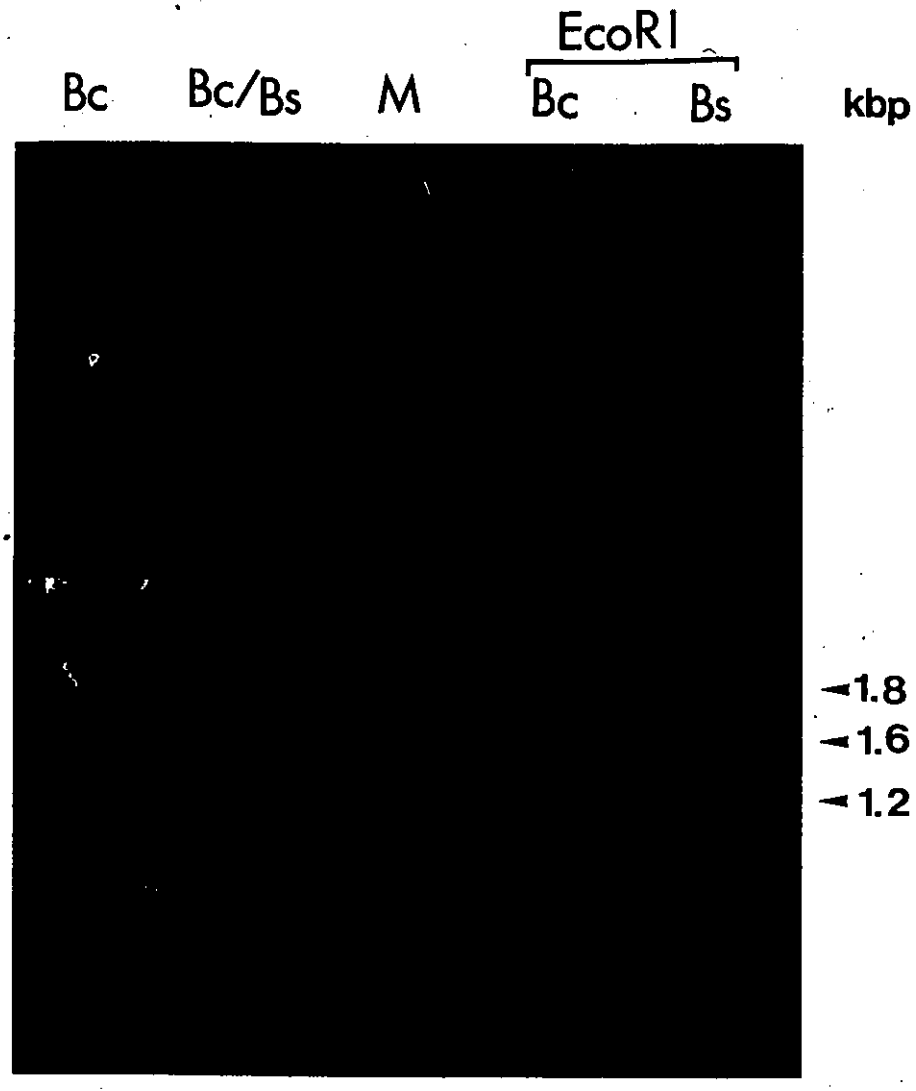

B. subtilis
P H
E
H P H
endoglucanase
B. circulans
P H
E
P H
B. polymyxa
H
E
P
E
P H
E
ORF-2
endoglucanase
1 Kbp

FIGURE 5. S1 NUCLEASE PROTECTION ASSAY OF *B. circulans* AND *B. subtilis* ENDOGLUCANASE GENES.

M13 clones were hybridized and digested with S1 nuclease as described in section C.4(c) of Material and Methods. In the 4th and 5th lane, the dsDNA products were cut with EcoRI after S1 nuclease digestion.

Bc = *B. circulans* M13 clone Bc-H4/*B. circulans* M13 clone Bc-HI
Bc/Bs = Bc-HI/ *B. subtilis* M13 clone SBO (see MacKay et al., 1986)
M = DNA size markers, see Appendix D
Bc (EcoRI) = Bc-H4/Bc-HI
Bs (EcoRI) = *B. subtilis* clones SBO/SB1 (see MacKay et al., 1986)


EcoR1
Bc
Bc/Bs
M
Bc
Bs
kbp
1.8
1.6
1.2

second fragment generated from the *B. subtilis* clone. These results are shown in Fig. 5 in the 4th and 5th lanes. In Fig. 3B a fragment that was 0.5 kbp in the *B. subtilis* HindIII digestion hybridized with the *B. circulans* probe while no such fragment existed in the *B. circulans* lane. This fragment corresponds to the difference in size between the major *B. circulans* and *B. subtilis* HindIII fragments.

D. Expression of the cloned genes in pUC vectors

Initial experiments with SDS-polyacrylamide gel electrophoresis of total proteins from the supernatants of overnight cultures of *E. coli* infected with M13 containing the endocellulase genes of either *B. polymyxa*, *B. circulans*, or *B. subtilis* showed no extra bands that could be assigned to the endoglucanase products (data not shown). To improve on the endoglucanase expression, the *B. polymyxa* gene was subcloned using the high copy number pUC vector; the clones from the other two bacilli were already subcloned in pUC plasmids. The *B. polymyxa* 1.6 kbp Pst I DNA fragment was ligated into the PstI site of three pUC 9 vectors (pUC9.0, pUC9.1, pUC9.2) which were designed to give the multiple cloning sites of each of the vectors a different reading frame relative to the ATG start codon of the α-peptide gene (Hanna et al., 1984). Recombinants expressing endoglucanase activity were identified by the Congo Red plate assay . Activity was only expressed by clones (e.g., pC2.2) produced from the pUC 9.0 ligation. This vector has exactly the same reading frame as that of M13 mp9 which means translation starts at the AUG transcribed from the *lac*Z α-peptide gene and transcription of the *B. polymyxa* endoglucanase gene is most likely due to the *lac* promoter.

The plasmids of each of the three *Bacillus* clones were transformed into *E. coli* HB101 instead of the standard JM series used for pUC vectors. This was done to avoid possible experimental error due to some unexplained activity for lowering the viscosity of a solution of CMC which has been found in *E. coli* JM103 (MacKay et al., 1986), but was not found in HB101 (A. Lo, personal communication).

The pUC clones were grown overnight in 2YT and the intracellular and

extracellular proteins were subjected to SDS-PAGE as shown in Fig. 6. Two similarly sized protein bands were produced by the *B. circulans* endoglucanase clone pCH.9 and the *B. subtilis* clone pC6.3 as indicated in Fig. 6A and 6C. The size of these bands are similar to those shown already for pC6.3 by Lo et al. (1988). Figure 6A and 6C shows that the plasmid pC2.2 encoding the *B. polymyxa* endoglucanase gene is responsible for the production of an approximately 37 kDa protein which is mainly cell associated. Figure 6B shows the proteins in the supernatant that have been precipitated with ETOH while Fig. 6C shows the extracellular proteins concentrated by lyophilization. The *B. circulans* endoglucanase was selectively precipitated by ETOH but the *B. polymyxa* endoglucanase was not. In Fig. 6B the endoglucanase of *B. circulans* was also selectively precipitated relative to *E. coli* proteins as the endoglucanase 52 kDa and 35 kDa bands appear relatively much darker by Coomassie staining than in Fig. 6C. The *B. subtilis* endoglucanase was precipitated by ETOH just as the *B. circulans* endoglucanase (data not shown). This might represent an overall greater hydrophilic nature for the *B. circulans* and *B. subtilis* endoglucanases than for the *E. coli* proteins or the *B. polymyxa* endoglucanase.

The amounts of expression of endoglucanase activity of the original organisms relative to that produced by *E. coli* cells with the respective genes in the pUC vectors are presented in Table 4. All measurements were done using the viscometric assay explained in the Material and Methods, section E.2. Activities were measured using enzymes from overnight cultures inoculated with an equivalent concentration of cells. Cells were grown in minimal media (see Material and Methods, section A) supplemented with various carbon sources. Glucose was used at two levels of initial concentration to see if any repression of transcription existed due to an increase of a readily usable carbon source. No induction/repression system similar to that of cellulolytic fungi was expected as none of the *Bacillus* endoglucanases studied have yet to be shown to be inducible. Cellobiose has previously been found to be a good carbon source for high endoglucanase production in

FIGURE 6. SDS PAGE OF INTRACELLULAR AND EXTRACELLULAR PROTEINS FROM *E. coli* CELLS TRANSFORMED WITH PUC CLONES CONTAINING THE *B. subtilis*, *B. circulans*, OR *B. polymyxa* ENDOGLUCANASE GENES.

Cellular and extracellular protein samples were prepared as described in section G of the Material and Methods. In (A), the sizes of the protein due to the endoglucanase genes (marked with dots in all panels) are marked on the left, the size of the protein size markers are indicated on the right. (B) shows proteins precipitated from the supernatants with ETOH and (C) shows the proteins recovered from lyophilization of the supernatants. Twenty times more cell culture was required for samples loaded in gels of (B) and (C) than (A).

M = High molecular weight BRL protein size markers
pC2.2 = pUC9 + *B. polymyxa* 1.6 kbp PstI fragment
pC6.3 = pUC8 + *B. subtilis* 3.1 kbp PstI fragment
pCH.9 = pUC8 + *B. circulans* 3.4 kbp HindIII fragment

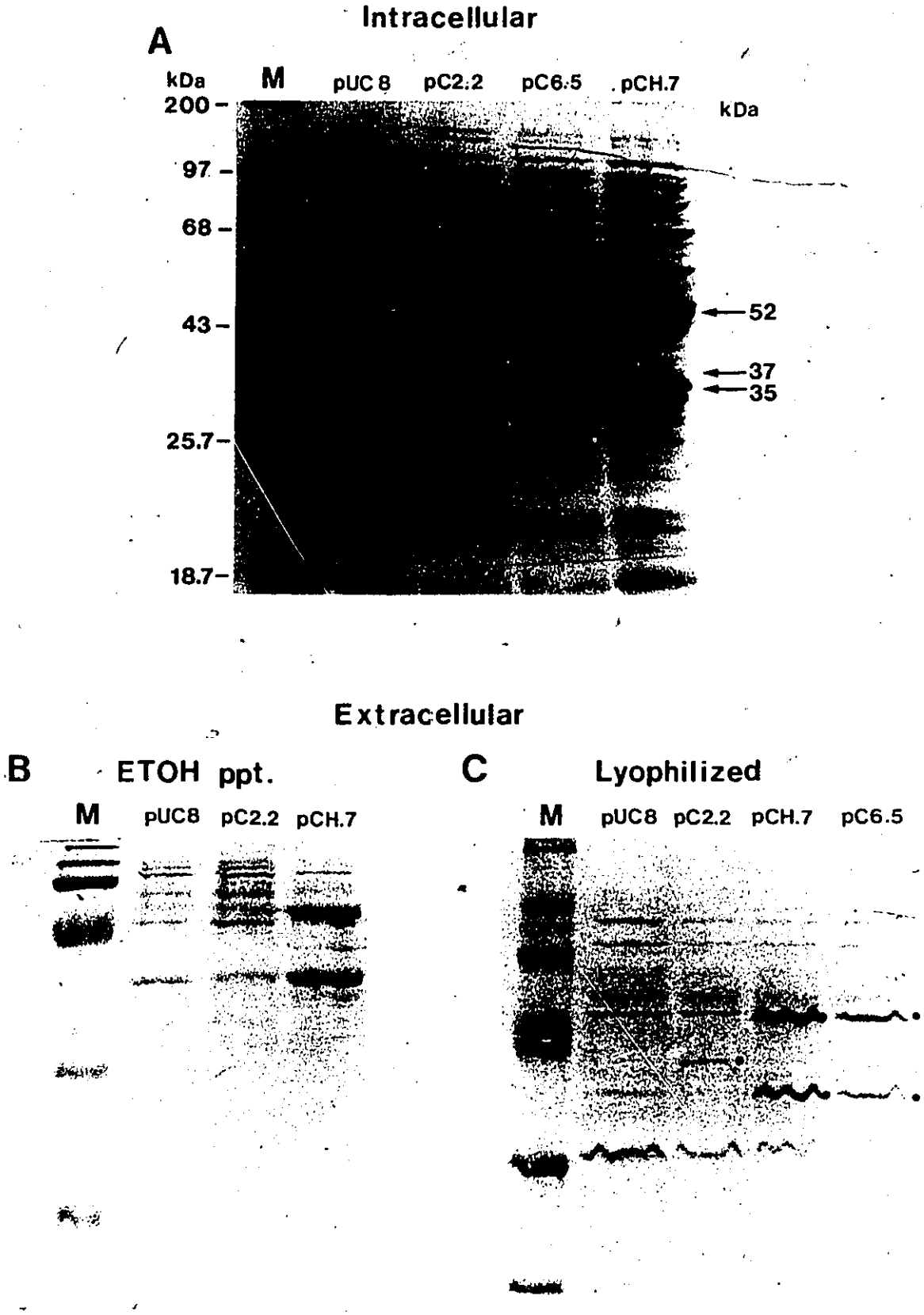

Intracellular
A
kDa
M
pUC 8
pC2.2
pC6.5
pCH.7
kDa
200
97
68
43
25.7
18.7
52
37
35
Extracellular
B
ETOH ppt.
M
pUC8
pC2.2
pCH.7
C
Lyophilized
M
pUC8
pC2.2
pCH.7
pC6.5

Table 4: Expression of endoglucanase activity by *B. subtilis*, *B. circulans*, and *B. polymyxa*, compared to their respective endoglucanase genes in *E. coli*.

	carbon source	enzyme units/ml X 10^2	specific activity units/mg	biomass O.D.$_{600nm}$	s.a. per O.D. biomass
B. circulans					
NRCC 9024	none	2.8	0.07	2.17	0.03
	1% glycerol	8.1	0.68	4.70	0.15
	1% glucose	3.2	0.15	2.23	0.07
	3% glucose	3.4	0.17	6.39	0.03
	1% cellobiose	0.4	0.05	1.23	0.04
B. polymyxa					
NRCC 2822	none				
	1% glycerol	3.5	0.21	4.86	0.04
B. subtilis					
PAP 115	none	0.8	0.11	3.62	0.03
	1% glycerol	8.7	0.44	5.22	0.08
	1% glucose	0.8	0.55	5.00	0.11
	3% glucose	2.3	0.19	6.84	0.03
	1% cellobiose	1.0	0.13	1.78	0.07
E.coli HB101					
pC2.2 (2822)	1% glucose	21.9	4.66	1.25	3.73
E. coli HB101 grown in yeast nitrogen base with:					
Plasmids					
pCH7 (9024)	1% glucose	4.5	4.30	0.39	11.05
	3% glucose	4.4	6.85	0.25	27.72
	1% arabinose	1.0	2.88	0.19	15.16
pC2.2 (2822)	1% glucose	0.2	0.27	0.47	0.57
	3% glucose	0.9	1.35	0.32	4.21
	1% arabinose	0.9	4.53	0.22	20.57
pC6.5 (PAP115)	1% glucose	2.7	2.90	0.44	6.63
	3% glucose	1.2	2.50	0.30	8.49
	1% arabinose	5.9	7.31	0.30	24.35

Enzyme units for Table 4 are expressed in ß-1,4 linkages broken per minute. The specific activity is given for the units of endoglucanase activity per mg of total extracellular protein. For the minimal media used, see Material and Methods, section A.

B. subtilis (Robson and Chambliss, 1984) and *B. polymyxa* (Fogarty and Griffin, 1973a). Fogarty and Griffin (1973b; also Griffin and Fogarty, 1971) have shown that more protease is produced by *B. polymyxa* as a result of growing it on arabinose as the sole carbon source. The increased production of proteases has been postulated as a cause for increased cellulase activity in the culture supernatants of hyperproducing mutants of *T. reesei* (Sheir-Neiss and Montenecourt, 1984). The protease was thought to be freeing more of the cellulase attached to the cell membrane into the liquid media. In contrast to arabinose, glycerol as the carbon source produced the lowest amount of protease but the highest biomass for *B. polymyxa* (Fogarty and Griffin, 1973b).

The minimal medium that the bacilli were grown on supported growth without the addition of any carbon source. Both *B. circulans* and *B. subtilis* produced more enzyme per OD of biomass at a concentration of 1% glucose than with either no added glucose or 3% glucose. This suggests that although these genes are constitutively expressed their transcription is effected by the different metabolic state of the microbe with the higher level of glucose in the medium.

In the *E. coli* clones, an increase of initial glucose concentration in the medium had produced the opposite effect to that shown by the bacilli. The increase of glucose produced an increase in the amount of endoglucanase in the supernatant. Arabinose had an even greater positive effect on endoglucanase production.

As expected, the transformed *E. coli* cells with the high gene copy number produced much more endoglucanase than the parent *Bacillus* species. Characterization of the *B. circulans* gene in the Results, section C showed a gene that appeared highly homologous to the *B. subtilis* endoglucanase gene but neither the *B. circulans* endoglucanase gene nor its clone in *E. coli* were expressed in the same manner as their *B. subtilis* counterparts.

The experiments in Table 4 were not completed as no important trends in the results were evident and the questions addressed were not fundamental to the objectives of this thesis. Analysis of the data in Table 4 must be qualified as each measurement was

taken only once.

E. Cloning the 5′ end of the *B. polymyxa* endoglucanase gene

It was evident from the subcloning results in section D and the first nucleotide sequencing results of the *B. polymyxa* PstI DNA fragment that part of the 5′end of the endoglucanase gene was missing and that transcription was due to the *lac*Z promoter and translation began with the AUG start codon from the *lac*Z gene. A large open reading frame (ORF) was found to be in frame with the 5′ end of the *lac*Z α-peptide ORF of the vectors mp9 and pUC9 (Vieira and Messing, 1982). These were the only two vectors that were capable of expressing the *B. polymyxa* endoglucanase. The fusion sequence is given in Fig. 7A. An example of increased endoglucanase production by inducing the *lac* operon with IPTG is shown in Fig. 7B.

To isolate the promoter and 5′end of the endoglucanase gene a HindIII library of *B. polymyxa* was created in mp19 using the 'restriction endonuclease site fill in' method of Zabarovsky and Allikmets, (1986) described in the Material and Methods, section D.1. From the genomic map in Fig. 4, a 6.2 kbp HindIII fragment contained all of the gene. Screening for 'positive' clones was done by the Congo Red plate assay. Two positives expressing plaques (clones Bp-C1, Bp-C2) were found in a library of approximately 10,000 clones. However, it was not possible to isolate a pure culture of either of the positively expressing phage, even after four screenings. A large population of CR^- phage were always present when grown in liquid media and purification of phage from plates still produced a large family of different sized phage DNA. These results suggested that this fragment was possibly unstable in M13.

As an alternative approach a 2.35 kbp EcoRI fragment which also would be expected to contain the 5′end of the endoglucanase gene was cloned with the objective to later fuse it with the already cloned portion of the gene. Genomic DNA was digested with EcoRI and electrophoresed on a 0.7% agarose gel. A portion of the DNA in the gel

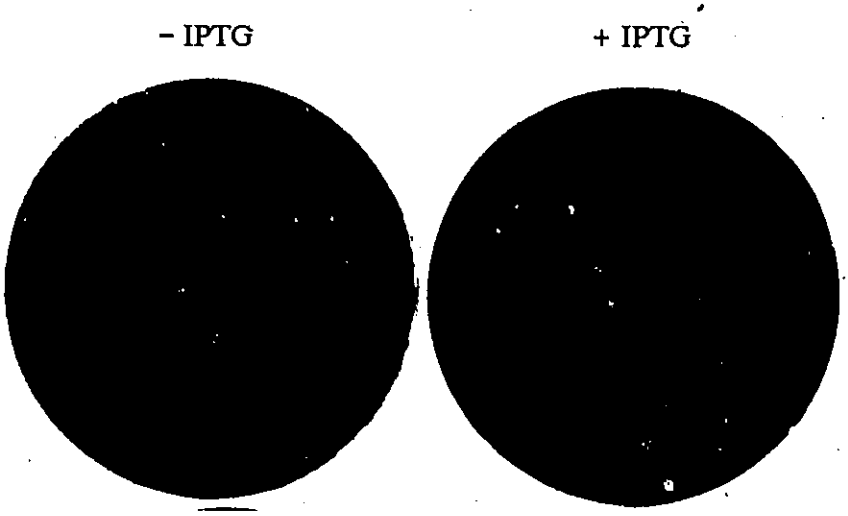
FIGURE 7. 5′ END FUSION OF *B. polymyxa* ENDOGLUCANASE TO *lacZ* α-PEPTIDE.

(A) The 5′ end of the α-peptide gene ends at the PstI site, the amino acids for this region are bold type. The *B. polymyxa* DNA which starts at the PstI site codes for a 370 amino acid ORF that is in frame with the 5′ end of the *lacZ* α-peptide gene. (B) Bp-P1 and *E. coli* JM101 were plated with B media and CMC with or without the gratuitous inducer of the *lac* operon, IPTG.

A

fMet Thr Met Ile Thr Pro Ser Leu Ala Ala Asp Ala Ala Ser Val ...
ATG ACC ATG ATT ACG CCA AGC TTG GCT GCA GAT GCA GCC AGT GTG ...
PstI
<——M13mp9 or pUC9——|——*B. polymyxa* DNA——>

B



corresponding to a size fraction of 2 to 2.8 kbp was excised and eluted from the gel slice by electrophoresis (see Material and Methods, section C.4(b)). This DNA was ligated with CIP treated RF of M13mp19 cut with EcoRI. After transfection in *E. coli* JM103 the plaques were screened with a radioactive DNA probe made from the 800 bp PstI/EcoRI fragment located at the 5′ end of the already cloned PstI fragment. Six clones gave strong positive signals from 1000 clones in the library. Two of the clones (E1.2, E2.1) were plaque purified and found to contain the 2.35 kbp EcoRI fragment.

The 2.35 kbp EcoRI fragment when cut with PstI gave an 820 bp fragment which contained most of the coding region of the endoglucanase gene and a 1.53 kbp piece of DNA which coded for 87 nucleotides of the 5′ region of the protein (see Fig.12, Results, section F.2) plus the putative promoter region of the gene. The 1.53 kbp PstI/EcoRI fragment was ligated to the original 1.6 kbp PstI fragment and fractionated by electrophoresis on a preparative gel in order to purify the desired 3.1 kbp product. This 3.1 kbp EcoRI/PstI fragment was further ligated into both mp18 and mp19 and used to transfect *E. coli*. As a control the 1.6 kbp PstI fragment used for this experiment was also ligated into pUC 9 to show that it could still confer activity.

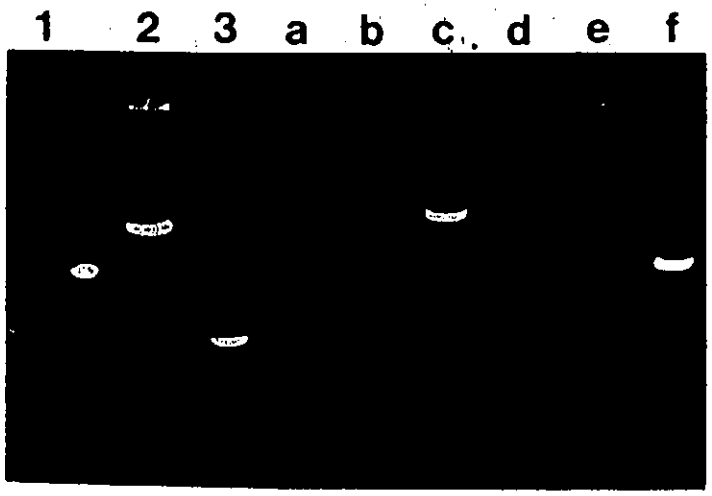
Although the pUC 9 -PstI clones exhibited activity, none of the approximately 1000 resulting clones from the 5′,3′ gene fusion expressed any endoglucanase activity. Single stranded DNA of four clones from each of the mp18 and mp19 ligations was purified. These clones were analyzed with the S1 nuclease protection assay using the 2.35 kbp EcoRI cloned fragment (i.e. E1.2 and opposite orientation E4.1) to show that the 5′ and 3′ fragments were indeed in the proper orientation to each other. As shown in Fig. 8, correct orientation gave only one 2.35 kbp double stranded piece of DNA. If the orientation of either the 3′ or 5′ fragments respective to each other was incorrect, no fragment or either the 1.53 kbp or 0.8 kbp, fragment would be recovered after S1 nuclease treatment. However the reconstituted gene did not express any activity in either orientation, with or

FIGURE 8. S1 NUCLEASE PROTECTION ASSAY OF *B. polymyxa* ENDOGLUCANASE 5′-3′ GENE FUSION CLONES.

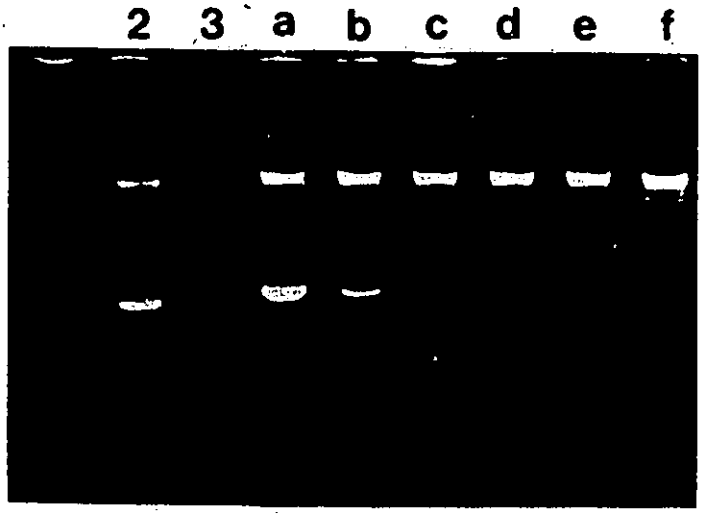
Two mp18 (lanes a-b) and four mp19 (lanes c-f) 5′-3′ fusion clones were hybridized to the *B. polymyxa* 2.35 kbp clones E1.2 (panel A) or E4.1 (panel B) and digested with S1 nuclease. The possible resulting sizes are shown in lanes 1 to 3.

1 = 1.5 kbp fragment used to construct 5′ regions of the gene
2 = S1 nuclease protection assay of E2.1/E4.1
3 = S1 nuclease protection assay of E2.1/Bp-op19

A



B



without the addition of IPTG. A possible explanation for this is given in the Results, section F.1.

F. DNA sequence of the complete *B. polymyxa* endoglucanase gene.

Both the 1.6 kbp PstI fragment and the non-overlapping section of the 2.35 EcoRI fragment were sequenced completely in both orientations. Most of the sequences were obtained from 3′ deletion clones constructed according to Dale et al. (1985). A map showing the deletion clones used for sequencing is presented in Fig. 9.

1. 5′ region of DNA fragment

The non-coding 5′ end of the putative *B. polymyxa* endoglucanase gene extending up to the putative ATG start codon is shown in Fig. 10. A typical Shine-Delgarno ribosome binding site GGAGG was found 11 bases upstream from the putative ATG start codon. The 300 bases extending from the start codon was searched for known -10 and -35 *E. coli* and *B. subtilis* RNA polymerase recognition sites using the search function of the *Microsoft WORD* program. The best -10 and -35 *E. coli* like sequences marked in Fig. 10 are most likely too far apart to be used by an *E. coli* promoter. The lack of typical -10 and -35 *E. coli* promoter sequences is in agreement with the lack of expression of the reconstituted *B. polymyxa* endoglucanase gene in the Results, section E. A sequence resembling the -10 and -35 regions of the *B. subtilis* σ^{32} recognition sites is also marked. It is not known whether the variation of the recognition sites that exist in the *B. polymyxa* sequence would allow a *B. subtilis* polymerase with the σ^{32} sigma factor to transcribe the gene.

Two other sequences, which are related to this endoglucanase and are known to be transcribed using their own promoters in *E. coli*, were also analyzed in Fig. 10. The *B. polymyxa* ß-amylase is the only known published *B. polymyxa* sequence; it shows little similarity to the endocellulase gene. However, the *Clostridium thermocellum cel*B gene has the highest amino acid sequence similarity to the *B. polymyxa* endoglucanase (see Results, section G.1). There has been no transcriptional mapping done on any of the genes used in Fig. 10. The 5′ UTR sequences of all three genes were searched for consensus -10 and

FIGURE 9. DELETION CLONES USED FOR SEQUENCING THE *B. polymyxa* ENDOGLUCANASE GENE.

The deletion clones made by the method of Dale et al. (1985) used for sequencing the overlapping 2.35 kbp EcoRI and 1.6 kbp PstI fragments are shown here graphically. The open reading frame corresponding to the endoglucanase gene is shown as a large black box on a line which symbolizes the total length of genomic DNA sequenced. The location, orientation and the number of bases sequenced of each clone is shown by arrows. The name of each clone is shown above the arrows. The clones E1.2 and E4.1 are the two orientations of the full length 2.35 kbp EcoRI fragment. OP19 (Bp-OP19) and Bp-P1 refer to the opposite orientation clones of the 1.6 kbp PstI fragment. Bp-A, Bp-B, Bp-C are primers used to sequence the full length clones in regions where sequence data had been ambiguous.

E = EcoRI
P = PstI

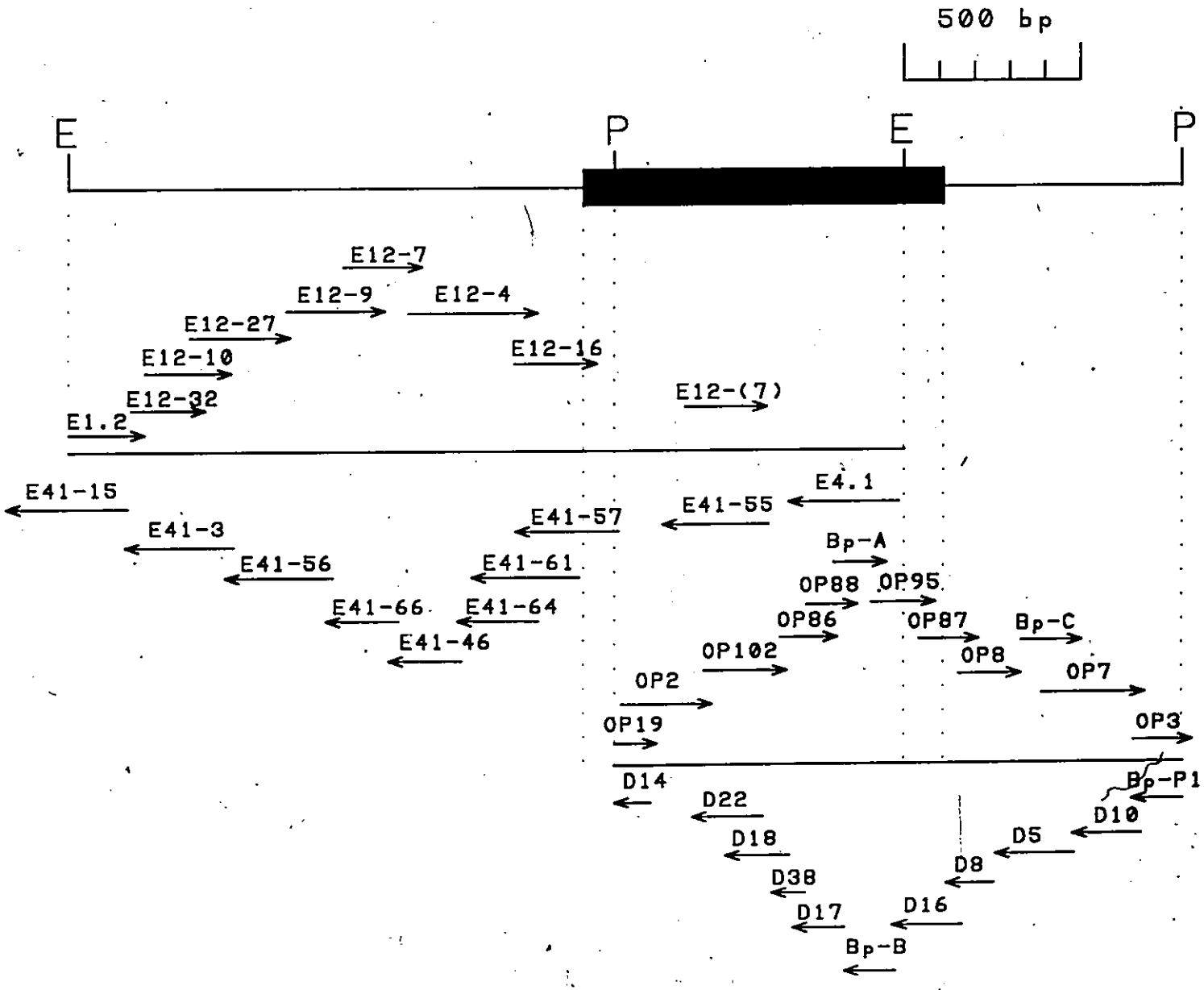
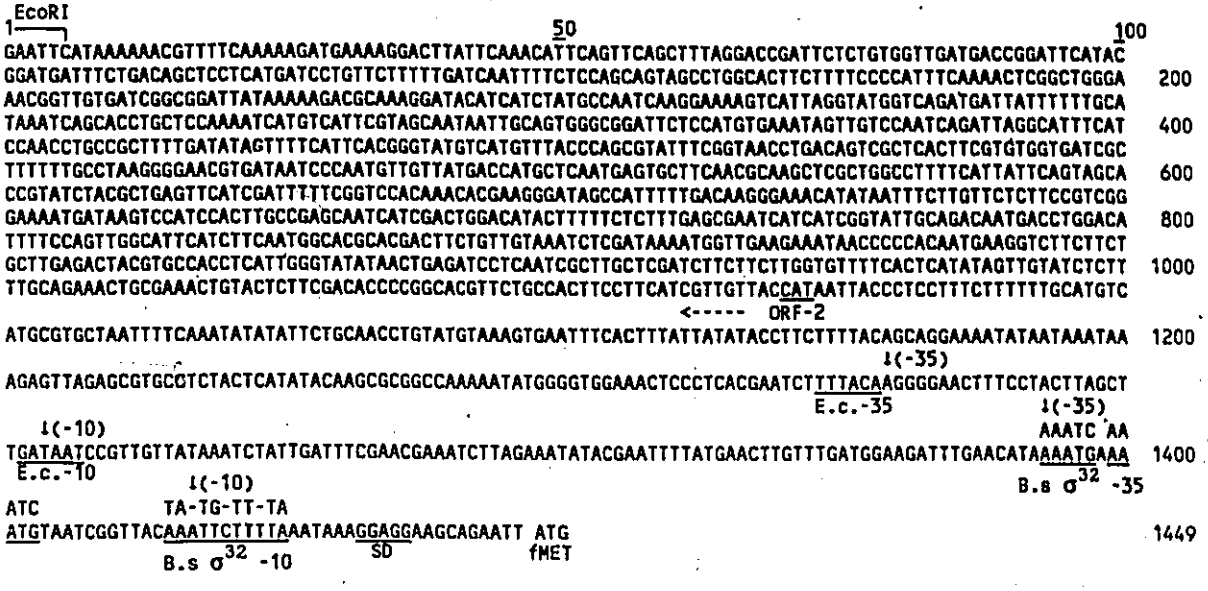

500 bp
E
P
E
P
E12-7
E12-9
E12-4
E12-27
E12-16
E12-10
E12-32
E12-(7)
E1.2
E41-15
E4.1
E41-55
E41-57
E41-3
Bp-A
E41-56
E41-61
OP88
OP95
E41-66
E41-64
OP86
OP87
Bp-C
E41-46
OP102
OP8
OP7
OP2
OP19
OP3
D14
Bp-P1
D22
D10
D18
D5
D8
D38
D17
D16
Bp-B

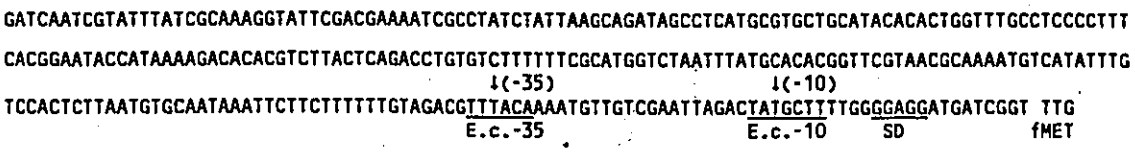
FIGURE 10. DNA SEQUENCE OF *B. polymyxa* DNA UPSTREAM OF THE PUTATIVE ENDOGLUCANASE CODING REGION OF THE ENDOGLUCANASE GENE AND THE -1 TO -300 DNA SEQUENCE OF TWO RELATED GENES, THE *B. polymyxa* ß-AMYLASE, AND THE *C. thermocellum celB*.

(A) The sequence of *B. polymyxa* DNA from the 5′ EcoRI site up to and including the putative start codon of the endoglucanase gene. The complementary sequence for the start codon of ORF2 is shown at nucleotide position 1072 with a dashed arrow showing the direction of transcription. Promoter sequences identical to the -10 and -35 region of known *E. coli* σ^{70} recognition sequences that are also spaced from 15-19 bases apart have been underlined and labelled underneath as E.c -10 and E.c -35 respectively. The consensus *E. coli* σ^{70} recognition sequences are the same as the consensus recognition sequences for *B. subtilis* σ^{43}, both are considered the major sigma factor in the cells and are responsible for the transcripts of vegetative proteins. Since the transcriptional start site (base position +1) is unknown in the sequences of this figure, the -35 point (↓-35) is measured from a consensus -10 position (↓-10) of the -10 promoter sequence . The only sequences similar to other consensus promoters of *B. subtilis* RNA polymerase sigma factors is σ^{32} which have been underlined and labelled 'B.s σ^{32}'. The consensus sequence for the -10 and -35 recognition sites for *B. subtilis* σ^{32} are written above the *B. polymyxa* sequence. SD refers to the consensus Shine-Delgarno nucleotide sequence (GGAGG) responsible for the correct binding of ribosomes to bacterial mRNA. The -1 to -300 5′sequence of two related genes *B. polymyxa* ß-amylase (Kawazu et al., 1987) and *C. thermocellum celB* (Grépinct and Béguin, 1986) are shown in (B) and (C) respectively. Both genes are transcribed in *E. coli* with their own promoters. Putative -10 and -35 *E. coli* like promoter sequences and their relative position to each other have been underlined and marked with vertical arrows as in (A). No *B. subtilis* like promoter sequences (other than *B. subtilis* σ^{43}) were found in either of these two sequences.

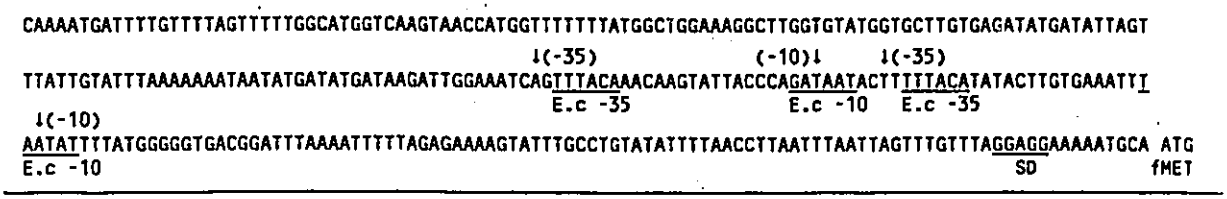
A Nucleotide sequence of DNA upstream of *B. polymyxa* endoglucanase start codon



B 5′ region upstream of *B. polymyxa* ß-amylase gene



C 5′region upstream of ATG start codon of *C. thermocellum celB*



-35 sequences for known promoters recognized by *E. coli* and *B. subtilis* RNA polymerases. Some of the core regions used to search the 5′ sequences are shown in Table 5. Variations of the -10 and -35 sequences also used in the search but not included in Table 5 are from 54 promoters recognized by *E. coli* sigma factors and compiled by Siebenlist et al.(1980; see Appendix B). No *B. subtilis* -10 and -35 recognition sites were found in either the ß-amylase or *C. thermocellum celB* sequences. *E. coli* σ^{70} recognition sites were found in the ß-amylase sequence, but the -10 and -35 like sequences found in the *celB* sequence were closer than has been found in functional *E. coli* promoters. It is possible that an *E. coli*-like promoter is further upstream or downstream from this 300 bp region.

Upstream of the endoglucanase gene there is a second ORF, ORF-2. The translation of this putative 187 amino acid protein would begin at the ATG start codon at nt position 1064 and end at nt position 511; the ORF-2 gene would be transcribed in the opposite direction to the endoglucanase gene. Figure 11 gives the DNA sequence, putative translation product and regulatory sequences of ORF-2. The amino acid sequence has also been used to search the NBRF and SWIS/PROT protein data base. A 28% similarity was found with the 1st 186 amino acids of the *E. coli lac* repressor protein. This suggests that this protein could be or could have been a DNA binding protein with some regulatory function for the endoglucanase gene or some other distal gene.

The ORF-2 sequence has a typical bacterial ribosome binding site 5' to the start codon and -10 and -35 sequences that are similar to the major vegetative promoter sequences of *E. coli* and *B. subtilis*. There is also a -10 and -35 sequence similar to the consensus recognition sequences for *B. subtilis* σ^{29}. In the 3′ region there is an imperfect inverted repeat that could possibly function as a *rho* independent transcription termination site.

2. Protein coding region

The nucleotide sequence of the coding region, the 3′ untranslated region and putative amino acid sequence of the *B. polymyxa* endoglucanase gene are given in figure 12.

Table 5. Sequences of -10 and -35 regions recognized by *E. coli* and *B. subtilis* RNA polymerases with various sigma factors.

promoter/ sigma factor	-35	-10	References
E. coli			
σ^{70}	* TTGACA	TATAAT	Travers, 1987
P_{cclA}	TTGTAT	TATAAT	Beguin et al., 1986
P_{N25}	TTGCTT	TATAAT	Brunner and Bujard, 1987
P_{a1}	TTGACT	GATACT	"
P_{tac}	TTGACA	TATAAT	"
$P_{lacUV25}$	TTTACA	TATAAT	"
P_L	TTGACA	GATACT	"
P_{bla}	TTCAAA	GACAAT	"
B. subtilis			
σ^{43}	* TTGACA	TATAAT	Johnson et al., 1983
	TTGACT	AATAAT	Pero et al., 1982
	TTGACT	CATAAT	"
	TTGAAA	TATATT	"
SPO1 σ^{gp28}	T-AGGAGA--A	TTTATTT	"
	T-AGGAGA--A	TTTGTTT	"
	T-AGGTGA--A	TTTCTTT	"
	T-ATGAGA--A	TGTTTTT	"
	T-AGGAGA--A	TTTATTT	"
σ^{28}	* CTAAA	CCGATA	Gilman et al., 1982
σ^{29}	* TT-AAA	CATATT	Johnson et al., 1983
σ^{32}	* AAATC	TA-TG-TT-TA	"
SPO1 $\sigma^{gp33-34}$	* CGTTAGA	GATATT	"
σ^{37}	* AGGTT	GG-ATTG T	"
		TAATGCTT	Pero et al., 1982
		AATTGATA	"
	GGTTTAAA	TATTGTTT	"

note: sequences marked with * are consensus sequences

FIGURE 11. NUCLEOTIDE AND PUTATIVE AMINO ACID SEQUENCE OF ORF-2.

Nucleotide position numbers are placed on top of the DNA sequence. The numbering used is with respect to the nucleotide positions of the DNA sequence figures for the complete piece of *B. polymyxa* genomic DNA (Fig 10, Fig.12). The numbers in brackets on the far right of the sequences are the amino acid positions of the last amino acid on each line of sequence. The complementary strand position of the ATG start codon for the endoglucanase has been underlined and labelled. Possible *B. subtilis* and *E. coli* -10 and -35 promoter sequences have been underlined. The positioning of the -10 and -35 sequences have been marked with vertical arrows as in Fig. 10. Since again, no transcriptional start site (base + 1) is known, the -35 (↓-35) and -10 (↓-10) position has been marked relative to the consensus position of the -10 recognition sequence. SD refers to the putative Shine-Delgarno ribosome binding site. A possible transcription termination site is shown by a pair of horizontal arrows which underline an imperfect inverted repeat in the sequence 3′ to the coding region. .

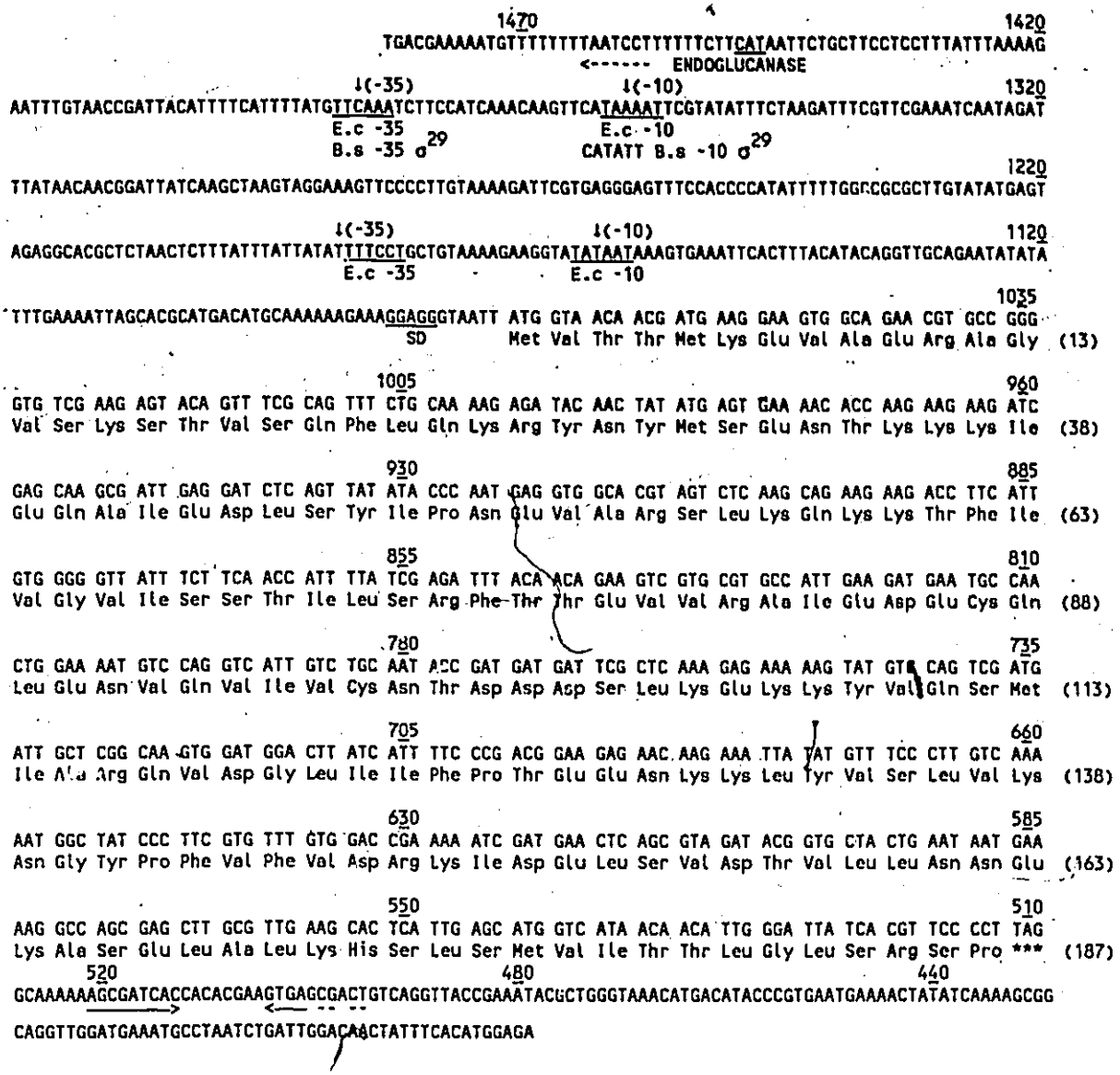

1470
1420
TGACGAAAAATGTTTTTTTTAATCCTTTTTTCTTCATAATTCTGCTTCCTCCTTTATTTAAAAG
<------ ENDOGLUCANASE
↓(-35)
↓(-10)
1320
AATTTGTAACCGATTACATTTTCATTTTATGTTCAAATCTTCCATCAAACAAGTTCATAAAATTCGTATATTTCTAAGATTTCGTTCGAAATCAATAGAT
E.c -35
E.c -10
B.s -35 σ29
CATATT B.s -10 σ29
1220
TTATAACAACGGATTATCAAGCTAAGTAGGAAAGTTCCCCTTGTAAAAGATTCGTGAGGGAGTTTCCACCCCATATTTTTGGCCGCGCTTGTATATGAGT
↓(-35)
↓(-10)
1120
AGAGGCACGCTCTAACTCTTTATTTATTATATTTTCCTGCTGTAAAAGAAGGTATATAATAAAGTGAAATTCACTTTACATACAGGTTGCAGAATATATA
E.c -35
E.c -10
1035
TTTGAAAATTAGCACGCATGACATGCAAAAAAGAAAGGAGGGTAATT ATG GTA ACA ACG ATG AAG GAA GTG GCA GAA CGT GCC GGG
SD Met Val Thr Thr Met Lys Glu Val Ala Glu Arg Ala Gly (13)
1005
960
GTG TCG AAG AGT ACA GTT TCG CAG TTT CTG CAA AAG AGA TAC AAC TAT ATG AGT GAA AAC ACC AAG AAG AAG ATC
Val Ser Lys Ser Thr Val Ser Gln Phe Leu Gln Lys Arg Tyr Asn Tyr Met Ser Glu Asn Thr Lys Lys Lys Ile (38)
930
885
GAG CAA GCG ATT GAG GAT CTC AGT TAT ATA CCC AAT GAG GTG GCA CGT AGT CTC AAG CAG AAG AAG ACC TTC ATT
Glu Gln Ala Ile Glu Asp Leu Ser Tyr Ile Pro Asn Glu Val Ala Arg Ser Leu Lys Gln Lys Lys Thr Phe Ile (63)
855
810
GTG GGG GTT ATT TCT TCA ACC ATT TTA TCG AGA TTT ACA ACA GAA GTC GTG CGT GCC ATT GAA GAT GAA TGC CAA
Val Gly Val Ile Ser Ser Thr Ile Leu Ser Arg Phe Thr Thr Glu Val Val Arg Ala Ile Glu Asp Glu Cys Gln (88)
780
735
CTG GAA AAT GTC CAG GTC ATT GTC TGC AAT ACC GAT GAT GAT TCG CTC AAA GAG AAA AAG TAT GTT CAG TCG ATG
Leu Glu Asn Val Gln Val Ile Val Cys Asn Thr Asp Asp Asp Ser Leu Lys Glu Lys Lys Tyr Val Gln Ser Met (113)
705
660
ATT GCT CGG CAA GTG GAT GGA CTT ATC ATT TTC CCG ACG GAA GAG AAC AAG AAA TTA TAT GTT TCC CTT GTC AAA
Ile Ala Arg Gln Val Asp Gly Leu Ile Ile Phe Pro Thr Glu Glu Asn Lys Lys Leu Tyr Val Ser Leu Val Lys (138)
630
585
AAT GGC TAT CCC TTC GTG TTT GTG GAC CGA AAA ATC GAT GAA CTC AGC GTA GAT ACG GTG CTA CTG AAT AAT GAA
Asn Gly Tyr Pro Phe Val Phe Val Asp Arg Lys Ile Asp Glu Leu Ser Val Asp Thr Val Leu Leu Asn Asn Glu (163)
550
510
AAG GCC AGC GAG CTT GCG TTG AAG CAC TCA TTG AGC ATG GTC ATA ACA ACA TTG GGA TTA TCA CGT TCC CCT TAG
Lys Ala Ser Glu Leu Ala Leu Lys His Ser Leu Ser Met Val Ile Thr Thr Leu Gly Leu Ser Arg Ser Pro *** (187)
520
480
440
GCAAAAAAGCGATCACCACACGAAGTGAGCGACTGTCAGGTTACCGAAATACGCTGGGTAAACATGACATACCCGTGAATGAAAACTATATCAAAAGCGG
CAGGTTGGATGAAATGCCTAATCTGATTGGACAACTATTTCACATGGAGA

FIGURE 12. DNA AND PUTATIVE AMINO ACID SEQUENCE OF *B. polymyxa* ß-1,4-ENDOGLUCANASE GENE AND 3′ UNTRANSLATED REGION.

The protein coding region is from nucleotide 1447 to 2640 of the 3135 base EcoRI - PstI fragment. The numbers on top of the sequences are the amino acid position while the numbers in brackets on the far right are the nucleotide position of the last nucleotide on the line. The nucleotide position numbers continue from the numbering of the 5′ region of the 3135 base EcoRI - PstI DNA fragment in Fig. 10. The 5′ PstI site that was initially fused to the *lac*Z promoter (Fig. 7) has been marked over amino acids 29-31. The 3′ EcoRI site of the 2.35 Kbp EcoRI fragment has been marked over amino acids 317-318. Three inverted repeats capable of forming a hairpin and loop structure in the 3′ UTR of the gene have been marked with arrows under the sequences.

```
 1               5                  10                  15                  20                  25
ATG AAG AAA AAA GGA TTA AAA AAA ACA TTT TTC GTC ATT GCC TCC CTC GTA ATG GGC TTC ACA CTG TAT GGC TAT   (1521)
met LYS LYS LYS GLY LEU LYS LYS THR PHE PHE VAL ILE ALA SER LEU VAL MET GLY PHE THR LEU TYR GLY TYR
                PstI
                35                  40                  45                  50
ACA CCC GTT TCT GCA GAT GCA GCC AGT GTG AAA GGA TAT TAT CAC ACC CAA GGA AAC AAG ATT GTA GAC GAA TCC   (1596)
THR PRO VAL SER ALA ASP ALA ALA SER VAL LYS GLY TYR TYR HIS THR GLN GLY ASN LYS ILE VAL ASP GLU SER
                55                  60                  65                  70                  75
GGG AAA GAA GCG GCA TTT AAC GGC CTG AAC TGG TTC GGT CTG GAA ACT CCT AAT TAC ACC TTG CAT GGA CTG TGG   (1671)
GLY LYS GLU ALA ALA PHE ASN GLY LEU ASN TRP PHE GLY LEU GLU THR PRO ASN TYR THR LEU HIS GLY LEU TRP
                80                  85                  90                  95                  100
AGC CGC TCA ATG GAC GAC ATG CTG GAT CAG GTG AAG AAA GAA GGC TAC AAT CTG ATT CGT CTG CCT TAC AGC AAT   (1740)
SER ARG SER MET ASP ASP MET LEU ASP GLN VAL LYS LYS GLU GLY TYR ASN LEU ILE ARG LEU PRO TYR SER ASN
                105                 110                 115                 120                 125
CAG TTG TTC GAT TCC AGT TCC CGT CCA GAC AGT ATT GAT TAT CAC AAA AAC CCT GAT CTG GTC GGA TTA AAC CCG   (1821)
GLN LEU PHE ASP SER SER SER ARG PRO ASP SER ILE ASP TYR HIS LYS ASN PRO ASP LEU VAL GLY LEU ASN PRO
                130                 135                 140                 145                 150
ATT CAA ATT ATG GAC AAG CTG ATC GAA AAA GCT GGA CAA CGC GGT ATT CAG ATT ATC CTT GAC CGT CAC CGT CCA   (1896)
ILE GLN ILE MET ASP LYS LEU ILE GLU LYS ALA GLY GLN ARG GLY ILE GLN ILE ILE LEU ASP ARG HIS ARG PRO
                155                 160                 165                 170                 175
GGC TCA GGT GGG CAA TCC GAG CTG TGG TAC ACA TCC CAG TAC CCT GAG TCT CGC TGG ATT AGT GAC TGG AAA ATG   (1971)
GLY SER GLY GLY GLN SER GLU LEU TRP TYR THR SER GLN TYR PRO GLU SER ARG TRP ILE SER ASP TRP LYS MET
                180                 185                 190                 195                 200
TTG GCT GAT CGT TAT AAA AAT AAC CCC ACC GTC ATT GGT GCG GAT TTG CAC AAC GAG CCA CAC GGT CAA GCA AGC   (2046)
LEU ALA ASP ARG TYR LYS ASN ASN PRO THR VAL ILE GLY ALA ASP LEU HIS ASN GLU PRO HIS GLY GLN ALA SER
                205                 210                 215                 220                 225
TGG GGT ACA GGC AAT GCC TCC ACA GAC TGG CGT CTG GCG GCA CAA CGT GCA GGG AAT GCG ATT CTG TCC GTG AAT   (2121)
TRP GLY THR GLY ASN ALA SER THR ASP TRP ARG LEU ALA ALA GLN ARG ALA GLY ASN ALA ILE LEU SER VAL ASN
                230                 235                 240                 245                 250
CCG AAT TGG CTG ATT CTC GTA GAA GGT GTA GAC CAC AAT GTA CAA GGC AAC AAT AGC CAA TAC TGG TGG GGT GGC   (2196)
PRO ASN TRP LEU ILE LEU VAL GLU GLY VAL ASP HIS ASN VAL GLN GLY ASN ASN SER GLN TYR TRP TRP GLY GLY
                255                 260                 265                 270                 275
AAC CTG ACA GGT GTA GCC AAC TAT CCT GTC GTT CTG GAC GTA CCG AAC CGT GTC GTA TAT TCT CCA CAC GAT TAC   (2271)
ASN LEU THR GLY VAL ALA ASN TYR PRO VAL VAL LEU ASP VAL PRO ASN ARG VAL VAL TYR SER PRO HIS ASP TYR
                280                 285                 290                 295                 300
GGC CCC GGT GTG TCT TCG CAG CCA TGG TTC AAC GAC CCG GCC TTC CCG TCC AAC CTG CCA GCG ATC TGG GAT CAA   (2346)
GLY PRO GLY VAL SER SER GLN PRO TRP PHE ASN ASP PRO ALA PHE PRO SER ASN LEU PRO ALA ILE TRP ASP GLN
                                                    EcoRI
                305                 310                 315                 320                 325
ACC TGG GGC TAC ATC AGC AAA CAA AAC ATA GCT CCG GTG CTG GTT GGT GAA TTC GGC GGC CGT AAT GTT GAT TTG   (2421)
THR TRP GLY TYR ILE SER LYS GLN ASN ILE ALA PRO VAL LEU VAL GLY GLU PHE GLY GLY ARG ASN VAL ASP LEU
                330                 335                 340                 345                 350
TCC TCC CCT GAG GGG AAA TGG CAA AAT GCG CTT GTT CAC TAT ATT GGT GCC AAC AAC CTG TAC TTT ACG TAC TGG   (2476)
SER SER PRO GLU GLY LYS TRP GLN ASN ALA LEU VAL HIS TYR ILE GLY ALA ASN ASN LEU TYR PHE THR TYR TRP
                355                 360                 365                 370                 375
TCC CTG AAT CCG AAT AGC GGC GAC ACA GGC GGT CTG CTG CTG GAT GAC TGG ACT ACC TGG AAT CGT CCG AAG CAA   (2571)
SER LEU ASN PRO ASN SER GLY ASP THR GLY GLY LEU LEU LEU ASP ASP TRP THR THR TRP ASN ARG PRO LYS GLN
                380                 385                 390                 395
GAT ATG CTG GGT CGA ATT ATG AAG CCT GTT GTT TCC GTA GCC CAG CAA GCG GAA GCA GCA GCC GAA TAG GCACAGG   (2647)
ASP MET LEU GLY ARG ILE MET LYS PRO VAL VAL SER VAL ALA GLN GLN ALA GLU ALA ALA ALA GLU ***

CCTTCAACTTCACTTAATTAAAAAAGCCGATTTTCCTCCTTCAGGCCAGATTACTGCTGGGCTATGAAGATGTGGAGAAACGGCTTTTTTTGTTCTTCCA   (2747)

GTTTTACACATTGTAAATTCCCCATTGAATGGGTCATTTTGTAAACTAAATCTCTCTTTCAACCGATAAATACATATGTTGGATTTTAATTTAGAAAGGA   (2847)

AGTATTCATGTGATTATTTGGAAGGGTTTTGGTATTTTAAATATTATTATCCCAGGGATTTTATTTGTTATTGTCGGTAGTTTAGTATCCGCTCTAGGGC   (2947)
TTGACTCCATAGATTCCCGGCTGCCAATGGCTTTTGTGTTCATTGTGTCGGGAGTTATTATCTGGTATTTGGGGAAAGCACTCAACTCCGACTCAAAGTG   (3047)
CTAGTTGATATGGAAACGGGACAACGCTATCGAATGGGAACCCAGCACAGTCTGTTTTTTATCCCTATGCATTATTGGGGCCCTGCAG                 (3135)
```

The PstI site which defines the 5′ end of the original expression clone (M13 Bp-P1 or pC2.2) is marked at amino acid position 33-34 in figure 12. The amino acid sequences 5′ to the PstI site appear to be part of a putative secretion signal sequence. Comparison of this sequence to the *B. polymyxa* ß-amylase secretion signal sequence is shown in figure 13. Both sequences exhibit a similar distribution of charged and hydrophobic amino acids comparable to signal sequences of *B. subtilis* (MacKay et al., 1986).

During the course of preparing 3′ deletions of the 1.6 kbp PstI fragment in mp9 for sequencing, each of the deletion clones was tested for endoglucanase activity to help localize the functional regions of the gene. Figure 14 shows that expression of CMCase activity is rapidly lost with removal of only a small amount of the coding region defining the carboxy end of the endoglucanase. These results are in marked contrast to those obtained for C-terminal deletions of the *B. subtilis* endoglucanase (Lo et al., 1988; Mackay et al., 1986). This same type of deletion analysis could not be as easily done with the 5′ end because there was no promoter for the gene when in the opposite orientation in the vector.

3. 3′ transcription termination

No simple transcription termination sequence of the inverted repeat type followed by 6 to 8 T's was found as described earlier by Platt and Bear (1983). Three inverted repeats were found in the 3′ UTR and have been marked in Fig. 12. There is a possibility that transcription is not terminated properly by any sequence of this DNA fragment when in *E. coli*. Figure 15 shows a plate assay for *lac*Z α-peptide formation in *E. coli* cells containing either pUC 9 or pUC 9 + 1.6 kbp PstI fragment (clone pC2.2). These results show that some α-peptide is made in the same colonies which express the endoglucanase gene. In Fig. 12 an ATG exists at nt position 3114 which could function as the initiation of translation for the truncated α-peptide. No consensus SD sequence appears upstream of this start codon. This would greatly effect the ability to translate the sequence and may be why the *E. coli* harbouring pC2.2 produce a lighter blue colored colony.

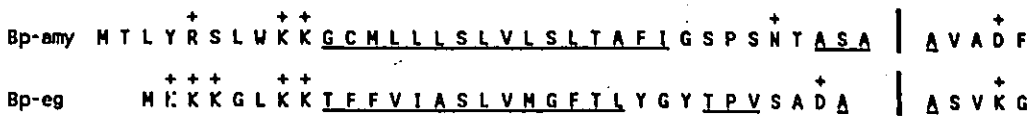


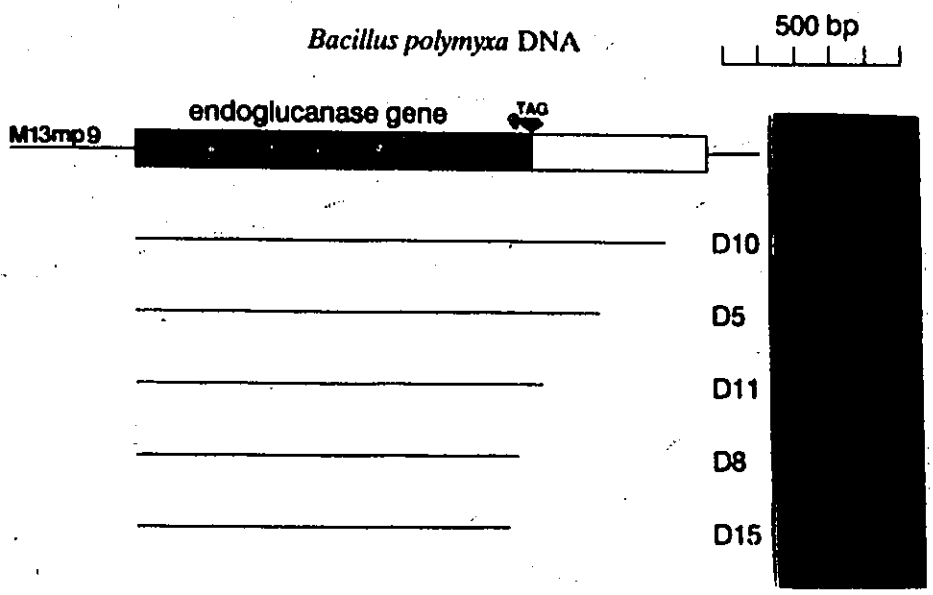
FIGURE 13. COMPARISON OF THE SECRETION SIGNAL SEQUENCE OF THE *B. polymyxa* ß-AMYLASE TO THE PUTATIVE SIGNAL SEQUENCE OF THE ß-1,4-ENDOGLUCANASE.

The hydrophobic regions underlined, were found using the parameters of Kyte and Doolittle (1982). Hydrophobic values for each amino acid were averaged out over 3 residues. Charged residues have been marked with a + sign. Vertical double bars mark the signal sequence protease cleavage site in the *B. polymyxa* ß-amylase (Kawazu et al., 1987) and the putative cleavage site in the *B. polymyxa* endoglucanase.

FIGURE 14. EXPRESSION OF 3′ DELETION CLONES OF THE *B. polymyxa* ENDOGLUCANASE GENE.

(A) Relative lengths of five 3′ end deletion clones are shown. The large box represents the size of the original PstI 1.6 kbp *B. polymyxa* DNA fragment isolated (Bp-P1). The activity of each clone was assayed by spotting 10 µl of culture supernatant on an agar plate containing 0.1% CMC, incubating for 0.5 hr at 50°C and then staining with Congo Red. Results are shown to the right of the DNA map. (B) The 3′ fusion between the *B. polymyxa* DNA and the vector DNA is shown for the shortest active clone and the longest inactive clones. The amino acids fused onto the truncated endoglucanase as a result of the fusion with the M13 DNA are shown in bold type.

A


Bacillus polymyxa DNA
500 bp
endoglucanase gene
TAG
M13mp9
D10
D5
D11
D8
D15

B

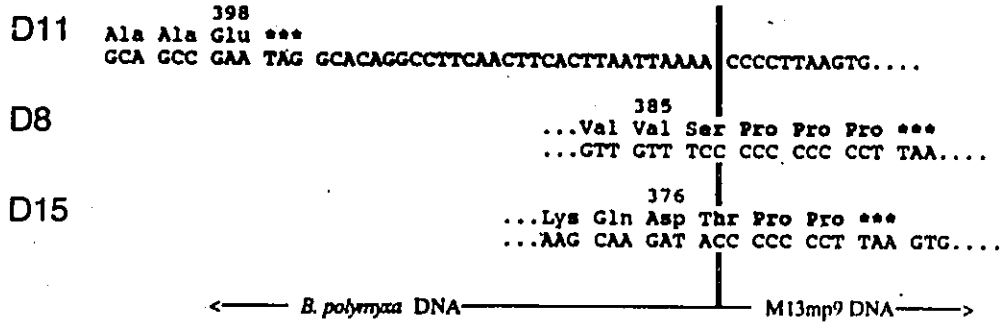
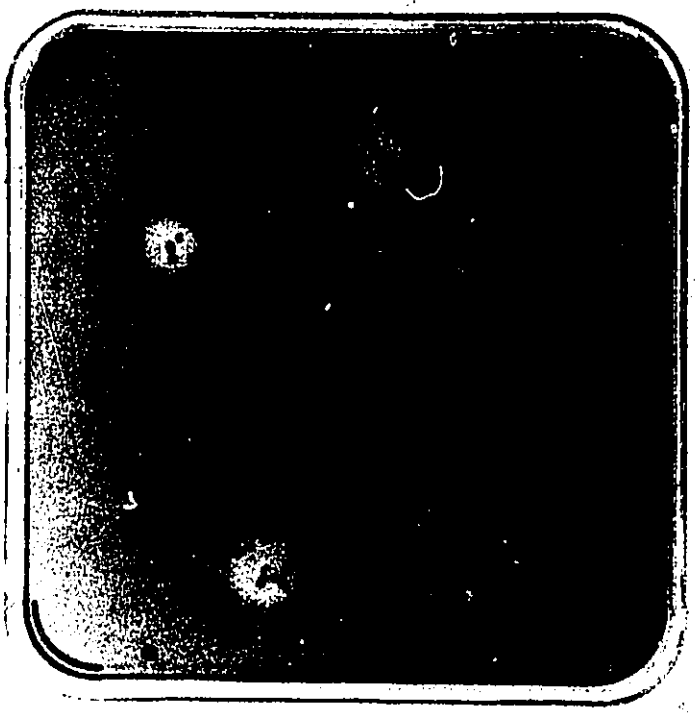

D11
398
Ala Ala Glu ***
GCA GCC GAA TAG GCACAGGCCTTCAACTTCACTTAATTAAAA CCCCTTAAGTG....
D8
385
...Val Val Ser Pro Pro Pro ***
...GTT GTT TCC CCC CCC CCT TAA....
D15
376
...Lys Gln Asp Thr Pro Pro ***
...AAG CAA GAT ACC CCC CCT TAA GTG....
B. polymyxa DNA
M13mp9 DNA

FIGURE 15. SIMULTANEOUS EXPRESSION OF THE *B. polymyxa* ENDOGLUCANASE AND THE *lac* α-PEPTIDE.

(A) *E. coli* JM101 transformed with either pUC9 or pC2.2 were grown on the same plate containing Bluo-gal and IPTG. (B) The same plate as in (A) after Congo Red assay. Black arrows denote *E. coli* colonies transformed with pC2.2. Open arrows point to a colony which produces no blue reaction

A

B



G. Amino Acid sequence comparisons

1. Initial comparisons of all published cellulase sequences

All published cellulase amino acid sequences studied to date were compared pairwise by the MATCH sequence alignment program. Each match was given a positive value of 10. Gaps were charged a value of 20 + 1 times (the number of residues in the gap). The list of the sequences used is presented in Table 6. The scores given by the MATCH program are presented in Table 7. The percentage similarity from amino acid alignments by the MATCH program are presented in Table 8. The DNA sequence of the *B. subtilis* strain IFO 3034 endoglucanase gene (Nakamura et al., 1987) was not used in this analysis since its predicted amino acid sequence was highly homologous (94% with no gaps in the comparison) to both the *B. subtilis* PAP115 and DLG endoglucanase.

2. "Jumble test"

The "jumble test" as termed by Doolittle (1986) is used to asses the significance of any similarity found between two proteins that are only marginally similar. The sequences of interest were randomized and recompared 20 times. This recomparison is done to check for any possible bias of amino acid residues which might give a higher than relevant % similarity value from the comparison of the original sequences. The value, sometimes called 'distance', of the comparisons of the randomized sequences were compared to the alignment values of the original sequences to determine the statistical relevance of the % similarity values given in Table 8. The values for "jumble tests" done on eight pairwise comparisons are given in Table 9. The *B. polymyxa* and *C. thermocellum* sequences are very significantly similar at 36 standard deviation units from the mean of the randomly compared sequences. The *S. commune* EG2 and *T. reesei* EG3 are also very similar. Both of these comparisons most likely demonstrate an evolutionary connection between the sequences. The "jumble test" results of *B. polymyxa* EG vs *S. commune* EG2 and the *B. subtilis* ß-1,3-1,4-EG vs the *C. thermocellum celD* are contrary to the results of Table 8

Table 6. List of compared cellulase amino acid sequences.

Organism	Enzyme/Gene	Abbrev.	Reference
B. polymyxa	endoglucanase	Bp-EG	(See Results, Section F2)
B. subtilis DLG	endoglucanase	Bs-DLG	Robson and Chambliss, 1987
B. subtilis PAP115	endoglucanase	Bs-PAP	MacKay et al., 1986
B. species strain 1139	endoglucanase	Bsp-1139	Fukumori et al., 1986a
B. species strain N-4	endoglucanase N1	Bsp-N1	Fukumori et al., 1986b
B. species strain N-4	endoglucanase N2	Bsp-N2	Fukumori et al., 1986b
B. subtilis	ß-1,3-1,4-endoglucanase	ß1,3-1,4	Murphy et al., 1984
C. thermocellum	endoglucanase *cel*A	Ct-celA	Béguin et al., 1985
C. thermocellum	endoglucanase *cel*B	Ct-celB	Grépinct and Béguin, 1986
C. thermocellum	endoglucanase celC	Ct-celC	Schwarz et al., 1988
C. thermocellum	endoglucanase *cel*D	Ct-celD	Joliff et al., 1986
Cellulomonas fimi	endoglucanase *cen*A	Cf-cen	Wong et al., 1986
Cellulomonas fimi	exoglucanase *cex*	Cf-cex	O'Neil et al., 1986
Cellulomonas uda CB4	endoglucanase	Cu-CB4	Nakamura et al., 1986
Erwinia chrysanthemi	endoglucanase *cel*Z	Ec-EG	Guiseppi et al., 1988
S.commune	endoglucanase EG2	Sc-EG2	M. Yaguchi, personal communication
T. reesei	endoglucanase EGI	Tr-EG1	Pentillä et al., 1986
T. reesei	endoglucanase EGIII	Tr-EG3	Salheimo et al., 1988
T. reesei	cellobiohydrolase CBHI	Tr-CBH1	Fagerstam et al., 1984
T. reesei	cellobiohydrolase CBHII	Tr-CBH2	Chen et al., 1987

note: all enzymes are ß-1,4-glycosidases unless indicated.

TABLE 7. MATCH SCORES FOR THE PAIRWISE COMPARISON OF ALL KNOWN PUBLISHED CELLULASE SEQUENCES. Scores over 700 have been printed in bolder lettering. The list of sequences and their references can be found in Table 6.

	BpEG	DLG	1139	PAP	N1	N2	1,3-1,4	celA	celB	celC	celD	cen	cex	CB4	Ec	Sc	TrE1	TrE3	TrC1	TrC2
Bp-EG	**3980**																			
Bs-DLG	571	**5080**																		
Bsp-1139	596	**1303**	**6400**																	
Bs-PAP	534	**4660**	**1270**	**4990**																
Bsp-N1	505	**2312**	**1350**	**2306**	**4880**															
Bsp-N2	531	**2226**	**1252**	**2232**	**3477**	**4090**														
β1,3-1,4	309	328	311	321	308	296	**2420**													
Ct-celA	516	549	632	585	552	515	314	**4770**												
Ct-celB	**1120**	**738**	**742**	**730**	643	602	308	**905**	**5630**											
Ct-celC	441	408	411	426	479	470	281	405	504	**3430**										
Ct-celD	519	618	**845**	606	607	556	341	**886**	**849**	439	**6490**									
Cf-cen	480	543	578	543	513	457	355	568	558	370	550	**4490**								
Cf-cex	510	608	621	616	621	544	310	615	532	382	611	580	**4840**							
Cu-CB4	481	436	453	417	420	392	284	618	453	370	471	441	494	**3590**						
Ec-EG	490	**1255**	**1017**	**1235**	**1249**	**1210**	287	556	569	430	573	516	623	468	**4280**					
Sc-EG2	488	473	450	473	500	495	297	384	451	441	455	434	426	387	467	**3180**				
Tr-EG1	499	556	549	583	527	492	320	681	590	346	549	585	588	407	501	404	**4590**			
Tr-EG3	483	515	544	519	517	494	301	542	493	406	519	567	517	395	503	**802**	568	**4180**		
Tr-CBH1	465	581	609	573	538	512	350	575	606	355	641	585	672	421	547	392	**1892**	555	**5130**	
Tr-CBH2	496	545	581	522	546	512	327	603	548	388	589	**824**	584	476	497	438	583	**825**	609	**4710**

TABLE 8. PERCENTAGE SIMILARITY FROM PAIRWISE COMPARISONS OF ALL PUBLISHED CELLULASE AMINO ACID SEQUENCES. Similarity is judged by the number of amino acids that occupy the same position when aligned using the Match program. The percentage is calculated relative to the protein with the smallest number of amino acids.

	BpEG	DLG	1139	PAP	N1	N2	13,14	celA	celB	celC	celD	cen	cex	CB4	Ec	Sc	TrE1	TrE3	TrC1	TrC2
Bp-EG	100																			
Bs-DLG	21	100																		
Bsp-1139	20	31	100																	
Bs-PAP115	19	93	30	100																
Bsp-N1	18	50	35	50	100															
Bsp-N2	18	57	38	56	88	100														
ß1,3-1,4	14	20	20	10	19	15	100													
Ct-celA	19	18	21	20	16	17	18	100												
Ct-celB	34	21	19	21	21	21	19	26	100											
Ct-celC	20	17	19	18	14	20	13	17	22	100										
Ct-celD	23	18	18	18	17	22	25	26	20	19	100									
Cf-cen	15	17	18	17	17	18	22	16	19	17	18	100								
Cf-cex	16	17	19	16	18	21	19	16	15	15	19	19	100							
Cu-CB4	18	18	18	15	17	17	11	24	18	15	19	18	19	100						
Ec-EG	15	35	32	34	35	35	18	19	18	18	17	15	22	21	100					
Sc-EG2	25	20	21	20	23	24	14	19	23	18	21	19	19	15	21	100				
Tr-EG1	17	18	18	18	16	18	17	19	15	15	15	18	18	18	17	16	100			
Tr-EG3	17	18	18	17	15	17	17	19	15	17	16	20	17	14	17	31	19	100		
Tr-CBH1	18	16	17	16	15	16	21	16	18	15	18	18	20	15	14	18	49	18	100	
Tr-CBH2	17	15	15	15	15	19	22	18	17	17	19	27	17	15	14	21	18	24	22	100

Table 9. "Jumble test" values for several pairwise comparisons of cellulase amino acid sequences.

Sequences	Distance D	Mean Dist. M +/- S	D-M / S	Probability
B.polymyxa EG vs. *C. thermocellum celB*	-391.0	-48.00+/-9.414	36.43	
B. polymyxa EG vs. *B. subtilis* PAP115 EG	-78.0	-45.00+/-14.99	2.17	1.49×10^{-2}
B. polymyxa EG vs. *S. commune* EG2	-52.0	-38.50+/-13.34	1.01	1.56×10^{-1}
B. subtilis PAP115 EG vs. *C. thermocellum celB*	-26.0	-37.85+/-8.72	-1.36	9.10×10^{-1}
B. subtilis PAP115 EG vs *S. commune* EG2	-82.0	-28.95+/-10.58	5.01	2.67×10^{-7}
B. subtilis ß-1,3-1,4 EG vs. *C. thermocellum celD*	-48.0	-42.90+/-10.31	0.49	3.10×10^{-1}
S. commune EG2 *T. reesei* EG3	-402.0	-33.45+/-7.47	49.37	
C. fimi cena vs *T. reesei* CBH2	-114.0	-44.70+/-15.80	4.39	5.73×10^{-6}

D = Distance value of comparison of initial sequences
M = Mean distance value of 20 comparisons after the initial sequences have been randomized
S = Standard deviation of the 20 comparisons
The 'Probability' column gives the calculated probability that the similarity between the sequences occurred by chance

but in agreement with the low MATCH values of Table 7. The comparison between *cen*A of *C. fimi* and *T. reesei* CBH2 has not been noted in the literature. The results from this comparison show that there is a 1 in 5.7×10^6 probability that the similarity that exists between these two proteins happened by chance. This comparison might not have any evolutionary significance, but it may show an analogous structural similarity. The same can probably be said also of the comparison of the *B. subtilis* PAP115 EG and *S. commune* EG2. Although the MATCH value and % similarity were low, the jumble test calculates that there is a 1 in 2.7×10^7 probability that the similarities occurred by chance.

3. Amino acid sequence alignments

The highest similarity of the putative *B. polymyxa* endoglucanase amino acid sequence was to the *C. thermocellum cel*B endoglucanase. An alignment was made by the MATCH program of these two sequences and is presented in Figure 16. Direct matches were boxed. A putative site connected with enzymatic function (substrate binding and/or catalytic activity) is surrounded by a double box. The choice of this site was based on further alignments seen in this section (Fig. 17) and reasons given in the discussion.

From the match and % similarity values in Tables 7 and 8, there is a group of enzymes which seem to be interrelated. They include all the *Bacillus* ß-1,4-endoglucanases, *C. thermocellum cel*B, *Erwinia chrysanthemi* endoglucanase and the two fungal sequences *S. commune* EG2 and *T. reesei* EG3. They are not all significantly similar (greater than 25% similarity) to each sequence in the group, but they do have a notable similarity to at least one other protein in the group. These eight sequences were aligned together by overlaying pairwise alignments made by the MATCH program of the sequences most alike. This multiple alignment of amino acid sequences is shown in Fig. 17. Several gaps were removed from the multiple alignments where no residues were aligned between at least four of the sequences. This was done in order to conserve space.

These results indicate that a consensus three residue sequence appears in all eight sequences. This Asn-Glu-Pro sequence is most likely the site of the proton donating

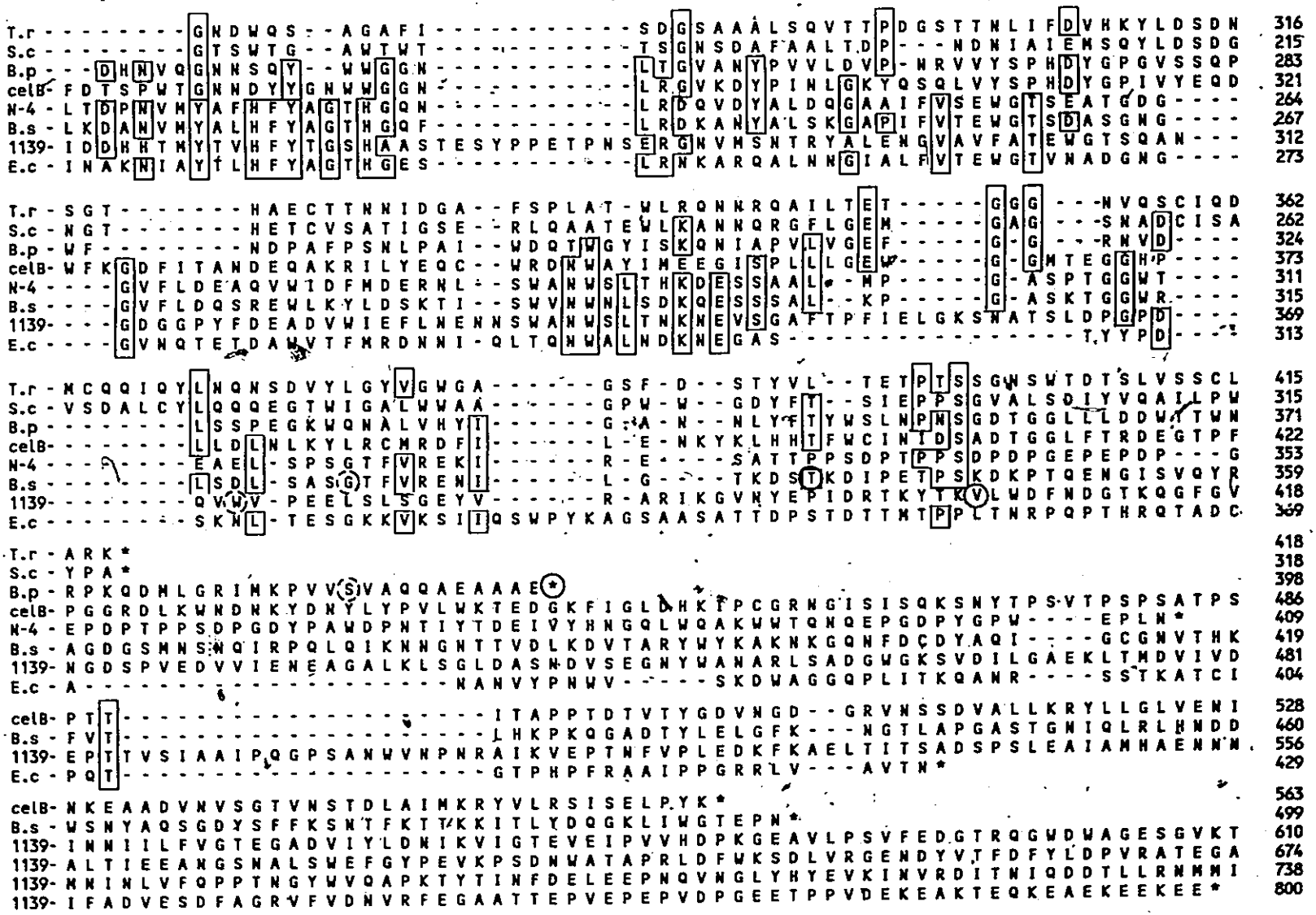
FIGURE 16. AMINO ACID SEQUENCE ALIGNMENT OF THE *Bacillus polymyxa* ENDOGLUCANASE VS. THE *Clostridium thermocellum celB* ENDOGLUCANASE.

The putative amino acid sequences from the *B. polymyxa* endoglucanase (EG) and *C. thermocellum celB* genes were aligned using the MATCH program. The optimized alignment did not match the carboxy end of the *C. thermocellum celB* sequence after residue number 434 to any of the *B. polymyxa* EG sequence. Amino acids in the same position in both sequences have been boxed in. A double box has been put around sequences believed to be part of the active site. "---" signifies gaps put in by the computer to optimize the alignment under the given constraints (see text section G.1 of Results).

```
B.p.-     MET LYS LYS LYS GLY LEU LYS LYS THR PHE PHE VAL ILE ALA SER LEU --- --- VAL MET GLY PHE THR
celb- MET LYS LYS PHE LEU VAL LEU LEU ILE ALA LEU ILE MET ILE ALA THR LEU LEU VAL VAL PRO GLY VAL GLN

B.p - LEU TYR GLY TYR THR PRO VAL SER ALA ASP ALA ALA SER VAL LYS GLY TYR TYR HIS THR GLN GLY ASN LYS
celB- THR SER ALA GLU GLY SER TYR ALA ASP LEU ALA GLU PRO ASP ASP ASP TRP LEU HIS VAL GLU GLY THR ASN

B.p - GLU SER GLY LYS GLU ALA ALA PHE ASN GLY LEU ASN TRP PHE GLY LEU GLU THR PRO ASN TYR THR LEU HIS
celB- LYS TYR GLY ASN LYS VAL TRP ILE THR GLY ALA ASN TRP PHE GLY PHE ASN CYS ARG GLU ARG MET LEU LEU

B.p - GLY LEU TRP SER ARG SER MET ASP ASP MET LEU ASP GLN VAL LYS LYS LYS GLU GLY TYR ASN LEU ILE ARG
celB- ASP SER TYR HIS SER ASP ILE ILE ALA ASP ILE GLU LEU VAL ALA ASP LYS --- GLY ILE ASN VAL VAL ARG

B.p - LEU PRO TYR SER ASN GLN LEU PHE ASP SER SER SER --- --- --- --- ARG PRO ASP SER ILE ASP TYR HIS
celB- MET PRO ILE ALA THR ASP LEU LEU TYR ALA TRP SER GLN GLY ILE TYR PRO PRO SER THR ASP THR SER TYR

B.p - LYS ASN PRO ASP LEU VAL GLY LEU ASN PRO ILE GLN ILE MET ASP LYS LEU ILE GLU LYS ALA GLY GLN ARG
celB- ASN ASN PRO ALA LEU ALA GLY LEU ASN SER TYR GLU LEU PHE ASN PHE MET LEU GLU ASN PHE LYS ARG VAL

B.p - GLY ILE GLN ILE ILE LEU ASP ARG HIS ARG PRO --- --- --- GLY SER GLY GLY GLN SER GLU LEU TRP TYR
celB- GLY ILE LYS VAL ILE LEU ASP VAL HIS SER PRO GLU THR ASP ASN GLN GLY HIS ASN TYR PRO LEU TRP TYR

B.p - THR SER GLN TYR PRO GLU SER ARG TRP ILE SER ASP TRP LYS MET LEU ALA ASP ARG TYR LYS ASN ASN PRO
celB- ASN THR THR ILE THR GLU GLU ILE PHE LYS LYS ALA TRP VAL TRP VAL ALA GLU ARG TYR LYS ASN ASP ASP

B.p - THR VAL ILE GLY ALA ASP LEU HIS ASN GLU PRO HIS --- --- --- --- --- --- --- --- --- --- GLY GLN
celB- THR ILE ILE GLY PHE ASP LEU LYS ASN GLU PRO HIS THR ASN THR GLY THR MET LYS ILE LYS ALA GLN SER

B.p - ALA SER TRP GLY THR GLY ASN ALA SER THR ASP TRP ARG LEU ALA HIS ASN VAL GLN GLY MET ARG LEU LEU
celB- ALA ILE TRP ASP ASP SER ASN HIS PRO ASN ASN TRP LYS ARG VAL ALA GLU GLU THR ALA LEU ALA ILE LEU

B.p - SER VAL ASN PRO ASN TRP LEU ILE LEU VAL GLU GLY VAL --- --- --- --- --- --- --- --- --- ---
celB- GLU VAL HIS PRO ASN VAL LEU ILE PHE VAL GLU GLY VAL GLU MET TYR PRO LYS ASP GLY ILE TRP ASP ASP

B.p - --- --- --- --- ASP HIS ASN VAL GLN GLY ASN ASN SER GLN TYR --- --- TRP TRP GLY GLY ASN LEU THR
celB- GLU THR PHE ASP THR SER PRO TRP THR GLY ASN ASN ASP TYR TYR GLY ASN TRP TRP GLY GLY ASN LEU ARG

B.p - GLY VAL ALA ASN TYR PRO VAL VAL LEU --- ASP VAL PRO ASN ARG VAL VAL TYR SER PRO HIS ASP TYR GLY
celB- GLY VAL LYS ASP TYR PRO ILE ASN LEU GLY LYS TYR GLN SER GLN LEU VAL TYR SER PRO HIS ASP TYR GLY

B.p - PRO GLY VAL SER SER GLN PRO TRP PHE --- --- --- --- --- --- --- --- ASN ASP PRO ALA PHE PRO SER
celB- PRO ILE VAL TYR GLU GLN ASP TRP PHE LYS GLY ASP PHE ILE THR ALA ASN ASP GLU GLN ALA LYS ARG ILE

B.p - ASN LEU PRO ALA ILE TRP ASP GLN THR TRP GLY TYR ILE SER LYS GLN ASN ILE ALA PRO VAL LEU VAL GLY
celB- LEU TYR GLU GLN CYS TRP ARG ASP ASN TRP ALA TYR ILE MET GLU GLU GLY ILE SER PRO LEU LEU LEU GLY

B.p - GLU PHE GLY GLY --- --- --- ARG ASN VAL ASP LEU SER SER PRO GLU GLY LYS TRP GLN ASN ALA LEU VAL
celB- GLU TRP GLY GLY MET THR GLU GLY GLY HIS PRO LEU LEU ASP LEU ASN LEU LYS TYR LEU ARG CYS MET ARG

B.p - HIS TYR ILE GLY ALA ASN --- --- ASN LEU TYR PHE THR TYR TRP SER LEU ASN PRO ASN SER GLY ASP THR
celB- ASP PHE ILE LEU GLU ASN LYS TYR LYS LEU HIS HIS THR PHE TRP CYS ILE ASN ILE ASP SER ALA ASP THR

B.p - GLY GLY LEU LEU LEU ASP ASP TRP THR THR TRP ASN ARG PRO LYS GLN ASP MET LEU GLY ARG ILE MET LYS
celB- GLY GLY LEU PHE THR ARG ASP GLU GLY THR PRO PHE PRO GLY GLY ARG ASP LEU LYS TRP ASN ASP ASN LYS

B.p - PRO VAL VAL SER VAL ALA GLN GLN ALA GLU ALA ALA ALA GLU ***          398
celB- TYR ASP ASN TYR LEU TYR PRO VAL LEU TRP LYS THR GLU ASP GLY LYS      450 of 563
```

FIGURE 17. AMINO ACID SEQUENCE ALIGNMENT OF EIGHT ENDOGLUCANASES.

Consensus residues that are in the same position of 4 to 6 of the 8 sequences have been boxed in with a single line. Amino acid residues in the same position of 7 or 8 sequences have been boxed in with a double line. Numbering of the amino acid sequence positions starts with the initiation codon except for the *S. commune* EG2 sequence which starts with the first amino acid of the mature protein. Sequence position for the last residue of each line is given on the far right. Dashes represent gaps made by the computer to optimize the alignment. A * signifies a stop codon. A dashed circle surrounding an amino acid shows the carboxy terminal residue of an inactive truncated protein produced by 3′ deletion of the gene. A solid circle around a residue deliniates the carboxy terminus of a truncated active protein. The sequences used in the alignment are: *T. reesei* EGIII (T.r), *S. commune* EG2 (S.c), *B. polymyxa* EG (B.p), *C. thermocellum celB* (celB), *B. subtilis* PAP115 EG (B.s), *Bacillus* sp.N-4 EG N2 (N-4), *Bacillus* sp. no.1139 EG (1139), *E. chrysanthemi celZ* (E.c). References for these sequences can be found in Table 6.



glutamic acid described in the lysozyme model of substrate hydrolysis (see Introduction, section D.4). Common polar residues like glycine appear throughout the sequences and are aligned in several spots, but consensus residues with at least 7 of the 8 sequences appears most often with hydrophobic residues.

After 220-250 residues from the NH_2-terminus, the number of aligned amino acids lessens significantly. No residues align in at least 7 of the enzyme sequences after this point. The next 70 to 100 amino acids in at least three of the enzyme sequences are known to be necessary to retain endoglucanase activity. Deletions of the 3′ end of the endoglucanase genes of both the *B. subtilis* PAP115 and alkalophilic *Bacillus* sp. No.1139 have been done to show how short the enzymes can be and remain active (Seligy et al., 1987; MacKay et al., 1986; Fukumori et al., 1987). The carboxy terminus of the longest deleted protein that produces an inactive enzyme, as well as the terminus of the shortest active enzyme for the *B. polymyxa*, *B. subtilis* and *B.* sp. no. 1139 endoglucanases, are shown on the sequences in Fig. 17. The 3′ endpoint of the deleted active clones are all within 40 amino acids of each other in the alignment. Within this region is the carboxy end of both of the fungal sequences. There is a large carboxy region of at least 100 amino acids after these endpoints in 4 of the sequences shown here (*B. subtilis*, *C. thermocellum celB*, *B.* sp. no. 1139, *B.* sp. N-4). *E. chrysanthemi* has a carboxy end only slightly longer than the fungal and *B. polymyxa* sequences. The additional 3′ sequence may have some functional significance. The possible function of the 3[redacted] sequences will be elaborated on in the Discussion, section F.

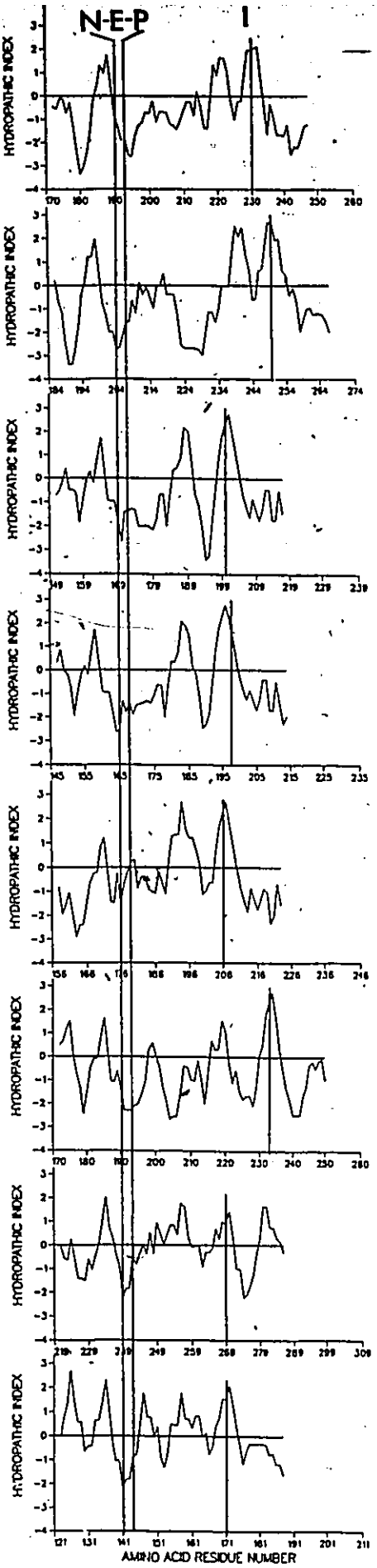
4. Hydropathy plots

In the comparisons and the alignments presented so far, no consideration was given to possible conservative amino acid changes between enzyme amino acid sequences. To attempt to do this hydropathy plots were made of each of the 8 sequences previously used in the multiple alignments in Fig. 17. No overall consensus pattern was observed in such plots. However the region surrounding the proposed active site region, the consensus Asn-

Glu-Pro sequence, had some similarity between all of the prokaryotic sequences, but not the fungal endoglucanase sequences. The hydropathy plots of this region for each sequence is shown in Figure 18. The plots are aligned together with the Asn-Glu-Pro sequences. The hydropathy plots for the whole protein of each of the eight endoglucanase sequences are included in Appendix C of this thesis.

FIGURE 18. HYDROPATHY PLOTS OF THE REGION SURROUNDING THE CONSENSUS ASN-GLU-PRO RESIDUES FOR EIGHT ENDOGLUCANASES.

These plots were generated separately by the HYDRO program of the COMPMOLOGEN sequence analysis package (see section I.2 of the Material and Methods) using the method of Kyte and Doolittle (1982). Values were averaged over 6 residues. The plots have been lined up using the Asn-Glu-Pro sequence whose position has been marked with a double vertical line. A single vertical line 30-45 residues towards the carboxy-terminus marks the consensus Ile shown in Fig. 17.



B. polymyxa EG

C. thermocellum celP

B. subtilis PAP115 EG

B. sp. strain N-4 N1

Erwinia chrysanthemi celZ

B. sp. strain no. 1139

T. reesei EG3

S. commune EG2

IV. DISCUSSION

A. Cloning of cellulase genes

Derivatives of the phage M13 have been widely used in secondary cloning techniques because of the many useful DNA manipulation techniques available to it. Some of these techniques include: site specific mutagenesis, single stranded probe preparation, C-test determination of fragment orientation without DNA purification, isolation of heterologous duplex DNA by S1 nuclease protection and the dideoxy chain termination method of sequencing. To my knowledge, this is the first report of the use of an M13 vector for cloning genes from a genomic bank by direct expression of its enzyme product. Genomic clones of *B. circulans* and *B. polymyxa* were successfully isolated using M13 as the primary vector. The two substrates employed for direct screening of endoglucanase expression on plates were easy to use and circumvented the need for replica plating. The use of OBR-cellulose for screening of EG clones is a new application. However, other researchers in the past have used the CR assay to clone and express EG genes by plasmid vectors in *E. coli* (Gijsegem et al., 1985; Koide et al., 1986; Park and Pack, 1986; Wolff et al., 1986; Romaniec et al., 1987; Kim et al., 1987; Seligy et al., 1987; Sharma et al., 1987; Barros and Thompson, 1987; Ohmiya et al., 1988; Faure et al., 1988) and once by the bacteriophage lambda (Kotoujansky et al., 1985). Lambda vector L47-1 was used to isolate cellulase genes from *Erwinia chrysanthemi* in *E. coli* because it would lyse the cells in the lytic cycle releasing enzymes that might not have been secreted by *E. coli* (Kotoujansky et al., 1985). Lysis of the *E. coli* cells was not necessary for detecting expression of this gene (Gijsegem et al., 1985) or any other cellulase genes expressing in *E. coli* as demonstrated by the use of plasmid vectors already mentioned.

The DNA fragments that were successfully cloned during the course of this work ranged from 1 to 3 kbp. These fragments appear to be small enough to be stably inserted into the DNA of the M13 phage. In contrast the 6.2 kbp Hind III fragment M13 clones

(Results section E) could not be purified in my hands to any homogeneity. The clones which were isolated with the presumed HindIII fragment were expressing CMCase activity. This is interesting because the final constructs of the 5' and 3' end of the endoglucanase gene did not express any activity even though there was 1.4 kbp of DNA upstream of the defined EG coding region. The original CR^+ clone could have been the result of a recombination event, but if so the recombinant must have been unstable. M13 is believed to be limited in including large heterologous fragments of DNA stably within its own DNA. This is possibly due to its requisite single stranded DNA stage of propagation which produces an environment open to deletions, insertion and recombination of DNA due to intra-strand interaction. Strains of *E. coli* have been developed trying to solve this problem (Yanisch-Perron et al., 1984).

B. DNA characterization

From the DNA-DNA hybridization results of Fig. 3C, it appears that both *B. circulans* and *B. subtilis* have very similar genomic organizations at least in the vicinity of their respective endoglucanase genes. The EG related DNA fragments generated by digesting the genomic DNAs with PstI and the EcoRI appear to be the same. The only difference is the extra HindIII site in the PAP115 genome at the 3' end of the endoglucanase gene. From the work of MacKay et al.(1986) this HindIII site is located 3 nt downstream of the *B. subtilis* endoglucanase gene TGA stop codon and 8 nt upstream of the inverted repeat which may serve as the gene's transcription termination site.

Data in Fig. 3C also show that there are sequences in the *B. circulans* 3.4 kbp HindIII probe that are partially homologous to sequences in other regions of the genome in both organisms. This could possibly be representative of other gene(s) similar to, but not duplicates of the genes already cloned from these two genomes. Firstly, only one type of endoglucanase gene for both the *B. subtilis* and *B. circulans* was found among the several clones. If more endocellulase genes did exist there is a good chance that they would have been also isolated. Secondly, it is known that the *B. subtilis* 2.9 kbp and 0.5 kbp HindIII

fragment counterparts of the *B. circulans* 3.4 kbp HindIII fragment used for probing the genomic DNA, contain only one complete functional, endoglucanase gene that is 1.6 kbp in length. In Fig. 3C the HindIII digests of *B. subtilis* and *B. circulans*, a single extra fragment of 0.7 kbp in length hybridizes to the probe which would truncate a duplicate endoglucanase gene. Since a truncated non-functional gene would unlikely be preserved in two different organisms, the fragments are more likely homologous to a region of the 1.8 kbp of extraneous DNA on the HindIII fragment. There was 1.5 kbp of extraneous DNA to the *B. subtilis* PAP115 endoglucanase gene previously characterized by MacKay et al.(1986). It had an ORF of an unknown size but at least 420 nucleotides in length. This ORF could have a homologous gene elsewhere in the genome.

Only one fragment hybridizes to the *B. polymyxa* endoglucanase gene probe in each lane with digested *B. polymyxa* DNA (Fig. 3C). Therefore the *B. polymyxa* endoglucanase gene is a single copy gene within the *B. polymyxa* genome and as far as DNA-DNA hybridization can show, it is very different to the genes of *B. subtilis* and *B. circulans*. The finding of only one copy of this gene in the *B. polymyxa* genome does not prove that it is the only endoglucanase gene in *B. polymyxa*. Since it is possible, as will be discussed later, that the RNA polymerase recognition sequences of *B. polymyxa* are not recognized by *E. coli* a second endoglucanase would not have been expressed unless a chance fusion with the *lacZ* gene occurred.

There has been only proof of a one base pair mutation difference between the *B. circulans* and *B. subtilis* endoglucanase genes so far. The similarity has been checked by REN site mapping, DNA-DNA hybridization, S1 nucleases protection assay, and SDS-PAGE of the protein products. Although the sensitivity of these assays are not equal to DNA sequencing, the *B. circulans* gene is most likely as close in homology to *B. subtilis* PAP115 as any of the other *B. subtilis* endoglucanases genes isolated so far. The *Bacillus* species isolated by Park and Pack (1986) has an endoglucanase gene isolated from it that has been described as being of "*B. subtilis* origin" (Lee and Pack, 1987), but the isolated

organism is not called *B. subtilis* most likely because the standard biochemical and morphological examinations have not been done. The tests to classify the *Bacillus* isolate of Robson and Chambliss (1986) are from a *Bacillus* handbook (Gorden et al., 1973) produced by the United States Department of Agriculture (USDA) and from Wolf and Barker (1968). None of these tests are as rigorous as those used to define species of higher order organisms. "There is no agreement among bacterial systematists as to a suitable definition of a 'bacterial species'" (Mandelstam and McQuillen, 1973). A genotypic based system of classification has been argued to have the problem of including organisms which may be able to exist in very different environments because of small differences in the genomes or extrachromosomal DNA (e.g., heat sensitive mutations in genes, ß-lactamase gene). This problem still exists in a non-genotypic system of classification and can be seen in medical environments where new antibiotic resistant mutants of infectious bacteria have not been reclassified even if they can exist in a new ecological niche. In Bergey's Manual of Determinative Bacteriology (Buchanan and Gibbons, 1974) if a microbe tests positive for 90% of the morphological and biochemical tests used for classification it is considered to be well classified. I believe the burgeoning amount of molecular genetic information resulting from studies like this thesis should be used for the reclassification of microorganisms.

A xylanase gene from the same strain of *B. circulans* used in thesis has recently been isolated and sequenced (Yang, 1988). I have compared the predicted amino acid sequence of this xylanase to one isolated from *B. subtilis* (Paice et al., 1986) and found very little difference between them (<0.5%, comparison not shown). Having two genes in this strain of *B. circulans* that are as homologous as the intraspecies variants of the genes of *B. subtilis* suggests that this strain of *B. circulans* should be reclassified as *B. subtilis*.

C. Expression and secretion

From the results of the protein-electrophoresis in Fig. 6A and 6B as well as the expression data of Table. 4, it would appear that the *B. circulans* clone, pCH.7 produces

more endoglucanase than the *B. subtilis* clone, pC6.3. This did not seem to be the case with the expression of these in genes in their parent *Bacillus* hosts. The results of Fig. 6 also suggests that most of the endoglucanase produced by both these clones is cell associated. This is contrary to the findings of Lo et al. (1988) but is in agreement with all other work of *Bacillus* species endoglucanases expressed in *E. coli* (Sashihara et al., 1984; Robson and Chambliss, 1986; Koide et al., 1986; Fukumori et al., 1986; Park and Pack, 1986). Lo and coworkers have shown quite clearly with endoglucanase specific antibodies to Western blots of the cytoplasmic, periplasmic and extracellular proteins that the bulk of the endoglucanase protein is extracellular. Their assay of endoglucanase enzymatic activity using cytoplasmic and periplasmic enzyme markers supports this result. The major difference between the work of Lo and coworkers and all the other studies including this thesis is the use of *E. coli* JM103 as the host. Most of the other studies including this thesis used *E. coli* HB101.

It is clear from Fig 6C that the endoglucanases from the *B. circulans* and *B. subtilis* clones are the major extracellular proteins. This does not seem to be the case for the *B. polymyxa* endoglucanase clone. While producing large amounts of cell associated endoglucanase, very little is secreted outside the cell. As DNA and predicted amino acid sequence analyses showed, the *B. polymyxa* gene responsible for expression in *E. coli* did not have a leader secretion sequence which is most likely needed for optimal extracellular expression.

D. Sequence analysis of the cloned *B. polymyxa* DNA

1. Endoglucanase gene

The endoglucanase gene codes for a putative translation product of 397 amino acids. The predicted molecular weight of this protein would be approximately 44 kDa. The estimated size of the protein in the SDS-PAGE gel of Fig. 6A is 37 kDa which is low considering that this protein lacks 22 amino acids due to the fusion with the *lac*Z gene. The fused protein should be 41 kDA. The discrepancy in theoretical and determined

molecular weight by SDS-PAGE is probably due to error in the latter procedure. No precursor form of the protein that might be transformed into the shorter form by proteolytic cleavage is visible in Fig. 6A or 6C. The ORF is most likely not shorter than that predicted because the 3′ end of the gene has been well defined by the deletion clones presented in Fig. 14.

No definite promoter or transcription termination sequences were found upstream of the coding region of the *B. polymyxa* endoglucanase gene. Cloned genes that are transcribed in *E. coli* must have sequences that are recognized by the DNA dependent RNA polymerase holoenzyme $E\sigma^{70}$. The RNA polymerase responsible for the transcription of genes during the logarithmic stage of growth in *B. subtilis*, $E\sigma^{43}$, recognizes similar sequences. Transcript mapping of the *B. subtilis* DLG endoglucanase gene indicated a transcript start 48 nt upstream of the translational start codon with sequences TAGACA found in the -35 position and TACAAT in the -10 position (Robson and Chambliss, 1987). The consensus sequences for *E. coli* $E\sigma^{70}$ are TTGACA (-35) and TATAAT (-10). For genes that do not have promoters like *E. coli*, they must have sequences that can be used for that purpose. This was seen by Béguin et al. (1986) when transcript mapping of the *C. thermocellum celA* gene in *C. thermocellum* and *E. coli* demonstrated that two different sites of transcript initiation were used by the different organisms. The major transcript in *C. thermocellum* started 56 bases before the start codon and had -35 and -10 sequences very similar to that required for *B. subtilis* σ^{28} which is a minor vegetative initiation factor. In *E. coli* the start of transcription was 135 bases upstream of the start codon and had -10 and -35 sequences like those recognized by $E\sigma^{70}$. The *cen*B gene of *C. fimi* was not transcribed in *E. coli* by its own promoter. It was found to have -10 and -35 sequences, GACGAT and TTGACC respectively, which are very similar to the promoter of the *Streptomyces erythraeus ermp2* gene and are known not to be used efficiently by *E. coli* RNA polymerase (Greenberg et al., 1987b). The *cen*A and *cex*A genes of *C. fimi* are expressed in *E. coli* (Gilkes et al., 1984) but it is not known whether the

promoters were supplied by the vector or the cloned sequences. Transcript mapping of mRNA from *C. fimi* found promoters that were again similar to that found in *Streptomyces* species (Greenberg et al., 1987a) and most likely not recognized by *E. coli* RNA polymerases.

The best *E. coli* -10 and -35 promoter-like sequences found 5′ to the *B. polymyxa* endoglucanase gene had a space of 23 nucleotides between them. The spacing most often found is 16 to 18 (Travers, 1987). The wide spacing could be the reason for the lack of expression when the 5′ and 3′ sections of the gene were fused together. In Fig. 10, the two related gene sequences had -10 and -35 *E. coli* promoter regions which conformed slightly better than the *B. polymyxa* sequences. The spacing between the two regions was 19 for the ß-amylase sequence and 15 for the *cel*B . The sequences outside of the -10 and -35 regions also have a strong effect on transcription (Travers, 1987). These sequences are not well understood and may have more of an effect on the lack of transcription of the *B. polymyxa* EG gene than the spacing between the -10 and -35 regions. The only bacillus promoter sequences found, were similar to those recognized by σ^{32}. The RNA polymerase holoenzyme $E\sigma^{32}$ is known to transcribe genes during the onset of spore formation in *B. subtilis* (Johnson et al., 1983). It is not known whether this holoenzyme is specific for sporulation genes. However, the endoglucanase is apparently produced in the greatest amounts when there is a readily usable carbon source not during times of limited nutrients (Fogarty and Griffin, 1973a). There have been no promoter sequences of *B. polymyxa* yet defined, so there is a possibility that they could be different from those of *B. subtilis*.

Some inverted repeats were found in the 3′ region of the *B. polymyxa* gene. However, they do not fit the consensus form of a *rho* independent transcription terminus (Platt and Bear, 1983) because they are too short and they are not surrounded by a poly T sequence. The typical hairpin loop transcription terminus has been found at the 3′ end of several bacterial cellulase genes (MacKay et al., 1986; Grépinct and Béguin, 1986; Greenberg, 1987a) suggesting that most of these sequences are monocistronic. The

termination of transcription of the *B. polymyxa* gene could be dependent on an additional factor like the *rho* protein. These complex forms of transcription terminations are not well known (Platt and Bear, 1983) least of all in *B. polymyxa*.

2. ORF-2

The putative protein sequence of ORF-2 has been found to have a 28% amino acid similarity to the first 186 amino acids of the *E.coli lac* repressor protein. The N-terminal region of the *lac* repressor protein is known to be responsible for binding to DNA. The lactose binding site and region needed for aggregation into the repressor's tetrameric form occurs in the latter half (Betz, 1987). This suggests that the ORF-2 protein is probably a DNA binding protein.

The arrangement of a repressor gene with a divergent direction of transcription relative to the gene it controls exists in the ß-lactamase locus in *B. licheniformis* (Salerno and Lampen, 1988). This arrangement could exist between the ORF-2 protein and the *B. polymyxa* endoglucanase gene, but none of the *Bacillus* endoglucanases studied have been shown to be inducible. From the work of Fogarty and Griffin (1973a) the *B. polymyxa* endoglucanase activity is expressed when the organism is grown in a wide variety of carbon sources including glucose and sucrose. Although some putative transcription and translation regulatory sequences have been identified in Fig.11, ORF-2 has not been examined in any detail to predict if any transcripts or protein products are produced from it in either *E. coli* or *B. polymyxa*. Since the possibility does exists that ORF-2 could be producing a repressor protein which might affect the production of endoglucanase, *E. coli* clones bearing the 5′-3′ reconstituted endoglucanase gene should be examined in closer detail for the production of a 20 Kda protein, the approximate weight of the putative protein produced by ORF-2. The cloned 5′-3′ reconstructed fragment should also be examined for endoglucanase expression in *E. coli* after deleting just the ORF-2 region.

E. Amino acid sequence comparisons

The results of the amino acid comparisons are presented in two forms. One as the

MATCH program value which has been maximized to achieve the optimum alignment of two sequences under the given constraints and the second is the percentage of direct matches of an amino acid residue in both sequences relative to the total number of amino acids in the smaller of the two sequences. This second value is universally familiar for the discussion of sequence comparisons. R.F. Doolittle has set a rule of thumb for deciding if the % sequence similarity is significant (Doolittle, 1986). Two sequences longer than 100 residues and greater than 25% in identity are very likely related. Random sequences can be 10 to 20% identical, but sequences with a similarity of 15 to 25% are in, what Doolittle termed, the "twilight zone". This region of similarity needs further statistical testing to establish a degree of confidence like the "jumble test" which has been described in the Material and Methods, section I.2.

The problem in using % similarity as a unit of comparison for discussion is that it does not represent the constraints that the sequences were under for the comparison. A small sequence compared to a large sequence will give a high % similarity if the penalty for inserting a gap is low. Gaps could be frequently placed between each residue, giving each residue of the small sequence the possibility of several residues to align with. The MATCH value takes into account the number of gaps and their lengths. In the case described above the small sequence might have a significant % similarity, but it could still have a very low MATCH value. An example of this is in Tables 7 and 8. The ß-1,3-1,4-endoglucanase from *B. subtilis* was also compared to the cellulases to see if any similarities exist. At 242 amino acids in length, it is much smaller than most of the cellulase sequences. The comparison to *C. thermocellum celD* (649 amino acids) gave a % similarity of 25% but the MATCH value for this comparison was very low at 341. Significant amounts of similarity from examining the scores and alignments have MATCH values of at least 700. The "jumble test" of these two sequences gave a calculated probability of 1 in 30 that this similarity happened by chance which seems quite probable.

The comparisons of all the published cellulases exhibited a distinctive group of

similar proteins which were used for the multiple alignment in Fig. 17. This group includes: all the *Bacillus* species, *C. thermocellum celB*, *Erwinia chrysanthemi celZ*, and to a lesser extent the two fungal sequences *T. reesei* EGIII, and *S. commune* EG2. The bacterial sequences have a relatively high percentage similarity within the group. *B. polymyxa* EG had a low % similarity and MATCH value to the other *Bacillus* sequences but a high amount of similarity of 34% to *C. thermocellum celB*. The *C. thermocellum celB* when compared to the *B. subtilis* sequences and the alkalophilic *Bacillus* sp. no. 1139 scored significant MATCH values. Some similarity was already been reported between *C. thermocellum celB* and *B. subtilis* PAP115 (MacKay et al.,1986). Similarity of the *B. subtilis* PAP115 EG to the *B.* sp. 1139 EG and the *Erwinia chrysanthemi* EGZ was reported by Guiseppi et al. (1988). The *Erwinia chrysanthemi* EGZ sequence also shows strong regions of homology to the *Bacillus* sp. N-4 endoglucanases N1 and N2 but does not with either *B. polymyxa* EG or *C. thermocellum celB*.

The jumble test of the comparison of *B. polymyxa* EG and the *C. thermocellum celB* sequence shows that the comparison is significant and that these two sequences are most likely related. Although both organisms are 'gram positive' and members of the *Bacillaceae* family, *C. thermocellum* is a thermophile and an anaerobe while *B. polymyxa* is an aerobe that cannot grow at temperatures above 30°C. Furthermore, the *B. polymyxa* endoglucanase has no cysteines but the *C. thermocellum celB* sequence has six. Perhaps these residues provide thermal stability by crosslinking domains of the *C. thermocellum* endoglucanase. There has not been much similarity found between the other *C. thermocellum* endoglucanase sequences and any of the *Bacillus* endoglucanases. The DNA sequence 5′ to the coding region of the *Clostridium* genes have been found to be very similar to that of *Bacillus* sequences (Béguin et al., 1986; Schwarz et al., 1988). As mentioned already, transcript mapping of the *C. thermocellum celA* gene suggests that the major transcript is produced from promoter sequences reminiscent of the sequences recognized by the *Bacillus subtilis* minor vegetative holoenzyme $E\sigma^{28}$ (Béguin et al., 1986).

1. Evolution

Although I have linked the eight different endoglucanase sequences together, some of these sequences share very few common regions of identity. In terms of relatedness through evolution, the results of the "jumble tests" and the MATCH values would suggest there are three separate groups within these sequences. The first include the two fungal sequences whose similarity has already been published (Saloheimo et al., 1988). It should be noted that the major shared sequences of *T. reesei* EG3 with *T. reesei* EG1 and CBHII are located in the amino terminal of *T. reesei* EG3 (Saloheimo et al., 1988) and not in the major portion of the enzyme. The next group of enzymes include the *C. thermocellum celB* and *B. polymyxa* endoglucanases. This is a fairly unusually connection since both organisms inhabit quite different environments. The third group would include the remaining *Bacillus* species endoglucanases and *Erwinia chrysanthemi celZ*. *Erwinia chrysanthemi* is a known plant pathogen that is truly cellulolytic. All eight of the organisms live in the soil, but only *C. thermocellum* would need an anaerobic environment.

Several of the comparisons by the MATCH programs gave high values but upon examination with the "jumble test" these values were statistically insignificant. The values assigned to different matches of amino acids are provided by the Dayhoff matrix (Dayhoff et al., 1978). This matrix was created from the analysis of several groups of well related proteins. Certain mutations from one amino acid to another, happen quite frequently and are given a low value if matched. The change between Asp and Glu occurs most frequently and is given a value of 3. If Asp did not change it would only be given a value of 4. In contrast, rare amino acids like Trp and Cys are given values of 17 and 12 if they stay the same. Therefore depending on the amino acid makeup of the proteins being compared, the "jumble test" value may be low even if there is a fair number of direct matches. This is most apparent with *C. thermocellum celB* being compared to *B. subtilis* EG. The frequent number of matches gave a significant MATCH value and allowed the two sequences to be aligned in Fig. 17, but the "jumble test" said this similarity was insignificant from the

viewpoint of evolution.

2. Active site

The multiple sequence alignment that exists in figure 17 shows a very strong consensus sequence of Asn-Glu-Pro within the 160-220 amino acid position in all the sequences. The similarity between some of the sequences used in this alignment is only 15%, which would suggest that an analogous structure exists between all the eight sequences more than any evident evolutionary relatedness.

This sequence appearing in all of these sequences is most likely not coincidental, because residues surrounding the Asn-Glu-Pro sequence are also very similar in all of the sequences. This region although very similar is still relatively small relative to the length of the endoglucanases and therefore is not large enough to give a high statistical probability of being related. The three residues to be conserved or used in analogous structures must be part of the active site. Either they are responsible for the binding of the substrate in a specific position or one of the residues contributes to catalysis.

A similar consensus sequence of Met-Asn-Glu-Pro has been reported near the proposed active site in other glycosidases. They include human lysosomal α-glucosidase, and both human and rabbit intestinal isomaltase and sucrase (Hoefsloot, 1988). The glutamic acid was not labelled a catalytic residue in these enzymes. The active site was assigned to a conserved aspartic acid 3 residues upstream of the Glu from studies of the sucrase and isomaltase enzymes (Quaroni and Semensa, 1976; Hunzicker et al., 1986). The catalytic Asp of the isomaltase and sucrase was found using affinity labelling (Braun et al., 1977). A reactive substrate analogue was allowed to bind into the active site and react with the active aspartic acid, covalently binding to it. This same procedure has been done several times with the ß-glucosidase of *Aspergillus wentii* (Legler et al., 1979; Roeser and Legler, 1981) also labelling the active aspartic acid. There was no similar sequence near the Asp of the ß-glucosidase (Legler et al., 1979). No consensus Asp residue exists in the same position in the endoglucanase sequences as in the α-glucosidase and

sucrase-isomaltase sequences, but another free carboxy group of an Asp or Glu is within 4 residues of the consensus Glu of all the eight endoglucanase. It would seem unlikely that both catalytic groups, the proton donor and stabilizing carboxy groups would be only 3 to 4 residues apart. One carboxy group must have no hydrogen atoms bound to its oxygen and the other must have one attached. A different micro-environment would be needed for each amino acid to produce this different ionization state which in turn suggests a greater separation than three or four amino acids. The Asn-Glu-Pro sequence when compared to the consensus sequence of α-glucosidase, sucrase-isomaltase complex seems more likely to be part of a similar binding site. All of these enzymes must bind to the same pyranose but they must break different bonds. The α-glucosidase and isomaltase must break α-1,4 and α-1,6 linkages while the sucrase cleaves a α-1,2 bond between a glucopyranoside and a fructofuranoside. The ß-amylase (Yang et al., 1983) and sucrase (Fouet et al., 1986) amino acid sequences of *B. subtilis* were also compared to the cellulase sequences. No notable similarity was found, and neither of the two sequences had the consensus sequence which surrounds the catalytic Asp residue of the mammalian α-glucosidase, isomaltase, and sucrase enzymes. It is possible that the Asn-Glu-Pro sequence in the mammalian glycosidases being near the catalytic site is just a coincidence.

In the lysozyme model of hydrolysis the aspartic acid is the residue responsible for creating a negative charged environment for the proton donation event during hydrolysis. The proton is donated by a Glu residue. The consensus Glu of the eight glucanases which is also similar to the consensus Glu of the mammalian glycosidases sequences mentioned, seem very likely to be the analogous proton donor of the active site to Glu-35 of HEWL or Glu-11 of T4 lysozyme. Steric positioning of the proton donor for the catalytic reaction is most likely more crucial than the groups that provide the negative charged environment for the reaction. It is therefore more likely to be properly preserved within a sequence. To test this hypothesis, experiments have been started which will mutate the Glu of the three residue consensus sequence by site directed mutagenesis to a Gln. If the mutation results

in a lower activity, the function of the residue is either connected with substrate binding or has a minor effect on the folding of the protein. If no activity is shown by the mutated enzyme, then the residue could be a catalytic residue or be very important for the structure of the active site. Unfortunately, site directed mutagenesis alone will not be definitive in determining this.

3. Hydropathy plots

The possibility that like amino acids will empirically provide the same function in similar proteins is hard to fully accept with any confidence. One place that like amino acids provide a similar function is in the signal sequence, where different hydrophobic residues seem to provide the same result of secreted proteins. I would suggest that a hydrophilic substrate like cellulose selects for specific hydrophilic sites on the protein to bind to and that these sites might be conserved over time, but according to the multiple alignment of Fig. 17, there was less variance with the hydrophobic residues. Hydropathy plots were done to look for common hydrophilic sites. No overall pattern is evident when looking at plots of the total sequences (see Appendix C for plots). However, when examining only the region surrounding the consensus Asn-Glu-Pro a similar profile exists between all the bacterial sequences. This is further evidence of sequence similarity in this region between the majority of the proteins. Active sites or binding sites might exist within this region and therefore dictate the conservation of several residues around the sites for a similar folding pattern and positioning of key residues.

F. *Bacillus* species and cellulases

One question that has dogged me since I started working with bacilli is 'why do these *Bacillus* species have only one cellulase gene, when they cannot grow on native/crystalline cellulose?' As more groups reported bacillus cellulases the more this question became more obvious. The only exception in the literature was with the *Bacillus* sp. strain N-4 which had two endoglucanases, but they were homologous enough to be definitely considered as a gene duplication event. To add to the conundrum, isolates in

Japan (Koide et al., 1986), Korea (Lee and Pack, 1987), Qubec (Seligy et al., 1987), and Wisconsin (Robson and Chambliss, 1987) had extremely homologous gene sequences for a secondary catabolic enzyme. If these organism cannot use cellulose then more accessible carbohydrates would be the carbon source. If these organism cannot grow on cellulose, why would they need a cellulase at all. There are now three pieces of information that make it easier for me to partially resolve or answer this question. The first has to do with the non-hydrolytic cellulose defibrillating enzyme domains found in the *Humicola* sp. endoglucanase and other cellulases (see section D.2 of the Introduction). There are domains of unknown purpose that are also preserved on the carboxy end of all the *B. subtilis* endoglucanases. Large unknown domains also occur on the carboxy end of the *B.* sp. strain 1139 endoglucanase and *B.* sp. strain N-4 endoglucanase N1 as well as one within the amino terminal sequences of *T. reesei* EG3. No extra domain seems to exists on the *B. polymyxa* endoglucanase. As this defibrillating action has not been examined for in these proteins, this putative function remains a possibility. During the purification of a CMCase from another strain of *B. subtilis* some hydrolytic activity towards crystalline cellulose existed in the fractions which had CMCase activity (Au and Chan, 1987). This small bit of activity may be a result of this defibrillating activity. The second piece of information is that H_2O_2 is produced in plant cells due to peroxidases in the cell wall and these peroxidases increase activity in some plant cells that are decaying (Miller, 1986). The hydrogen peroxide produced with naturally occurring Fe^{2+} might provide an exoglucanase function for *Bacillus* species similar to the use of these components postulated for brown-rot fungi. During the ripening of fruit there is an increase in transcription of plant cellulase genes (Christoffersen et al., 1984). These enzymes could also aid bacillus in breaking down cellulose. The third piece of information was that some organisms only would break down highly hydrated cellulose (i.e. non-crystalline) and this type of cellulose exists in leaf material of plants (Wedkind et al., 1988). This would mean that there is an abundant source of cellulose for these soil bacteria to use without the need of a complete cellulase

system. The *Bacillus* sp. mentioned in this thesis can grow on cellobiose. They may be able to grow on larger oligosaccharides, but the larger oligomers are either hard to make or expensive to buy and therefore have not been tested as a carbon source. The endoglucanases may not have to produce cellobiose as the end product and therefore exoglucanase activity would not be needed. This group of enzymes might form one type or subclass of endoglucanases which would fill a specific need for cellulose degradation outside of the hydrolysis of crystalline cellulose.

On the basis of the discussion and data provided so far, I would like to propose that all the ß-1,4-endoglucanases aligned in Fig. 17 are in one functional subclass of endocellulases. As mentioned in the introduction, Dr. T. Wood has postulated that more than one endoglucanase is needed for the different steric position of cellulose. It is also possible that other types of endoglucanase activity are needed for non-crystalline cellulose. Whatever the function, the similarity in all the sequences around the Asn-Glu-Pro residues to the consensus Isoleucine suggests a commonality in these sequences even though it is recognizable that some of the sequences are not related through evolution. At the International Congress of Genetics (Toronto, Canada, Aug. 22-30, 1988) a ninth sequence was presented that also shared the same consensus sequences. This sequence was an endoglucanase from *Cellulomonas flavigna* (Al-Tawheed and McConnel, 1988) that shared over 30% amino acid similarity with both *C. thermocellum cel*B and the *B. polymyxa* endoglucanase. This sequence links *Cellulomonas flavigna, C. thermocellum* and *B. polymyxa* as related organisms and reinforces that a preservation of a certain type of endoglucanase sequence exists.

V. Appendix A

DNA Clones and Constructs used in this Thesis

(T+ = Direction of transcription of cloned gene is the same as the *lac*Z gene of the vector, T- = opposite of T+, # is a variable for numbers)

Clone	Description
Bc-H4	M13mp19 + 3.4 kbp HindIII fragment with *B. circulans* EG gene, T+
Bc-HI	same as Bc-H4 with opposite orientation insert, T-
Bc-P1	M13mp9 + 3.1 kbp PstI fragment with *B. circulans* EG gene, T+, original clone
Bc-P2	same as Bp-P1
Bc-P3	same as Bp-P1
Bp-op19	M13mp19 + 1.6 kbp PstI fragment with *B. polymyxa* EG gene, T-
Bp-P1	M13mp9 + 1.6 kbp PstI fragment with *B. polymyxa* EG gene fused with 5′ of *lac*Z gene, T+, original clone
Bp-P2	as above with 0.2 kbp PstI fragment included on the 3′ end of the insert
Bs-P3.2	M13mp19 + 3.1 kbp PstI fragment with *B. subtilis* PAP115 EG gene, T+
Bs-P3.3	same as Bs-P3.2 with insert in the opposite orientation, T-
D#	deletion clones of Bp-P1, T+, see Fig. 9
E1.2	M13mp19 + 2.35 kbp EcoRI fragment 5′ of *B. polymyxa* EG gene, T-
E12-#	deletion clones of E1.2, T-,see Fig. 9
E2.1	same as E1.2
E4.1	M13mp19 + 2.35 kbp EcoRI fragment 5′ of *B. polymyxa* EG gene, T+
E41-#	deletion clones of E4.1, T+, see Fig. 9
op#	deletion clones of Bp-op19, T-, see Fig. 9
pC2.2	pUC9 + 1.6 kbp PstI fragment with *B. polymyxa* EG gene, fused with 5′ of *lac*Z gene, T+
pC6.3	pUC8 + 3.1 kbp PstI fragment with *B. subtilis* PAP115 EG gene, T+, see Seligy et al., 1987
pC6.5	opposite orientation clone of pC6.3, T-
pCH.7	pUC8 + 3.4 kbp HindIII with *B. circulans* EG gene
pCH.9	same as pCH.7
SB0	M13mp9 + 2.8 kbp HindIII fragment with *B. subtilis* PAP115 EG gene, T+. see Seligy et al., 1987
SB1	opposite orientation clone of SB0, T-

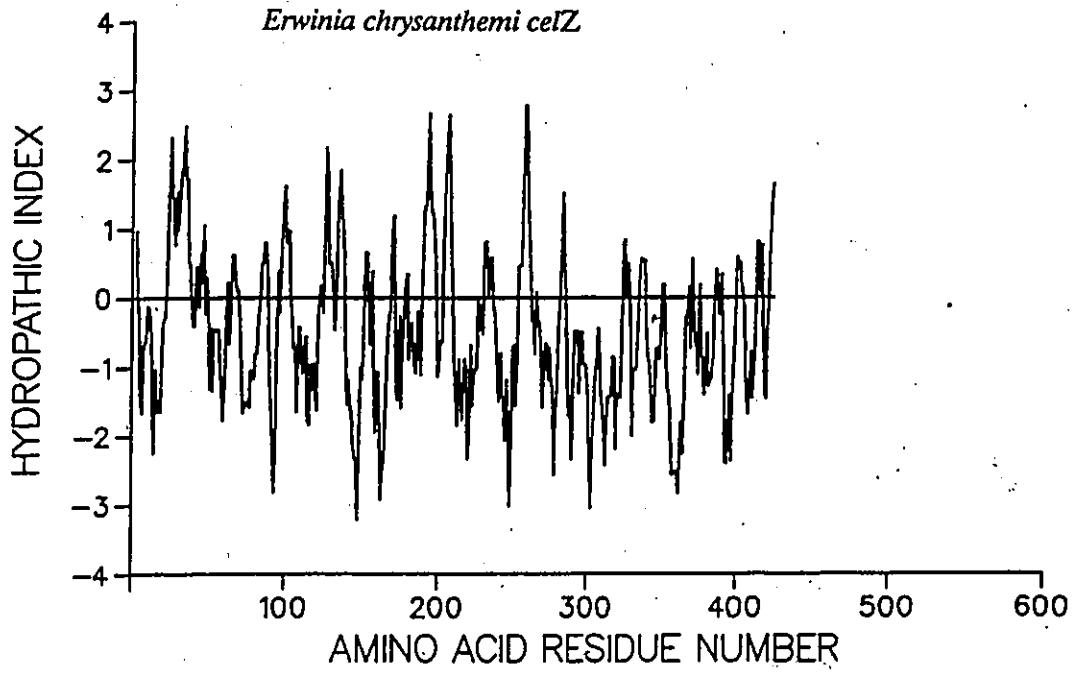
Appendix B

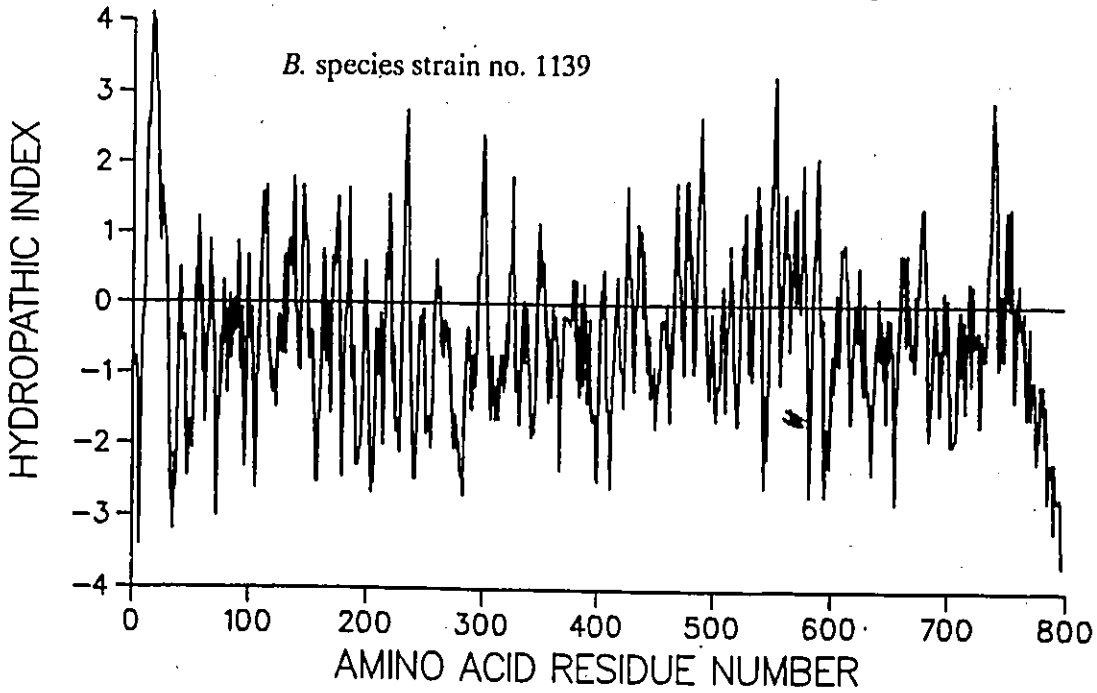
Varients of the -10 and -35 recognition sequences of *E. coli* σ^{70} (Sienbenlist, 1980)

Sequences in each column are in alphabetical order and are not matched together.

-10	-35
CATAAT	ATGAGG
CATGAT	ATGCAA
GAAAAT	ATGTCA
GACAAT	CACACT
GATAAT	CTGACG
GATACT	CTTCCG
TAAAAT	GTGCAA
TAAATT	GTGTTT
TAAGAT	TACCCA
TAAGCT	TACTCA
TAATAT	TAGACA
TACAAT	TAGACT
TACACT	TAGATA
TACAGT	TCGAAA
TACGAT	TTCAAA
TACTAT	TTGAAT
TACTCT	TTGACA
TACTGT	TTGACG
TAGACT	TTGACT
TAGATT	TTGAGT
TAGGTT	TTGCAA
TAGTCT	TTGCAC
TATAAT	TTGCGG
TATGAT	TTGCTT
TATGCT	TTGTAA
TATGGT	TTGTAC
TATTAT	TTGTAT
TATTGT	TTGTCA
TGTCAT	TTGTCT
TTAACT	TTGTGC
TTTAAT	TTGTTA
TTTCAT	TTTACA
	TTTACG
	TTTCCT
	TTTGCA

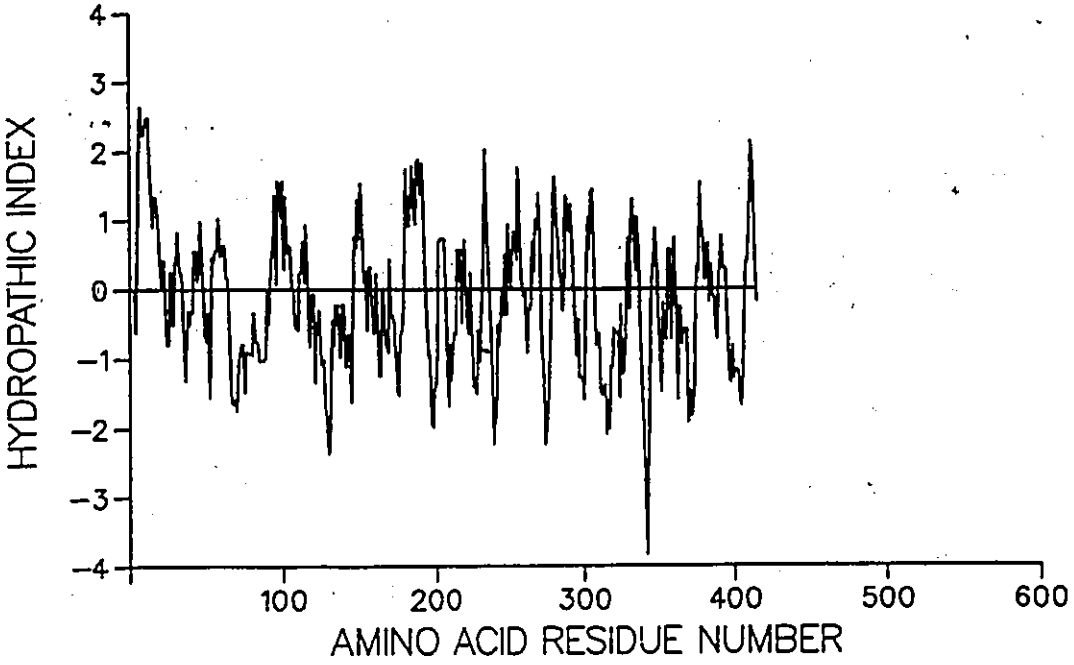
Appendix C


Erwinia chrysanthemi celZ
HYDROPATHIC INDEX
4
3
2
1
0
−1
−2
−3
−4
100
200
300
400
500
600
AMINO ACID RESIDUE NUMBER

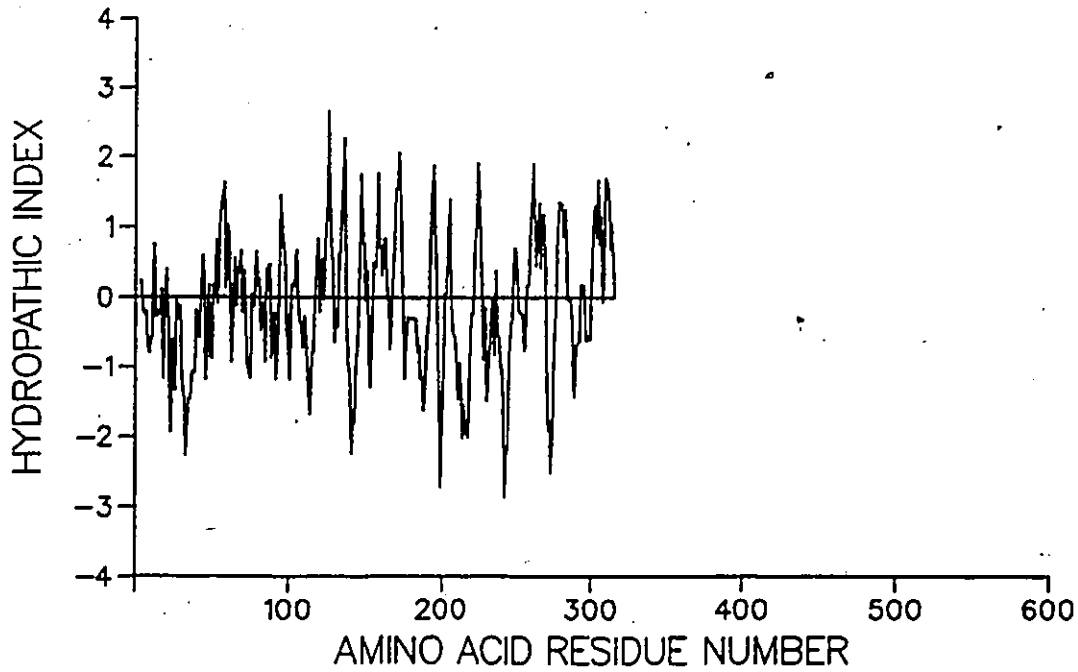

B. species strain no. 1139
HYDROPATHIC INDEX
4
3
2
1
0
−1
−2
−3
−4
0
100
200
300
400
500
600
700
800
AMINO ACID RESIDUE NUMBER

Appendix C

T. reesei EG3

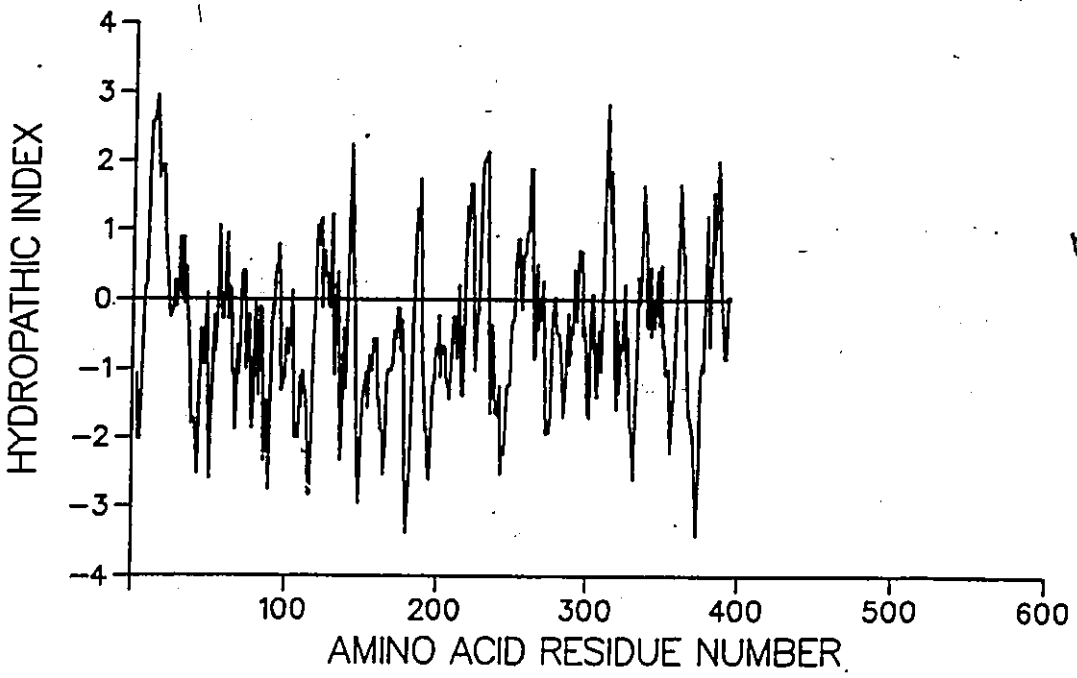


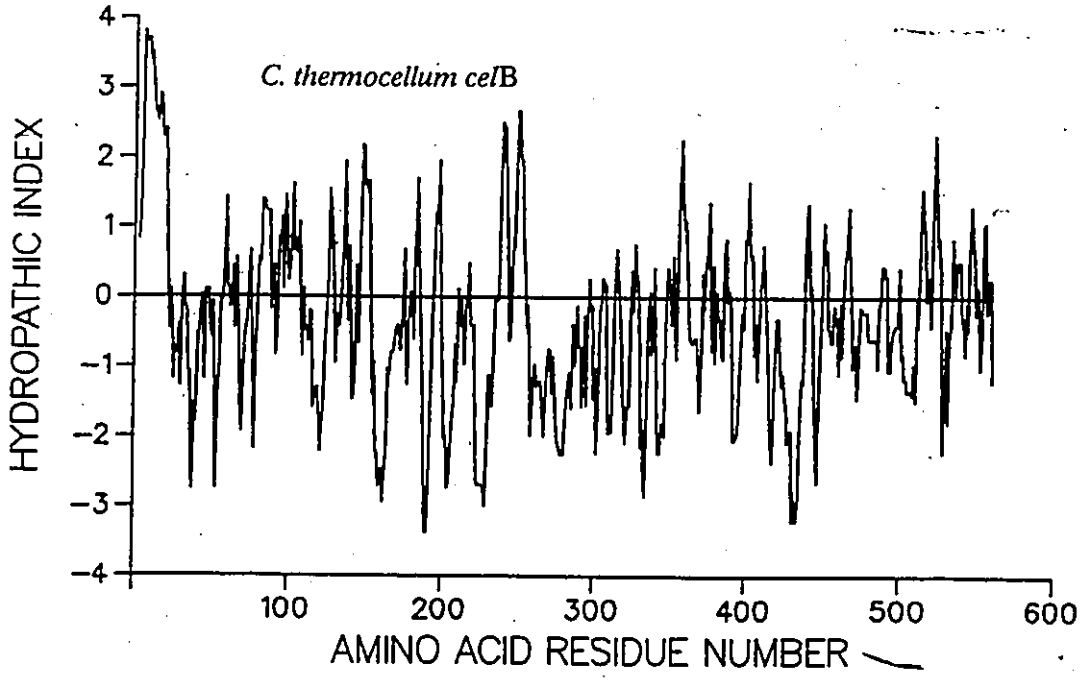
S. commune EG2



Appendix C

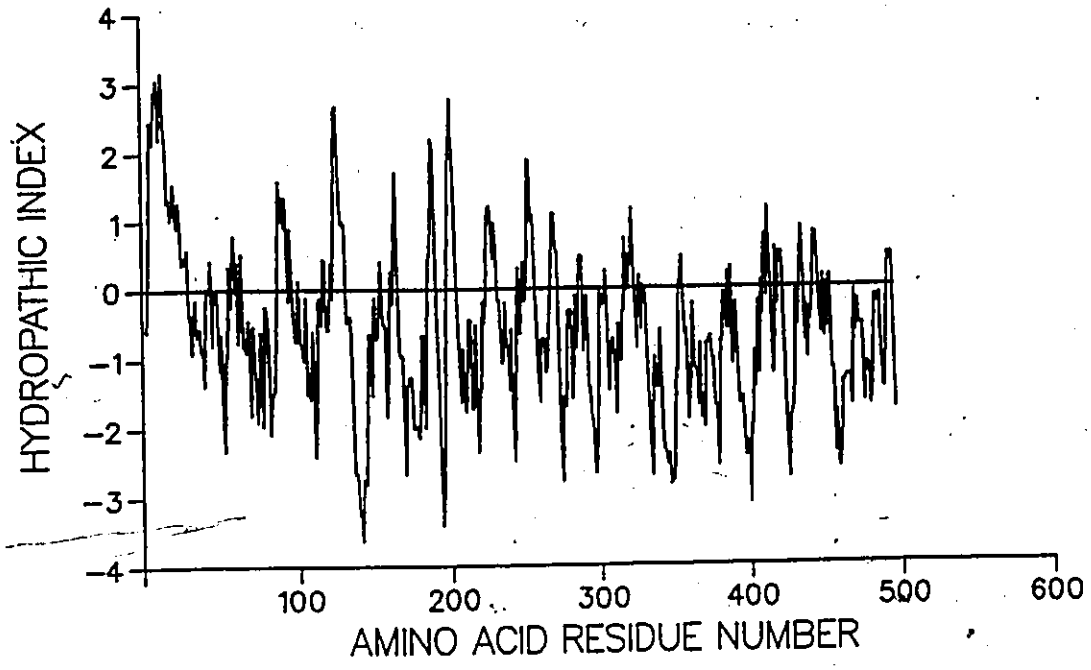
B. polymyxa EG



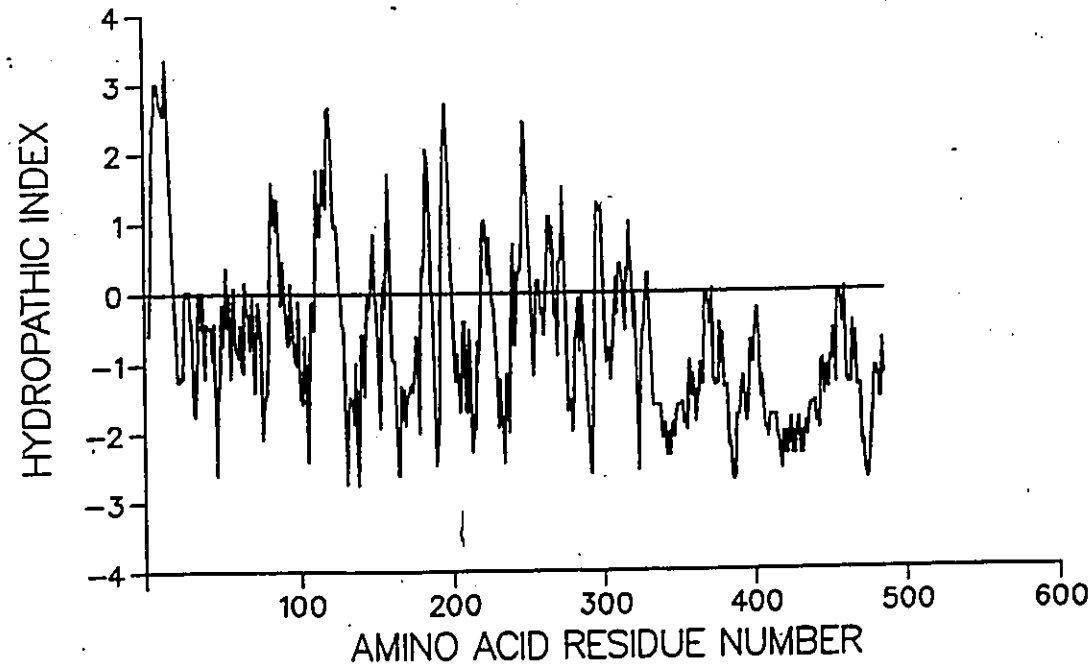


Appendix C

B. subtilis PAP115 EG



B. species N-4



Appendix D

DNA size markers for Fig. 3B and Fig. 5

Size (bp)	DNA Type	REN
7250	M13mp7	PstI
5996	pBR325	PstI
4906	pBR328	PstI
3584	pBR322	PstI/HindIII
2686	pUC8	PstI
2319	pBR322	BglI
1809	"	"
1444	pBR322	TaqI
1307	"	"
779	pBR33	PstI/HindIII
475	pBR322	TaqI
368	"	"
315	"	"
313	"	"
234	"	BglI
141	"	TaqI

note: some bands due to partial digests of the DNA with TaqI of around 2400 to 2500 bp and 830 to 930 bp ranges are evident in the pictures of the stained DNA in the gels.

VI. REFERENCES

ALLEN, A. 1983. Enzymatic hydrolysis of cellulose to fermentable sugars. In Liquid Fuels Developments, CRC Series in Bioenergy Systems, pp. 49-64. Edited by D. L. Wise. Boca Raton, Florida: CRC Press.

AU, K.-S., AND K.-Y. CHAN. 1987. Purification and properties fo the endo-1,4-ß-glucanase from *Bacillus subtilis*. Journal of General Microbiology **133**, 2155-2162.

BARNETT, C. AND S.P. SHOEMAKER. 1988. Expression of endoglucanase genes of *Trichoderma reesei* in *Aspergillus*. Poster Presentation, see Biochemistry and Genetics of Cellulose Degradation, pp. 412. Edited by J.-P. Aubert, P. Béguin and J. Millet. London: Academic Press.

BARROS, M.E., AND J.A. THOMPSON. 1987. Cloning and expression in *Escherichia coli* of a cellulase gene from *Ruminococcus flavefaciens*. Journal of Bacteriology **169**, 1760-1762.

BÉGUIN, J. MILLET, O. GRÉPINET, A. NAVARRO, M. JUY, A. AMIT, R. POLJÁK AND J.-P. AUBERT. 1988. The *cel* (cellulose degradation) genes of *Clostridium thermocellum*. In Biochemistry and Genetics of Cellulose Degradation, pp. 267-282. Edited by J.-P. Aubert, P. Béguin and J. Millet. London: Academic Press.

BÉGUIN, P., M. ROCANCOURT, M.-C. CHEBRO AND J.P. AUBERT. 1986. Mapping of mRNA encoding endoglucanase A from *Clostridium thermocellum*. Molecular and General Genetics **202**, 251-254.

BÉGUIN, P., N.R. GILKES, D.G. KILBURN, R.C. MILLER,JR, G. O'NEILL, R.A. WARREN. 1987. Cloning of cellulase genes. CRC Critical Reviews of Biotechnology. **6**, 129-162.

BÉGUIN, P., P. CORNET AND J.-P. AUBERT. 1985. Sequence of a cellulase gene of the thermophilic bacterium *Clostridium thermocellum*. Journal of Bacteriology. **162**, 102-105.

BELDMAN, G., M. SEARLE-VAN LEEUWEN, F. ROMBOUTS AND F..VORGAN. 1985. The cellulase of *Trichoderma viride*: Purification, characterization, and comparison of all detectable endoglucanases, exoglucanases and ß-glucosidases. European Journal of Biochemistry **146**, 301-308.

BETZ, J. 1987. Affinities of tight-binding lactose repressors for wild-type and pseudo-operators. Journal of Molecular Biology **195**, 495-504.

BIELY, P., D. MISLOVICOVA, R. TOMAN. 1985. Soluble chromogenic substrates for the assay of endo-1,4-ß-xylanases and endo-1,4-ß-glucanases. Analytical Biochemistry **144**, 142-146.

BIGGIN, M.D., T.J. GIBSON AND G.F. HONG. 1983. Buffer gradient gels and ^{35}S label as an aid to rapid DNA sequence determination. Proceedings of the National Academy of Science of the U.S.A. **80**, 3963-3965.

BISARIA, V.S. AND T.K. GHOSE. 1981. Biodegradation of cellulosic materials: substrates, microorganisms, enzymes and products. Enzyme and Microbial Technology **3**, 90-104.

BRADFORD, M. 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Analytical Biochemistry **92**, 248-254.

BRAUN, H., G. LEGLER, J. DESHUSSES AND G. SEMENZA. 1977. Stereospecific ring opening of conduritol-B-epoxide by and active site aspartate residue of sucrase-isomaltase. Biochimica et Biophysica Acta **483**, 135-140.

BROWN, D.E., AND M. HASAN. 1988. Fat production by *Trichoderma reesei.* Biotechnology Letters **10**, 249-254.

BRUNNER, M. AND H. BUJARD. 1981. Promoter recognition and promoter strength in the *Escherichia coli* system. The EMBO Journal **6**, 3139-3144

BUCHANAN, R.E., AND N.E. GIBBONS (ED.). 1974. Bergey's Manual of Determinative Bacteriology, eighth edition. Baltimore: Williams and Wilkins Co.

CANTWELL, B.A., P.M. SHARP, E. GORMLEY AND D.J. MCCONNELL. 1988. Molecular cloning of *Bacillus* ß-glucanases. In Biochemistry and Genetics of Cellulose Degradation, pp. 181-201. Edited by J.-P. Aubert, P. Béguin and J. Millet. London: Academic Press.

CARLSON, J. AND J. MESSING. 1984. Efficiency in cloning and sequencing using the single-stranded bacteriophage M13. Journal of Biotechnology **1**, 253-264.

CHAHAL, D.S. 1982. Enzymatic Hydrolysis of Cellulose "State of the Art". National Research Council of Canada.

CHEN, C.M., M. GRITZALI AND D.W. STAFFORD. 1987. Nucleotide sequence and deduced primary structure of cellobiohydrolase II from *Trichoderma reesei.* Bio/Technology **5**, 274-278.

CHRISTOFFERSEN,R.E., M.L. TUCKER AND G. LATIES. 1984. Cellulase gene expression in ripening avocado fruit: the accumulation of cellulase mRNA and protein as demonstrated by cDNA hybridization and immunodetection. Plant Molecular Biology **3**, 385-391.

CLARK, A., AND M. YAGUCHI. 1985. The role of carboxyl groups in the function of endo-ß-1,4-glucanase from *Schizophyllum commune*. European Journal of Biochemistry **149**, 233-238.

CLAUSEN, E.C. AND J.L. GADDY. 1983. Production of ethanol from biomass. Biochemical Engineering III **413**, 435-447.

COLLMER, A. AND D.B. WILSON. 1983. Cloning and expression of a *Thermonospora* YX endocellulase gene in *Escherichia coli*. Bio/Technology **1**, 594-601.

COUGHLAN, M.P. 1985. The properties of fungal and bacterial cellulases with comment on the production and application. Biotechnology and Genetic Engineering Reviews **3**, 39-109.

COUGHLAN, M.P AND L.G. LUNGJDAHL. 1988. Comparative biochemistry of fungal and bacterial cellulolytic enzyme systems. In Biochemistry and Genetics of Cellulose Degradation, pp. 11-30. Edited by J.-P. Aubert, P. Béguin and J. Millet. London: Academic Press.

COWLING, E.B., AND W. BROWN. 1969. Structural features of cellulosic materials in relation to enzymatic hydrolysis. In Cellulases and Their Applications. Advances in Chemistry Series **95**, pp. 152-187. Edited by R.F. Gould. Washington, D.C: American Chemical Society Publications.

CROSBY, B., B. COLLIER, D.Y. THOMAS, R.M. TEATHER AND J.D. ERFLE. 1984. Cloning and expression in *Escherichia coli* of cellulase genes from *Bacteroides succinogenes*. In Proceedings of the 5th Bioenergy R and D seminar, pp. 573-576. Barking, England: Elsevier Applied Science Publishing Ltd.

DALE, R.M.K., B.A. MCCLURE AND J.P. HOUCHINS. 1985. A rapid single stranded cloning strategy for producing a sequential series of overlapping clones for use in DNA sequencing: application to sequencing the corn mitochondrial 18S rDNA. Plasmid **13**, 31-40.

DAYHOFF, M.O., R.M. SHWARTZ AND B.C. ORCUTT. 1978. A model of evolutionary change in proteins. In Atlas of Protein Sequence and Structure, **5**(supplement 3), pp. 345-352. Washington, D.C.: National Biomedical Research Foundation.

DEL SAL, G. AND C. SCHNEIDER. 1987. A simple and fast method for preparing single stranded DNA template suitable for sequencing. Nucleic Acids Research **15**, 10047.

DHILLON, N., S. CHHIBBER, M. SAXENA, S. PAJNI AND D.V. VADEHRA. 1985. A constitutive endoglucanase (CMCase) from *Bacillus licheniformis*-1. Biotechnology Letters **7**, 695-697.

DOOLITTLE, R.F. 1986. Of URFs and ORFs: a primer on how to analyze derived amino acid sequences. Mill Valley, CA.: University Science Books.

DRETZEN, G., M. BELLARD, P. SASSONE-CORSI AND P. CHAMBON. (1981). A reliable method for the recovery of DNA fragments from agarose and acrylamide gel. Analytical Biochemistry **112**, 295-298.

EMERT, G.H., R. KATZEN, R.E. FREDRICKSON AND K.F. KAUPISCH. 1980. Economic update of the Gulf cellulose alcohol process. Chemical Engineering Processes **79**, 47-52.

ENARI, T.-M. 1983. Microbial cellulases. In Microbial Enzymes and Biotechnology, pp. 183-223. Edited by W.M. Fogarty. New York: Applied Science Publisher.

ERIKSSON, K.E. 1982. Advances in enzymatic degradation of lignocellulosic material. In Proceedings of the International Symposium on Ethanol from Biomass, pp. 345-369. Edited by H. Duckworth. Winnipeg: Royal Society of Canada.

FAGERSTAM, L.G., L.G. PETTERSSON AND J.A. ENGSTROM. 1984. The primary structure of a 1,4-ß-glucan cellobiohydrolase from the fungus *Trichoderma reesei* QM9414. FEBS Letters **167**, 309-315.

FAURE, E., C. BAGNARA, A. BELAICH AND J.-P. BELAICH. 1988. Cloning and expression of two cellulase genes of *Clostridium cellulolyticum* in *Escherichia coli*. Gene **65**, 51-58.

FEINBERG, A.P. AND B. VOGELSTEIN. 1983. A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Analytical Biochemistry **132**, 6-13.

FINCH, P., AND J.C. ROBERTS. 1985. Enzymatic degradation of cellulose. In Cellulose Chemistry and its Application, pp. 312-343. Edited by T.P. Nevell and S.H. Zeronian. Chichester, England: Ellis Horwood Limited.

FOGARTY W.M., AND P. J. GRIFFIN. 1973a. Some Preliminary Observations on the Production and Properties of a Cellulolytic Enzyme Elaborated by *Bacillus Polymyxa*. Biochemical Society Transaction **1**, 1297-1298.

FOGARTY, W.M., AND P. GRIFFIN. 1973b. Production and purification of the metalloprotease of *Bacillus polymyxa*. Applied Microbiology **26**, 185-190.

FOUET, A., A. KLIER AND G. RAPPORT. 1986. Nucleotide sequence of the sucrase gene of *Bacillus subtilis*. Gene **45**, 221-225.

FOWLER, M.W. 1983. Commercial applications and economic aspects of mass plant cell culture. In Plant Biotechnology, Society for Experimental Biology Seminar Series **18**, pp. 3-37. Cambridge: Cambridge University Press.

FRIEDRICH, J., A. CIMERMAN AND A. PERDIH. 1987. Mixed culture of *Aspergillus awamori* and *Trichoderma reesei* for bioconversion of apple distillery waste. Applied Microbiology and Biotechnology **26**, 299-303.

FUKUMORI, F., T. KUDO AND K. HORIKOSHI. 1987. Truncation analysis of an alkaline cellulase from an alkalophilic *Bacillus* species. FEMS Microbiology Letters **40**, 311-314.

FUKUMORI, F., T. KUDO, Y. NARAHASHI AND K. HORIKOSHI. 1986a. Molecular cloning and nucleotide sequence of the alkaline cellulase gene from the alkalophilic *Bacillus* sp. Strain 1139. Journal of General Microbiology **132**, 2329-2335.

FUKUMORI, F.,N. SASHIHARA, T. KUDO AND K. HORIKOSHI. 1986b. Nucleotide sequences of two cellulase genes from alkalophilic *Bacillus* sp. strain N-4 and their strong homology. Journal of Bacteriology **168**, 479-485.

GHANGAS, G.S. AND D.B. WILSON. 1987. Expression of a *Thermomonospora fusca* cellulase gene in *Streptomyces lividans* and *Bacillus subtilis*. Applied and Environmental Microbiology **53**, 1470-1475.

GHOSE, T.K., AND P. GHOSH. 1979. Cellulase production and cellulose hydrolysis. Process Biochemistry **14**(11), 20-24.

GIJSEGEM, F.V., A. TOUSSAINT, AND E. SCHOONEJANS. 1985. *In vivo* cloning of the pectate lyase and cellulase genes of *Erwinia chrysanthemi*. The EMBO Journal **4**, 787-792.

GILKES, N.R., M.L. LANGSFORD, R.C. MILLER JR., W.W. WAKARACHUK, R.A. Warren, D.J. Whittle, and W.K. Wong. 1984. Isolation and characterization of *Escherichia coli* clones expression cellulase genes from *Cellulomonas fimi*. Journal of General Microbiology **130**, 1377-1384

GILMAN, M.Z., J.L. WIGGS AND M.J. CHAMBERLAIN. 1982. Isolation and sequencing of two promoter sites used by *Bacillus subtilis* sigma-28 RNA polymerase. In Promoters: Structure and Function, pp. 293-304. Edited by R.L. Rodriguez and M.J. Chamberlain. New York: Praeger Publications.

GOKSØYR, J. AND J. ERIKSEN. 1980. Cellulases. In Microbial Enzymes and Bioconversions, Economic Microbiology Series 5, pp. 283-330. Edited by A.H.Rose. London: Academic Press.

GORDEN, R.E., W.C. HAYNES AND D.H.-N. PANG. 1973. The genus Bacillus. Agriculture Handbook no.427. Washington, D.C.: U.S. Department of Agriculture.

GREAVES, H. 1971. The effect of substrate availability on cellulolytic enzyme production by selected wood-rotting organisms. Australian Journal of Biological Science **24**, 1169-1180.

GREENBERG, N.M., R.A.J. WARREN, K.G. KILBURN AND R.C. MILLER JR. 1987a. Regulation, initiation and termination of the *cen*A and *cex*A transcripts of *Cellulomonas fimi*. Journal of Bacteriology **169**, 646-653.

GREENBERG, N.M., R.A.J. WARREN, K.G. KILBURN AND R.C. MILLER JR. 1987b. Regulation and initiation of *cen*B transcripts of *Cellulomonas fimi*. Journal of Bacteriology **169**, 4674-4677.

GRÉPINCT, O., AND P. BÉGUIN. 1986. Sequence of the cellulase gene of *Clostridium thermocellum* coding for endoglucanase B. Nucleic Acids Research **14**, 1791-1799.

GRIFFIN, H, F.R. DINTZIS, L. KRULL AND G.L. BAKER. 1984. A microfibril generating factor from the enzyme complex of *Trichoderma reesei*. Biotechnology and Bioengineering **26**, 296-300.

GRIFFIN, P.J., AND W.M. FOGARTY. 1971. Some properties of a protease from *Bacillus polymyxa*. Proceedings of the Biochemical Society **125**, 109p.

GUISEPPI, A., B. CAMI, J.-L. AYMÈRIC, G. BALL AND N. CREUZET. 1988. Homology between endoglucanase Z of *Erwinia chrysanthemi* and endoglucanases of *Bacillus subtilis* and alkalophilic *Bacillus*. Molecular Microbiology **2**, 159-164.

HAIGLER, C.H. 1985. The functions and biogenesis of native cellulose. In Cellulose Chemistry and its Application, pp. 31-83. Edited by T.P. Nevell and S.H. Zeronian. Chichester, England: Ellis Horwood Limited.

HANAHAN, D. 1983. Studies on transformation of *Escherichia coli* with plasmids. Journal of Molecular Biology **166**, 557-580.

HANAHAN, D. 1985. Techniques for Transformation of *E. coli*. In DNA Cloning Vol. I a practical approach, pp. 109-135. Edited by D.M. Glover. Oxford, England: IRL Press Ltd.

HANNA, Z., C. FREGEAU, G. PREFONTAINE AND R. BROUSSEAU. 1984. Construction of a family of universal expression plasmid vectors. Gene **30**, 247-250.

HAYASHIDA, S., AND K. MO. 1986. Production and characteristics of Avicel-disintegrating endoglucanase from a protease-negative *Humicola grisea* var. *thermoidea* mutant. Applied and Enviromental Microbiology **51**, 1041-1046.

HAYASHIDA, S., AND K. MO AND A. HOSADA. 1988. Production and Characteristics of Avicel-digesting and non-Avicel-digesting cellobiohydrolases from *Aspergillus ficum*. Applied and Enviromental Microbiology **54**, 1523-1529.

HOEFSLOOT, L.H., M. HOOGEVEEN-WESTERVELD, M. A. KROOS, J. BEEUMEN, A.J. REUSER and B. OOSTRA. 1988. Primary structure and processing of lysosomonal α-glucosidase; homology with the intestinal sucrase-isomaltase complex. The EMBO Journal **7**, 1697-1704.

HORIKOSHI, K., AND F. FUKUMORI. 1988. Modification and expression of alkaline cellulase genes of alkalophilic *Bacillus* strains. In Biochemistry and Genetics of Cellulose Degradation, pp. 203-217. Edited by J.-P. Aubert, P. Béguin and J. Millet. London: Academic Press.

HUFF, G.F. 1981. Ethanol from biomass. In Alternative Energy Sources, Part A pp. 331-353. Edited by J.T. Manassah. New York: Academic press.

HULME, M.A. 1971. Viscometric determination of carboxymethylcellulase in Standard International units. Archives of Biochemica and Biophysica **147**, 49-54.

HUNZICKER, W., M. SPIESS, G. SEMENZA AND H.F. LODISH. 1986. The sucrase-isomaltase complex: primary structure, membrane-orientation, and evolution of a stalked, intrinsic brush border protein. Cell **46**, 227-234.

JAWETZ, P. 1981. The common sense approach in developing fuel alcohols. In Alternative Energy Sources II, Hydrocarbon Conversion 7, pp. 2861-2865. Edited by T. Veziroglu. Miami, Florida.: Hemisphere Publishing Co.

JESSEE, J. 1986. New subcloning efficiency competent cells: > 1 X 10^6 transformants/μg. Focus **9**, 9.

JOHNSON, W.C., C.P. MORAN JR. AND R. LOSICK. 1983. Two RNA polymerase sigma factors from *Bacillus subtilis* discriminate between overlapping promoter for a developmentally regulated gene. Nature(London) **302**; 800.

JOLIFF G., P. BEGUIN AND J.-P. AUBERT. 1986. Nucleotide sequence of the cellulase gene *celD* encoding endoglucanase D of *Clostridium thermocellum*. Nucleic Acids Research **14**, 8605-8613.

KANEHISA, M. 1982. Los Alamos sequence analysis package for nucleic acids and proteins. Nucleic Acids Research **10**, 183-196.

KAWAZU T., Y. NAKANISHI, N. UOZUMI, T. SASACKI, H. YAMAGATA, N. TSUKAGOSHI AND S. UDAKA. 1987. Cloning and nucleotide sequence of the gene coding for enzymatically active fragments of the *Bacillus polymyxa* ß-amylase. Journal of Bacteriology **169**, 1564-1570.

KIM, J-M., I-S. KONG, J-H. YU. 1987. Molecular cloning of an endoglucanase gene from an alkalophilic *Bacillus* sp. and its expression in *Escherichia coli*. Applied and Environmental Microbiology **53**, 2656-2659.

KNØSEL, D. 1971. Continued investigation for pectolytic and cellulolytic activity of different *Bacillus* species. Zentrabl. Bakteriol. Parasitenkd. Infecktionskr. Hyg. Abt. 2 **126**, 604-609.

KOENIGS J.W. 1975. Hydrogen Peroxide and iron: a microbial cellulolytic system. Biotechnology and Bioengineering Symposium **5**, 151-159.

KOIDE, Y., A. NAKAMURA, T. UOZUMI AND T. BEPPU. 1986. Molecular cloning of a cellulase gene from *Bacillus subtilis* and its expression in *Escherchia coli*. Agricultural Biology and Chemistry **50**, 233-237.

KOTOUJANSKY, A., A. DIOLEZ, M. BOCCARA, U. BERTHEAU, T. ANDRO AND A. COLENO. 1985. Molecular cloning of *Erwinia chrysanthemi* pectinase and cellulase structural genes. The EMBO Journal **4**, 781-785.

KYTE,J. AND R.F. DOOLITTLE . 1982. A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology **157**, 105-132.

LABUDOVA, I, AND V. FARKAS. 1983. Multiple enzyme forms in the cellulase system of *Trichoderma reesei* during its growth on cellulose. Biochimica et Biophysica Acta. **744**, 135-140.

LADISH, M.R., K.W. LIN, M. VOLDOCH AND G.T. TSAO. 1983. Process considerations in the enzymatic hydrolysis of biomass. Enzyme and Microbiological Technology **5**, 82-102.

LAEMMLI, U.K. 1970. Cleavage of Structural proteins during the assembly of the head of bacteriophage T4. Nature **4**, 680-685.

LAMED, R., J. NAIMARK, E. MORGENSTERN AND E. BAYER. 1987. Specialized cell surface structures in cellulolytic bacteria. Journal of Bacteriology **169**, 3792-3800.

LANGSFORD, M.L., N.R. GILKES, W.W. WAKARCHUK, D.G. KILBURN, R.C. MILLER JR. AND R.A.J. WARREN. 1984. The cellulase system of *Cellulomonas fimi*. Journal of General Microbiology **130**, 1367-1376.

LEATHERWOOD, J.M. 1969. Cellulase complex of *Ruminococcus* and a new method of cellulose degradation. In Cellulases and their Applications, Advances in Chemistry Series **95**, pp. 53-59. Edited by R.F. Gould. Washington, D.C.: American Chemical Society Publications.

LEE, D.S.,AND M.Y. PACK. 1987. Use of bacilli for overproduction of exocellular endo-ß-1,4-glucanase encoded by cloned gene. Enzyme and Microbial Technology **9**, 10, 594-497.

LEGLER, G., K.-R. ROESER AND H.-K. ILLIG. 1979. Reaction of ß-D-glucosidase A_3 from *Aspergillus wentii* with D-glucal. European Journal of Biochemistry **101**, 85-92.

LEWIS, C. 1977. Fuels from biomass-energy outlay versus energy returns: a critical appraisal. Energy **2**, 241-248.

LIPMAN, D.J. AND W.R. PEARSON. 1985. Rapid and sensitive protein similarity searches. Science **227**, 1435-1441.

LIZARDI, P.M., R. BINDER AND S. SHORT. 1984. Preparative isolation of DNA and biololgically active mRNA from diethylaminoethyl membrane. Gene Analysis Techniques 1, 33-39.

LJUNGDAHL, L.G., B. PETTERSSON, K.-E. ERIKSSON AND J. WEIGEL. 1983. A yellow affinity substance involved in the cellulolytic system of *Clostridium thermocellum*. Current Microbiology **9**, 195-200.

LJUNGDAHL, L.G. AND K.-E. ERIKSSON. 1985. Ecology of microbial cellulose degradation. Advances in Microbial Ecology **8**, 237-299.

LO, A., R.M. MCKAY, V.L. SELIGY AND G. WILLICK. 1988. *Bacillus subtilis* ß-1,4-endoglucanase products from intact and truncated genes are secreted into the extracellular medium by *Escherichia coli*. Applied and Enviromental Microbiology, **54**(9), in press.

LUCKNOW, V.A., AND M.D. SUMMERS. 1988. Trends in the development of Baculovirus expression vectors. Bio/Technology **6**, 47-55.

MACDONALD, A. 1982. Political economy of energy. From transcripts of *Ideas*. Canadian Broadcasting Corporation.

MACKAY R.M., A. LO, G. WILLICK, M. ZUKER, S. BAIRD, M.DOVE, F. MORANELLI AND V. SELIGY. 1986. Structure of a *Bacillus subtilis* endo-ß-1,4-glucanase gene. Nucleic Acids Research **14**, 9159-9170.

MANDELS, M. 1985. Applications of cellulase. Biochemical Society Transactions **13**, 414-416.

MANDELSTAM, J., AND K. MCQUILLEN. 1973. Biochemistry of Bacterial Growth 2nd edition. New York: John Wiley and Sons, Inc.

MANIATIS, T., E.F. FRITSCH AND J. SAMBROOK. 1982. Molecular Cloning: a Laboratory Manual. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory.

MARGARITIS, A. 1986. Thermostable cellulases from thermophilic microorganisms. Critical Reviews in Biotechnology **4**:327-377.

MATHESON, N.K. AND B.V. MCCLEARY. 1985. Enzymes metabolizing polysaccharides and their application to the analysis of structure and function of glycans. In The Polysaccharides Vol. 3, pp. 1-105. Edited by G. Aspinall. London: Academic Press.

MAYAUX, J.-F., F. SOUBRIER AND M. LATTA. 1987. Cloning *E. coli* genes by oligonucleotide hybridization. Nucleic Acids Research **15**, 10593-10594.

MESSING, J. 1983. New M13 vectors for cloning genes. Methods in Enzymology. **101**, 20-78.

MILLER, A.R. 1986. Oxidation of cell wall polysaccharides by hydrogen peroxide: a potential mechanism for cell wall breakdown in plants. Biochemical and Biophysical Research Communications **141**, 238-244.

MILLER, G.L. 1959. Use of dinitrosalicylic acid reagent for determination of reducing sugar. Analytical Chemistry **31**, 426-428.

MILLER, R.C., N.R. GILKES, N.M. GREENBERG, D.G. KILBURN, M.L. LANGSFORD AND R.A.J. WARREN. 1988. *Cellulomonas fimi* cellulases and their genes in Biochemistry and Genetics of Cellulose Degradation, pp. 235-248. Edited by J.-P. Aubert, P. Béguin and J. Millet. London: Academic Press.

MULLER, P.C., AND W.G. GLASSER. 1983. Lignin's potential contribution to the feasibility of biomass conversion of ethanol. In Proceedings of the Fifth Symposium on Biotechnology for Fuels and Chemicals, pp. 481-494. Edited by C.D. Scott. Salem, MA: John Wiley and Sons.

MURPHY, N., D.J. MCCONNELL AND B.A. CANTWELL. 1984. The DNA sequence of the gene and genetic control sites for the excreted *B. subtilis* enzyme ß-glucanase. Nucleic Acids Research **12**, 5355-5367.

NAKAMURA, K., N. MISAWA AND K. KITMURA. 1986. Sequence of a cellulase gene of *Cellulomonas uda* CB4. Journal of Biotechnology **4**, 247-254.

NAKAMURA, A., T. UOZUMI AND T. BEPPU. 1987. Nucleotide sequence of a cellulase gene of *Bacillus subtilis*. European Journal of Biochemistry **164**, 317-320.

NEEDLEMAN, S.B., AND C.D. WUNSCH. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology **48**, 443-453.

NEN. 1983. Protocols for eletrophoretic and capillary transfer of DNA, DNA hybridization, DNA rehybridization. New England Nuclear, Dupont, document no. NEF-976.

NEVELL, T.P., AND S.H. ZERONIAN. 1985. Cellulose chemistry fundamentals. In Cellulose Chemistry and its Applications, pp. 15-29. Edited by T.P.Nevell and S.H. Zeronian. Chichester, England: Ellis Horwood Limited.

NIKU-PAAVOLA, M.-L., A. LAPPALAINEN, T.-M. ENARI AND M. NUMMI. 1985. A new appraisal of the endoglucanases of the fungus *Trichoderma reesei*. Biochemical Journal **231**, 75-81.

O'NEILL, G., S.H. GOH, R.A.J. WARREN, D.G. KILBURN AND R.C. MILLER JR. 1986. Structure of the gene encoding the exoglucanase of *Cellulomonas fimi*. Gene **44**, 325-330.

ODELSON, D.A., AND J. BREZNAK. 1985. Nutrition and growth characteristics of *Trichomitopsis termopsidis*. Applied and Environmental Microbiology, **49**, 614-621.

OHMIYA, K. K. NAGASHIMA, T. KAJINO, E. GOTO, A. TSUKADA AND S. SHIMIZU. 1988. Cloning of the cellulase gene from *Ruminococcus albus* and its expression in *Escherichia coli*. Applied and Enviromental Microbiology **54**, 1511-1515.

ORNSTEIN, D.L., AND M.A. KASHDAN. 1985. Sequencing DNA using ^{35}S-labeling: a troubleshooting guide. BioTechniques **3**, 476-483.

OTJEN, L., AND R. BLANCHETTE. 1984. *Xylobolus frustulatus* decay of oak: patterns of selective delignification and subsequent cellulose removal. Applied and Enviromental Microbiology **47**, 670-676.

PAICE, M.G., AND L. JURASEK. 1979. Structural and mechanistic comparisons of some ß-1,4-glycoside hydrolases. In Hydrolysis of Cellulose: Mechanisms of Enzymatic and Acid Catalysis, Advances in Chemistry Series **181**, pp. 361-374. Edited by R.D. Brown Jr. and L. Jurasek. Washington, D.C.: American Chemical Society.

PAICE, M.G., R. BOURBONNAIS, M. DESROCHER, L. JURASEK AND M. YAGUCHI. 1986. A xylanase gene from *Bacillus subtilis*: nucleotide sequence and comparison with *B. pumilus* gene. Archives of Microbiology **144**, 201-206.

PAICE, M.G., M. DESROCHER, D. RHO, L. JURASEK, C. ROY, C.F. ROLLIN, E. MIGUEL AND M. YAGUCHI. 1984. Two forms of endoglucanase from the basidiomycete *Schizophyllum commune* and their relationship to other ß-1,4-glycoside hydrolases. Bio/Technology **2**, 533-539.

PARK, S. H., AND M.Y. PACK. 1986. Cloning and expression of a *Bacillus* cellulase gene in *Escherichia coli*. Enzyme and Microbial Technology **8**, 725-728.

PARKER, A.J., T.J. TIMBARIO AND J.A. MULLONEY. 1981. Ethanol from municipal cellulosic wastes. In Alternative Energy Sources II, Hydrocarbon Conversion 7, pp. 2909-2922. Edited by T. Veziroglu. Miami, Florida: Hemisphere Publishing Co.

PENTILLÄ, M., P. LEHTOVAARA, H. NEVALAINEN, R. BHIKHABBAI AND J. KNOWLES. 1986. Homology between cellulase genes of *Trichoderma reesei*: complete nucleotide sequence of the endoglucanase I gene. Gene **45**, 253-263.

PENTTILÄ, M.E., L. ANDRÉ, P. LEHTOVAARA, M. BAILEY, T.T. TEERI AND J.K. KNOWLES. 1988. Efficient secretion of two fungal cellobiohydrolases by *Saccharomyces cerevisiae* Gene **63**, 103-112.

PERO, J., G. LEE, C.P. MORAN, N. LANG AND R. LOSICK. 1982. Promoters controlled by novel sigma factors in *Bacillus subtilis*. In Promoters: Structure and Function, pp. 283-292. Edited by R.L. Rodriguez and M.J. Chamberlain. New York: Praeger Pub.

PETRE, J., R. LONGIN AND J. MILLET. 1981. Purification and properties of an endo-ß-1,4-glucanase from *Clostridium thermocellum*. Biochemie **63**, 629-639.

PETROLEUM MONITORING AGENCY. 1988. Canadian Petroleum Industry 1987 Monitoring Survey. Ministry of Supply and Services Canada.

PETROLEUM MONITORING AGENCY. 1987. Canadian Petroleum Industry 1986 Monitoring Survey. Ministry of Supply and Services Canada.

PLATT, T. AND D. BEAR. 1983. Role of RNA polymerase, *rho* factor, and ribosomes in transcription termination. In Gene Function in Prokaryotes, pp. 123-161. Edited by J. Beckwith, J. Davies and J. Gallant. New York: Cold Spring Harbour Laboratory.

PRESTON, R.D. 1986. Natural celluloses. In Cellulose: structure, modification and hydrolysis, pp. 3-27. Edited by R.A. Young and R. M. Rowell. New York: John Wiley & Sons, Inc.

PRIEST, F.G. 1977. Extracellular enzyme synthesis in the Genus *Bacillus*. Bacteriological Reviews, **41**, 3, 711-753.

PUSTELL, J., AND F. KAFATOS. 1982. A convenient and adaptable package of DNA sequence analysis programs for microcomputers. Nucleic Acids Research **10**, 51-59.

PUSTELL, J., AND F. KAFATOS. 1984. A convenient and adaptable package of computer programs for DNA and protein sequence management, analysis and homology determination. Nucleic Acids Research **12**, 643-655.

QUARONI, A., AND G.SEMENZA. 1976. Partial amino acid sequences around the essential cargosylate in the active sites of the intestinal sucrase-isomaltase complex. Journal of Biological Chemistry **251**, 3250-3253.

REED, K.C., AND D.A. MANN. 1985. Rapid transfer of DNA from agarose gels to nylon membranes. Nucleic Acid Research. **13**, 7207-7221.

REESE, E.T., AND M MANDELS. 1984. Rolling with the times: Production and applications of *Trichoderma reesei* cellulase. In Annual Reports on Fermentation Processes, vol.7, pp. 1-20. Edited by G. Tsao, M. Flickner,and R. Finn. New York: Academic Press.

REESE, E.T.,R. SIU AND H. LEVINSON. 1950. The biological degradation of soluble cellulose derivatives and its relationship to the mechanism of cellulose hydrolysis. Journal of Bacteriology **59**, 485-497.

ROBSON, L. AND G. CHAMBLISS. 1984. Characterization of the cellulolytic activity of a *Bacillus* isolate. Applied and Environmental Microbiology **47**, 1039-1046.

ROBSON, L.M., AND G.H. CHAMBLISS. 1986. Cloning of the *Bacillus subtilis* DLG ß-1,4-glucanase gene in *Escherichia coli* and *B. subtilis*. Journal of Bacteriology **165**, 612-619.

ROBSON, L.M., AND G.H. CHAMBLISS. 1987. Endo-ß-1,4-glucanase gene of *Bacillus subtilis* DLG. Journal of Bacteriology **169**, 2017-2025.

ROESER, K.R., AND G. LEGLER. 1981. Role of sugar hydroxyl groups in glycoside hydrolysis: cleavage mechanism of deoxyglucosides and related subtrates by ß-glucosidase A_3 from *Aspergillus wentii*. Biochimica et Biophysica Acta **657**, 321-333.

ROMANIEC, M.P., N. CLARKE AND G.P. HAZLEWOOD. 1987. Molecular cloning of *Clostridium thermocellum* DNA and the expression of further novel endo-ß-1,4-glucanase genes in *Escherichia coli*. Journal of General Microbiology **133**, 1297-1307.

ROMANIEC, M.P., T. KIESER AND G.P. HAZLEWOOD. 1988. Insertion and expression of *Clostridium thermocellum cel* genes in *Streptomyces albus*. Poster Presentation see Biochemistry and Genetics of Cellulose Degradation, pp. 411. Edited by J.-P. Aubert, P. Béguin and J. Millet. London: Academic Press.

SACCO, M., J. MILLET AND J.P. AUBERT. 1984. Cloning and expression in *Saccharomyces cerevisiae* of a cellulase gene from *Clostridium thermocellum*. Ann. Microbiol. Inst. Pasteur **135A**, 485-488.

SADDLER, J.N. 1986. Factors limiting the efficiency of cellulase enzymes. Microbiological Sciences **3**, 84-87.

SAITO, H., AND K. MIURA. 1963. Preparation of transforming deoxyribonucleic acid by phenol treatment. Biochemica et Biophysica Acta. 72, 619-629.

SALERNO, A.J., AND J.O. LAMPEN. 1988. Differential transcription of the *bla* regulatory region during induction of ß-lactamase in *Bacillus licheniformis*. FEBS Letters **227**, 61-65.

SALOHEIMO, M., P. LEHTOVAARA, M. PENTTILÄ, T.T. TEERI, J. STAHLBERG, G. JOHANSSON, G. PETTERSSON, M. CLAEYSSENS, P. TOMME AND J. K. KNOWLES. 1988. EGIII, a new endoglucanase from *Trichoderma reesei*: the characterization for both gene and enzyme. Gene **63**, 11-21.

SASHIHARA, N., T. KUDO AND K. HORIKOSHI. 1984. Molecular cloning and expression of cellulase genes of alkalophilic *Bacillus* sp. strain N-4 in *Escherichia coli*. Journal of Bacteriology **158**, 503-506.

SCHWARZ, W. S., S. SCHIMMING, K.P.RCKNAGEL, S. BURGSCHWAIGER, G. KREIL AND W. STAUDENBAUER. 1988.Nucleotide sequence for the *celC* gene encoding endoglucanase C of *Clostridium thermocellum*. Gene **63**, 23-30.

SEILLIERE, G. 1906. Compte Rendu de la Société de Biologie **61**, 205.

SELIGY, V.L., S. BAIRD, R. BERNIER, M. DESROCHER, M.J.DOVE, A. LO, R.M. MACKAY AND G. E. WILLICK. 1987. Functional mapping and expression of a *Bacillus subtilis* endo-1,4-ß-D-glucanase gene in Escherichia coli and the Yeasts, Saccharomyces cerevisiae and Schwanniomyces alluvius. In Biological Research on Industrial Yeast, Volume II, pp. 91-110. Edited by G.G Stewart, I. Russell, R. Klein, R. Hiebsch. Boca Raton, Florida: CRC Press, Inc.

SHARMA, P., J.K. GUPTA, D.V. VADEHRA AND D.K. DUBE. 1987. Molecular cloning and expression in *Escherichia coli* of a thermophilic *Bacillus* sp. PDV endo-ß-1,4-glucanase gene. Enzyme and Microbial Technology **9**(10), 62-66.

SHEIR-NEISS, G. AND B.S. MONTENECOURT. 1984. Characterization of the secreted cellulases of *Trichoderma reesei* wild type and mutants during controlled fermentations. Applied Microbiology and Biotechnology **20**, 46-53.

SIEBENLIST, U.,R.B. SIMPSON, AND W. GILBERT. 1980. *E. coli* RNA polymerase interacts homologously with two different promoters. Cell **20**, 269-281.

Siu, R.G.H. 1951. Microbial Decomposition of Cellulose. New York, New York: Reinhold Publishing

SKIPPER, N., M. SUTHERLAND, R.W. DAVIES, D. KILBURN, R.E. MILLER JR., A. WARREN AND R. WONG. 1985. Secretion of a bacterial cellulase by yeast. Science **230**,pg. 958-960.

SMITH, S.N., R.A. ARMSTRONG AND M.W. FOWLER. 1985. Breakdown of plant cell biomass be filamentous fungi. Enzyme and Microbial Technology 7(2), 61-65.

SOUTHERN, E.M. 1975. Detection of specific sequences amoung DNA fragments separated by gel electrophoresis. Journal of Molecular Biology **92**, 503-521.

SPECIAL COMMITTEE ON ALTERNATIVE ENERGY AND OIL SUBSTITUTION. 1981. Energy Alternatives. Ottawa: Ministry of Supply and Services Canada.

STATISTICS CANADA. 1987. Quarterly report on energy supply-demand in Canada 1986 III. Ottawa: Ministry of Supply and Services Canada.

STEPHENS, R.M. 1985. A sequencers' sequence analysis package for the IBM PC. Gene Analysis Techniques **2**, 67-75.

STONE, J. AND ASSOCIATES LTD. 1982. Analysis of Ethanol Production Potential from Cellulosic Feedstock. Ottawa: Energy Mines and Resources Canada.

STUTZENBERGER, F. 1985. Regulation of Cellulolytic Activity. In Annual Reports on Fermentation Processes, vol.9, pp.111-154. Edited by G. Tsao, M. Flickner and R. Finn. New York: Academic Press.

SZAKACS-DOBOZI, M., A. HALASZ, E. KOZMA-KOVACS AND G. SZAKACS. 1988. Enhancement of mustard oil yield by cellulolytic pretreatment. Applied Micobiology and Biotechnology **29**, 39-43.

TEATHER, R. M. & WOOD P. J.(1982) Use of Congo red-polysaccharide interactions in enumeration and characterization of cellulolytic bacteria from bovine rumen. Applied and Environmental Microbiology **43**, 777-780.

THAYER, D.W. 1978. Carboxymethylcellulase produced by facultative bacteria from the hind-gut of the termite *Reticulitermes hesperus*. Journal of General Microbiology **106**, 13-18.

TILBEURGH, H.V., AND L.G. PETTERSSON. 1985. Studies of the cellulolytic system of *Trichoderma reesei* QM 9414. European Journal of Biochemistry **148**, 329-334.

TOYAMA, N. 1969. Applications of cellulases in Japan. In Cellulases and their Applications, Advances in Chemistry Series **95**, pp. 359-390. Edited by R.F. Gould. Washington, D.C.: American Chemical Society Publications.

TRAVERS, A.A. 1987. Structure and function of *E. coli* promoter DNA. CRC Critical Reviews in Biochemistry **22**, 181-219.

VAN ARDSELL, J.N., S. KWOK, V.L. SCHWEICKART, M.B. LADNER, D.H. GELFAND AND M.A. INNIS. 1987. Cloning, characterization and expression in *Saccharomyces cerevisiae* of endoglucanase I from *Trichoderma reesei*. Bio/Technology **5**, 60-64

VANCE, I., S. BLAYDEN AND J. TAMPION. 1981. Cellulose degradation and carboxymethylcellulase production by *Sporocytophaga myxococcides* grown in an air-lift fermenter. Biotechnology Letters **3**, 45-48.

VERNON, C.A., and B. Banks. 1963. The enzymic hydrolysis of glycosides. Biochemical Journal **86**, 7p

VIEIRA, J., AND J.MESSING. 1982. The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. Gene **19**, 259-268.

WAHL, G.M., M. STERN AND G.R. STARK. 1979. Efficient transfer of large DNA fragments from agarose gels to diazobenzyloxymethyl-paper and rapid hybridization by using dextran sulfate. Proceedings of the National Academy of Science of the USA **76**, 3683-3687.

WARWICKER, J.O. 1985. Cellulose as a matrix material. In Cellulose and its Derivatives, Chemistry, Biochemistry, and Applications, pp. 372-385. Edited by Kennedy, J.F., G.O. Phillips, D.J. Wedlock and P.A. Williams. New York: Ellis Harwood Limited Publishers.

WEDEKIND, K.J., H. MANSFIELD AND L. MONTGOMERY. 1988. Enumeration and isoation of cellulolytic and hemicellulolytic bacteria from human feces. Applied and Enviromental Microbiology **54**, 1530-1535.

WILLICK, G. AND V.L. SELIGY. 1985. Multiplicity in cellulases of *Schizophyllum commune*; derivation partly from heterogeneity in transcription and glycosylation. European Journal of Biochemistry **151**, 89-96.

WOLFF, B.R., T. MUDRY, B. GLICK AND J. PASTERNAK. 1986. Isolation of endoglucanase genes from *Pseudomonas fluorescens* subsp. *cellulosa* and a *Pseudomonas* sp. Applied and Environmental Microbiology **51**, 1367-1369.

WOLF, J., AND N. BARKER. 1968. The genus *Bacillus*: aids to the identification of its species. In Identification methods for Microbiologists: Part B, Society for Applied Bacteriology Technical Series no. 2, pp. 93-109. Edited by B.M. Gibbs and D.A. Shapton. London: Academic Press.

WOLIN, M.J. AND T.L. MILLER. 1983. Interactions of microbial populations in cellulose fermentation. Federation Proceedings **42**, 109-113.

WONG, W.K., C. CURRY, R.S. PAREKH, S.R. PAREKH, M. WAYMAN, R.W. DAVIES, D.G. KILBURN AND N. SKIPPER. 1988. Wood hydrolysis by *Cellulomonas fimi* endoglucanase and exoglucanase coexpressed as secreted enzymes in *Saccharomyces cerevisiae*. Bio/Technology **6**, 713-719.

WONG, W.K.R., B. GERHARD, Z.-M. GUO, D.G. KILBURN, R. ANTHONY, J. WARREN AND R.C. MILLER JR. 1986. Characterization and structure of an endoglucanase gene *cen*A of *Cellulomonas fimi*. Gene **44**, 315-324.

WONG Y.-S., G. FINCHER AND G.A. MACLACHLAN. 1977. Kinetic properties and substrate specificities of two cellulases from auxin-treated pea epicotyls. The Journal of Biological Chemistry **252**, 1402-1407.

WOOD, T. M. 1985. Aspects of the biochemistry of cellulose degradation. In Cellulose and its Derivatives, Chemistry, Biochemistry, and Applications, pp. 173-188. Edited by Kennedy, J.F., G.O. Phillips, D.J. Wedlock and P.A. Williams. New York: Ellis Harwood Limited Publishers.

WOOD, T.M. AND S.I.MCCRAE. 1977. Cellulase from *Fusarium solani*: purification and properties of the C_1 component. Carbohydrate Research **57**, 177-133.

WOOD, T.M, S.I MACCRAE, C.A. WILSON, K.M. BHAT AND L. GOW. 1988. Aerobic and anaerobic fungal cellulases, with special reference to their mode of attack on crystalline cellulose.In Biochemistry and Genetics of Cellulose Degradation, pp. 31-52. Edited by J.-P. Aubert, P. Béguin and J. Millet. London: Academic Press.

WOODWARD, J. AND S.L. ARNOLD. 1981. The inhibition of ß-glucosidase activity in *Trichoderma reesei* C30 cellulase by derivatives and isomers of glucose. Biotechnology and Bioengineering **23**, 1153-1562.

WU, D. AND A. DEMAIN. 1988. Proteins of the *Clostridium thermocellum* cellulase complex responsible for degradation of crystalline cellulose. In Biochemistry and Genetics of Cellulose Degradation, pp. 117-131. Edited by J.-P. Aubert, P. Béguin and J. Millet. London: Academic Press.

YAGUCHI, M., C. ROY, C.F. ROLLINS, M.G. PAICE AND L.JURASK. 1983. A fungal cellulase shows sequence homology with the active site of hen egg-white lysozyme. Biochemical and Biophysical Research Communications **116**, 408-411.

YAMADA, K. 1977. Recent advances in industrial fermentation in Japan. Biotechnology and Bioengineering **19**, 1563-1621.

YANG, M., A., GALIZZI, AND D. HENNER. 1983. Nucleotide sequence of the amylase gene from *Bacillus subtilis*. Nucleic Acids Research **11**, 237-249.

YANG, R.C., C.R. MACKENZIE AND S.A. NARANG. 1988. Nucleotide sequence of a *Bacillus circulans* xylanase gene. Nucleic Acids Research **16**, 7187.

YANISH-PERRON, C., J. VIEIRA AND J. MESSING. 1984. Improved M13 phage cloning vectors and host strains: nucleic sequences of M13mp18 and pUC19 vectors. Gene **33**, 103-119.

YONEDA, Y. 1982. Regulation of α-amylase production in *B. subtilis*. In Molecular Cloning and Gene Regulation in Bacilli, pp. 111-120. Edited by A.T. Ganesan, S. Chang and J.A. Hoch. New York: Academic Press.

YOON, K.H., S.H. PARK AND M.Y.PACK. 1988. Transfer of a *Bacillus subtilis* endo-ß-1,4-glucanase gene into *Zymomonas anaerobia*. Biotechnology Letters **10**, 213-216.

ZABAROVSKY, E.R. AND R.L. ALLIKMETS. 1986. An improved technique for the efficient construction of gene libraries by partial filling in of cohesive ends. Gene **42**, 119-123.

ZUBAY, G.L. 1983. Biochemistry. Reading, Mass.: Addison-Wesley Publishing Company.