Sequence characteristics within nuclear genes from *Sordaria macrospora*

Stefanie Pöggeler- Ruhr-Universität, 44780 Bochum, Germany

This paper reports sequence features within nuclear genes from *Sordaria macrospora*. Eight nuclear gene sequences were analyzed for codon usage, GC content, intron regulatory sequences and translation initiation sites.

The homothallic ascomycete *Sordaria macrospora* is an excellent model system to study not only meiotic pairing and recombination (Zickler 1977 Chromosoma **61**:29-316) but also fruiting body development (Esser and Straub 1958 Z. Vererbungslehre **89**:729-746). Recently, these studies have been extended to a molecular level (Walz and Kück 1995 Curr. Genet. **29**:88-95) and knowledge about sequence features would be a helpful tool in sequence analysis. Until now, sequence information from *S. macrospora* was only available from a single nuclear gene (LeChevanton and Leblon 1989 Gene **77**:39-49). Here we compile sequence data from eight recently sequenced genes to determine common features of nuclear genes from *S. macrospora*. We provide a consensus sequence for the translation initiation site (Table 1), a codon usage table (Table 2), and consensus sequences for intron regulatory sequences (Table 3). Comparison of the data presented here with sequence features from the well studied ascomycete *Neurospora crassa* (data taken from Brucherez *et al*. 1993 Fungal Genet. Newsl. **40**:85-95; and Edelman and Staben 1994 Exp Mycol **18**:70-81) shows that *S. macrospora* sequence characteristics are very similar to those determined for *N. crassa* genes.

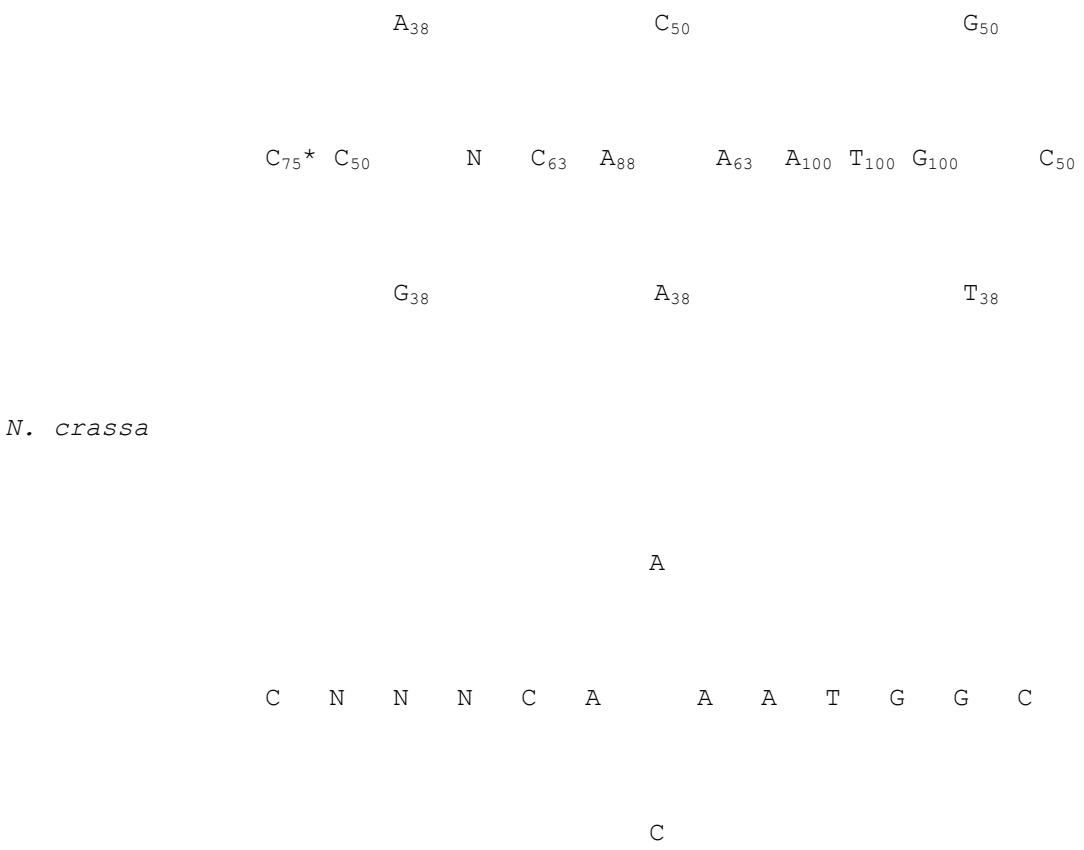
Table 1. Translation initiation context

Gene	Gene product	Translation initiation	Reference[a]
EF1-	EF1- translation elongation factor	CCGTCAAA**ATG**GG	1
tuba	-tubulin	CATACAAA**ATG**CG	2
ura3	orotidine phosphoribosyl transferase	CCGCCACA**ATG**TC	3
ura5	orotidine monophosphate decarboxylase	CCAGCACA**ATG**GC	4
SmtA-1	mating-type protein	GAAGTACG**ATG**TC	5
SmtA-2	mating-type protein	CGACTGAC**ATG**GA	5
SmtA-3	mating-type protein	CTTTCAGC**ATG**TC	6
Smta-1	mating-type protein	TCGAAACA**ATG**GA	5

[a] (1)Gagny, Koll and Silar, unpublished (Accession # X96615) (2) Pöggeler *et al*., submitted (Accession # Z70290) (3) Nouwrousian, unpublished (Accession # Z70291) (4) Nouwrousian, unpublished (5) Pöggeler *et al*., submitted (Accession # Y10616) (6) Pöggeler, unpublished

Consensus translation initiation

S. macrospora



*The subscript number indicates the percentage occurrence of the particular nucleotide.

The *S. macrospora* consensus for initiation of translation shows a high degree of identity to the *N. crassa* translation initiation consensus sequence and, as *N. crassa*, a prevalence of GC following the ATG which means that an alanine (GCN) is found at the amino terminus of most proteins studied so far.

Table 2. Codon usage analysis based upon 2497 codons

TTT-Phe 24(25.0%)[a]	TCT-Ser 22(16.4%)	TAT-Tyr 21(26.9%)	TGT-Cys 3(8.6%)
TTC-Phe 72(75.0%)	TCC-Ser 42(31.3%)	TAC-Tyr 57(73.1%)	TGC-Cys 32(91.4%)
TTA-Leu 3(1.7%)	TCA-Ser 12(9.0%)	TAA-Ter 4(57.1%)	TGA-Ter 1(14.3%)
TTG-Leu 21(11.6%)	TCG-Ser 28(20.9%)	TAG-Ter 2(28.6%)	TGG-Trp 32(100.0%)
CTT-Leu 45(24.9%)	CCT-Pro 41(30.6%)	CAT-His 23(30.7%)	CGT-Arg 34(27.6%)
CTC-Leu 80(44.2%)	CCC-Pro 68(50.7%)	CAC-His 52(69.3%)	CGC-Arg 57(46.3%)

```
CTA-Leu   3(  1.7%)    CCA-Pro  13(  9.7%)    CAA-Gln  24( 23.8%)    CGA-Arg
7(  5.7%)
CTG-Leu  29( 16.0%)    CCG-Pro  12(  9.0%)    CAG-Gln  77( 76.2%)    CGG-Arg
6(  4.9%)
ATT-Ile  47( 32.6%)    ACT-Thr  25( 19.2%)    AAT-Asn  17( 17.3%)    AGT-Ser
5(  3.7%)
ATC-Ile  95( 66.0%)    ACC-Thr  71( 54.6%)    AAC-Asn  81( 82.7%)    AGC-Ser
25( 18.7%)
ATA-Ile   2(  1.4%)    ACA-Thr  15( 11.5%)    AAA-Lys  12(  7.2%)    AGA-Arg
7(  5.7%)
ATG-Met  63(100.0%)    ACG-Thr  19( 14.6%)    AAG-Lys 154( 92.8%)    AGG-Arg
12(  9.8%)
GTT-Val  40( 24.2%)    GCT-Ala  70( 30.7%)    GAT-Asp  61( 42.4%)    GGT-Gly
61( 32.3%)
GTC-Val 101( 61.2%)    GCC-Ala 115( 50.4%)    GAC-Asp  83( 57.6%)    GGC-Gly
94( 49.7%)
GTA-Val   5(  3.0%)    GCA-Ala  21(  9.2%)    GAA-Glu  33( 19.0%)    GGA-Gly
24( 12.7%)
GTG-Val  19( 11.5%)    GCG-Ala  22(  9.6%)    GAG-Glu 141( 81.0%)    GGG-Gly
10(  5.3%)
```

[a]The percent shown by each codon represents the percent of the time that the amino acid is encoded by that codon.

The GC content in a coding region of 7491 nucleotides is 56.7%. For comparison in *N. crassa* the GC content is 58.6% in the coding region (GC content in total DNA 54.1%). In cases where amino acids are represented by more than one codon, *S. macrospora*, as many other organisms, does not use synonym codons equally (Table 2).

In *S. macrospora*, as in *N. crassa*, codons are preferred with a C in the third position and in four codon families the codon ending in T is usually preferred to those ending in A or G. The stop codon TAA is more frequently used than TAG or TGA, respectively. The six least used codons for *S. macrospora* are ATA (Ile), TTA (Leu), CTA (Leu), TGT (Cys), GTA (Val), and AGT (Ser). All of these six codons are belonging to low-usage codons in *N. crassa* as well. As reported by Zhang *et al*. (1991 Gene **105**:61-67) in many organisms, low-usage codons are clearly avoided in abundant proteins and therefore may affect translation rates.

Table 3. Intron regulatory sequences and intron length

Intron	5' Intron Donor	Branch Site	Distance to 3' Splice-Site/nt[b]	3' Intron Acceptor	Intron Length / bp
SmtA-1/1[a]	T^GTAAGT	ACTGATT	-19-	TTCAG^	58
SmtA-1/2[a]	G^GTTAGT	ACTCGTG	-21-	GGCAG^	60
SmtA-2/1[a]	G^GTAACA	ACTGATG	-14-	GCCAG^	57
SmtA-2/2[a]	G^GTGAGT	ACTGACA	-12-	GATAG^	71
SmtA-2/3[a]	T^GTAAGA	ACTAATA	-12-	GACAG^	47
SmtA-2/4[a]	G^GTTTGC	GCTAACA	-16-	GACAG^	55

SmtA-3/1[a]	C^GTGAGT	ACTGACT	-12-	GTTAG^	54
Smta-1/1[a]	A^GTAAGT	ACTGACC	-15-	TTTAG^	53
Smta-1/2[a]	T^GTAGGT	ACTAACC	-12-	CTTAG^	57
tuba/1	G^GTACGT	GCTAACG	-22-	TCTAG^	256
tuba/2	G^GTAGGT	GCTAACC	-15-	ATTAG^	149
tuba/3	G^GTAAGC	GCTAACC	-17-	TACAG^	80
tuba/4	G^GTACAT	GCTTACA	-18-	CACAG^	60
tuba/5	G^GTATGT	ACTAACT	-16-	CTTAG^	64
tuba/6	T^GTAAGT	GCTAACT	-14-	CCTAG^	57
ef1/1	G^GTAATG	GCTAACG	-14-	AACAG^	100
ef1/2	G^GTTAGT	ACTGACT	-15-	AACAG^	243
ef1/3	G^GTATGT	GCTAACT	-17-	AAAAG^	60

[a] positions of intron splice sites inferred from cDNA sequences
[b] distance between the C of the intron branch point and the G of the 3' intron acceptor

Consensus 5' Intron-Donor

S. macrospora	G_{67}^	G_{100}	T_{100}	A_{72}	A_{61}	G_{83}	T_{72}
N. crassa	G^	G	T	A	A	G	T

Consensus Intron Branch Site

		A_{56}				A_{56}			
S. macrospora			C_{100}	T_{100}		A_{94}	C_{78}	N	
		G_{44}				G_{33}			
		A				A		C	
N. crassa			C	T			A		C
		G				G		A	

Consensus 3' Intron-Acceptor

	G_{33}	A_{39}	C_{56}		
S. macrospora				A_{100}	G_{100}
	A_{27}	T_{39}	T_{44}		
	A	A	T		
N. crassa				A	G
	T	T	C		

In *S. macrospora* genes the intron length lies between 47 bp and 256 bp, the average length is 88 bp and the median length is 60 bp. Intron length in *N. crassa* ranges from 46 to 856 bp with a tendency toward 60 to 70 bp. Among the eight genes analyzed so far, two genes, *ura3* and *ura5*, do not contain introns. In *S. macrospora* introns the distance from the C of the splice branch site to the G of the 3' splice site is between 12 nt and 22 nt. This distance varies in *N. crassa* from 14 to 30 nucleotides. The *S. macrospora* intron signals (5' donor site, intron branch site and 3' intron acceptor site) are very similar to the *N. crassa* intron consensus sequences.

Acknowledgments

I would like to thank Prof. Dr. U. Kück (Bochum) for his generous support, M. Nouwrousian for providing the unpublished sequence from the *S. macrospora ura3* and *ura5* genes, and D. Hahn for critical reading of the manuscript. This work was supported by a grant of the Deutsche Forschungsgemeinschaft (Bonn-Bad Godesberg).

Return to the FGN 44 Table of Contents

Return to the FGSC home page