

Regulatory sequences involved in the translation of *Neurospora crassa* mRNA: Kozak sequences and stop codons

Jon J.P. Bruchez, J. Eberle and V.E.A. Russo - Max Planck Institut für Molekulare Genetik, Ihnestr. 73, D-14195 Berlin, Germany

We have analyzed the sequences of 77 nuclear genes of *N. crassa* thought to be transcribed by RNA polymerase II (References 1-72) which should represent virtually all of the presently published nuclear gene sequences for this fungus. Kozak (1988, Nucl. Acids Res. 15:8125-) analyzed 699 vertebrate genes leading to identification of the vertebrate consensus sequence for initiation of translation, or Kozak Sequence:

G44C39C53(A61/G36)(C49/A27)C55A100T100G100G46

We show here that the *N. crassa* Kozak sequence is:

C57NNNC77A81(A44/C43)"T"3A99T100G99G51C53

where the subscript number indicates the % occurrence of the particular nucleotide and "T" indicates the conserved absence of that particular nucleotide.

We arbitrarily decided that a nucleotide was to be included in the consensus only if it was present in at least 50% of all the sequences analyzed. If two nucleotides, each represented at less than 50%, gave a summed total of at least 75% representation for a single position, then both are shown in brackets.

Table I. Consensus for initiation of translation and stop codons in *Neurospora crassa*

No.	Ref.	Gene	Distance from +1 to ATG (bases)	Kozak Sequence	Stop codon
				Consensus : CNNNCAATGGC	
1	1	acp	46	AATATCACAATGGCG	TAA
2	2	acu-3	-	CTGCCCATCATGGCT	TAG
3	3	acu-5	103	ATACGAGTTATGGCG	TAA
4	4	acu-8	-	TCACCAACCATGGCG	TAA
5	5	acu-9	60	CTTTTCACAATGGCT	TAA
6	6	al-1	-	ACAGACAAAATGGCT	TAG
7	7	al-3	90	CACGTCACCATGGCC	TGA
8	8	alc	54	TCCCTCACCATGACC	TAA
9	9	am	109	ACCTTCAAAATGTCT	TAA
10	10	arg-2	118	CAAGTCAAGATGTTC	TAA
11	11	atp-1	90	CTCCACAACATGTTC	TAA
12	11	atp-2	58	ATCGTCAAGATGTTC	TAA
13	12	bli-7	110	ACCGCCAAAATGCAG	TAA
14	13	Bml	-	ACCGTCAAGATGCGT	TAA
15	14	chs-1	69	TCCGCAACCATGGCG	TGA

16	15	cmt	127	TCTATCAAAATGGGT	TAA
17	16	con-8	221	ACAATAACCATGGAT	TGA
18	17	con-10	91	ATCGTCAACATGGCT	TAG
19	18	con-13	86	CGTCGCAAGATGCCC	TGA
20	19	cot-1	-	GGTACCAAGATGGAC	TAA
21	20	cpc-1	622	TCCATCAAGATGCGT	TAA
22	21	cpi	-	TTAGTGAAAATGTTT	TAA
23	22	crp-1	-	GCAGACAACATGGTA	TAA
24	23	crp-2	62	ACCGTCAAGATGCCC	TGA
25	24	crp-3	58	GCCGGCAAAATGGGT	TAA
26	25	cya-4	146	GCCGCCACCATGCTT	TAA
27	26	cys-3	30	CATGGCACAATGTCT	TAA
28	27	cys-14	32	GACACTCAGATGGCT	TAA
29	28	cyt-2	-	TCAGTCGCAATGGGT	TAA
30	29	cyt-18	-	TCACATCAAATGCTG	TAA
31	30	cyt-20	57	GTCCTCTGGATGCCG	TAA
32	31	cyt-21	125	CGGTCCAACATGGTT	TGA
33	32	for	66	TCAGTCACCATGTCT	TAA
34	33	frq	-	GAAACCTGAGTTGGA	TGA
35	34	grg-1	89	TCAACCAAAATGGAT	TAA
36	35	H3	-	ACCATCACAATGGCC	TAA
37	35	H4	-	CATATCAAAATGACT	TAA
38	36	his-3	124	GAAAACACCATGGAG	TAA
39	37	hsp30	120	AAGTCAAAAATGGCG	TAA
40	38	ilv-2	-	TCCATCACAATGGCC	TAA
41	39	laccase	190	TTTATCACCATGAAA	TAG
42	40	leu-5	146	CACAACGCGATGCCT	TAG
43	41	leu-6	220	TAAACAAACATGGCC	TAA
44	42	lox	123	TCATACAAGATGAAG	TGA
45	43	met-7	98	ATCACAGCCATGCTT	TGA
46	44	mrp-3	-	CCTCTCACCATGATC	TAA
47	45	mta-1	-	ACCGAAACAATGGAC	TGA
48	46	mtA-1	-	AGAAACACGATGTCG	TAG
49	47	nac	162	CCGGTGACAATGACG	TAA
50	48	ncypt1	-	TTGCCCATCATGAAC	TAA
51	49	nit-2	284	TGTGCGACAATGGCG	TAA
52	50	nit-3	110	AGCATCATCATGGAG	TGA
53	51	nit-4	39	CCCCGGCAGATGAAC	TGA
54	52	nuc-1	-	GCGGGCGTGATGAAC	TAA
55	53	nur22	-	ACCGTCAAGATGGCG	TGA
56	54	nur40	-	ACTCACAAGATGGCT	TGA
57	55	nur49	-	CAAACAACAATGGCG	TAA
58	56	pho-4	145	TCGTTCAAGATGGTT	TGA
59	57+58	pma-1	56	ATAACGCCAATGGCG	TAA
60	59	preg	-	GGATTTGTGATGCTG	TAA
61	60	pyr-4	61	ACAGCCAACATGTCG	TAG
62	61	qa-1F	330	AATCCCAACATGCCG	TAG
63	61+62	qa-1S	346	GCCGCCATCATGAAC	TGA
64	61	qa-2	85	CCAAACACAATGGCG	TGA
65	61	qa-3	83	TATATCACCATGTCG	TGA
66	61+63	qa-4	190	CCTTTTCGCCATGCCG	TAA
67	61	qa-x	84	TCAGCAGCCATGACA	TGA
68	61	qa-y	133	CGCGTCAAGATGACT	TAA
69	64	sod-1	-	TCCGTCAAAATGGTC	TAA
70	65	spe-1	535	TCTTGGGATATGGTT	TAA
71	66	T	94	GCAGCAACCATGAGC	TGA
72	67	trp-1	29	CCAATCACAATGTCG	TAA

73	68	trp-3	147	TCATACACAATGGAG	TAA
74	69	Ubi	-	ACCCCCATCATGCAG	TAA
75	70	ucr	-	ACCGACACAATGGCG	TAA
76	71	vma-1	-	TCGCCCCAAGATGGCT	TGA
77	72	vma-2	-	TCTTCCACAATGGCC	TAA

Key: - in the Distance from +1 to ATG (bases) means that the authors had not determined the +1 position

The reason why the methionine start codon (ATG) is not 100% perfectly conserved within the Kozak consensus is that, for reasons unknown, the gene *frq* (Ref 33) starts its protein sequence with a valine (GTT).

It is also interesting to note that the choice of the second codon appears to be limited in that about half of the second codons have a guanosine in the first position and another half have a cytosine in the second position.

On the whole, our consensus shows a good resemblance to the mammalian Kozak sequence with a similar hierarchy of nucleotide preference for a given position, although the degree of preference may be shifted. An exception is the nucleotide position immediately preceding the initiator methionine codon (ATG) where *N. crassa* exhibits a definite suppression of thymine in contrast to a positive preference for any other nucleotide.

Fifty genes among the 77 analyzed have a determined mRNA 5' end. When several 5' ends were presented, +1 was taken to be the most distal from the ATG except when given by the authors themselves. In this way the mRNA sequences before the ATG have lengths between 30 and 622 bases.

The stop codon, determined by computer analysis by the authors, TAA in 62% of the cases, TGA in 27% and TAG in 11%

REFERENCES

- 1 Arends and Sebald, 1984, EMBO J., 3:377-382
- 2 Gainey et al. 1992, Curr. Genet., 21:43-47
- 3 Connerton et al. 1990, Molec. Microbiol., 4:451-460
- 4 Marathe et al. 1990, Mol. Cell. Biol., 10:2638-2644
- 5 Sandeman et al. 1991, Mol. Gen. Genet., 228:445-452
- 6 Schmidhauser et al. 1990, Mol. Cell. Biol., 10:5064-5070
- 7 Carrattoli et al. 1991, J. Biol. Chem., 266:5854-5859
- 8 Lee et al. 1990, Biochemistry 29:8779-8787
- 9 Kinnaird and Fincham, 1983, Gene, 26:253-260
- 10 Orbach et al. 1990, J. Biol. Chem., 265:10981-10987
- 11 Bowman and Knock, 1992, Gene, 114:157-163
- 12 Eberle and Russo, 1992, DNA Sequence, 3:131-141
- 13 Orbach et al. 1986, Mol. Cell. Biol., 6:2452-2461
- 14 Yarden and Yanofsky, 1991, Genes Dev., 4:2420-2430
- 15 Munger et al. 1985, EMBO J., 4:2665-2668

- 16 Roberts and Yanofsky, 1989, *Nucl. Acids Res.*, 17:197-214
- 17 Roberts et al. 1988, *Mol. Cell. Biol.*, 8:2411-2418
- 18 Hager and Yanofsky, 1990, *Gene*, 96:153-159
- 19 Yarden et al. 1992, *EMBO J.*, 11:2159-2166
- 20 Paluh et al. 1988, *Proc. Natl. Acad. Sci. USA*, 85:3728-3732
- 21 Tropschug, 1990, *Nucl. Acids Res.*, 18:190
- 22 Kreader and Heckman, 1987, *Nucl. Acids Res.*, 15:9027-9042
- 23 Tyler and Harrison, 1990, *Nucl. Acids Res.*, 18:5759-5766
- 24 Shi and Tyler, 1991, *Nucl. Acids Res.*, 19:6511-6517
- 25 Sachs et al. 1989, *Mol. Cell. Biol.*, 9:566-577
- 26 Fu et al. 1989, *Mol. Cell. Biol.*, 9:1120-1127
- 27 Ketter et al. 1991, *Biochemistry*, 30:1780-1787
- 28 Drygas et al. 1989, *J. Biol. Chem.*, 264:17897-17906
- 29 Akins and Lambowitz, 1987, *Cell*, 50:331-345
- 30 Kubelik et al. 1991, *Mol. Cell. Biol.*, 11:4022-4035
- 31 Kuiper et al. 1988, *J. Biol. Chem.*, 263:2840-2852
- 32 McClung et al. 1992, *Mol. Cell. Biol.*, 12:1412-1421
- 33 McClung et al. 1989, *Nature*, 339:558-562
- 34 McNally and Free, 1988, *Curr. Genet.*, 14:545-551
- 35 Woudt et al. 1983 *Nucl. Acids Res.*, 11:5347-5366
- 36 Legerton and Yanofsky, 1985, *Gene*, 39:129-140
- 37 Plesofsky-Vig and Brambl, 1990, *J. Biol. Chem* 265:15432-15440
- 38 Sista and Bowman, 1992, *Gene*, 120:115-118
- 39 Germann et al. 1988, *J. Biol. Chem.* 263:885-896
- 40 Chow et al. 1989, *Mol. Cell. Biol.*, 9:4631-4644
- 41 Chow and RajBhandary, 1989, *Mol. Cell. Biol.*, 9:4645-4652
- 42 Niedermann and Lerch, 1990, *J. Biol. Chem.*, 265:17246-17251
- 43 Crawford et al. 1992, *Gene*, 111:265-266
- 44 Kreader et al. 1989, *J. Biol. Chem.*, 264:317-327
- 45 Staben and Yanofsky, 1990, *Proc. Natl. Acad. Sci. USA* 87:4917-4921
- 46 Glass et al. 1990, *Proc. Natl. Acad. Sci. USA* 87:4912-4916
- 47 Kore-eda et al. 1991, *Jap. J. Genet.*, 66:317-334
- 48 Heintz et al. 1992, *Mol. Gen. Genet.*, 235:413-421
- 49 Fu and Marzluf, 1990, *Mol. Cell. Biol.*, 10:1056-1065
- 50 Okamoto et al. 1991, *Mol. Gen. Genet.*, 227:213-223
- 51 Yuan et al. 1991, *Mol. Cell. Biol.*, 11:5735-5745
- 52 Kang and Metzenberg, 1990, *Mol. Cell Biol.*, 10:5839-5848
- 53 Nehls et al. 1991, *Biochim. Biophys. Acta*, 1088:325-326
- 54 Rohlen et al. 1991, *FEBS*, 278:75-78
- 55 Preis et al. 1990, *Curr. Genet.*, 18:59-64
- 56 Mann et al. 1989, *Gene*, 83:281-290
- 57 Aaronson et al. 1988, *J. Biol. Chem.*, 263:14552-14558
- 58 Hager et al. 1986, *Proc. Natl. Acad. Sci. USA* 83:7693-7697
- 59 Kang and Metzenberg, 1993, *Genetics*, 133:193-202
- 60 Glazebrook et al. 1987, *Mol. Gen. Genet.*, 209:399-402
- 61 Geever et al. 1989, *J. Mol. Biol.*, 207:15-34

- 62 Huiet and Giles. 1986, Proc. Natl. Acad. Sci. USA, 83:3381-3385
- 63 Rutledge, 1984, Gene, 32:275-287
- 64 Chary et al. 1990, J. Biol. Chem., 265:18961-18967
- 65 Williams et al. 1992, Mol. Cell Biol., 12:347-359
- 66 Kupper et al. 1989. J. Biol. Chem 264:17250-17258
- 67 Schechtman and Yanofsky, 1983, J Molec. Appl. Genet., 2:83-99
- 68 Burns and Yanofsky, 1989, J. Biol. Chem., 264:3840-3848
- 69 Taccioli et al. 1989, Nucl. Acids Res., 17:6153-6166
- 70 Harnish et al. 1985, Eur. J. Biochem., 149:95-99
- 71 Bowman et al. 1988, J. Biol. Chem. 263:13994-14001
- 72 Bowman et al. 1988, J. Biol. Chem., 263:14002-14007