Nucleotide Sequence of a Member of the Napin Storage Protein Family from *Brassica napus*  

(Received for publication, December 19, 1986)

Steven R. Scofield and Martha L. Crouch  
From the Department of Biology, Indiana University, Bloomington, Indiana 47405

We have begun the molecular characterization of genes encoding napin, the 1.7 S embryo-specific storage protein of *Brassica napus*. Genomic Southern blot analysis indicates that napin is encoded by a multigene family comprised of a minimum of 16 genes. Two DNA fragments containing single napin genes have been recovered from *B. napus* genomic libraries. We have determined the nucleotide sequence of one member of the napin gene family, gNa. The gene has a simple structure lacking introns and containing the canonical features expected for genes transcribed by RNA polymerase II. The site of the initiation of transcription was determined to be 37 base pairs upstream of the initiation codon by S1 and primer extension analyses. A gene-specific hybridization probe from the 3' non-translated portion of gNa was used to demonstrate transcription of gNa.

As the sequences of seed proteins from different plants become known, homologies between proteins with drastically different properties are being detected. For example, several of the diverse 2 S proteins found in seeds have been shown to share sequence homology: the methionine-rich Brazil nut storage protein,¹ the allergenic storage protein in castor bean endosperm (Sharief and Li, 1982), the very basic 1.7 S storage protein in rapeseed embryos (Crouch *et al.*, 1983), and a trypsin inhibitor from barley (Odani *et al.*, 1985). Also, these proteins are related to the prolamin storage proteins such as γ-galactin from rye (Kreis *et al.*, 1985) and α-gliadin from wheat (Kasadara *et al.*, 1984), even though the prolamins are much larger and are hydrophobic rather than hydrophilic. In many cases, the properties of the specific proteins are the result of repeated sequences that differ between them (Higgins, 1984). Despite the different physical properties conferred by these repeats, all of the proteins accumulate to high levels during seed development, are stored during the period of developmental arrest separating embryogeny from germination, and are then degraded during seedling growth. Thus, the basic pattern of temporal expression has been retained. This class of storage proteins is particularly important for animal nutrition, since they usually have higher levels of the sulfur-containing amino acids than the other abundant seed proteins (Youle and Huang, 1981).

We have been studying the expression of the genes for the 1.7 S storage proteins from *Brassica napus* L. (rapeseed), the napins. Using a cloned cDNA probe from one of the napin family members, transcripts can first be detected early in embryogenesis, just after the major tissue systems have been delineated (Crouch *et al.*, 1985). Levels of napin mRNA increase until they constitute about 8% of the total mRNA at the end of cell division,² stay high for 15 days, and then decrease to barely detectable levels in dry seeds. Napin transcripts cannot be detected at any other time in development. However, this pattern of expression reflects the average of several napin genes. In order to study regulation of napin gene expression in detail, it is necessary to analyze family members individually.

In this paper, we begin an analysis of the napin gene family by determining the minimum number of napin genes and by cloning and sequencing one member of the family. From S1 protection and primer extension experiments, we have determined where in the sequence transcription begins and that this family member is expressed.

**Materials and Methods**³

**Results**

**Napin Gene Family—**It is clear from genomic Southern blots that napin is encoded by a family of genes. At least 14 fragments, ranging from 2 to 23 kb in size, hybridize with different intensities to a napin cDNA probe pN1 when genomic DNA is restricted with EcoRI (Fig. 1A). EcoRI does not cleave within any cloned napin sequence. The hybridization pattern observed is the same whether the probed DNA is made from a single plant or from a population, indicating that this pattern is not due to population polymorphism (data not shown). The hybridization pattern is also unchanged when probes representing the 5' and 3' halves of the pN1 coding sequence are tested, indicating that all the bands are due to homology with the napin coding sequence and not a repeated sequence in one portion of the cDNA clone pN1 (data not shown).

Fig. 1B is a genomic reconstruction experiment. The genomic clone ABnNa, described later, was digested with EcoRI, containing amino acids than the other abundant seed proteins (Youle and Huang, 1981).

* This work was supported in part by National Science Foundation Grant PCM-8318403 (to M. L. C.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

¹ The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EMBL Data Bank with accession number(s) J02782.

² Portions of this paper (including "Materials and Methods" and Figs. 2 and 3) are presented in miniprint at the end of this paper. Miniprint is easily read with the aid of a standard magnifying glass. Full size photocopies are available from the Journal of Biological Chemistry, 9050 Rockville Pike, Bethesda, MD 20814. Request Document No. 86M4334, cite the authors, and include a check or money order for $3.20 per set of photocopies. Full size photocopies are also included in the microfilm edition of the Journal that is available from Waverly Press.


The abbreviations used are: kb, kilobase(s); bp, base pair(s).
and dilutions representing 1, 3, 5, and 10 copies/haploid genome were electrophoresed beside EcoRI-digested genomic DNA. We conclude that the fragments which have the least intense signals contain single napin genes, and the stronger signals represent two or more genes. By this analysis there are at least 16 napin genes/haploid genome. The more intense signals result either from fragments of similar size that contain single genes or linkage of two or more napin genes on an EcoRI fragment.

**Isolation of Genomic Napin Clones**—A genomic library was constructed in the λ vector EMBl6 from *B. napus* DNA digested partially with *Sau3A*. Two unique napin genomic clones, designated λBnNa and λBnNb, were isolated when $4 \times 10^5$ recombinant phage were screened by plaque hybridization with a nick-translated pN1 napin cDNA probe (Crouch et al., 1983).

The napin genomic clones were analyzed by restriction nuclease mapping and Southern blot hybridizations. Each phage contains just one napin gene, and only the napin gene region hybridizes to cDNA made from embryo RNA, indicating that no other abundant embryo transcripts are encoded by the cloned DNA (data not shown). Comparison of restriction maps derived for genomic napin subclones with those of the cDNA clones pN1 and pN2 shows that these genes do not encode the messages represented by the cDNA clones (Fig. 2). λBnNa was chosen for more thorough examination.

**Nucleotide Sequence of Napin Gene**—The 3.3-kb EcoRI fragment containing the λBnNa napin gene was subcloned in pUC8 (Sieira and Messing, 1982) and designated pgNa. One kb to the right of the first EcoRI site, as drawn in Fig. 3, has been sequenced.

**Fig. 4.** Nucleotide sequence and deduced amino acid sequence of gNa. The boxed sequences labeled RY are alternating purine-pyrimidine elements 7 bp or longer. Also labeled are: CAAT sequence at 186; the TATA box at 295; the initiating ATG at 295; and three sequences with homology to the consensus polyadenylation processing sequence at 954, 979, and 1004 (underlined).
been sequenced by the method of Maxam and Gilbert (Maxam and Gilbert, 1980). This sequence is comprised of 561 coding bp, 292 5’ and 172 3’ flanking bp (Fig. 5).

The napin reading frame is the only open reading frame of significant length on either strand. The 5’ end of this sequence is very AT-rich (64%) and is marked by many blocks of 4–6 consecutive A or T residues. A TATA box closely matching the consensus is found 70 bp upstream from the ATG codon initiating the napin precursor. This is the first ATG codon downstream of the TATA sequence. Forty-two bp upstream of the TATA box is the sequence CAAT (position 186, Fig. 4). Though in the expected position, this sequence shows only 4 bp of homology to the 9-bp consensus element shown to be important for efficient promoter recognition (Benoist et al., 1980). Three regions of alternating purine-pyrimidine residues occur upstream of the TATA box: between positions 90 and 112 are three 7-bp alternating purine-pyrimidine units, at position 167 a block of 21 consecutive purine-pyrimidine residues occurs, and at position 193 an 8-bp unit is found.

The 3’ untranslated region is high in AT content (67%). Plant genes frequently are found to contain multiple sequences resembling the consensus element associated with polyadenylation of mRNA (Fitzgerald and Shenk, 1981), and three of these elements are present in the gNa sequence, occurring at nucleotides 954, 979, and 1004 (Fig. 4).

Comparison of the coding sequences of gNa and the cDNA clones pN1 and pN2 indicates that there are no introns and that all three coding sequences terminate with a single TAG codon. Within the coding sequence there is some divergence between the genomic and cDNA clones. For example, when the gNa sequence is aligned for maximum homology with the pN2 sequence it is observed that the genomic coding sequence is 24 bp longer than the cDNA. Excluding insertions, the two sequences are 90% homologous at the nucleotide level, with 57% of the nucleotide substitutions occurring in the third base of the codon. Alignment of the gNa- and pN2-deduced peptide sequences shows 18 amino acids have been substituted excluding the gNa insertions but that only five of the substitutions are conservative (hydrophobic to hydrophobic, for example).

Expression of gNa—Demonstrating the expression of a particular gene family member by hybridization requires a gene-specific probe. Since the nontranslated portions of genes often provide such probes, the 0.4-kb XhoI-BamHI fragment of gNa complementary to the 3’ nontranslated portion gNa transcripts was nick-translated and used to probe duplicate genomic Southern blots of EcoRI-digested B. napus DNA (Fig. 5A, lanes 2 and 3). This probe hybridizes to just two napin genes at Tm = −8 °C (Fig. 5A, lane 2) and specifically to the 3.3-kb gNa EcoRI fragment at Tm = −3 °C (Fig. 5A, lane 3).

Duplicate blots of size-fractionated B. napus embryo RNA were hybridized and washed in parallel with the DNA filters. Under the conditions that gave gene-specific DNA/DNA hybridization, a signal is detectable on the Northern blot corresponding to a napin-sized transcript (Fig. 5B, lane 2). No hybridization was evident when the DNA and RNA blots were washed at Tm = +3 °C, however (data not shown).

Mapping the 5’Terminus of the gNa Transcript—Our first studies of the initiation site of gNa transcripts employed S1 nuclease digestion analysis (Fig. 6). The 0.38-kb SalI-EcoRI fragment of gNa was 5’ end-labeled at the SalI site, and the labeled strand was purified on a polyacrylamide gel to use as a probe. The same fragment was sequenced to provide accurate electrophoretic size standards. Aliquots of this probe were hybridized at either Tm = −25 °C or Tm = −4 °C with 100 µg of B. napus embryo RNA. After digestion of the resulting hybrids the longest protected probe fragment was 138 bp, indicating initiation at the T number 258 in Fig. 4. Also apparent are strong signals corresponding to cleavage in the 2 blocks of dA residues located 4 and 10 bp downstream from the initiation site, but the reason for cleavage at these sites is unclear. Local denaturation in the AT-rich regions seems unlikely as these signals are generated under nonstringent S1 nuclease digestion conditions. It is possible that these signals represent other initiation points for the same gene or different 5’ end structures of transcripts from other napin genes which are able to hybridize with the probe, which does contain 88 bp of coding sequence.

Primer extension analysis employing a synthetic oligonucleotide primer without coding sequence was undertaken to more specifically define the 5’ end of the gNa transcript (Fig. 7). An oligomer was synthesized that was complementary to the 15 bases immediately 5’ to the gNa codon. When hybridized to embryo RNA this primed the reverse transcription of a product extending 22 bases beyond the oligomer indicating RNA initiation at the dA numbered 259 in Fig. 4, just one
base short of the 5' end mapped by S1 nuclease protection. Since eucaryotic mRNAs are capped at purine residues (Cory and Adams, 1975), we expect the authentic RNA initiation site of gNa transcripts to be the dA, position 259 (Fig. 4), indicated by primer extension analysis. The gNa transcript thus has a 5' nontranslated leader 37 nucleotides long.

The primer extension experiments were also used to address the expression of gNa by performing the reverse transcription in the presence of deoxyribonucleotides to determine the sequence of the primer extension product (Fig. 7). A sequence consistent with the expression of gNa can be detected, although the extent to which this portion of the gNa sequence is conserved among the napin genes is not yet known. The presence of heterogeneous signals in the sequencing ladder indicates that the primer hybridized with other napin transcript(s) as well.

Fig. 7. Primer extension analysis of gNa. The nucleotide sequence upstream of the napin initiation codon is shown at the top of the figure. A 15-mer complementary to the gNa sequence 5' to the initiation codon was kinased and annealed to 100 μg of total embryo RNA. Hybrids were extended by avian myeloblastosis virus reverse transcriptase. Lane 1, primer only. Lane 2, mock primer extension with no RNA. Lane 3, primer extension using 10-fold less primer (3 ng) than lanes 4-7. Lanes 4, 5, 6, and 7, deoxyribonucleotide sequencing of the extension product. Lane 4, G reaction. Lane 5, A reaction. Lane 6, T reaction. Lane 7, C reaction. Dots have been placed in the sequencing ladder where bands should occur if gNa transcripts serve as template in this experiment.

**DISCUSSION**

Of the approximately 16 napin genes in *B. napus*, one has now been sequenced by us, gNa, and another, napA, by L.-G. Josefsson. In addition we previously reported the sequences
Fig. 8. One kb of the gNa genome sequence has been aligned for maximum homology with napA and two cDNA clones, pN1 and pN2. To emphasize the close homology between the cDNAs and napA only the cDNA bases that differ from napA have been displayed. Dots indicate positions where gaps have been introduced into a sequence for alignment purposes. Conserved features which have been designated are: the alternating purine-pyrimidine blocks (RY), the TATA boxes, the initiation and termination codons, and the 12 bp of homology shared at the most downstream consensus sequence associated with polyadenylation.
of two different cDNA clones representing transcripts from other genes (Crouch et al., 1983). Thus, four members of the family have been examined, although their relative levels of expression are not known. Comparison of all four coding sequences (Fig. 8) indicates that the cDNAs and napA are greater than 95% homologous. The gNa sequence with insertions at positions 521, 588, 714, and 734 of Fig. 8 is likely to represent a minor class of napins, perhaps one of the four discrete species fractionated by Lonnerdal and Janson (1972).

The 3′ nontranslated regions of the cDNAs and napA are as highly conserved as the coding regions. Such high homology would preclude the use of these sequences for gene-specific hybridization as was possible for gNa. One of the distinctive features of this portion of gNa is the presence of three sequences resembling the consensus associated with polyadenylation of mRNAs. It is striking that although the gNa 3′ nontranslated region is divergent, all four napin sequences are perfectly homologous for 12 bp around the most downstream consensus polyadenylation element, suggesting that this is the authentic polyadenylation signal for the genomic clones.

The nucleotide sequence of the genomic clone gNa and its flanking regions contain the canonical features expected of plant genes transcribed by RNA polymerase II (Messing et al., 1983). There are no introns, which is characteristic of genes for many of the other 2S seed proteins and related cereal prolamins. In the 5′ flanking region of gNa are several blocks of alternating purine-pyrimidine nucleotides, which have been observed in viral enhancer (Lusky et al., 1983). Their significance in napin genes remains to be tested.

Alignment of the two napin genomic clones for maximum homology (Fig. 8) shows the coding sequence of gNa is 24 bp longer than napA with the extra sequence occurring as three additions of single codons and two insertions of two codons. However, the 5′ RNA leader region of gNa is deleted by 10 nucleotides relative to napA. As already mentioned, the two genomic sequences diverge sharply past the coding sequence termination codons. In contrast, the 5′ flanking region is highly conserved overall, including the regions of alternating purine-pyrimidine residues. Since the entire 5′ flanking region is so highly conserved, it is difficult to single out regions by comparative homology that might be involved in the temporal or spatial regulation of napins.

As mentioned earlier, napin is evolutionarily related to some of the cereal prolamin storage proteins. However, there is no evidence in napin genomic sequences of homology to the short upstream sequences found to be conserved in the genes for α-gliadin, β-hordein, and the (unrelated) zeins (Forde et al., 1985). If the conserved prolamin sequence is functionally significant, its absence in napin may be related to the difference in spatial expression; napin is synthesized in the embryo, whereas prolamins are restricted to endosperm cells.

Acknowledgments—We wish to thank Dr. Lars-Goran Josefsson, Wallenberg Laboratory, Uppsala, Sweden, for exchanging napin nucleotide sequences prior to publication. We appreciate the synthesis of oligonucleotides by Lawrence Washington, Institute for Cell and Molecular Biology, with funds from Indiana Corporation for Science and Technology, and the expert advice of Karen Tenbarge, Jerome Cane, and Lorraine Solberg. Most helpful discussions were provided by Drs. Keith Blundy and Vic Knauf, Calgene Inc., Davis, CA. We are grateful for Karen Parr’s swift assistance in the preparation of this manuscript.

REFERENCES


Continued on next page.
Napin Nucleotide Sequence

Supplementary material for:

NUCLEOTIDE SEQUENCE OF A MEMBER OF THE NAPIN STORAGE PROTEIN FAMILY FROM PRAETICA MAPT

Steven R. Keilholz and Martha W. Crouch

MATERIALS AND METHODS

Plants

Napin seeds from 'Praetica' were grown in a sand-peat mixture under light and temperature conditions simulating those found in the field. The plants were grown as described above. Nucleic acids were isolated as described in detail in Materials and Methods.

Genomic DNA isolation

Genomic DNA was isolated from 2- to 5-week-old seeds by the method described in Materials and Methods. The DNA was isolated from 2- to 5-week-old seeds by the method described in Materials and Methods.

DNA isolation

DNA isolation was performed using 15 M DNA per gel lane. The DNA isolation was performed using 15 M DNA per gel lane. The DNA isolation was performed using 15 M DNA per gel lane. The DNA isolation was performed using 15 M DNA per gel lane. The DNA isolation was performed using 15 M DNA per gel lane.