THE JOURNAL OF BIOLOGICAL CHEMISTRY

Vol. 262, No. 25, Issue of September 5, pp. 12196-12201, 1987
Printed in U.S.A.

Structure of a Gene Encoding the 1.7 S Storage Protein, Napin, from *Brassica napus**

(Received for publication, December 22, 1986)

Lars-Göran Josefsson, Marit Lenman, Mats L. Ericson, and Lars Rask

From the Department of Cell Research, The Swedish University of Agricultural Sciences, Box 596, S-751 24 Uppsala, Sweden

A rapeseed chromosomal region containing a gene (*napA*), which encodes the 1.7 S seed storage protein (napin), was isolated in several overlapping recombinant clones from a phage λ genomic library. Following restriction enzyme mapping of the genomic region, a subclone containing the *napA* coding region as well as some 1.1 and 1.4 kilobases of DNA from the 5′ and 3′ regions, respectively, was mapped and sequenced. The gene turned out to lack introns. Southern blotting analyses utilizing a napin cDNA clone as a probe revealed the presence of on the order of 10 napin genes in the rapeseed genome. The major polyadenylated transcript encoded by these genes was shown to be an 850-nucleotide species, the initiation site of which was mapped onto the *napA* gene. The major initiation site for transcription is located some 33 nucleotides downstream from a sequence perfectly conforming to the consensus sequence of a TATA box. Further analyses of the sequence revealed several features that may be of relevance for the expression of the napin genes.

Napin, or the 1.7 S protein, is one of the major seed storage proteins in *Brassica napus*. It is expressed in a tissue-specific manner, apparently under the influence of abscissic acid (Crouch and Sussex, 1981; Crouch *et al.*, 1983). The mature protein, which is rather basic, consists of two subunit polypeptides that are linked by disulfide bridges (Ericson *et al.*, 1986; Lönnerdal and Janson, 1972). Comparison of amino acid sequences of the subunits with the sequence of a cDNA clone has shown that the initial translation product, a 20-kDa precursor, contains both the subunit polypeptides as well as polypeptide stretches that are removed during the maturation of the protein (Ericson *et al.*, 1986). By analogy with other storage proteins, the final product is thought to reside in specialized organelles, protein bodies, within the seed cells (Larkins and Hurkman, 1978). As far as is known, the sole function of napin is to serve as a nutrient source during germination and initial development of the seedling. Confirmatory evidence that napin, like other storage proteins, possesses minor heterogeneities in the amino acid sequence stems from protein separation data (Lönnerdal and Janson, 1972) as well as protein sequencing (Ericson *et al.*, 1986) and the analysis of cDNA clones (Crouch *et al.*, 1983; Ericson *et al.*, 1986). As an initial step toward an increased understanding of the regulation of napin genes, we have isolated and sequenced a member of what turns out to be a small gene family.

MATERIALS AND METHODS AND RESULTS[1]

DISCUSSION

We have isolated and sequenced a gene encoding napin. The gene is a member of a small family with some 10 genes. Transcription of an as yet unknown number of these genes yields an 850-nucleotide-long mRNA, the cap site of which was mapped onto the *napA* sequence. We have compared our sequence with that of another napin gene, pGNA, as well as with previously sequenced cDNA clones (Crouch *et al.*, 1983; Ericson *et al.*, 1986). The *napA* sequence is completely identical to the pNAP1 cDNA clone that we have previously sequenced (Ericson *et al.*, 1986). This makes us rather confident that we have sequenced an expressed copy of the napin gene family, although we have no formal proof that this is the case.

Comparison with the pGNA gene sequence revealed that, apart from single nucleotide changes, a quite frequently occurring divergence in the coding region is insertions of one or two triplets in pGNA relative to *napA*. These occur in four and two instances, respectively (data not shown). Apart from one previously reported triplet deletion in the pN1 cDNA clone (Crouch *et al.*, 1983). These are the first examples of differences that affect the length of the primary sequence of the translated napin product. The number of nucleotide changes in the coding region is also higher when comparing *napA* with pGNA than with any of the previously sequenced cDNA clones (data not shown). It is interesting to speculate whether these observations may be related to the fact that *B. napus* is an amphidiploid of *Brassica campestris* and *Brassica oleracae*. It might be expected that the genes derived from one of the respective parental species would be more homologous to each other than when comparing across the parental border. We are presently attempting to assign parentalship of isolated napin genes by comparison with Southern blots of genomic DNA from the three species. Preliminary data[2] indicate that the *napA* gene most likely is derived from *B. oleracae*.

* This work was supported by The Swedish Research Council for Natural Sciences, The Swedish Research Council for Forestry and Agriculture, and the Stiftelsen Brinkgården. 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.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank™/EMBL Data Bank with accession number(s) J02798.

[1] Portions of this paper (including "Materials and Methods," "Results," and Figs. 3 and 4) are presented in miniprint at the end of this paper. The abbreviations used are: SDS, sodium dodecyl sulfate; kb, kilobase; dNTP, deoxynucleotide triphosphate; AMV, avian myeloblastosis virus; hn, heterogenous nuclear. Miniprint is easily read with the aid of a standard magnifying glass. Full-size photocopies are available from the Journal of Biological Chemistry, 9650 Rockville Pike, Bethesda, MD 20814. Request Document No. 86 M-4366, 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.

[2] M. L. Ericson, unpublished data.

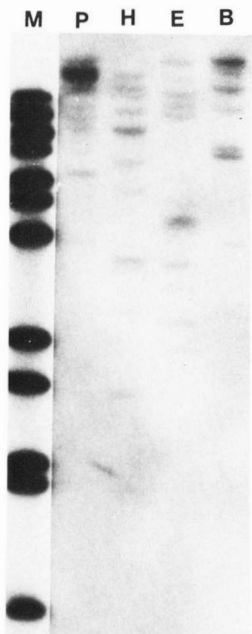

FIG. 1. **Genomic restriction fragments hybridizing with napin cDNA sequences.** Genomic DNA was cut with restriction enzymes. The generated fragments were separated and blotted onto nitrocellulose filters as described under "Materials and Methods." Nick-translated pNAP1 cDNA was used as a probe in hybridization to these filters. The enzymes used were *B, Bam*HI; *E, Eco*RI; *H, Hin*dIII; and *P, Pvu*II. The size marker (*M*) used was an end-labeled *Bst*EII digest of phage λ DNA. Sizes of the marker bands were (from *top* to *bottom*): 8454, 7242, 6369, 5687, 4822, 4324, 3675, 2323, 1929, 1371, 1264, and 702 base pairs.

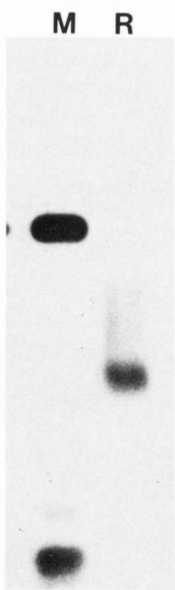

FIG. 2. **Northern blotting and hybridization of rapeseed mRNA to pNAP1 cDNA.** mRNA was purified and separated on denaturing agarose gels as described under "Materials and Methods." After transfer to nitrocellulose filters the immobilized mRNA was hybridized to a nick-translated cDNA probe. *R* denotes the RNA lane; *M*, the marker lane. The marker used was a denatured *Hin*fI digest of pBR322. The autoradiogram reveals the marker bands hybridizing to nick-translated pUC19. The sizes of the bands are 1631 and 517/506 nucleotides, respectively.

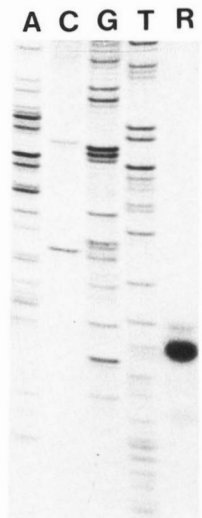

FIG. 5. **Transcript cap site mapping of napin mRNA.** An 18-mer oligonucleotide, complementary to a napin sequence just downstream from the initiation condon, was synthesized. This synthetic oligonucleotide, ^{32}P end-labeled and unlabeled in the respective cases, was annealed to either mRNA or M13 DNA covering this region on the minus strand. In separate reactions the primer was allowed to be elongated to the 5′ end of the napin transcripts or to prime a standard set of sequencing reactions. The products were separated on a gradient sequencing gel. *Lane R* shows the terminated forms that were elongated on the mRNA, *lanes A, C, G,* and *T*, the respective sequencing reactions.

With regard to the primary translation product, comparisons of all the known sequences have made us aware of an interesting repeated structure in the removed parts of the napin polypeptide. All of the previously sequenced cDNA clones and the two genomic clones discussed here conform to this structure. It consists of a stretch of 7 or 8 amino acids, *X-X*---(−)*X*, where *X* denotes hydrophobic and − negatively charged amino acids, respectively. These sequences in *napA* are shown *boxed* in Fig. 6. The negatively charged amino acid in *brackets* is only present in the first copy of the repeat which occurs in the amino-terminal part of the precursor sequence, before the small subunit. The second copy of the repeat occurs within the removed sequence which is present between the small and large subunits. These two repeats in fact carry almost all of the negative charges that are contained in the processed parts of the precursor (Ericson *et al.*, 1986). It is possible that these repeats are involved in processes relevant for the translocation, intracellular transport, and/or deposition of napin into protein bodies. Alternatively, they could serve as signals in the proteolytic processing steps necessary for the generation of mature napin. However, confirmation of a possible role of these repeats in the above processes will have to await experiments directly aimed at these points.

We have noted several interesting features in the sequence of *napA* (and pGNA) that may be of relevance to different aspects of gene regulation. It is tempting to speculate that the 5′ hairpin region and the TACACAT repeat region may be directly involved in the transcriptional activation of the gene and that the 3′ hairpin region may be involved in the termination of transcription. There is ample precedence in the literature for the former point, *i.e.* degenerate (or non-degenerate) repeats as well as alterations in DNA topology (possibly manifesting itself in cruciform structures) have been implied in gene regulation in several systems (Gidoni *et al.*, 1985; Hall *et al.*, 1982; Harland *et al.*, 1983; Serfling *et al.*, 1985). It appears more doubtful what role hairpin loops may play in

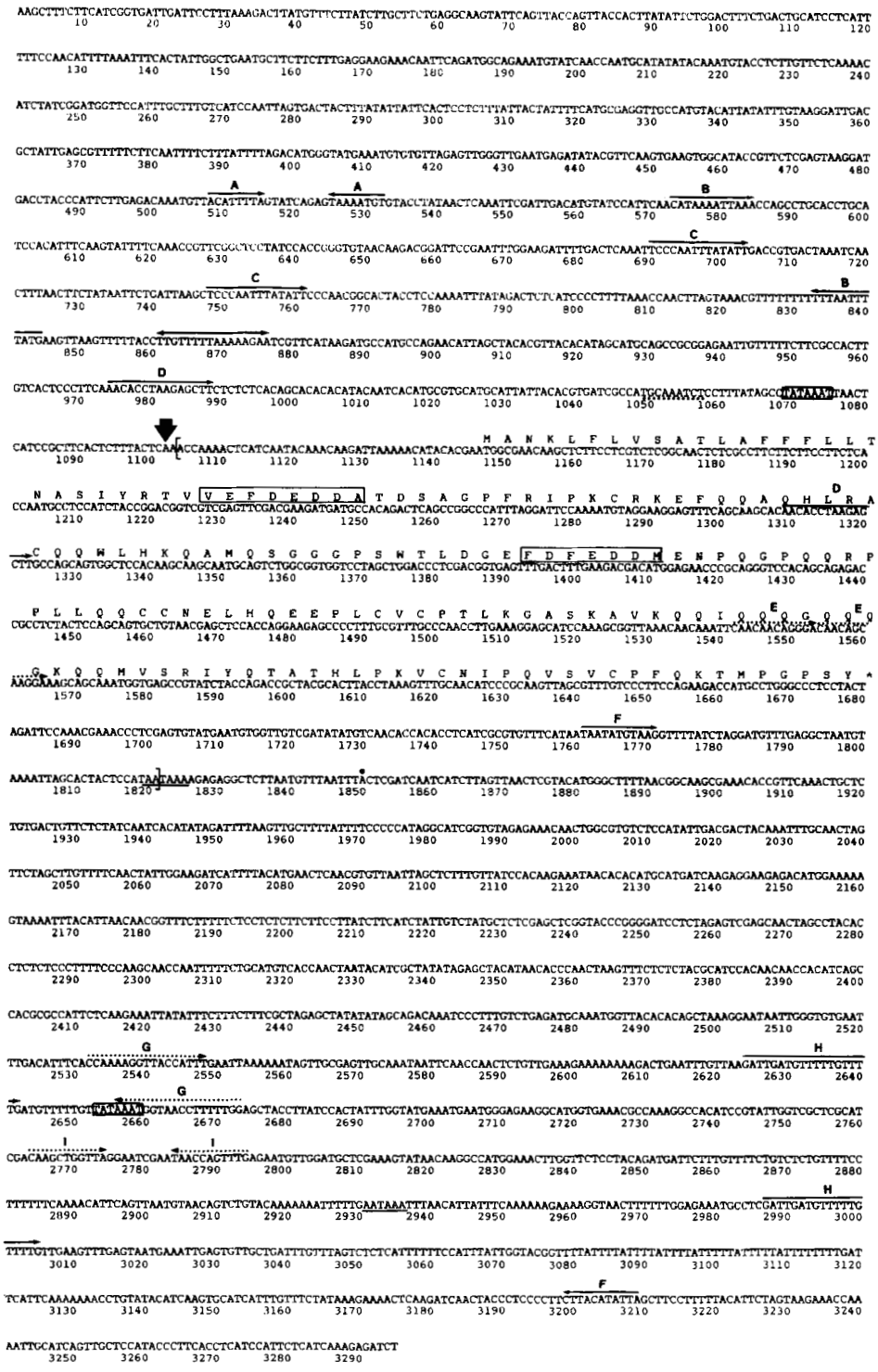

FIG. 6. **Sequence of the *napA* gene.** The figure shows the sequence of the 3.3-kilobase *Hin*dIII-*Bgl*II fragment. The symbols used are all described and discussed in the text.

termination of RNA polymerase II transcripts (Birnstiel *et al.*, 1985), although they may be involved in the termination of specific sets of genes (Hentschel and Birnstiel, 1981). In this context it is worth noting that the *napA* gene has several A/T-rich clusters downstream of the poly(A) addition site. As an alternative, these could fulfill a function as terminator signals.

The determination and analysis of the nucleotide sequence of the *napA* gene have revealed features which we suggest may be related to gene regulation. Still, an increased understanding of gene regulation in the case of napin will undoubtedly have to await data regarding (*a*) co-regulated genes (*e.g.* cruciferin (Simon *et al.*, 1985)), (*b*) a functional definition of the cis sequences by *in vitro* mutagenesis and transformation studies, (*c*) a definition of transacting factors either by the study of regulatory mutants or by studying DNA binding proteins, and (*d*) studies on how the abscissic acid response is mediated. The isolation and characterization of the napin gene described in this paper facilitate studies aimed at solving some of these questions.

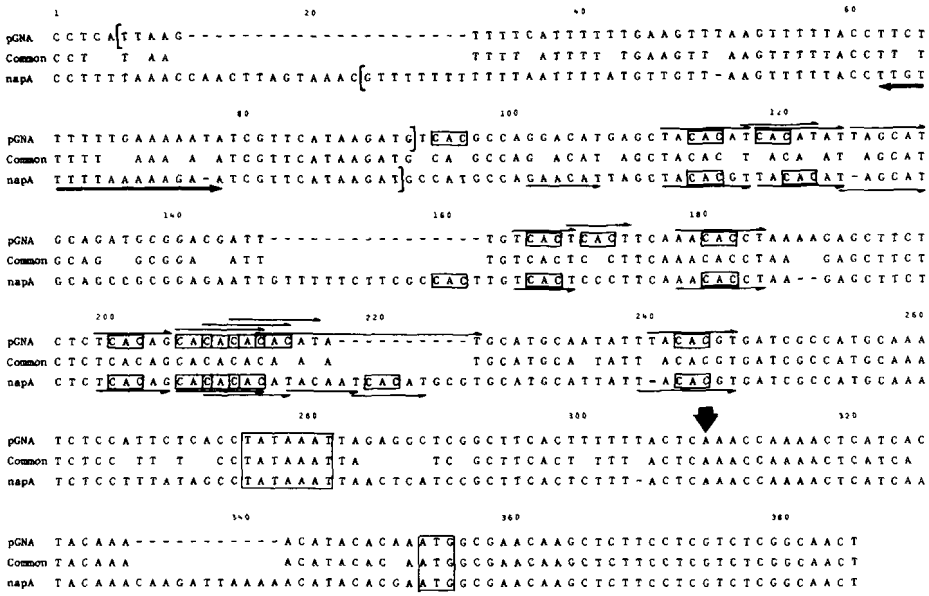

FIG. 7. **Alignment of the *napA* promoter region and the promoter region of the pGNA napin gene.** The nucleotide sequences of the promoter regions of *napA* and the pGNA napin gene were aligned by use of the ALIGN program (Dayhoff *et al.*, 1979) run with the UN matrix, a break penalty of 2 and 100 random runs. CAC trinucleotides are *boxed* and perfect or degenerate versions of the TACACAT repeats are indicated by *arrows*. The TATA box and initiation ATG are *boxed* for reference. The major transcription cap site is indicated by an *arrow*. *Brackets* at the 5′ end encompass sequences with a tendency to form hairpin loops.

Acknowledgment—Dr. Steve R. Scofield is gratefully acknowledged for making his sequence of the pGNA napin gene available to us prior to publication.

REFERENCES

Bankier, A. T., and Barrell, B. G. (1983) in *Techniques in Nucleic Acid Biochemistry* (Flavell, R. A., ed) Vol. 85, pp. 1–73, Elsevier Scientific Publishers Ltd., Limerick, Ireland

Birnstiel, M., Busslinger, M., and Strub, K. (1985) *Cell* **41,** 349–359

Breathnach, R., and Chambon, P. (1981) *Annu. Rev. Biochem.* **50,** 349–383

Courey, A. J., and Wang, J. C. (1983) *Cell* **33,** 817–829

Crouch, M. L., and Sussex, I. M. (1981) *Planta (Berl.)* **153,** 64–74

Crouch, M. L., Tenbarge, K. M., Simon, A. E., and Ferl, R. (1983) *J. Mol. Appl. Genet.* **2,** 273–283

Dayhoff, M. O., Schwarts, R. M., and Orcutt, B. C. (1979) in *Atlas of Protein Sequence and Structure* (Dayhoff, M. D., ed) Vol. 5, Suppl. 3, pp. 345–362, National Biomedical Research Foundation, Washington, D. C.

Ericson, M. L., Rödin, J., Lenman, M., Glimelius, K., Josefsson, L.-G., and Rask, L. (1986) *J. Biol. Chem.* **261,** 14576–14581

Frischauf, A.-M., Lehrach, H., Poustka, A., and Murray, N. (1983) *J. Mol. Biol.* **170,** 827–842

Gidoni, D., Kadonaga, J. T., Barrera-Saldana, H., Takahashi, K., Chambon, P., and Tjian, R. (1985) *Science* **230,** 511–517

Hall, B. D., Clarkson, S. G., and Tocchini-Valentini, G. (1982) *Cell* **29,** 3–5

Harland, R. M., Weintraub, H., and McKnight, S. L. (1983) *Nature* **302,** 38–43

Hentschel, C. C., and Birnstiel, M. L. (1981) *Cell* **25,** 301–313

Hohn, B. (1979) *Methods Enzymol.* **68,** 299–309

Kamlen, H., Schell, J., and Kreuzaler, F. (1986) *EMBO J.* **5,** 1–8

Larkins, B. A., and Hurkman, W. J. (1978) *Plant Physiol.* **62,** 256–263

Lönnerdal, B., and Janson, J.-C. (1972) *Biochim. Biophys. Acta* **278,** 175–183

Maniatis, T., Fritsch, E. F., and Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual,* Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Morelli, G., Nagy, F., Fraley, R. T., Rogers, S. G., and Chua, N.-H. (1985) *Nature* **315,** 200–204

Mizuuchi, K., Mizuuchi, M., and Gellert, M. (1982) *J. Mol. Biol.* **156,** 229–243

Proudfoot, N. J., and Brownlee, G. G. (1976) *Nature* **263,** 211–214

Rackwitz, H.-R., Zehetner, G., Frischauf, A.-M., and Lehrach, H. (1984) *Gene (Amst.)* **30,** 195–200

Sanger, F., Nicklen, S., and Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5463–5467

Serfling, E., Jasin, M., and Schaffner, W. (1985) *Trends Genet.* **1,** 224–230

Simon, A. E., Tenbarge, K. M., Scofield, S. R., Finkelstein, R. R., and Crouch, M. L. (1985) *Plant Mol. Biol.* **5,** 191–201

Sinden, R. R., Carlson, J. O., and Pettijohn, D. E. (1980) *Cell* **21,** 773–783

Staden, R. (1980) *Nucleic Acids Res.* **8,** 3673–3694

Staden, R. (1982) *Nucleic Acids Res.* **10,** 4731–4751

Weiher, H., König, M., and Gruss, P. (1983) *Science* **219,** 626–631

Yanisch-Perron, C., Vieira, J., and Messing, J. (1985) *Gene (Amst.)* **33,** 103–119

Continued on next page.

Supplementary material to

Structure of a Gene Encoding the 1.7S Storage Protein,Napin, from Brassica napus

by

Josefsson,L-G., Lenman,M., Ericson,M.L., and Rask,L.

MATERIALS AND METHODS

Plants

B.napus seeds of a dihaploid variety of "Svensk Karat" were generously provided by Dr. Lena Bengtsson, Svalöv AB, Sweden. This rapeseed line was used throughout these studies.

Isolation of DNA

100 g quantities of etiolated, frozen leaf tissue were homogenized along with solid CO_2 in a Waring blender. When the powder was starting to thaw, 100 ml of 50 mM Tris-HCl, pH 8.0/ 10 mM EDTA/1% SDS (sodium dodecyl sulphate) were added and the suspension incubated at 60°C for 20 min with gentle agitation. This was followed by two gentle extractions with a 24:1 mixture of chloroform:isoamylalcohol. The aqueous phase was retained and the DNA precipitated . The precipitate was collected by centrifugation, rinsed with 70% ethanol, dried lightly and resuspended in 8 ml of TE (10 mM Tris-HCl, pH 7.5/1 mM EDTA). RNA was degraded by an incubation for 30 minutes at 37°C after the addition of RNAse A to a final concentration of 100 µg/ml. 0.2 volumes of a solution containing 50 mM Tris-HCl, pH 7.5/0.4 M EDTA/0.5% SDS/1 mg/ml proteinase K were then added and the mixture incubated at 50°C for 30 minutes. The mixture was then extensively dialysed against TE. After dialysis the DNA was extracted twice with chloroform:isoamylalcohol and then precipitated. The DNA was rinsed with 70% ethanol, lightly dried and resuspended in TE. This procedure yielded easily restrictable DNA with a mean size of some 70 kb as determined by low voltage electrophoresis in a soft agarose gel.

Southern blotting

10 µg portions of rapeseed DNA were digested to completion with different restriction enzymes and loaded on 0.7% agarose gels run with the TBE (Tris/Borate/EDTA) buffer system (Maniatis et al 1982). After light staining with ethidium bromide the gel was immersed in 0.25 M HCl for 5 minutes. After the depurination the DNA in the gel was denatured and transferred to nitrocellulose filters as described (Maniatis et al,1982).The subsequent treatment of the filters was also according to Maniatis et al (1982).

Isolation of mRNA and Northern blotting

mRNA was isolated as described by Ericson et al(1986). Denaturing agarose gels were prepared and run according to Maniatis et al(1982). 2 µg of denatured mRNA were loaded on a 1% agarose/formaldehyde gel and subjected to electrophoresis. Transfer of the mRNA to nitrocellulose filters and the subsequent treatment of the filters was according to standard procedures (Maniatis et al, 1982).

Nick-translation and hybridization to Southern blots, Northern blots and screening filters

0.1-0.2 µg portions of pNAP1 cDNA were nick-translated to obtain radioactively labelled probe. Prehybridizations and hybridizations were done with formamide-containing solutions according to standard protocols (Maniatis et al,1982). Washing of filters was done at high stringency, i.e. 3 mM sodium citrate-HCl, pH 7.1/30 mM NaCl/0.5% SDS at 65°C, two times 1 h. Filters were exposed on X-ray film with intensifying screens at -70°C.

Construction of genomic library and screening for napin clones

Rapeseed DNA (120 ug) was partially degraded with MboI under conditions that predominantly yielded fragments in the size range of 15-25 kb. DNA molecules of this size class were further purified by fractionation on 5-40% sucrose gradients in 1M NaCl that were run for 6 h at 39,000 rpm in a Beckman SW40 rotor. The fractions containing 15-25 kb DNA were pooled, the DNA precipitated, resuspended and phosphatase treated to further reduce the risk of insert concatemerization during ligation. After removal of the phosphatase by extraction the DNA was precipitated, pelleted, rinsed and resuspended in TE to a concentration of 1.25 µg/µl. The EMBL3 vector DNA was double cleaved with BamHI and EcoRI and the small linker-pieces removed by isopropanol precipitation of the DNA (Frischauf et al, 1983). It was resuspended to a concentration of 0.4 µg/µl. Extracts for packaging of phage lambda in vitro were prepared according to Hohn (1979).

Conditions of ligation and packaging of lambda particles in vitro were investigated on a small scale prior to preparation of the library. The conditions finally chosen for the library construction were the following: 4 ug of vector and 12.5 ug of 15-25 kb insert DNA were ligated and packaged in vitro after dividing into 10 mixes containing each: 35 µl buffer A (20 mM Tris-HCl, pH 8.0/5 mM $MgCl_2$/0.05% β-mercaptoethanol/1 mM EDTA); 1.65 ug DNA (in 10 µl); 5 µl Q1 mix (6 mM Tris-HCl, pH7.5/60 mM spermidine/18 mM $MgCl_2$/15 mM ATP/0.2% β-mercaptoethanol); 17.5 µl sonication extract and 25 µl freeze/thaw lysate. This yielded a total of $2.2x10^6$ plaque forming units (pfu). The library was subsequently applified on 10 large screening plates of 23x23 cm. Some $1.5x10^5$ clones of the amplified library were screened by spreading $7.5x10^4$ pfu on each screening plate. Replica nitrocellulose filters were prepared, and pretreated as described (Maniatis et al,1982).Conditions for hybridization are given above. Recombinant phages were purified by two consecutive rescreenings on regular 10 cm LB/agar plates. Growth of recombinant phages and purification of phage DNA were according to established protocols (Maniatis et al,1982)

Mapping of genomic clones and subcloning

Lambda recombinant clones were mapped by combining the procedure of Rackwitz et al (1984) with a set of complete digestions with either SalI alone or with SalI together with either of six other restriction enzymes. Southern blots from gels on which the latter digestions were analyzed were prepared and hybridized with labelled pNAP1 cDNA. Subcloning of a fragment containing the napA gene was done by purification of the fragment on an agarose gel cast with low gelling temperature agarose. The purified fragment was subcloned into pUC19 (Yanisch-Perron et al, 1985).The subclone was mapped by conventional digestion/double digestion techniques (Maniatis et al,1982).

Nucleotide sequencing

Nucleotide sequencing was performed according to Sanger et al (1977) with [^{35}S] α-thio-dATP as the labelled nucleotide. The M13 vectors used were mp18 and mp19 (Yanisch-Perron et al,1985).Both the shotgun procedure (Bankier and Barrel,1983) and directed subcloning into M13 of fragments derived by restriction enzyme digestion were used. The sequence data was handled by use of the DB system (Staden,1980;Staden,1982).

Mapping of the transcription start site

An 18-mer oligonucleotide, 5'AGGAAGAGCTTGTTCGCC 3',was [^{32}P] end-labelled with polynucleotide kinase (Maniatis et al,1982) Approximately 0.2 pmoles (70.000 cpm) of the labelled oligonucleotide were added to 1 ug of mRNA in a 10 µl mix that in addition contained: 35 u Human placental RNAse inhibitor; 34 mM Tris-HCl, pH 8.3 (measured at 42°C); 25 mM NaCl and 6 mM $MgCl_2$. After annealing for 1 h at room temperature, unlabelled dNTPs to a final concentration of 200 µM each and 1.5 units of AMV reverse transcriptase were added. The sample was then incubated at 42°C for 20 min and subsequently treated as a regular sequencing gel sample. Approximately 1 µl of the mixture (5000 cpm) was loaded onto the gel and run alongside a reference set of sequencing reactions.

Databases

The three major data bases (NBRF, EMBL and GENBANK) were used in the sequence comparisons.

RESULTS

Southern and Northern blotting analyses

As an initial step towards defining the complexity of the rapeseed genome with regard to napin genes we decided to use pNAP1, a cDNA clone which encodes napin (Ericson et al,1986), as a radioactive probe in Southern blotting analyses. 10 ug portions of total rapeseed DNA were in separate reactions digested to completion with four different restriction enzymes. Following separation of the generated DNA fragments on agarose gels, the fragments were denatured and transferred to nitrocellulose filters. Hybridization to the filters of nick-translated pNAP1 cDNA yielded the pattern shown in Figure 1. The different enzymes yielded between 8 and 13 hybridizing bands. Since it is not known to what extent the enzymes may cut within individual napin genes, there is no way of deducing an exact gene number. Nevertheless, considering the data as a whole it appears reasonable to assume that there are in the order of 10 genes for napin. How many of these hybridizing bands that represent expressed napin genes is at present not clear.
Irrespective of the fact that several genes may be expressing napin, one well defined, major napin mRNA species was evident when rapeseed embryonal mRNA was subjected to Northern blotting with the cDNA probe (Figure 2). In addition to the major 850 nucleotides transcript, a diffuse population of RNA species is also evident. This ranges in size from approximately 900 to 1500 nucleotides, and as a whole constitutes quite a significant fraction of the total hybridizing material. We cannot at present determine whether these larger RNAs represent a vast population of differently polyadenylated species of napin transcripts or simply are contaminating hnRNA which has not yet been polyadenylated. In light of the fact that polyadenylation appears to be a much more site specific process compared to that of transcriptional termination (Birnstiel et al,1985) we favour the latter explanation.

Isolation and restriction mapping of napin genomic clones

A genomic phage library was constructed with DNA from a dihaploid line of B.napus. Screening of $1.5x10^5$ recombinants with the pNAP1 cDNA clone as the probe yielded eight positive clones. DNA was prepared from these clones after they had been purified by two consecutive rescreenings.Mapping of the genomic clones showed that four of the positive recombinants were overlapping clones containing the same gene, which we have designated napA. Figure 1 displays the restriction map of this region, as well as the individual clones that cover the region. A 3.3 kb HindIII - BglII fragment hybridizing to the pNAP1 cDNA probe was subcloned into plasmid pUC19 (Yanish-Perron et al,1985),and further mapped by conventional techniques (Maniatis et al,1982). Figure 4 shows the map that was obtained and a comparison with the pNAP1 cDNA restriction map.

It has been shown in other plant gene systems that the cis signals involved in regulating transcriptional initiation usually are contained within sequences that are located reasonably close to the transcribed part of the gene (Kamlen et al,1986;Morelli et al,1985).Thus, we considered it likely that all the linked sequences involved in transcriptional regulation were contained in this subclone and consequently decided to sequence the whole insert of the subclone.

Sequencing of the napA gene

The entire sequence of the 3.3 kb fragment was determined in overlapping sequence reactions on both strands by a combination of "shotgun" sequencing and sequencing of individual, restriction enzyme-derived M13 subclones. Both the universal 17-mer sequencing primer and synthetic oligonucleotides (18-mers) complementary to sequences within the subclones were used to obtain the complete sequence. The sequencing strategy is represented in a schematic fashion below the restriction map in Figure 4. This represents a minimal estimate of sequence data that were collected. Sequences that were well represented in the "shotgun" clones, the transcribed region in particular, were determined with a lot higher frequency than is apparent from the figure. In addition, many individual reactions were performed more than once.

Mapping of the initiation site for transcription

The transcription cap-site of napin mRNA was determined by mRNA directed primer extension. A synthetic oligonucleotide, complementary to mRNA sequences close to the initiation ATG, was [^{32}P] end-labelled, annealed to mRNA and subsequently elongated to the 5' end of napin mRNAs by the incorporation of unlabelled nucleotides mediated by AMV reverse transcriptase. Figure 5 shows the elongated and terminated primer alongside the sequence reactions obtained by letting the same oligonucleotide, unlabelled in this case, prime sequencing reactions on an M13 shotgun clone that covered this region on the minus strand. When mapped onto the sequence of the napA gene the major initiation site is at the A in position 1102. The minor bands correspond to positions 1098, 1112 and 1113. Thus, the major site of transcriptional initiation appears to be located 33 nucleotides downstream from a sequence which conforms to the consensus of a TATA box (see below).

General features of the sequence

Figure 6 shows the sequence of the 3295 nucleotides of the HindIII - BglII subclone insert. The translated sequence of the coding region is also shown above the nucleotide sequence in one letter code. The sequence that is contained in the pNAP1 cDNA clone (Ericson et al,1986) is shown within brackets, and is absolutely identical to that of napA. In this context it is worth noting that the pNAP1 cDNA clone was isolated from a cDNA library constructed from the same dihaploid rapeseed line that was used for these studies. A thick arrow indicates the major transcription start site. This is preceded by an enboxed TATA conforming sequence (Breathnach and Chambon,1981). A dotted line shows an imperfect CAT box (Breathnach and Chambon,1981) which is (if it is at all functional) located unusually close to the TATA box. On the 3' side of the coding region one poly A addition signal (Proudfoot and Brownlee,1976) is found (underlined by a solid line). A dot above nucleotide 1850 indicates the actual site where the poly A tail is added, as deduced from a comparison with the pN1 and pN2 cDNA clones (Crouch et al, 1983). Figure 6 also shows a second set of TATA/poly A

addition signals (enboxed/underlined) at nucleotides 2653 and 2931, respectively. We presently do not know whether this part of the sequence represents an expressed portion of the genome. Considering the size of a hypothetical transcript and the relative positions of ATG's and termination codons within it, we think it is less likely that this sequence is expressed (at least at the protein level). However, we are presently settling this point by a direct examination.
On the opposite strand there are several TATA boxes and polyA addition signals. An A/T-rich sequence between positions 3078-3117 would on the opposite strand correspond to 6 closely spaced polyA addition signals. One additional polyA addition signal occurs at position 1963 (plus strand numbering). Towards the 3' end of the minus strand three TATA boxes occur at 695, 753 and 783 (plus strand numbering), the two first of which are part of a 14 bp direct repeat. Slightly further downstream on the minus strand are two additional polyA addition signals (306 and 388; plus strand numbering). Although we seriously doubt whether any of the above sequences constitute functional signals, we can at present not strictly rule it out.

Hairpins, repeats and palindromes

The major direct and inverted repeats of the sequence are indicated by arrows, pairwise connected as indicated by the lettering. Solid arrows indicate perfect, dotted arrows imperfect repeats. The E repeat has been observed and discussed previously (Ericson et al,1986). In addition to the repeats shown in Figure 6, the region 3078-3117 has several overlapping direct repeats. A two-headed arrow indicates a slightly imperfect palindromic sequence present some 210 nucleotides upstream of the TATA box. Two different regions, nucleotides 824-889 and 2139-2208, display features which appears to result in quite a strong tendency to form hairpins (and possibly cruciform structures). Several other regions in the sequence exhibit some tendency to form hairpins. However, the unique feature of the regions that are discussed here is that within a rather short stretch of nucleotides (60-70) several different hairpin structures can be generated simply by sliding the hairpin arms relative to each other. The above sequences are able to form 6 and 7 different hairpin structures, respectively, with between 9 and 15 base pairs in the stems (data not shown). With regard to the former of these sequences a corresponding region in the pGNA napin gene (S.Scofield; personal communication) is, although the sequence differs somewhat from napA, able to form 5 different alternative structures with between 9 and 13 base pairs in the stem. In addition, these regions in napA and in the pGNA gene are both very A/T-rich. This strengthens the suggestion that the sequences may be involved in a local perturbation of the DNA structure.The sequences involved are shown within brackets in Figure 7. Both of the sequences in the napA gene are rather suggestively placed, one (positions 824-889) upstream and the other (positions 2139-2208) downstream from the transcribed part of the gene. No doubt, the possibility to form several alternative hairpins could be of importance in stabilizing a non B-DNA structure, particularly if the regions are under negative superhelical stress (Mizuuchi et al,1982). However, it has been argued that the kinetics of cruciform formation may restrict the importance of such reactions in vivo (Courey and Wang,1983). The question whether transcriptional activation of eukaryotic chromatin is at all influenced by torsional stress in vivo is also controversial. Data favouring both opinions have been put forward (Harland et al,1983;Sinden et al,1980).
Another interesting feature of the promoter region is that within a sequence 30-160 base pairs upstream of the TATA box the trinucleotide CAC is repeated 11 times. Most copies of the CAC trinucleotide occur as part of a tandemly repeated, degenerate heptamer, which in turn is repeated 12 times. Figure 7 shows the region of interest with the sequences of napA and the pGNA napin gene aligned to maximize the homologies. The CAC trinucleotides and the degenerate repeats are indicated in the figure. The consensus sequence of the repeat considering all the different copies is TACACAT. The TATA box and initiation ATG are enboxed for reference. The major cap-site is also indicated by an arrow.

Comparison with other nucleotide sequences

A search with the napA 5' region sequence against the three major data bases as well as against recently published (and not yet entered) sequences of some storage protein genes from other species failed to reveal any features that we could tentatively identify as being related to gene regulation. We were also unable to find sequences in napA related to SV40 enhancer core sequences (Weiher et al,1983) unless allowing for 3 or more mismatches.

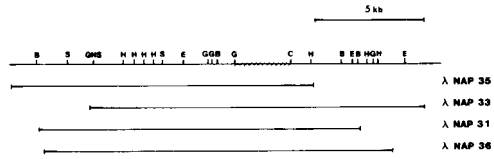

Figure 3: Restriction map of the genomic region containing the napA gene. Individual lambda recombinant clones were mapped as described in Materials and Methods. The figure shows the map of the genomic region and the parts contained in different recombinants. The measuring bar corresponds to 5 kb of DNA. The enzymes used were B=BamHI; C=SacII; E=EcoRI; G=BglII; H=HindIII; S=SalI and N=NruI. The hatched area indicates the part that hybridized to pNAP1.

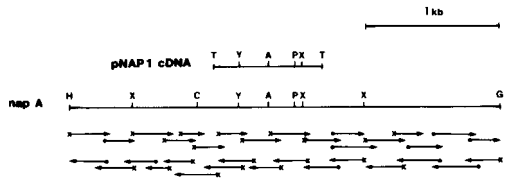

Figure 4: Restriction map of napA and sequencing strategy. The 3.3 kb HindIII - BglII subclone in pUC19 was mapped with conventional techniques. The figure shows the map obtained for the insert and how it compared to the map previously obtained for pNAP1 cDNA. Measuring bar corresponds to 1 kb of DNA. The enzymes used were: A=SacI; C=SacII; G=BglII; H=HindIII; P=ApaI; T=PstI; X=XhoI and Y=NaeI. Below the map is a schematic representation of the sequencing strategy, as discussed in the text. X denotes reactions primed by the universal 17-mer primer on either shotgun clones or restriction enzyme derived M13 subclones. • denotes reactions primed by synthetic 18-mer primers within different subclones.