Identification of putative regulatory signals including the HAP1 binding site in the upstream sequence of the *Aspergillus nidulans* cytochrome *c* gene (*cycA*).

Linda J. Johnson and Rosie E. Bradshaw - Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand.

We speculate that a HAP1-like protein, similar to those which regulate oxygen transcriptional activation of many yeast respiratory genes, will probably also regulate the *A. nidulans* cytochrome c *(cycA)* gene. As part of a study to investigate the significance of a putative HAP1 (Haem Activator Protein) binding site in the regulatory region of the *cycA* gene, routine sequencing revealed an error in the published sequence (Raitt *et al.* 1994 *Mol. Gen. Genet.* **242:** 17-22). Examination of the corrected sequence, including RT-PCR analysis of *cycA* mRNA, showed that an extra intron was present, and that the published translational start site was incorrect. This meant that the putative HAP1-binding site proposed by Raitt *et al.* could not be a regulatory element. However, further sequence analysis of the upstream sequence of the corrected *cycA* gene revealed putative regulatory signals, including possible HAP1 binding sites which are a closer match to recently reported yeast consensus sequences (Ha *et al.* 1996 *Nucl. Acids Res.* **24:** 1453-1459).

During the construction of a reporter vector containing the *A. nidulans cycA* promoter, a sequencing discrepancy was found when compared to the published *cycA* sequence (Raitt *et al.* 1994 *Mol. Gen. Genet.* **242:** 17-22). Confirmation that the published sequence was incorrect was obtained by sequencing a sub-clone of the *cycA* gene from an *A. nidulans* genomic library. These results confirmed that two additional thymidine bases were present in the coding region of the *cycA* gene, bringing into question the validity of the open reading frames published by Raitt *et al.* Indeed, the published translational start codon could not be correct, because it was no longer in the correct reading frame to produce the highly conserved cytochrome *c* protein. Further examination of the *cycA* gene sequence showed that both the sequencing error and the published translational start codon fall within a previously undetected intron region. To determine if an additional (third) intron was present and thus provide further confirmation that the published translational start point was incorrect, RT-PCR analysis was performed on *cycA* mRNA.

Two amplification products of 596 bp and 298 bp, indicative for the splicing of three introns, were produced by RT-PCR using two different pairs of primers (positioned at -266 and 867 nt, and at -266 and 415 nt, respectively, relative to the translational start site at +1). Subsequent sequencing of both these products confirmed that the *cycA* gene contained three introns instead of the published two. Consequently, the published ATG initiation codon fell within the region of the previously undetected intron (Intron I).

We propose a new translational start site, which has the correct reading frame to produce the conserved cytochrome c protein after the splicing of the new intron. This ATG initiation codon is preceded by a strong Kozak sequence for initiation of translation, and provides the first AUG in the mRNA. In addition the new predicted N-terminal region is very similar in sequence to the N-terminal region of the *S. cerevisiae CYC7* gene, and the position of the additional intron (Intron I) is conserved, being found at an identical position to that of an intron in the *Neurospora crassa* cytochrome *c* gene.

Due to the re-location of the translational start codon, the putative HAP1 binding site proposed by Raitt *et* was found to be situated in the coding region of the gene, and hence is not an upstream regulatory element. To determine if alternative HAP1 binding sites and/or other upstream regulatory elements are present, the upstream (5') sequence of the *cycA* gene was examined.

A 2.1 kb *Eco*RI fragment containing the 5' region of the *cycA* gene was identified from an *Aspergillus nidulans* l library (kindly provided by Michael Hynes, University of Melbourne) and an additional 1297 bp of *cycA* sequence was obtained upstream of the coding sequence (Figure 1). Analysis of this region revealed pos·· le consensus sequences for the binding of regulatory proteins.

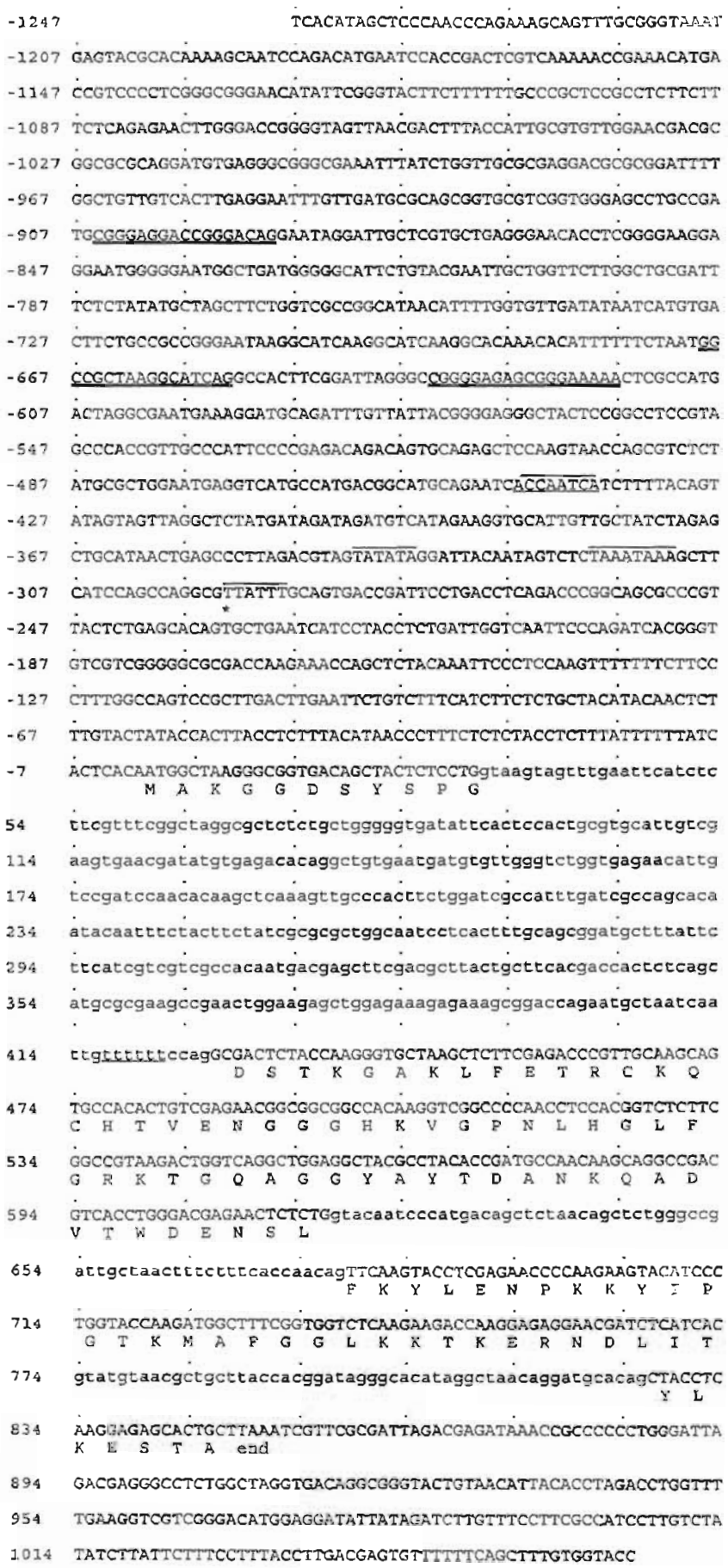
The CCAAT motif, which is a recognition site for the *A. nidulans* AnCF complex (*A. nidulans* CCAAT binding Factor), was found at position -446 nt (Figure 1). This complex is required to set the basal level of *amdS* transcription in *A. nidulans* (Bonnefoy *et al.* 1995 *Mol. Gen. Genet.* **246:** 223-227). If the AnCF complex acts via the CCAAT sequence in the *cycA* promoter, it seems likely that it will probably affect the expression of the *cycA* gene by setting a basal level of transcription.

In addition, three candidate HAP1 binding sites were found in the *cycA* upstream region which were similar to the 'optimal' HAP1 binding site CGG N_3 TA N CGG N_3 TA (Ha *et al.* 1996 *Nucl. Acids Res.* **24:** 1453-1459). These are aligned with the known yeast HAP1 binding sites in Figure 2. The study by Ha *et al.* showed that HAP1 will only bind to direct CGG repeats with a 6-bp spacer. If the CGG repeats are not conserved (ie. are degenerate forms), the TA repeats positioned asymmetrically in the spacer region are then essential for HAP1 binding.

The *A. nidulans* HAP1 site at -669 of the *cycA* sequence is a strong candidate for a HAP1 binding site as it has a direct repeat of GGC and has the TA sequence to stabilise protein binding. The other putative HAP1 site at -634 is also a good candidate since it has a direct repeat of the optimal CGG triplet, but does not have the TA sequence. The sequence at -905 has some features of a HAP1 binding site but is situated 900 bp upstream from the ATG start codon, whereas most regulatory binding sites are usually found closer to the ATG. Thus the -634 HAP1 region is proposed to contain the most likely binding site for HAP1, but the possibility exists that both the -634 and -669 sites, only 35 bp apart, may both be HAP1 sites.

Given these results, we propose that a HAP1-type gene protein will be involved in the oxygen induced transcriptional activation of the *cycA* gene, and that the *A. nidulans* AnCF complex, which is analogous to the yeast HAP2/3/4/5 complex, will regulate the *cycA* gene by setting the basal level of transcription. Examination of the functional significance of the proposed *cycA* regulatory motifs is underway in our laboratory.

Figure 1. (following page) The nucleotide sequence of the *cycA* gene. The nucleotide sequence of the *cycA* gene published by Raitt *et al.* has been presented here in its corrected form, along with the additional upstream sequence (-1247 to -248 nt) obtained from this study. Nucleotides are numbered from the A of the newly proposed translational start codon (+1). The major transcriptional start site (revealed by primer extension) is indicated by an asterisk. Three intervening regions (Introns I, II and III) are displayed in lower case letters. The predicted amino acid sequence is shown below the coding strand. The position of the observed sequencing error within Intron I is underlined. The HAP1 consensus sequences are double underlined, the putative AnCF complex binding site is both over and underlined, and putative TATA motifs are overlined.


-1247 TCACATAGCTCCCAACCCAGAAAGCAGTTTGCGGGTAAAT
-1207 GAGTACGCACAAAAGCAATCCAGACATGAATCCACCGACTCGTCAAAAACCGAAACATGA
-1147 CCGTCCCCTCGGGCGGGAACATATTCGGGTACTTCTTTTTTGCCCGCTCCGCCTCTTCTT
-1087 TCTCAGAGAACTTGGGACCGGGGTAGTTAACGACTTTACCATTGCGTGTTGGAACGACGC
-1027 GGCGCGCAGGATGTGAGGGCGGGCGAAATTTATCTGGTTGCGCGAGGACGCGCGGATTTT
-967 GGCTGTTGTCACTTGAGGAATTTGTTGATGCGCAGCGGTGCGTCGGTGGGAGCCTGCCGA
-907 TGCGGGAGGACCGGGACAGGAATAGGATTGCTCGTGCTGAGGGAACACCTCGGGGAAGGA
-847 GGAATGGGGGAATGGCTGATGGGGGCATTCTGTACGAATTGCTGGTTCTTGGCTGCGATT
-787 TCTCTATATGCTAGCTTCTGGTCGCCGGCATAACATTTTGGTGTTGATATAATCATGTGA
-727 CTTCTGCCGCCGGGAATAAGGCATCAAGGCATCAAGGCACAAACACATTTTTTCTAATGG
-667 CCGCTAAGGCATCAGGCCACTTCGGATTAGGGCCGGGGAGAGCGGGAAAAACTCGCCATG
-607 ACTAGGCGAATGAAAGGATGCAGATTTGTTATTACGGGGAGGGCTACTCCGGCCTCCGTA
-547 GCCCACCGTTGCCCATTCCCCGAGACAGACAGTGCAGAGCTCCAAGTAACCAGCGTCTCT
-487 ATGCGCTGGAATGAGGTCATGCCATGACGGCATGCAGAATCACCAATCATCTTTTACAGT
-427 ATAGTAGTTAGGCTCTATGATAGATAGATGTCATAGAAGGTGCATTGTTGCTATCTAGAG
-367 CTGCATAACTGAGCCCTTAGACGTAGTATATAGGATTACAATAGTCTCTAAATAAAGCTT
-307 CATCCAGCCAGGCGTTATTTGCAGTGACCGATTCCTGACCTCAGACCCGGCAGCGCCCGT
-247 TACTCTGAGCACAGTGCTGAATCATCCTACCTCTGATTGGTCAATTCCCAGATCACGGGT
-187 GTCGTCGGGGGCGCGACCAAGAAACCAGCTCTACAAATTCCCTCCAAGTTTTTTTCTTCC
-127 CTTTGGCCAGTCCGCTTGACTTGAATTCTGTCTTTCATCTTCTCTGCTACATACAACTCT
-67 TTGTACTATACCACTTACCTCTTTACATAACCCTTTCTCTCTACCTCTTTATTTTTTATC
-7 ACTCACAATGGCTAAGGGCGGTGACAGCTACTCTCCTGgtaagtagtttgaattcatctc
M A K G G D S Y S P G
54 ttcgtttcggctaggcgctctctgctgggggtgatattcactccactgcgtgcattgtcg
114 aagtgaacgatatgtgagacacaggctgtgaatgatgtgttgggtctggtgagaacattg
174 tccgatccaacacaagctcaaagttgcccacttctggatcgccatttgatcgccagcaca
234 atacaatttctacttctatcgcgcgctggcaatcctcactttgcagcggatgctttattc
294 ttcatcgtcgtcgccacaatgacgagcttcgacgcttactgcttcacgaccactctcagc
354 atgcgcgaagccgaactggaagagctggagaaagagaaagcggaccagaatgctaatcaa
414 ttgttttttccagGCGACTCTACCAAGGGTGCTAAGCTCTTCGAGACCCGTTGCAAGCAG
D S T K G A K L F E T R C K Q
474 TGCCACACTGTCGAGAACGGCGGCGGCCACAAGGTCGGCCCCAACCTCCACGGTCTCTTC
C H T V E N G G G H K V G P N L H G L F
534 GGCCGTAAGACTGGTCAGGCTGGAGGCTACGCCTACACCGATGCCAACAAGCAGGCCGAC
G R K T G Q A G G Y A Y T D A N K Q A D
594 GTCACCTGGGACGAGAACTCTCTGgtacaatcccatgacagctctaacagctctgggccg
V T W D E N S L
654 attgctaactttctttcaccaacagTTCAAGTACCTCGAGAACCCCAAGAAGTACATCCC
F K Y L E N P K K Y I P
714 TGGTACCAAGATGGCTTTCGGTGGTCTCAAGAAGACCAAGGAGAGGAACGATCTCATCAC
G T K M A F G G L K K T K E R N D L I T
774 gtatgtaacgctgcttaccacggatagggcacataggctaacaggatgcacagCTACCTC
Y L
834 AAGGAGAGCACTGCTTAAATCGTTCGCGATTAGACGAGATAAACCGCCCCCCTGGGATTA
K E S T A end
894 GACGAGGGCCTCTGGCTAGGTGACAGGCGGGTACTGTAACATTACACCTAGACCTGGTTT
954 TGAAGGTCGTCGGGACATGGAGGATATTATAGATCTTGTTTCCTTCGCCATCCTTGTCTA
1014 TATCTTATTCTTTCCTTTACCTTGACGAGTGTTTTTTCAGCTTTGTGGTACC

Known Yeast UASs of HAP1:

CYC1	TGGC CGG GGTTTA CGG ACGATGA
CYC7	CCCT CGC TATTAT CGC TATTAGC
CTT1	GGAA TGG AGATAA CGG AGGTTCT
CYB2	GGCA AGG AGATAT CGG CAGGCTT
CYT1	CCGC CGG AAATAC CGG CCGCCCA
CYT1 (reverse)	CGGC CGG TATTTC CGG CGGCCAA
KlCYC1 (reverse)	ATTT CGG GAACAT CGG TCAAGAC

A. nidulans putative HAP1 UAS:

CYCA (-634)	CCGC CGG GGAGAG CGG GAAAAGG
CYCA (-669)	TAAT GGC CGCTAA GGC ATCAGGC
CYCA (-905)	GATG CGG GAGGAC CGG GACAGGA
OPTIMAL	CGG NNNTAN CGG NNNTA

Genes with HAP1 sites (Ha *et al.* 1996 *Nucleic Acids Res* **24**: 1453-1459):

CYC1: Iso-1-cytochrome *c* from *S. cerevisiae*
CYC7: Iso-2-cytochrome *c* from *S. cerevisiae*
CTT1: Catalase T from *S. cerevisiae*
CYB2: Cytochrome b_2 from *S. cerevisiae*
CYT1: Cytochrome c_1 from *S. cerevisiae*
KlCYC1: Cytochrome *c* from *Kluyveromyces lactis*

Figure 2. Comparison of known yeast and putative *A. nidulans* HAP1 UASs.
The known (functional) HAP1 binding sites from yeast are aligned with the putative HAP1 binding sites from the *A. nidulans cycA* gene. Characters underlined indicate nucleotides which match the conserved CGG triplets or TA repeats given in the 'optimal' HAP1 sequence.