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 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 oleracea. 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 indicate that the napA gene most likely is derived from B. oleracea.
Structure of a Napin Gene

12197

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 pNAPl cDNA was used as a probe in hybridization to these filters. The enzymes used were B, BamHI; E, EcoRI; H, HindIII; and P, PvuII. The size marker (M) used was an end-labeled BstEII digest of phage X 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 pNAPl 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 HindIII 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 codon, was synthesized. This synthetic oligonucleotide, 32P 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
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 has 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., Buslinger, M., and Strub, K. (1985) Cell 41, 349-359

Breathnach, R., and Chambon, P. (1981) Annu. Rev. Biochem. 50, 499-524

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

Crouch, M. L., and Sussex, I. M. (1981) Planta (Bert.) 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., Schwartz, R. M., and Orcutt, B. C. (1979) in Atlas of Protein Sequence and Structure (Dayhoff, M. O., ed) Vol. 5, Suppl. 3, pp. 345-362, National Biomedical Research Foundation, Washington, D. C.

Ericson, M. L., Rodin, J., Lenman, M., Giimelius, K., Josefson, 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

Kamien, 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

Mizunuki, K., Mizunuki, 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

Scherling, E., Jasim, M., and Schaffner, W. (1985) Trends Genet. 1, 224-230

Simon, A. E., Tenbarge, K. M., Scofield, S. R., Finkeinstein, R. R., and Crouch, M. L. (1985) Plant Mol. Biol. 8, 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, 4731-4731

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 Napin Gene**

**Encodmg the 1.75 Storage Protein, Nap 1, from**

**Rape Butter**

**Johnefson, L.-G., Levanon, H., Ericson, H.-M., and Bass, I.**

**MATERIALS AND METHODS**

**Flasks**

* L.1 mass seeds of a diploid variety of "Svensk Karat" were generously provided by Dr. Lars Bengtsson, NAL, Sweden. This rape seed was used throughout these studies.

**Isolatlon of DNA**

100-g quantities of etiolated, frozen leaf tissue were homogenized along with solid CO₂ in a Waring Blender. The resulting slurry was filtered through cheesecloth, and the suspension precipitated at 6°C for 20 min with sodium sulfate. The DNA was then extracted with 0.4 N NaOH and the suspension incubated at 60°C for 20 min with sodium dodecyl sulfate. The DNA was then precipitated with 2.2 volumes of aqueous ethanol. The DNA was then rinsed lightly and resuspended in TE (1 mM Tris-Cl, pH 7.5, 0.1 mM EDTA). The volume of a solution containing 1 mg/ml DNA was adjusted to 10 ml with TE. Washing of filters was done with 1.5 ml/plate of 0.1% SDS, 1 M NaCl, 0.1 M Tris-Cl, pH 7.5, and 20 mM EDTA, pH 8.0. DNA was resuspended in 2 ml of TE (1 mM Tris-Cl, pH 7.5, 0.1 mM EDTA) and digested with HindIII restriction enzyme. The DNA was then treated to further digestion with either HindIII or BamHI restriction enzymes. The digestion products, DNA digests, were isolated by electrophoresis on 10 large gels. Gel sections containing each fragment and their positions were excised and cut out into 10 ml volumes of TBE (0.89 M Tris-Cl, pH 8.3, 0.89 M boric acid, pH 8.3) and 0.02 M EDTA. The gels were incubated at 65°C for 30 min. After further digestion with HindIII, the DNA was denatured and transferred to nitrocellulose filters. Hybridization was done with nick-translated pNAPl DNA yielded the pattern shown in Figure 5. This pattern contains positive bands. The gel was closed in microtromex in a soft agarose gel.

**Southern blotting**

Cloning of pNAPl DNA was digested to completion with different restriction enzymes (e.g., HindIII) and loaded on 0.7% agarose gel (Maniatis et al., 1982). After further digestion with HindIII, the DNA was denatured and transferred to nitrocellulose filters. Hybridizations were done in a solution containing 50% formamide, 0.4 M NaCl, 10 mM Tris-Cl, pH 7.5, 50 mM EDTA, 0.1% SDS, 0.1 mg/ml poly d(A-T)20, and 0.1 mg/ml poly d(C-G)20. The filters were washed at 5°C for 30 min. After further digestion with HindIII, the DNA was denatured and transferred to nitrocellulose filters. Hybridizations were done in a solution containing 50% formamide, 1 M NaCl, 10 mM Tris-Cl, pH 7.5, 50 mM EDTA, 0.1% SDS, 0.1 mg/ml poly d(A-T)20, and 0.1 mg/ml poly d(C-G)20. The filters were washed at 5°C for 30 min. After further digestion with HindIII, the DNA was denatured and transferred to nitrocellulose filters. Hybridizations were done in a solution containing 50% formamide, 1 M NaCl, 10 mM Tris-Cl, pH 7.5, 50 mM EDTA, 0.1% SDS, 0.1 mg/ml poly d(A-T)20, and 0.1 mg/ml poly d(C-G)20. The filters were washed at 5°C for 30 min. After further digestion with HindIII, the DNA was denatured and transferred to nitrocellulose filters. Hybridizations were done in a solution containing 50% formamide, 1 M NaCl, 10 mM Tris-Cl, pH 7.5, 50 mM EDTA, 0.1% SDS, 0.1 mg/ml poly d(A-T)20, and 0.1 mg/ml poly d(C-G)20. The filters were washed at 5°C for 30 min.
addition signals (boxed/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 TATA and transcription start sites, we tentatively identify it as being related to gene regulation. We also note the presence of two slightly imperfect palindromic sequences (210 bases in total). The putative transcription start sites, 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 tentatively 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 G repeat has been observed and discussed previously (Ericson et al., 1982). In addition to the repeats shown in Figure 4, the region 3078-3117 has several slightly imperfect, palindromic sequences some 2.5 kb upstream of the TATA box. The different regions, nucleotides 214-489 and 215-220B, display features which appear 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 (600-700) several different hairpin structures are generated simply by shifting the hairpin some relative to each other. To our knowledge this is the first time different hairpin structures, respectively, with between 9 and 15 base pairs in the stems idea shown, with regard to the former of these sequences a corresponding region in the psa1 gene (psm11/psm12; personal communication) is, although the sequence differs somewhat from psa1, able to form 3 different alternative structures with between 9 and 15 base pairs in the stems. For the latter sequences the same holds true but the formation of the psa1 gene are both very A/T-rich. This strengthens the suspicion that such sequences may be involved in a local perturbation of the DNA structure. The sequences involved are shown within brackets in the plus strand. In the regions of the psa1 gene are rather repetitively placed; one (nucleotides 214-489), upstream and the other one (nucleotides 215-220B) downstream from the transcribed part of the gene. The possibility to form several alternative hairpin structures could be due to the presence of direct repeats, particularly if the regions are under negative superhelical stress (Auchtuna, 1983). However, it has been argued that the kinetics of cruciform formation may restrict the importance of cruciform reactions (Chory and Wang, 1983). The question whether transcriptional activation of superhelical DNA is involved in the generation of the new transcriptional stress is not clear. The sequences of the TATA box, the transcriptional start site and the nucleotides 215-220B have been numbered (as indicated in Figure 4 and Figure 7). Figure 7 shows the region of interest with the sequences of psa1. The psa1 gene is aligned to explain the region was described (see Figures 4 and 7). The repeats are indicated in the figure. The consensus sequence of the repeats does not contain the different regions TATAAT. The TATA box and initiation K04 are boxed for reference. The major repeat is also indicated by an arrow. 

Comparison with other nucleotide sequences

A search with the psa1 5' region sequence against the three major data bases as well as against recently published (and yet not indexed) sequences of some storage protein genes from various plant species did not reveal any similarity. This implies that the regulatory sequences of the psa1 gene are not generally conserved with other storage protein genes.