Curr Genet (1993) 24:205–211

Current Genetics

# Cloning and sequence analysis of the glucoamylase gene of *Neurospora crassa*

**P. J. Stone**[1], **A. J. Makoff**[2], **J. H. Parish**[1], **A. Radford**[3]

[1] Department of Biochemistry and Molecular Biology, The University of Leeds, Leeds LS2 9JT, UK
[2] Department of Cell Biology, The Wellcome Research Laboratories, Beckenham, Kent BR3 3BS, UK
[3] Department of Genetics, The University of Leeds, Leeds LS2 9JT, UK

Received: 14 December 1992 / Accepted: 19 March 1993

**Abstract.** A 1.0-kb DNA fragment, corresponding to an internal region of the *Neurospora crassa* glucoamylase gene, *gla-1*, was generated from genomic DNA by the polymerase chain reaction, using oligonucleotide primers which had been deduced from the known N-terminal amino-acid sequence or from consensus regions within the aligned amino-acid sequences of other fungal glucoamylases. The fragment was used to screen an *N. crassa* genomic DNA library. One clone contained the gene together with flanking regions and its sequence was determined. The gene was found to code for a preproprotein of 626 amino acids, 35 of which constitute a signal and propeptide region. The protein and the gene are compared with corresponding sequences in other fungi.

**Key words:** Glucoamylase – *Neurospora crassa* – Extracellular protein – Signal sequence

## Introduction

The filamentous fungi secrete substantial amounts of protein, notably hydrolytic enzymes. Many of these enzymes are used in industrial processes such as the production of antibiotics and organic acids, the saccharification of starch, glucose isomerisation, the processing of wines and fruit juices and the degradation of cellulose and lignin (Bennett 1985; Bu'Lock and Kristiansen 1987). The promoter and signal sequences of the genes of such enzymes represent targets for manipulation for developing the filamentous fungi as hosts for heterologous gene expression. The potentials have been reviewed with particular reference to the genus *Aspergillus* (Van den Hondel et al. 1991).

The genus *Neurospora* has several advantages for study with a view to its possible exploitation as a host for heterologous gene expression. The literature on the genetics of *Neurospora crassa* is very detailed (reviewed by Perkins 1992) and the organism is the most-thoroughly studied and characterised of all the filamentous fungi. It is fast growing, with simple growth requirements, and is more acceptable than many alternatives as a host for producing proteins for human use as it generates no toxic secondary metabolites.

Glucoamylases (exo-1,4-α-D-glucan glucohydrolase, EC 3.2.1.3) are secreted in large amounts by a variety of filamentous fungi. They catalyse the removal of single glucose units from the non-reducing ends of starch, and other poly- and oligo-saccharides. Their use in industrial processes includes the production of glucose syrups from starch (Kennedy et al. 1988), and the fermentation of sake (rice wine) in Japan. Heterologous expression systems in filamentous fungi commonly use their glucoamylase promoters to drive expression, their signal sequences to secrete foreign peptides, and their 3′ flanking regions to direct termination (Archer et al. 1990; Ward et al. 1990, 1992). Glucoamylases have been cloned and characterised from several fungi: *Aspergillus awamori* (Nunberg et al. 1984), *A. awamori* var. *kawachi* (Hayashida et al. 1989), *A. niger* (Boel et al. 1984), *A. oryzae* (Hata et al. 1991), *A. shirousami* (Shibuya et al. 1990), *Humicola grisea* var. *thermoidea* (Berka et al., personal communication), *Rhizopus oryzae* (Ashikari et al. 1986), *Saccharomyces cerevisiae* (Pardo et al. 1988), *S. diastaticus* (Yamashita et al. 1985), *S. fibuligera* (Itoh et al. 1987) and *S. occidentalis* (Dohmen et al. 1990).

Koh-Luar et al. (1989) analysed culture supernatants of *N. crassa*, growing on a variety of carbon sources, and showed that the protein present in the largest amount was a glucoamylase of approximately 69 kDa. This protein was purified and the N-terminal sequence of the glucoamylase determined. Here we report the DNA sequence of the glucoamylase gene, *gla-1*, of *N. crassa* together with flanking sequences and compare its amino-acid sequence with other glucoamylases.

*Correspondence to:* J. H. Parish

## Materials and methods

*Strains, plasmids and media. N. crassa* 74-OR23-1A was grown in Vogel's sucrose medium (Davis and de Serres 1970) and DNA was extracted by the method of Azevedo et al. (1990). *E. coli* strain TG2 was used for cloning and sequencing. The *N. crassa* genomic DNA library screened was in the vector λJ1 (Orbach et al. 1986).

*DNA manipulation and sequencing.* DNA was labelled in vitro by the random hexamer method. This and other routine techniques followed Sambrook et al. (1989). The complete DNA sequence (see Fig. 2) was determined on both strands.

*Polymerase chain reaction (PCR).* PCR amplification was carried out on 50 ng of template DNA in a total volume of 40 μl containing 10 mM Tris-HCl (pH 8.4), 50 mM KCl, 1.5 mM $MgCl_2$, 0.01% gelatin, 10 mM each of dATP, dCTP, dTTP, dGTP, 100 pmol of each primer and 0.5 units of *Thermus aquaticus* (*Taq*) DNA polymerase. The PCR was conducted for 40 cycles: denaturation at 95 °C for 1 min, annealing at 55 °C for 1 min and extension at 72 °C for 2 min.

## Results and discussion

### *Amplification of the coding sequence of gla-1 by using the PCR and identification of a genomic clone*

A PCR primer was synthesised (Fig. 1) using the N-terminal amino-acid sequence of *N. crassa* glucoamylase (Koh-Luar et al. 1989). Several fungal glucoamylase amino-acid sequences were aligned and, using their conserved regions to predict the *Neurospora* sequence, three more PCR primers were synthesised (Fig. 1). The three pairwise combinations of these primers were employed to amplify *Neurospora* genomic DNA. The combination of primers P1 : P4 and P3 : P4 gave amplified products of the expected size (1020 and 450 bp respectively). Heterologous Southern-blot hybridisation using a probe from the *A. awamori* glucoamylase gene confirmed the identity of these products.

The 1020-bp product was used as a probe to screen approximately 150,000 plaques from the *Neurospora* genomic DNA library in λJ1. Of 14 positive plaques, six were chosen for secondary screening. After plaque purification and restriction analysis, a recombinant was selected for further investigation. Southern blotting showed the coding sequence to be located within an internal 5.0-kb *Cla*I fragment and this was subcloned into pBluescript for subsequent sequencing.

### *Sequence of the gla-1 gene*

The sequence of the 3.8-kbp fragment was determined for both strands and is shown in Fig. 2. The coding sequence was identified by analogy to the aligned amino-acid sequences of known fungal glucoamylases. The gene encodes a deduced protein of 626 amino acids, with an unglycosylated molecular weight of 66,574 Da. This includes a leader peptide of 35 amino acids when compared to the known N-terminus of the secreted protein (Koh-Luar et al. 1989).

```
N-terminal N. crassa  (37)       V   D   S   Y   I   Q   T   E   T
PRIMER   P1                   5'-GTCGAATTCTACATCCAGACCGAGAC-3'


A. niger              (132)      K   F   N   V   D   E   T   A   Y   T
A. oryzae             (134)      K   F   N   V   D   E   T   A   F   T
H. grisea             (125)      K   F   N   V   D   L   T   A   F   T
                              5'-AAGTTCAACGTCGACCTGACCGCCTTCAC-3'
PRIMER   P2                   3'-TTCAAGTTGCAGCTGGACTTTCGAAAGTG-5'


A. niger              (197)      T   G   Y   D   L   W   E   E   V
A. oryzae             (199)      S   G   F   D   L   W   E   E   V
H. grisea             (190)      T   G   F   D   L   W   E   E   V
PRIMER   P3                   5'-ACCGAATTCGACCTCTGGGAGGAGGT-3'


A. niger              (335)      Y   Y   N   G   N   P   W   F   L   C
A. oryzae             (337)      Y   Y   N   G   N   P   W   F   L   T
H. grisea             (332)      Y   Y   N   G   N   P   W   Y   L   A
                              5'-TACTACAACGGCAACCCCTGGAAGCTTGCC-3'
PRIMER   P4                   3'-ATGATGTTGCCGTTGGGGACCTTCGAACGG-5'
```

**Fig. 1.** Derivation of PCR primers (P1 – P4) by known N-terminal sequence or by alignment of homologous sequences from within three closely-related fungal glucoamylases, showing the corresponding predicted DNA sequence. The *numbers in brackets* refer to the N-terminal amino acid in each sequence. The primers were designed by using the *Neurospora* codon bias table (Gurr et al. 1987). Restriction sites are *underlined in italics*; the primer names (*P1 – P4*) are written against the appropriate strand of DNA sequence used as the PCR primer, which is shown in *bold*

The coding region contains only one 62-bp intron (Fig. 2): this is in contrast to *Aspergillus* glucoamylases that contain up to five introns (reviewed by Gurr et al. 1987), and *Humicola* which has three. This results in two exons of 243 bp and 1635 bp respectively. The splice sites (Fig. 2) correspond to the fungal consensus sequences [in brackets; from Gurr et al. (1987)] as follows: 5′ splice site GTAAGT (GTANGT), intron internal consensus AGCTCAC (YGCTAAC), and 3′ splice site TAG (YAG). The intron location is exactly conserved with respect to the position of intron I in the *Aspergillus* genes (Boel et al. 1984; Nunberg et al. 1984; Hayashida et al. 1989).

The GC content of the 1881-bp coding region is 61%, compared to 49% in the non-coding region. Gurr et al. (1987) reviewed nuclear genes of filamentous fungi in general, and showed that they have a marked codon bias, indicated by a strong preference for a pyrimidine, especially C, in the 3rd nucleotide position. This is most striking in highly expressed genes. This pattern is true for the *Neurospora gla-1* gene: over 78% of the codons end in T or C, and less than 2% end in A (data not shown).

### *Promoter*

We have sequenced 938 bp upstream of the translation initiation codon. There is a TATA box at position −101 with respect to the ATG codon. The actual sequence (Fig. 2) is TATATAA and the eukaryotic consensus TATA (A/T) TA. There are several potential, although no perfect, CAAT boxes upstream of the TATA box, the most likely one to function being at −133 to the ATG start codon (CATCAATAT). The eukaryotic consensus sequence is GG (C/T) CAATCT (Breathnach and Chambon 1981).

```
atcgatggca gccaccattc atttctcgat gcgacggtaa acgacgcccg cggcagatta ggtcattgcc -869
gaacggattg aagctctctc catcttggat ccattcccgg ccaatcccgt ctcggccaac cacactgtcc -799
actcgcccag gtcagcagct caggactctc tcctggtttg gtaccgctta gtgtagagca taccgctctc -729
agtccccata gaccaaccat aacaccgcac gttctctttc actcaagatg cttatcatgt cccctctttc -659
tgctccaatg attcggactg gtcgaatacc aatgagacaa gcgagagcgc agtgcgagca agcgttcctg -589
cagatagagc agtgggactg ccgcgccaca aaggaagagg atcgtgacgt gacgtgacca gtgaccagaa -519
agcagaagat ccaaaagagt caaaaggacc gagcctcacc tacagtaatg gcccggatgg cactcaagac -449
cgtcctctcg gccctttctc caactcttct ccttccataa ttcacctagg tacatacacg gcctacgctt -379
ccgcctcatc ccatcccatc ccatcccatc ccatcccatc gacgactcta acccgcccgc gagtgcaaac -309
ctcgtccacg aacggacacc ccggctctcc tccgaagccc ttgcaagtgg aagctgaggt tgccgaactt -239
agacgaccag gttcaccagc cggaccgcaa ctcgaacgtc agaatacagc ctcagcctcc aaaggggggtt -169
aacgccaagc gagagcaaga caagatcgtc gcccatcaat atcctggaca agacaacatg gacgcaatat  -99
ataacctcaa gcaagtcctc ctcagcaacc atgatttcac caccagcctg gtctccaacg caacagactt  -29
ctcgacaagt cccttgacct acttcgcc ATG CAT CTC GTC TCT TCG CTC CTC GTC GTG GGC    33
                               met his leu val ser ser leu leu val val gly   (11)

GCC GCC TTC CAG GCC GTG CTC GGT CTG CCG GAT CCT CTG CAT GAA AAG AGG CAC AGC    90
ala ala phe gln ala val leu gly leu pro asp pro leu his glu lys arg his ser  (30)
                             ↑
GAC ATC ATC AAG CGG TCT GTC GAC TCG TAT ATC CAG ACC GAG ACT CCC ATT GCG CAG   147
asp ile ile lys arg ser val asp ser tyr ile gln thr glu thr pro ile ala gln  (49)

AAG AAC CTT CTG TGC AAC ATC GGT GCT TCT GGA TGC AGA GCC TCC GGT GCT GCC TCT   204
lys asn leu leu cys asn ile gly ala ser gly cys arg ala ser gly ala ala ser  (68)

GGT GTT GTG GTT GCC TCC CCT TCC AAG TCG AGC CCT GAC T gtaagtgga aattgcaca     262
gly val val val ala ser pro ser lys ser ser pro asp t                        (82)

gtgtgtctca tctctcatgg cagcatagct cacagtgtcg atag AC TGG TAT ACC TGG ACT CGT   326
                                                 yr trp tyr thr trp thr arg  (88)

GAT GCC GCC CTT GTC ACC AAG CTT ATT GTC GAC GAA TTC ACC AAC GAC TAC AAC ACC   383
asp ala ala leu val thr lys leu ile val asp glu phe thr asn asp tyr asn thr (107)

ACT CTT CAG AAC ACC ATT CAG GCT TAT GCT GCT GCA CAG GCC AAG CTT CAG GGC GTT   440
thr leu gln asn thr ile gln ala tyr ala ala ala gln ala lys leu gln gly val (126)

AGC AAC CCG TCC GGT TCC CTC TCC AAC GGG GCC GGT CTT GGT GAG CCC AAG TTC ATG   497
ser asn pro ser gly ser leu ser asn gly ala gly leu gly glu pro lys phe met (145)

GTC GAC CTC CAG CAG TTC ACC GGT GCC TGG GGC CGC CCC CAG AGG GAT GGC CCT CCC   554
val asp leu gln gln phe thr gly ala trp gly arg pro gln arg asp gly pro pro (164)

CTT CGC GCC ATT GCC CTG ATC GGC TAT GGC AAG TGG CTC GTC AGC AAC GGT TAT GCT   611
leu arg ala ile ala leu ile gly tyr gly lys trp leu val ser asn gly tyr ala (183)

GAT ACG GCC AAG AGC ATC ATC TGG CCC ATT GTG AAG AAC GAC CTT GCC TAC ACT GCC   668
asp thr ala lys ser ile ile trp pro ile val lys asn asp leu ala tyr thr ala (202)

CAG TAC TGG AAC AAC ACT GGC TTC GAT CTC TGG GAG GAG GTT AAC AGC TCT TCT TTC   725
gln tyr trp asn asn thr gly phe asp leu trp glu glu val asn ser ser ser phe (221)

TTC ACC ATC GCC GCC TCC CAC CGT GCT CTC GTT GAG GGT TCT GCT TTT GCC AAG TCC   782
phe thr ile ala ala ser his arg ala leu val glu gly ser ala phe ala lys ser (240)

GTC GGC AGC TCT TGC AGC GCT TGC GAT GCC ATT GCC CCC CAA ATT CTG TGC TTC CAG   839
val gly ser ser cys ser ala cys asp ala ile ala pro gln ile leu cys phe gln (259)

CAG AGC TTC TGG TCC AAC AGC GGC TAC ATC ATC TCC AAC TTT GTC AAC TAC CGC AGC   896
gln ser phe trp ser asn ser gly tyr ile ile ser asn phe val asn tyr arg ser (278)

GGC AAG GAC ATC AAC TCC GTC TTG ACT TCC ATC CAC AAC TTC GAC CCC GCT GCC GGT   953
gly lys asp ile asn ser val leu thr ser ile his asn phe asp pro ala ala gly (297)

TGC GAT GTC AAC ACC TTC CAG CCC TGC AGC GAC CGG GCT CTT GCC AAC CAC AAG GTT  1010
cys asp val asn thr phe gln pro cys ser asp arg ala leu ala asn his lys val (316)

GTC GTT GAC TCC ATG CGC TTC TGG GGT GTC AAC TCC GGT CGC ACT GCC GGT AAG GCC  1067
val val asp ser met arg phe trp gly val asn ser gly arg thr ala gly lys ala (335)

GCC GCT GTC GGT CGC TAC GCT GAG GAT GTC TAC TAC AAC GGT AAC CCG TGG TAC CTC  1124
ala ala val gly arg tyr ala glu asp val tyr tyr asn gly asn pro trp tyr leu (354)

GCT ACT CTC GCC GCC GCC GAG CAG CTC TAC GAC GCC GTC TAC GTC TGG AAG AAG CAG  1181
ala thr leu ala ala ala glu gln leu tyr asp ala val tyr val trp lys lys gln (373)

GGT TCT ATC ACT GTC ACC TCC ACC TCC CTC GCC TTC TTC AAG GAC CTC GTT CCC TCC  1238
gly ser ile thr val thr ser thr ser leu ala phe phe lys asp leu val pro ser (392)

GTC AGC ACC GGC ACC TAC TCC AGC TCT TCC TCC ACC TAC ACC GCC ATC ATC AAC GCC  1295
val ser thr gly thr tyr ser ser ser ser ser thr tyr thr ala ile ile asn ala (411)
```

**Fig. 2.**

```
GTC ACC ACC TAT GCC GAC GGC TTC GTC GAC ATC GTT GCC CAG TAC ACT CCC TCC GAC  1352
val thr thr tyr ala asp gly phe val asp ile val ala gln tyr thr pro ser asp (430)

GGC TCC CTG GCC GAG CAG TTC GAC AAG GAT TCG GGC GCC CCC CTC AGC GCC ACC CAC  1409
gly ser leu ala glu gln phe asp lys asp ser gly ala pro leu ser ala thr his (449)

CTG ACC TGG TCG TAC GCC TCC TTC CTT TCC GCC GCC GCC CGC CGC GCC GGC ATC GTC  1456
leu thr trp ser tyr ala ser phe leu ser ala ala ala arg arg ala gly ile val (468)

CCT CCC TCG TGG GGC GCC GCG TCC GCC AAC TCT CTG CCC GGT TCC TGC TCC GCC TCC  1523
pro pro ser trp gly ala ala ser ala asn ser leu pro gly ser cys ser ala ser (487)

ACC GTC GCC GGT TCA TAC GCC ACC GCG ACT GCC ACC TCC TTT CCC GCC AAC CTC ACG  1580
thr val ala gly ser tyr ala thr ala thr ala thr ser phe pro ala asn leu thr (506)

CCC GCC AGC ACC ACC GTC ACC CCT CCC ACG CAG ACC GGC TGC GCC GCC GAC CAC GAG  1637
pro ala ser thr thr val thr pro pro thr gln thr gly cys ala ala asp his glu (525)

GTT TTG GTA ACT TTC AAC GAA AAG GTC ACC ACC AGC TAT GGT CAG ACG GTC AAG GTC  1694
val leu val thr phe asn glu lys val thr thr ser tyr gly gln thr val lys val (544)

GTC GGC AGC ATC GCT CGG CTC GGC AAC TGG GCC CCC GCC AGC GGG CTC ACC CTG TCG  1751
val gly ser ile ala arg leu gly asn trp ala pro ala ser gly leu thr leu ser (563)

GCC AAA CAG TAC TCT TCC AGC AAC CCG CTC TGG TCC ACC ACT ATT GCG CTG CCC CAG  1808
ala lys gln tyr ser ser ser asn pro leu trp ser thr thr ile ala leu pro gln (582)

GGC ACC TCG TTC AAG TAC AAG TAT GTC GTC GTC AAC TCG GAT GGG TCC GTC AAG TGG  1865
gly thr ser phe lys tyr lys tyr val val val asn ser asp gly ser val lys trp (601)

GAG AAC GAT CCT GAC CGC AGC TAT GCT GTT GGG ACG GAC TGC GCC TCT ACT GCG ACT  1922
glu asn asp pro asp arg ser tyr ala val gly thr asp cys ala ser thr ala thr (620)

CTT GAT GAT ACG TGG AGG TAA       atcgcttgc ttcgtactag gtagtaagta gtgattggga 1982
leu asp asp thr trp arg ***                                                (626)

aaaggaaatg agagaacggg aacgggaacg ggaacgggaa tttgtgatta caaagtgtaa aattaataggg 2052
cccgggattt tggttagatg cataaggggg gcaggggggg ctaggaaacg gaaggttgca tatcaaccga 2122
ggaagaatgg gaagaaaggg aagaaagaca gaaagaagga acaacaggac ttcattctct cacatcgaca 2192
tgagctacct gggcatcagc tacctgggca tcttgatttc ctttttagaa gattgttttg tatccttttt 2262
tcttcctccc ttttctttc ttgtccgtct cttacaccta cctattttta gccaaagtcc acacacacac 2332
aaactttttg ttagatattc tctgtatcaa aattgacaag tttcaatgtt atacagtacc ttgccaagtt 2402
taatacacat tcaaatcaat caaccacaca cacacaagtt ttattgtgca gaaatggagt gaagaagaaa 2472
catgtttggg attatgatga caagcttctc aacaaaattt caacgagtta agcttcaaag gtccgctggc 2542
tcaatggcag agcgtctgac tacgaatcag gaggttccag gttcgacccc tgggtggatc gagttgcaaa 2612
ttggtacttt gagtaccaaa gttccttttt tttttcgtt tggctctctg cttttcgaca gttcactgag 2682
tcatgtgcaa gacacccctg atcgggtacg tactgaactg cttttggtgc agtgcaatgg ttctcgagtg 2752
caagggatga aaggaagata tgtcttgg                                             2780
```

**Fig. 2.** Sequence of 3871 bp of the 5-kb *Cla*I insert in pPS8. The *gla-1* reading frame is shown in *upper case*, together with its translation below. Untranslated regions are shown in *lower case*. The numbering for the nucleotides is based on the A of the ATG being +1. The numbering for the amino acids are in brackets. The putative promoter elements are *underlined in bold*. The functional domains of the intron are in *bold*. The leader sequence of the protein is in *bold*, with the signal splice shown as an *arrow*. The Lys-Arg (Kex2) propeptide processing sites are *underlined*. The putative polyadenylation signal is also *underlined in bold*. The EMBL accession number of this sequence is X67291

### *Initiation of translation*

The initiation points of translation have been shown to have a very strong requirement for a purine at position −3 with respect to the initiating AUG (Cavener and Ray 1991). The *gla-1* gene has a G at this position relative to the putative start, which is consistent with the general pattern for eukaryotes. The surrounding sequence, TCGCCAUGC, is close to the consensus sequence for eukaryotes CC(A/G)CCAUGG (Kozak 1984).

### *Termination*

The 3′ end of the transcript has not yet been mapped. There is no consensus AATAAA polyadenylation sequence in the 837 bp which have been sequenced downstream from the TAA stop codon. Gurr et al. 1987 showed that this sequence was not a necessary feature, but does appear in several filamentous fungal genes. However, there is an AT-rich region (TAAAATTAATA) approximately 100 bp downstream from the stop codon, which may act as a polyadenylation signal.

### *Comparison of the glucoamylase amino-acid sequence with other amylases*

The deduced protein sequence shows most homology with the glucoamylases of *H. grisea* var. *thermoidea* (63%), *A. oryzae* (62%), *A. niger* (54%), *A. shirousami* (54%), *A. awamori* (54%), and *R. oryzae* (21%), with lower homology for the yeast glucoamylases. The *Neurospora* amino-acid sequence was aligned with the *H. grisea*, *A. oryzae* and *A. niger* sequences (Fig. 3). The catalytic region includes the $Trp_{155}$ of the WGRPQ region (Fig. 3), and has been shown to be essential for enzyme activity (Sierks et al. 1989). The carboxylic-acid residues of $Asp_{211}$, $Glu_{214}$ and $Glu_{215}$ in the sequence DLWEEV, have been shown by Svensson et al. (1990) to participate in catalysis and substrate binding. These residues are conserved in the *Neurospora* sequence and serve to define the catalytic domain.

Data reviewed by Svensson et al. (1989) showed that the putative raw starch-binding domain of the *Aspergillus* glucoamylases was contained within the C-terminal end of the protein. Figure 3 shows that there is a high level of homology between the *Neurospora* glucoamylase and the

```
                                       ↓             P1 -------->
N. crassa   1-62     MHLVSSLLVVGAAFQAVLGLPDPLHEKRHSDIIKRSVDSYIQTETPIAQKNLLCNIGASGCR
A. niger    1-52     M-SFRSLLALSGLVCTGLAN--VISK-------RATLDSWLSNEATVARTAILNNIGADGAW
A. oryzae   1-55     MVSFSSCLRALALGSSVLAVQPVLRQ-------ATGLDTWLSTEANFSRQAILNNIGADGQS
H. grisea   1-58     MHTFSKLLVLGSAVQSALGRPHGSSRLQE----RAAVDTFINTEKPIAWNKLLANIGPNGKA

N. crassa   63-123   ASGAASGVVVASPSKSSPDYWYTWTRDAALVTKLIVDEFTNDYNTTLQNTIQAYAAA-QAKL
A. niger    53-112   VSGADSGIVVASPSTDNPDYFYTWTRDSGLVLKTLVDLFRNG-DTSLLSTIENYISA-QAIV
A. oryzae   56-115   AQGASPGVVIASPSKSDPDYFYTWTRDSGLVMKTLVDLFRGG-DADLLPIIEEFISS-QARI
H. grisea   59-109   APGAAAGVIIASPSRTDPPCTWWHGMDPRDYFFTWTPDAALVLTGIIESLGHNYNTTLQQVS

                                        <--------- P2
N. crassa   124-185  QGVSNPSGSLSNGAGLGEPKFMVDLQQFTGAWGRPQRDGPPLRAIALIGYGKWLVSNGYADT
A. niger    113-174  QGISNPSGDLSSGAGLGEPKFNVDETAYTGSWGRPQRDGPALRATAMIGFGQWLLDNGYTST
A. oryzae   116-176  QGISNPSGALSSG-GLGEPKFNVDETAFTGAWGRPQRDGPALRATAMISFGEWLVENSHTSI
H. grisea   110-167  NPSGTFADGSGLGEALGEAKFNVDLTAFTGEWGRPQRDGPPLRAIALIQYAKWLIANGYS-T
                                                    ^

                                       P3 -------->
N. crassa   186-247  AKSIIWPIVKNDLAYTAQYWNNTGFDLWEEVNSSSFFTIAASHRALVEGSAFAKSVGSSCSA
A. niger    175-236  ATDIVWPLVRNDLSYVAQYWNQTGYDLWEEVNGSSFFTIAVQHRALVEGSAFATAVGSSCSW
A. oryzae   177-238  ATDLVWPVVRNDLSYVAQYWSQSGFDLWEEVQGTSFFTVAVSHRALVEGSSFAKTVGSSCPY
H. grisea   168-229  AKSVVWPVVKNDLAYTAQYWNETGFDLWEEVPGSSFFTIASSHRALTEGAYLAAQLDTECPP
                                              ^  ^^

N. crassa   248-309  CDAIAPQILCFQQSFWSNSGYIISNFVNYRSGKDINSVLTSIHNFDPAAGCDVNTFQPCSDR
A. niger    237-297  CDSQAPEILCYLQSFWTGS-FILANFDSSRSGKDANTLLGSIHTFDPEAACDDSTFQPCSPR
A. oryzae   239-299  CDSQAPQVRCYLQSFWTGS-YIQANFGGGRSGKDINTVLGSIHTFDPQATCDDATFQPCSAR
H. grisea   230-294  CTTVAPQVLCFQQAFWNSKGNY-STAGEYRSGKDANSILASIHNFDPEAGCDNLTFQPCSER

                                                          <--------- P4
N. crassa   310-370  ALANHKVVVDSMR-FWGVNSGRTAGKAAAVGRYAEDVYYNGNPWYLATLAAAEQLYDAVYVW
A. niger    298-359  ALANHKEVVDSFRSIYTLNDGLSDSEAVAVGRYPEDTYYNGNPWFLCTLAAAEQLYDALYQW
A. oryzae   300-361  ALANHKVVTDSFRSIYAINSGRAENQAVAVGRYPEDSYYNGNPWFLTTLAAAEQLYDALYQW
H. grisea   295-356  ALANHKAYVDSFRNLYAINKGIAQGKAVAVGRYSEDVYYNGNPWYLANFAAAEQLYDAIYVW

N. crassa   371-432  KKQGSITVTSTSLAFFKDLVPSVSTGTYSSSSSTYTAIINAVTTYADGFVDIVAQYTPSDGS
A. niger    360-421  DKQGSLEVTDVSLDFFKALYSDAATGTYSSSSSTYSSIVDAVKTFADGFVSIVETHAASNGS
A. oryzae   362-423  DKIGSLAITDVSLPFFKALYSSAATGTYASSTTVYKDIVSAVKAYADGYVQIVQTYAASTGS
H. grisea   357-418  NKQGSITVTSVSLPFFRDLVSSVSTGTYSKSSSTFTNIVNAVKAYADGFIEVAAKYTPSNGA

N. crassa   433-494  LAEQFDKDSGAPLSATHLTWSYASFLSAAARRAGIVPPSWGAASANSLPGSCSASTVAGSYA
A. niger    422-483  MSEQYDKSDGEQLSARDLTWSYAALLTANNRRNSVVPASWGETSASSVPGTCAATSAIGTYS
A. oryzae   424-485  MAEQYTKTDGSQTSARDLTWSYAALLTANNRRNAVVPAPWGETAATSIPSACSTTSASGTYS
H. grisea   419-481  LAEQYDRNTGKPDSAADLTWSYSAFLSAIDRRAGLVPPSWRASVAKSLPSTCSRIEVAGTYV

N. crassa   495-532  TATATSFPA----------------------NLTPASTTVTPPTQ--TGCAADHEVLVTFNE
A. niger    484-545  SVTVTSWPSIVATGGTTTTATPTGSGSVTSTSKTTATASKTSTSTSSTSCTTPTAVAVTFDL
A. oryzae   486-518  SVVITSWPTISGYPGA-----------------------------PDSPCQVPTTVSVTFAV
H. grisea   482-521  AATSTSFPS----------------------KQTPNPSAAPSPSPYPTACADASEVYVTFNE

N. crassa   533-593  KVTTSYGQTVKVVGSIARLGNWAPASGLTLSAKQYSSSNPLWSTTIAL-PQGTSFKYKYVVV
A. niger    546-606  TATTTYGENIYLVGSISQLGDWETSDGIALSADKYTSSDPLWYVTVTL-PAGESFEYKFIRI
A. oryzae   519-579  KATTVYGESIKIVGSISQLGSWNPSSATALNADSYTTDNPLWTGTINL-PAGQSFEYKFIRV
H. grisea   522-583  RVSTAWGETIKVVGNVPALGNWDTSKAVTLSASGYKSNDPLWSITVPIKATGSAVQYKYIKV
                        *  *      *    ** *                 * *               *

N. crassa   594-626  NSDGSVKWENDPDRSYAVGTDCAS----TATLDDTWR
A. niger    607-640  ESDDSVEWESDPNREYTVPQACGTS---TATVTDTWR
A. oryzae   580-612  Q-NGAVTWESDPNRKYTVPSTCGVK---SAVQSDVWR
H. grisea   584-620  GTNGKITWESDPNRSITLQTASSAGKCAAQTVNDSWR
                            *    *
```

**Fig. 3.** An alignment of the amino-acid sequences of the glucoamylase of *A. niger*, *A. oryzae*, *H. grisea* var. *thermoidea* and *N. crassa*. The *numbers* refer to the appropriate amino acid at the ends of each line. The amino acids which are identical or similarly conserved are in *bold*. The putative leader sequences are *underlined*. The putative signal sequence splice sites are marked by a *vertical arrow* above the aligned sequences. The amino acids involved in catalysis and substrate binding are in *bold with ^ underneath* the particular columns (Sierks et al. 1989), and the amino acids which are invariant in all starch-binding domains are in *bold with * underneath* the relevant columns (Svensson et al. 1989). The consensus regions from where the four PCR primers were designed are shown as *horizontal arrows* above the aligned sequences, indicating their length and direction

residues 538–640 of the *A. niger* glucoamylase (numbers refer to the unprocessed *A. niger* glucoamylase). Eleven conserved amino acids in the starch-binding domains of other glucoamylases are also conserved in the same region of the *Neurospora* protein, and are shown in Fig. 3.

The alignment in Fig. 3 clearly demonstrates that the *Neurospora* glucoamylase has the same overall domain structure as the *Aspergillus* and *Humicola* proteins, i.e., an N-terminal catalytic region, followed by a C-terminal starch-binding domain and not that of the *R. oryzae* protein, which has the starch-binding domain at the N-terminal end, followed by the catalytic region (Jespersen et al. 1991).

### *Signal peptide*

Comparison of the deduced sequence with the N-terminal sequence of the purified protein (Koh-Luar et al. 1989) revealed the presence of a leader peptide of 35 amino acids. The length of this peptide suggests that it is not only a signal peptide, but probably contains a pro-region as well. In order to estimate the position of the signal cleavage site, the matrix supplied in von Heijne (1986) was used. This predicted two dipeptides which gave equivalent scores: $Gln_{15}$ ↓ $Ala_{16}$ and $Gly_{19}$ ↓ $Leu_{20}$. However, the scores for the important $-3$ and $-1$ positions of the latter sequence are much higher and this cleavage would result in a signal peptide of 19 aa similar to those of the secreted glucoamylases of other filamentous fungi (see alignment in Fig. 3). The overall structure of this *Neurospora* signal sequence corresponds well with the model reviewed by von Heijne (1990), i.e., an amino-terminal positively charged region (n-region) $His_2$, followed by a central hydrophobic region (h-region) of 17 aa (12 of which are hydrophobic), followed by a more polar carboxy-terminal region (c-region) that contains the potential signal peptide cleavage site.

This signal sequence is only the second such sequence of *Neurospora* which has been published, the other being that of the extracellular laccase. This laccase has a 21 amino-acid signal sequence (which has a similar structure to that of the glucoamylase), and a putative propeptide of 27 amino acids (Germann et al. 1988).

### *Propeptide*

Cleavage at the putative signal peptide would generate a propeptide of 16 amino acids. Its function may be to assist the folding of the protein (Winther and Sorensen 1991). The propeptide processing site, by comparison with the exported protein, is on the C-terminal side of the $Lys_{34}$-$Arg_{35}$ dipeptide. This dibasic site is the same propeptide cleavage site seen in the *A. niger* (Boel et al. 1984) and *Saccharomyces fibuligera* (Itoh et al. 1987) glucoamylases. The site itself was first identified in the processing site of the *S. cerevisiae* pre-pro-α-factor which is cleaved by the protein encoded by the *kex2* gene. The Kex2 protein is a membrane-bound endopeptidase which specifically cleaves on the carboxyl side of pairs of basic residues that contain arginine, i.e., -Lys-Arg-↓ and -Arg-Arg-↓ (Julius et al. 1984). This type of processing is common in filamentous fungi, and has been exploited in the production of heterologous proteins in *A. nidulans* by using the Kex2 site as a linker between two fused proteins (Contreras et al. 1991).

Examination of the distribution of amino acids in this propeptide region, reveals a high proportion of charged amino acids ($Asp_{22}$, $His_{25}$, $Glu_{26}$, $Lys_{27}$, $Arg_{28}$, $His_{29}$, $Asp_{31}$, $Lys_{34}$, $Arg_{35}$) plus five turn-promoting amino acids ($Pro_{21}$, $Asp_{22}$, $Pro_{23}$, $Ser_{30}$, $Asp_{31}$) (Levitt 1978). The length of the propeptide (16 aa) is much longer than the corresponding propeptide in *A. niger* (6 aa). The reasons for this extended length and the biased amino-acid distribution are as yet unknown. In addition to the Lys-Arg dipeptide at the C-terminus of the propeptide there is also an internal dibasic pair ($Lys_{27}$-$Arg_{28}$). It is therefore possible that cleavage occurs at both Lys-Arg pairs which would yield fragments of 9 and 7 aa. A similar arrangement has been observed in the gene encoding the killer toxin α-subunit of *Kluyveromyces lactis* (Stark and Boyd 1986).

Thus, the promoter and signal sequences of the *gla-1* gene of *Neurospora* can now be used as an alternative for the design of an expression/export cassette for the study of heterologous expression in this fungus.

*Acknowledgements.* P. J. S. is an SERC-CASE research student (Grant number 89504101). We are grateful to Dr. S. J. Gurr for helpful advice and comment. We would like to thank Mr. H. Spence, Wellcome, for synthesising some of the oligonucleotides used in this work.

## References

Archer DB, Jeenes DJ, MacKenzie DA, Brightwell G, Lambert N, Lowe G, Radford SE, Dobson CM (1990) Bio/Technology 8:741–745
Ashikari T, Nakamura N, Tanaka Y, Kiuchi N, Shibano Y, Tanaka T, Amachi T, Yosizumi H (1986) Agric Biol Chem 50:957–964
Azevedo MD, Felipe MSS, Astolfifilho S, Radford A (1990) J Gen Microbiol 136:2569–2576
Bennett JW (1985) Molds, manufacturing and molecular genetics. In: Timberlake WE (ed) Molecular genetics of filamentous fungi. A. R. Liss, New York, pp 345–367
Boel H, Hjort I, Svensson B, Noriss F, Norris KE, Fiil NP (1984) EMBO J 3:1097–1102
Breathnach R, Chambon P (1981) Annu Rev Biochem 50:349–383
Bu'Lock J, Kristiansen B (1987) Basic biotechnology. Academic Press, London
Cavener DR, Ray SC (1991) Nucleic Acids Res 19:3185–3192
Contreras R, Carrez D, Kinghorn JR, van den Hondel CAMJJ, Fiers W (1991) Bio/technology 9:378–381
Davis RH, de Serres FJ (1970) Methods Enzymol 27A:79–143
Dohmen JR, Strasser AW, Dahlems UM, Hollenberg CP (1990) Gene 95:111–121
Germann UA, Mueller G, Hunziker PE, Lerch K (1988) J Biol Chem 263:885–896
Gurr SJ, Unkles SE, Kinghorn JR (1987) The structure and organization of nuclear genes of filamentous fungi. In: Kinghorn JR (ed) Gene structure in eukaryotic microbes. IRL Press, Oxford, pp 93–139
Hata Y, Kitamoto K, Gomi K, Kumagi C, Tamura G, Hara S (1991) Agric Biol Chem 55:941–949
Hayashida S, Kuroda K, Ohta K, Kuhara S, Fukuda K, Sakaki Y (1989) Agric Biol Chem 53:923–929
Heijne G von (1986) Nucleic Acids Res 14:4683–4690
Heijne G von (1990) J Membrane Biol 115:195–201
Itoh T, Ohtsuki I, Yamashita I, Fukui S (1987) J Bacteriol 169:4171–4176
Jespersen HM, Macgregor EA, Sierks MR, Svensson B (1991) Biochem J 280:51–55
Julius D, Brake A, Blair L, Kunisawa R, Thorner J (1984) Cell 37:1075–1089
Levitt M (1978) Biochemistry 17:4277–4285
Kennedy JF, Cabalda VM, White CA (1988) Trends Biotechnol 6:184–189
Koh-Luar SI, Parish JH, Bleasby AJ, Pappin DJC, Ainley K, Johansen F-E, McPherson MJ, Radford A (1989) Enzyme Microbiol Technol 11:692–695
Kozak M (1984) Nucleic Acids Res 12:857–872
Nunberg JH, Meade JH, Cole G, Lawyer FC, McCabe P, Schweickart V, Tal R, Wittman VP, Flatgaard JE, Innis MA (1984) Mol Cell Biol 4:2306–2315
Orbach MJ, Porro EB, Yanofsky C (1986) Mol Cell Biol 6:2452–2461
Pardo JM, Ianez E, Zalacain M, Claros MG, Jimenez A (1988) FEBS Lett 239:179–184
Perkins DD (1992) Genetics 130:687–700
Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: A laboratory Manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York
Shibuya I, Gomi K, Iimura Y, Takahashi K, Tamura G, Hara S (1990) Agric Biol Chem 54:1905–1914
Sierks MR, Ford C, Reilly PJ, Svensson B (1989) Prot Eng 2:621–625
Stark MJR, Boyd A (1986) EMBO J 5:1995–2002
Svensson B, Jespersen H, Sierks MR, Macgregor EA (1989) Biochem J 264:309–311

Svensson B, Clarke AJ, Svendsen I, Moller H (1990) Eur J Biochem 188:29–38

Van den Hondel CAMJJ, Punt PJ, Van Gorcom RFM (1991) Heterologous gene expression in filamentous fungi. In: Bennett JW, Lasure LL (eds) More gene manipulation in fungi. Academic Press, San Diego, pp 396–428

Ward M, Wilson LJ, Kodama KH, Rey MW, Berka RM (1990) Bio/Technology 8:435–438

Ward PP, Lo J-Y, Duke M, May GS, Headon DR, Connelly OM (1992) Bio/Technology 10:784–789

Winther JR, Sorensen P (1991) Proc Natl Acad Sci USA 88:9330–9334

Yamashita I, Suzuki K, Fukui S (1985) J Bacteriol 161:567–573

Communicated by H. Bertrand