5'-Untranslated sequences of two structural genes in the *qa* gene cluster of *Neurospora crassa*

(*qa* gene coding sequence/nuclease S1 mapping/promoter sequence)

N. KIRBY ALTON*, FRANK BUXTON, VIRGINIA PATEL, NORMAN H. GILES, AND DANIEL VAPNEK*

Department of Molecular and Population Genetics, University of Georgia, Athens, Georgia 30602

Contributed by Norman H. Giles, December 21, 1981

ABSTRACT The coding regions of two genes (*qa-2* and *qa-3*) in the *qa* gene cluster of *Neurospora crassa* have been localized by nucleotide sequence analysis combined with data on previously determined NH₂-terminal amino acid sequences for the proteins that these genes encode. The start point of transcription for each of these genes has been determined by nuclease S1 mapping experiments with poly(A)⁺ RNA isolated from quinic acid-induced cultures of *N. crassa*. The sequences of ≈200 nucleotides 5' to the start point of transcription have been compared with each other and with those of other eukaryotes. The results show that neither of these regions for the *qa-2* nor the *qa-3* genes share any significant homology with sequences apparently conserved in higher eukaryotic promoters (–25 and –70 regions). However, the *qa-2* and *qa-3* sequences do show homology with each other in these regions. Comparison of the 5'-flanking regions of these *Neurospora* genes with those of several *Saccharomyces cerevisiae* genes reveals a number of similarities in the region preceding the translation initiation codons.

In *Neurospora crassa*, the ability to use quinic acid as a sole carbon source is due to the presence of a group of tightly linked genes, the *qa* cluster, located on the right arm of linkage group VII. Three of the genes are structural genes encoding the enzymes necessary for the conversion of quinic acid to protocatechuic acid. These genes and the enzymes they encode are *qa-2*, catabolic dehydroquinase (3-dehydroquinase hydro-lyase, EC 4.2.1.10); *qa-3*, quinate (shikimate) dehydrogenase (quin-ate:NAD⁺ 3-oxidoreductase, EC 1.1.1.24); and *qa-4*, dehydroshikimate dehydratase. A fourth gene, *qa-1*, is a regulatory gene encoding a protein that, when combined with the inducer quinic acid, exerts positive control over expression of the three structural genes. The order of the four genes has been established as *qa-1*, *qa-3*, *qa-4*, and *qa-2* (1).

Previously, we reported the molecular cloning on recombinant plasmids and functional expression in *Escherichia coli* of the structural gene for catabolic dehydroquinase (*qa-2*) (2). These plasmids were selected in *E. coli* by their ability to complement an *aroD6* auxotroph. One of these plasmids, pVK88, contained a 7.2-kilobase (kb) *N. crassa* DNA fragment cloned in the *Pst* I site of pBR322 (3). By using this plasmid, an efficient transformation system for *Neurospora* was developed (4). This allowed all of the genes of the cluster to be cloned in *E. coli* and identified by retransformation back into *Neurospora* (5). The results of these experiments showed that, in addition to the *qa-2* gene, pVK88 also carried the entire *qa-4* gene and at least part of the *qa-3* gene. The *qa-1* regulatory gene was shown to be >5 kb distal to the *qa-3* gene. None of the *qa* cluster genes other than *qa-2* is functionally expressed in *E. coli* (5).

By using recombinant plasmids carrying individual *qa* genes, it has been possible to demonstrate by hybridization with *Neurospora* poly(A)⁺ RNA (i.e., reverse Southern gel analysis) that each of the *qa* genes is transcribed independently. Furthermore, these experiments showed that regulation in the system occurs at the level of transcription (6).

In this communication, we report the nucleotide sequences of the coding and 5'-untranslated regions of the *qa-2* and *qa-3* genes, together with nuclease S1 mapping experiments that localize the start point of transcription for both of these genes. These 5'-untranslated regions are compared with each other and with those of yeast and higher eukaryotes.

MATERIALS AND METHODS

Strains and Plasmids. The *N. crassa* strains used have been described (5). The *E. coli* strains were SK1572 (F'*aroD6*, *argE3*, *his4*, *hsdR4*) containing plasmid pVK88 (3) and JM101 (a *traD36* derivative of 71-18) [Δ(lac-proAB), *supE*, *thi*, F'*lacI*^qZ ΔM15 *proA*⁺B⁺] (7).

Materials. Reagents were obtained from the following sources: DNA polymerase I (Klenow subfragment) and T4 polynucleotide kinase, New England Nuclear; restriction endonucleases, Bethesda Research Laboratories; [α-³²P]dATP (400 Ci/mmol; 1 Ci = 3.7 × 10¹⁰ becquerels) and [γ-³²P]ATP (3000 Ci/mmol) Amersham; ultrapure urea, Schwarz/Mann; nuclease S1, Sigma; T4 DNA ligase and *Eco*RI were the gift of M. Bittner. All other chemicals were of reagent grade.

DNA Cloning and Sequence Analysis. Plasmid DNA preparation, molecular cloning reactions, and gel electrophoresis of DNA were carried out as described (3, 8). Transformation of *E. coli* strain K-12 was carried out by a modification of the low pH procedure (9). Rapid screening of strains harboring putative recombinant plasmids was carried out using the alkaline procedure of Birnboim and Doly (10). DNA sequence analysis was carried out by either the chain-termination technique (11) as described (12) or the chemical modification technique as described by Maxam and Gilbert (13). Single-stranded templates for use in the chain-termination method were obtained by molecular cloning of subfragments of the region of interest in the single-stranded bacteriophage vectors M13mp2 or M13mp7 as described by Messing *et al.* (14). A universal primer for use with this system was supplied by Roberto Crea (Genentech, San Francisco, CA). For cloning in the *Eco*RI site of either phage, synthetic *Eco*RI "linkers" were added to restriction fragments as described by Goodman and MacDonald (15).

Nuclease S1 Mapping. Mapping of the 5' termini of *N. crassa* mRNAs was carried out by a modification (16) of the original Berk and Sharp (17) method as follows. Thirty micrograms of

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

Abbreviations: kb, kilobase(s); bp, base pair(s).

*Present address: Applied Molecular Genetics, Inc., 1892 Oak Terrace Lane, Newbury Park, CA 91320.

total poly(A)⁺RNA was mixed with 2×10^5 cpm of end-labeled DNA fragment (1×10^6 cpm/pmol of 5' termini) in a final volume of 30 μ l of hybridization buffer (80% formamide/0.04 M Pipes, pH 6.4/0.4 M NaCl/1 mM EDTA). The solution was heated to 90°C to denature the DNA and immediately placed at 55°C to allow hybridization of the DNA and RNA strands. After 17 hr of incubation, the resulting DNA-RNA hybrids were treated with 100 units of nuclease S1 for 30 min at 37°C in 400 μ l of S1 buffer (0.4 M NaCl/0.2 M NaOAc, pH 4.6/2 mM ZnCl₂/2 mM EDTA containing denatured salmon sperm DNA at 20 μ g/ml). After nuclease S1 digestion, the DNA was precipitated with ethanol, dried at reduced pressure, and suspended in 10 μ l of 90% formamide/10 mM EDTA, pH 7.0/0.3% xylene cyanol/0.3% bromophenol blue. A 5- μ l aliquot was subjected to electrophoresis on an 8% polyacrylamide/7 M urea gel (40 cm $\times$ 20 cm $\times$ 0.4 mm) for 1.5 hr at 35 W constant power. The gel was fixed in 10% acetic acid and autoradiographed overnight at room temperature.

RESULTS

Localization of the *qa-2* and *qa-3* Genes. As noted above, the *qa-2*, *qa-4*, and at least part of the *qa-3* gene are contained on a 7.2-kb *Pst* I fragment. A restriction map of this fragment is presented in Fig. 1. The relative location of the *qa-2* gene in this fragment was determined by a series of subcloning experiments. When the *Hind*III/*Bam*HI fragments were subcloned, only the fragment spanning the *Eco*RI site at position 2618 in Fig. 1 complemented an *aroD6* auxotroph. However, when the *Eco*RI/*Hind*III fragments were subcloned, none was capable of complementing the *aroD6* auxotroph (unpublished results). This result showed that the *Eco*RI site at position 2618 is either located within the structural gene or separates the structural gene from its promoter.

The relative location of the *qa-3* gene within this fragment was determined by nuclease S1 mapping and DNA sequence analysis. The exact location of the *qa-4* gene is not known, but genetic mapping data (18) place it between the *qa-2* and *qa-3* genes, as indicated in Fig. 1. This location has been confirmed by transformation experiments (5).

Nucleotide Sequence of the *qa-2* Gene Region. Since cleavage at the *Eco*RI site at position 2618 inactivates the *qa-2* gene, the nucleotide sequence of ≈ 1 kb of DNA around the *Eco*RI site (from the *Hae* III site at position 1983 to the *Bam*HI site at position 2970) was determined by using the chain-termination method of DNA sequence analysis (11). The sequence of the region from the *Hae* III site at position 1983 to just past the *Eco*RI site at position 2618 is presented in Fig. 2. An open translational reading frame beginning with a methionine codon and proceeding through the *Eco*RI site can be predicted from the DNA sequence. The NH₂-terminal amino acid sequence determined from catabolic dehydroquinase isolated from *E. coli* (unpublished results) is identical to amino acid residues 7–14 predicted by the DNA sequence (Fig. 2). In addition, a partial amino acid sequence of the enzyme isolated from *N. crassa* (unpublished results) is identical to amino acid residues 89–106 predicted by the DNA sequence. Both of these partial amino acid sequences are in the same translational frame as the polypeptide predicted from the DNA sequence. There are several possibilities that could explain the different NH₂-terminal amino acid sequences determined for catabolic dehydroquinase isolated from the two organisms. The most likely explanation is differential proteolytic cleavage of the protein during isolation (unpublished). We conclude that the polypeptide shown in Fig. 2 is the first 124 amino acids of *N. crassa* catabolic dehydroquinase.

Nucleotide Sequence of the *qa-3* Gene Region. Nuclease S1 mapping experiments showed that the 5' end of a quinic acid-induced mRNA in *N. crassa* was located 163 base pairs (bp) from the *Sst* I site at position 6432 (Fig. 1). Transcription was predicted to be in the direction indicated in Fig. 1. Based on genetic analysis, it was assumed that this transcript was the mRNA for quinate (shikimate) dehydrogenase (*qa-3*). Accordingly, the DNA sequence of 590 bp around the *Eco*RI site at position 6255 was determined by using the chemical modification technique of Maxam and Gilbert (12). The nucleotide sequence of this region from very close to the *Bgl* II site at position 5935 to position 6525 is presented in Fig. 2. An open translational reading frame beginning with a methionine codon at position 6354 continues through the available sequence. The NH₂-terminal amino acid

FIG. 1. Restriction endonuclease cleavage maps of the 7.2-kb *Pst* I fragment of pVK88. Numbers below the upper line refer to distances in kb pairs. Positions of the *qa-2* and *qa-3* coding regions are indicated. The exact location of the *qa-4* gene between *qa-2* and *qa-3* is not known. Arrows at bottom indicate direction and extent of sequence analysis runs.

QA-2

GGCCTTTTGAATGGTCATCAGTGCAGTGAACGAGCAATGTACGGAAGCGGGGACGGGTAATCGCTTATCGCTCGTGCAGACAACTTCGTCGCTGATTAGAGATGGGAATGATGAGGGAACCGTGATTAAACAA
 CAAAACATAAACACACTTCAATTCAACCTTCTGGCTGTGAGTGTGGGTATAGTGGGGGCGATCTTTGGGAAGCATTCCTCTGTGGGCCATCTCCACAAGCCCATCGCAACCAACAGAGGTACCAACACAA
 met ala ser pro arg his ile leu leu ile asn gly pro asn leu asn leu leu gly thr arg glu pro gln ser thr ala gln ser thr leu his asp ile glu
 ATG GCG TCC CCC CGT CAC ATT CTC CTC ATC AAT GGC CCC AAT CTC AAC CTC CTC GGC ACC CGG GAG OCC CAA TCT ACG GCT CAA TCA ACC CTC CAT GAC ATT GAG
 gln ala ser gln thr leu ala ser ser leu gly leu arg leu thr thr phe gln ser asn his glu gly ala ile ile asp arg ile his gln ala ala gly phe
 CAA GGC TCC CAG ACT CTG GCG TCC TCG CTA GGT CTT CGT CTT ACA ACC TTC CAG TCC AAC CAT GAA GGA GCC ATC ATC GAC CGT ATC CAT CAA GCA GCG GGA TTC
 val pro ser pro pro ser pro ser pro ser ala ala thr thr thr glu ala gly leu gly pro gly asp lys val ser ala ile ile ile asn pro gly ala
 GTC CCG TCT CCA CCG TCA CCG TCG CCG TCA AGT GCC GCA ACC ACG ACG GAG GCA GGA TTG GGT CCC GGA GAC AAA GTG TCG GCC ATC ATC ATT AAC CCC GGC GCT
 tyr thr his thr ser ile gly ile arg asp ala leu leu gly thr gly ile pro phe
 TAT ACG CAC ACG AGT ATA GGC ATC CCG GAC GCG CTT CTG GGG ACA GGA ATT CCG TTT

QA-3

GTTGAAGAGAGCTCTCTCTTAATCACTGAAATCCCGAGCATGCCCAAGGCAAAAGCACAACCTGAATCTTTTCCCTATGACGAGCAAAATCGGACCATAGGCAGCCGCTTTCCCAAAATTCGCCGAGG
 ACATTCGAGGGTTCTCATTCATCTCTCTCCACGAAGTGTCTTGTACGATGGAGATGCTTAGTGAATGAACCTGTCAACTCCAGCGCCCATGTAGTAATGAAAATGGGGGAATACATTATAGCCACGCTTATG
 CTCTCTCTCCGAGTTAGACGATCTCGGGAATTCCTTAGGTCTCTCTATTTTCATTCGCGTCTCTGTGCAATCTTGATTTTGGAGTACTGTGACTTCTCATAGCCAGATACACCACACAATCAAGCATATATCACC
 met ser thr ala thr thr thr thr ser ala thr thr thr met ser val val gln pro arg gln gln arg ala his leu thr ser thr pro asp ile thr pro tyr
 ATG TCG ACA GCA ACC ACC ACA ACA TCA GCG ACG ACG ACG ATG TCG GTC GTC CAG CCC CGA CAG CAA AGA GCT CAC CTC ACC AGC ACA CCC GAC ATC ACC CCC TAC
 thr arg his gly tyr leu phe gly gln asp gly pro ser pro pro leu his arg leu thr pro thr
 ACC AGA CAT GGC TAT CTC TTC GCG CAG GAT GGC CCC TCT CTT CCA CTC CAT CCG TTA ACT CCC ACC TC

FIG. 2. Partial nucleotide sequences of the *qa-2* and *qa-3* genes. The *qa-2* sequence shown is from the *Hae* III cleavage site at position 1983 to slightly beyond the *Eco*RI site at position 2618. The first 124 amino acids of catabolic dehydroquinase are indicated above their respective codons. Amino acids underlined (7–11 and 89–106) were identical with partial amino acid sequences determined from catabolic dehydroquinase isolated from *E. coli* and *N. crassa*, respectively. The *qa-3* sequence shown is from close to the *Bgl* II site at position 5935 to position 6525. The first 57 amino acids of quinate (shikimate) dehydrogenase are indicated above their respective codons. Amino acids underlined were essentially identical with partial amino acid sequences determined from quinate dehydrogenase isolated from *N. crassa*. For each gene, the 5'-terminal nucleotide of the predominant transcript is indicated by a dot.

sequence of quinate dehydrogenase isolated from *N. crassa* has been determined (19). The amino acid sequence predicted from the DNA sequence of residues 26–43 agrees with the published NH₂-terminal amino acid sequence with five exceptions. These include substitutions of a proline for asparagine (residue 34), an arginine for proline (residue 37), and a histidine for tyrosine (residue 38). The threonine and serine residues predicted by the DNA sequence at amino acid positions 27 and 28 were not detected in the protein sequence analysis. In addition to the methionine codon beginning at position 6354, a second in-phase methionine codon occurs beginning at position 6393 (amino acid 13). Both of these are potential start codons for the quinate dehydrogenase protein. However, if *Neurospora* follows the general rule in eukaryotes of initiating translation at the first AUG from the 5' end of the mRNA (20), the ATG beginning at position 6354 would correspond to the initiation codon of the protein. Based on this analysis, we conclude that the amino acid sequence presented in Fig. 2 is the first 57 amino acids of quinate dehydrogenase.

Determination of the 5' Termini of the *qa-2* and *qa-3* mRNAs. The *qa-2* DNA sequence contains a *Sma* I restriction endonuclease cleavage site in the region corresponding to the NH₂-terminal portion of the *qa-2* structural gene and a *Kpn* I restriction endonuclease cleavage site 73 bp upstream from the *Sma* I site (Fig. 1). The 5' terminus of the *qa-2* mRNA isolated from *N. crassa* was determined by a method similar to that described by Berk and Sharp (17). Restriction fragments uniquely labeled with [γ -³²P]ATP at the 5' end of either the *Sma* I or the *Kpn* I cleavage sites were hybridized to total *N. crassa* poly(A)⁺RNA isolated from strains that had been induced with quinic acid. Hybridization conditions used favored DNA-RNA hybridization over DNA-DNA hybridization. Treatment of the resulting hybrids with the single-strand-specific nuclease S1 generates duplex DNA-RNA molecules devoid of single-strand tails (17). When these hybrid molecules are denatured and sub-

jected to electrophoresis on an 8% polyacrylamide/7 M urea gel, specific radioactive oligonucleotides should appear at a position in the gel corresponding to the distance from the unique radioactive label to the 5' end of the mRNA.

The oligonucleotides resulting from nuclease S1 digestion of a duplex formed between a DNA fragment uniquely labeled at the *Sma* I site and poly(A)⁺RNA isolated from *N. crassa* induced with quinic acid are shown in Fig. 3A, lane c. Although a number of bands are visible, the major band is 146 $\pm$ 2 bases upstream from the *Sma* I site. The radioactive oligonucleotide resulting from nuclease S1 digestion of a duplex formed between a DNA fragment uniquely labeled at the *Kpn* I site and poly(A)⁺RNA from *N. crassa* should be 73 $\pm$ 2 bases long because the *Kpn* I site is 73 bases before the *Sma* I site. A major band of 72 bases was observed (data not shown), confirming the position of the first nucleotide of the predominant transcript of the quinate dehydrogenase mRNA. This nucleotide is numbered +1 in Fig. 4A. Whether the other bands observed are minor transcripts from the region or artifacts of the method used cannot be determined from these experiments.

As mentioned above, the *qa-3* structural gene was localized by DNA sequence analysis after mapping the 5' terminus of the *qa-3* mRNA. The *Sst* I site at position 6432 and the *Sal* I site at position 6368 (64 bp before the *Sst* I site) were uniquely labeled at their 5' ends with [γ -³²P]ATP, hybridized to poly(A)⁺RNA isolated from quinic acid-induced *N. crassa*, and treated with nuclease S1. The radioactive oligonucleotide resulting from nuclease S1 digestion of a duplex formed between a DNA fragment uniquely labeled at the *Sst* I site and poly(A)⁺RNA is shown in Fig. 3B, lane b. The major band visible is 163 $\pm$ 2 bases long, which demonstrates that the predominant transcript from this region in *N. crassa* begins 163 $\pm$ 2 bases before the *Sst* I site. That this nucleotide is the first nucleotide of the predominant transcript of the *qa-3* mRNA was confirmed by labeling at the *Sal* I site. The radioactive oligo-

FIG. 3. Nuclease S1 mapping of the 5' termini of the *qa-2* and *qa-3* mRNAs. (A) *qa-2* mRNA. Lanes: a, ϕ X174 *Hae* III fragments terminally labeled with 32 P; b, terminally labeled *Sma* I DNA probe incubated in hybridization buffer; c, *Sma* I DNA probe hybridized with *N. crassa* poly(A)⁺RNA and treated with 100 units of nuclease S1; d, *Sma* I DNA probe incubated in hybridization buffer and treated with 100 units of nuclease S1. (B) *qa-3* mRNA. Lanes: a, terminally labeled *Sst* I DNA probe incubated in hybridization buffer and treated with 100 units of nuclease S1; b, *Sst* I DNA probe hybridized with *N. crassa* poly(A)⁺RNA and treated with 100 units of nuclease S1; c, *Sst* I DNA probe incubated in hybridization buffer; d, pBR322 *Hind*II fragments terminally labeled with 32 P. Nuclease S1 digestion products were subjected to electrophoresis on 8% polyacrylamide gels followed by autoradiography.

nucleotide resulting from nuclease S1 digestion of a duplex formed between a fragment uniquely labeled at the 5' end of the *Sal* I site and *N. crassa* poly(A)⁺RNA was 99 ± 2 bases long (data not shown).

DISCUSSION

We have determined the nucleotide sequence of and characterized the regions 5' to the coding sequences of *N. crassa* structural genes. These genes, *qa-2* (catabolic dehydroquinase) and *qa-3* (quinate dehydrogenase), are coordinately regulated at the transcriptional level by the product of the *qa-1* gene. Therefore, comparison of the primary sequences of these 5'-flanking regions could provide an insight into the molecular

mechanism of regulation in this system. Furthermore, comparison of these *Neurospora* regions with comparable regions of other eukaryotes and prokaryotes might prove interesting from an evolutionary perspective.

Several conserved sequences in the region of transcription initiation have been identified in higher eukaryotes (21). One such region is an A/T-rich sequence, the so-called Hogness box, centered 25 bp before the mRNA start point (consensus sequence T-A-T-A-A-T-A). Another region conserved in many higher eukaryotic promoters is located 70–80 bp before the mRNA start point (consensus sequence G-G-py-C-A-A-T-C-T) (22). Comparison of the *qa-2* and *qa-3* sequences in these regions shows no significant homology to the canonical sequences.

The observation that neither the *qa-2* nor the *qa-3* 5'-untranslated sequences share any apparent similarity with higher eukaryotic promoters may reflect the requirement for the product of the *qa-1* regulatory gene to obtain expression of these genes. If the regulatory protein exerts its positive control by binding to a site in the *N. crassa* 5'-untranslated region, then a common binding site should be present in both sequences in the region preceding the start point of transcription. A computer search of ≈ 200 bp before the start point of transcription in both the *qa-2* and *qa-3* 5' regions shows no common dyad symmetries or repeated sequences. This result suggests that the *qa-1* regulatory protein may not recognize a common symmetrical sequence, as has been proposed for *E. coli* regulatory proteins such as the *lac* repressor (23) and the cAMP receptor protein (24). However, it is also possible that the binding site for the *qa-1* product is located upstream to the ≈ 200 bp compared.

If the two *qa* 5' regions are aligned for maximum homology within 4 bp of the mRNA start points (Fig. 4A), then two regions of homology become apparent. Both of these regions correspond in location, but not in primary sequence, to the conserved sequences common in most higher eukaryotic promoters. One region of homology is centered ≈ 25 bp before the mRNA start point (Fig. 4A). In this region, between -20 and -30 , there is 70% homology. Another region of striking homology between the two 5' sequences is centered 80 bp before the start point of transcription. In this region, the two sequences are 52% homologous and there is a marked preference for purine residues (i.e., 80% purine). Since no data are available on comparable *Neurospora* 5' regions, it is not known whether these structural features are common to *Neurospora* or unique to the coordinately regulated *qa* cluster.

In light of the fact that *N. crassa* is a lower eukaryote, it is not surprising that the *qa* 5' regions do not share sequence features common in higher eukaryotes. *A priori*, one might expect the sequence to be similar to that of other lower eukaryotes such as the yeast, *S. cerevisiae*. Comparison of the *qa-2* and *qa-3* 5' regions with several yeast 5' regions does show similarities. If

FIG. 4. Sequence homologies between the 5'-flanking regions of the *qa-2* and *qa-3* genes. (A) *qa-2* and *qa-3* 5'-untranslated regions are aligned for maximum homology within 4 bp of the 5'-terminal nucleotide (indicated by +1) of each mRNA. Numbering is relative to the *qa-2* sequence. (B) *qa-2* and *qa-3* sequences are aligned with respect to the translation initiation codon of each. +1, Adenine initiation codons.

the *qa-2* and *qa-3* sequences are aligned with respect to the ATG translation initiation codon (Fig. 4B), several similarities with each other and with *S. cerevisiae* genes are apparent. The 25 nucleotides preceding the ATG initiation codon in the yeast genes compared [i.e., iso-1-cytochrome *c* (25), iso-2-cytochrome *c* (26), two enolase isozymes (27), two nontandemly repeated glyceraldehyde 3-phosphate dehydrogenases (28), *trp1* (29), and actin (30)] are extremely adenine rich in the strand with the same polarity as the mRNA. The average base composition of this strand in the yeast genes in this region is A₁₄C₆T₄G₁. This region in the *qa-5'* sequences is very A/C rich. The base composition of the *qa-2* strand in this region is A₁₁C₁₀T₁G₃ while that for *qa-3* is A₁₁C₉T₄G₁.

In a number of yeast genes, the sequence C-A-C-A-C-A is present in the 25 nucleotides preceding the initiation codon. The proximity of the sequence to the initiation codons in the yeast genes has prompted suggestions that the sequence has a role in the initiation of translation (29, 31). An identical sequence is present in the *qa-3* gene (C-A-C-A-C-A at position -18 to -23; Fig. 4B). Three of these six nucleotides are conserved in the same position of the *qa-2* sequence (Fig. 4B). As noted above, a second in-phase methionine codon begins at nucleotide 6393. Interestingly, if translation were initiated at this second methionine codon, the above comparison would still be valid because the 25 nucleotides preceding this ATG are also extremely A/C rich (A₉C₁₀T₁G₅) and show ≈50% homology to the corresponding region in the *qa-2* gene.

Other conserved nucleotides in the region preceding the translation initiation codon are the adenine residues at positions -3 and -14. The adenine at -3 is conserved in all the yeast genes examined and in both *Neurospora* sequences, while the adenine at -14 is conserved in all except the yeast actin gene.

The presence of pyrimidine clusters of length ≥4 in the non-transcribed strand preceding the initiation codon by 150 nucleotides has been noted for several yeast genes (26). Examination of the *Neurospora* sequences in this region shows the presence of similar clusters. The *qa-3* sequence has 11 such pyrimidine clusters while the *qa-2* sequence has 5. It has been suggested that the presence of a high content of clustered pyrimidine residues correlates with high gene activity (26). Clearly, it will be necessary to determine transcriptional and translational efficiencies of the *qa* genes before a functional role for these sequences can be determined.

The *qa-2* gene is efficiently expressed in *E. coli*. A sequence, G-A-G-G, complementary to the 3' end of the 16S rRNA of *E. coli* occurs 12–15 nucleotides before the ATG codon of the *qa-2* gene. By analysis of a series of deletion mutations, it has been shown that expression of the *qa-2* gene in *E. coli* is dependent on this sequence (unpublished results). Interestingly, the *qa-3* gene, which is not expressed in *E. coli*, does not possess a comparable region of homology with 16S rRNA. Whether this is the only reason for its lack of expression in *E. coli* remains to be determined.

The results reported here should provide a basis for elucidating the molecular mechanisms of regulation in this eukaryotic system. Analysis of the effects produced by single base-pair mutations and by deletions in the 5'-untranslated sequences described here and, ultimately, the establishment of an *in vitro* transcription system for *N. crassa* should lead to understanding

the regulatory role of the *qa-1* gene product and how the primary sequences described here are involved in the expression of these *qa* genes.

We thank Sonya Leach for excellent technical assistance, Michael Bittner for helpful discussions, and Fred Sherman and Gerald Fink for critical reading of the manuscript. This research was supported in part by National Institutes of Health Grants GM28777 (to N.H.G.) and GM27973 (to D.V.).

- Giles, N. H., Alton, N. K., Case, M. E., Hautala, J. A., Jacobson, J. W., Kushner, S. R., Patel, V. B., Reinert, W. R., Støman, P. & Vapnek, D. (1978) *Stadler Genet. Symp.* **10**, 49–63.
- Vapnek, D., Hautala, J. A., Jacobson, J. W., Giles, N. H. & Kushner, S. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 3508–3512.
- Alton, N. K., Hautala, J. A., Giles, N. H., Kushner, S. R. & Vapnek, D. (1978) *Gene* **4**, 241–259.
- Case, M. E., Schweizer, M., Kushner, S. R. & Giles, N. H. (1979) *Proc. Natl. Acad. Sci. USA* **74**, 5259–5363.
- Schweizer, M., Case, M. E., Dykstra, C. C., Giles, N. H. & Kushner, S. R. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 5086–5090.
- Patel, V. B., Schweizer, M., Dykstra, C. C., Kushner, S. R. & Giles, N. H. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 5783–5787.
- Messing, J., Gronenborn, B., Muller-Hill, B. & Hofschneider, P. H. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 3642–3646.
- Alton, N. K. & Vapnek, D. (1978) *Plasmid* **1**, 388–404.
- Enea, V., Vovis, G. F. & Zinder, N. D. (1975) *J. Mol. Biol.* **96**, 495–509.
- Birnboim, H. C. & Doly, J. (1979) *Nucleic Acids Res.* **7**, 1513–1523.
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
- Alton, N. K. & Vapnek, D. (1979) *Nature (London)* **282**, 864–869.
- Maxam, A. & Gilbert, W. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 560–564.
- Messing, J., Crea, R. & Seeburg, P. H. (1981) *Nucleic Acids Res.* **9**, 309–321.
- Goodman, H. M. & MacDonald, R. J. (1979) *Methods Enzymol.* **68**, 75–90.
- Ingolia, T. D., Craig, E. A. & McCarthy, B. S. (1980) *Cell* **21**, 669–679.
- Berk, A. S. & Sharp, P. A. (1977) *Cell* **2**, 721–732.
- Case, M. E. & Giles, N. H. (1976) *Mol. Gen. Genet.* **147**, 83–89.
- Strøman, P., Reinert, W. R., Case, M. E. & Giles, N. H. (1979) *Genetics* **92**, 67–74.
- Kozak, M. (1981) in *Protein Biosynthesis in Eukaryotes*, ed. Perez-Bercoff, R. (Plenum, New York), in press.
- Corden, J., Wasylyk, B., Buchwalder, A., Sassone-Corsi, P., Keding, C. & Chambon, P. (1980) *Science* **209**, 1406–1414.
- Benoist, C., O'Hare, K., Breathnach, R. & Chambon, P. (1980) *Nucleic Acids Res.* **8**, 127–142.
- Gilbert, W. & Maxam, A. (1973) *Proc. Natl. Acad. Sci. USA* **70**, 3581–3584.
- Majors, J. (1975) *Nature (London)* **256**, 672.
- Smith, M., Leung, D. W., Gillam, S., Astell, C. R., Montgomery, D. L. & Hall, B. D. (1979) *Cell* **16**, 753–761.
- Montgomery, D. L., Leung, D. W., Smight, M., Shalit, P., Faye, G. & Hall, B. D. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 541–545.
- Holland, M. S., Holland, J. P., Thill, G. P. & Jackson, K. A. (1981) *J. Biol. Chem.* **256**, 1385–1395.
- Holland, J. P. & Holland, M. J. (1980) *J. Biol. Chem.* **255**, 2596–2605.
- Tschumper, G. & Carbon, J. (1980) *Gene* **10**, 157–166.
- Ng, R. & Abelson, J. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 3912–3916.
- Stiles, J. I., Szostak, J. W., Young, A. T., Wu, R., Consaul, S. & Sherman, F. (1981) *Cell* **25**, 277–284.