Volume 11 Number 20 1983 Nucleic Acids Research

Nucleotide sequence of soybean chloroplast DNA regions which contain the *psb* A and *trn* H genes and cover the ends of the large single copy region and one end of the inverted repeats

Albert Spielmann and Erhard Stutz*

Laboratoire de Biochimie, Université de Neuchâtel, Chantemerle 18, CH-2000 Neuchâtel, Switzerland

Received 11 August 1983; Revised and Accepted 26 September 1983

ABSTRACT

The soybean chloroplast psb A gene (photosystem II thylakoid membrane protein of Mr 32 000, lysine-free) and the trn H gene ($tRNA^{His}_{GUG}$), which both map in the large single copy region adjacent to one of the inverted repeat structures (IR1), have been sequenced including flanking regions. The psb A gene shows in its structural part 92% sequence homology with the corresponding genes of spinach and N. debneyi and contains also an open reading frame for 353 aminoacids. The aminoacid sequence of a potential primary translation product (calculated Mr, 38 904, no lysine) diverges from that of spinach and N. debneyi in only two positions in the C-terminal part. The trn H gene has the same polarity as the psb A gene and the coding region is located at the very end of the large single copy region. The deduced sequence of the soybean chloroplast $tRNA^{His}_{GUG}$ is identical with that of Zea mays chloroplasts. Both ends of the large single copy region were sequenced including a small segment of the adjacent IR1 and IR2.

INTRODUCTION

We have recently mapped the psb A gene on the soybean (Glycine max.) chloroplast genome in close vicinity of one of the inverted repeat regions (1). This gene codes for the so-called Mr 32 000 rapidly labeled photosystem II thylakoid membrane protein (2,3) which may be involved in the binding of urea and triazine herbicides (4,5). Zurawski et al. (6) sequenced the psb A gene region from Spinacia oleracea and Nicotiana debneyi chloroplast DNA and they observed in both cases identical open reading frames for 353 aminoacids equivalent to a protein of Mr 38 950. The size of this primary translation product surpasses the size of the rapidly labeled thylakoid membrane protein (Mr 32 000 to 36 000); the authors therefore argued that translation might start at a second ATG in the same reading frame, reducing thereby the size of the translation product by about 4 000. On the other hand, the total sequence identity of the reading frame in both types of chloroplast DNA strongly suggested that the entire reading frame must be functional.

In view of these open questions and considering the functional importance of this gene product it seemed warranted to sequence the psb A region of the soybean chloroplast genome. Representatives of the legume family are known to contain chloroplast genomes which have undergone relative to other angiosperm chloroplast genomes considerable DNA rearrangements (7,8). Therefore, size and fine anatomy of the soybean chloroplast psb A gene region might be different from those reported and answer some of the open questions.

Swamy and Pillay (9) recently identified soybean chloroplast $tRNA^{His}$ without, however, mapping the corresponding gene. In case of spinach the trn H gene maps between the psb A gene and the inverted repeat (10). We included in our sequence studies the corresponding DNA segment and furthermore determined the beginning of the inverted repeats by sequencing the corresponding DNA region on the other side of the large single copy region. This allowed to exactly position both the psb A and trn H gene relative to one of the inverted repeats, which are structural hallmarks of most higher plant chloroplast genomes.

MATERIALS AND METHODS

Isolation of soybean chloroplast DNA, restriction sites analysis and mapping of the psb A gene on the circular chloroplast genome have been described (1). For a more detailed restriction site mapping and sequencing of the relevant region, we cloned HindIII fragments of total chloroplast DNA into pBR322 and tested the clones with both, a nick-translated (11) HpaII fragment obtained from the clone pSoc B511 (1,3) which carries 330 nucleotides of the spinach chloroplast psb A gene (probe a) and a HpaII-SalI fragment (850 bp) which carries the 3' end of the psb A gene including about 200 bases of the adjacent inverted repeat (probe b). A HindIII fragments of 2.8 kb (HindIII-J) interacted only with probe a, a HindIII fragment of 1.4 kb (HindIII-0) interacted with probe a and b and a HindIII fragment of 9.0 kb (HindIII-B) interacted only with probe b. Further mapping experiments allowed to place the three HindIII fragments as shown in Fig. 1A,D. For the sequencing experiments we used the entire fragments HindIII-0, the HindIII-SmaI subfragment of HindIII-J (0.8 kb) and a HindIII-PstI subfragment (3.9 kb) of HindIII-B (Fig. 1,A,D). DNA fragments were isolated and purified as described (12). For DNA sequencing we used both current methods (13,14) and as specified (12). Enzymes were purchased from Boehringer-Mannheim and New

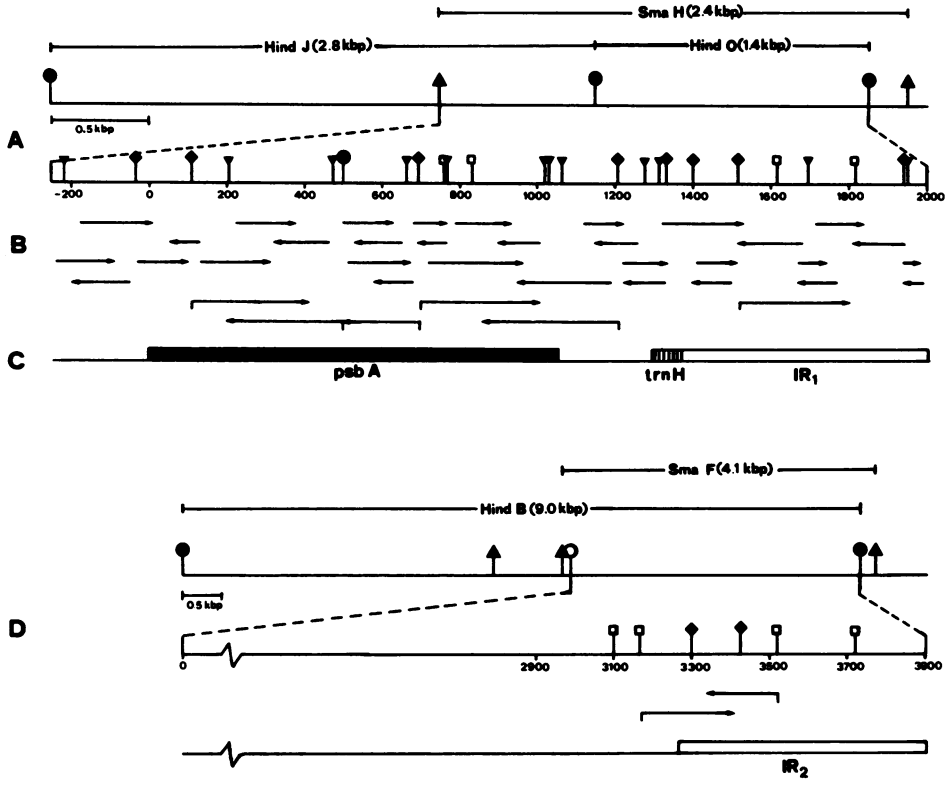


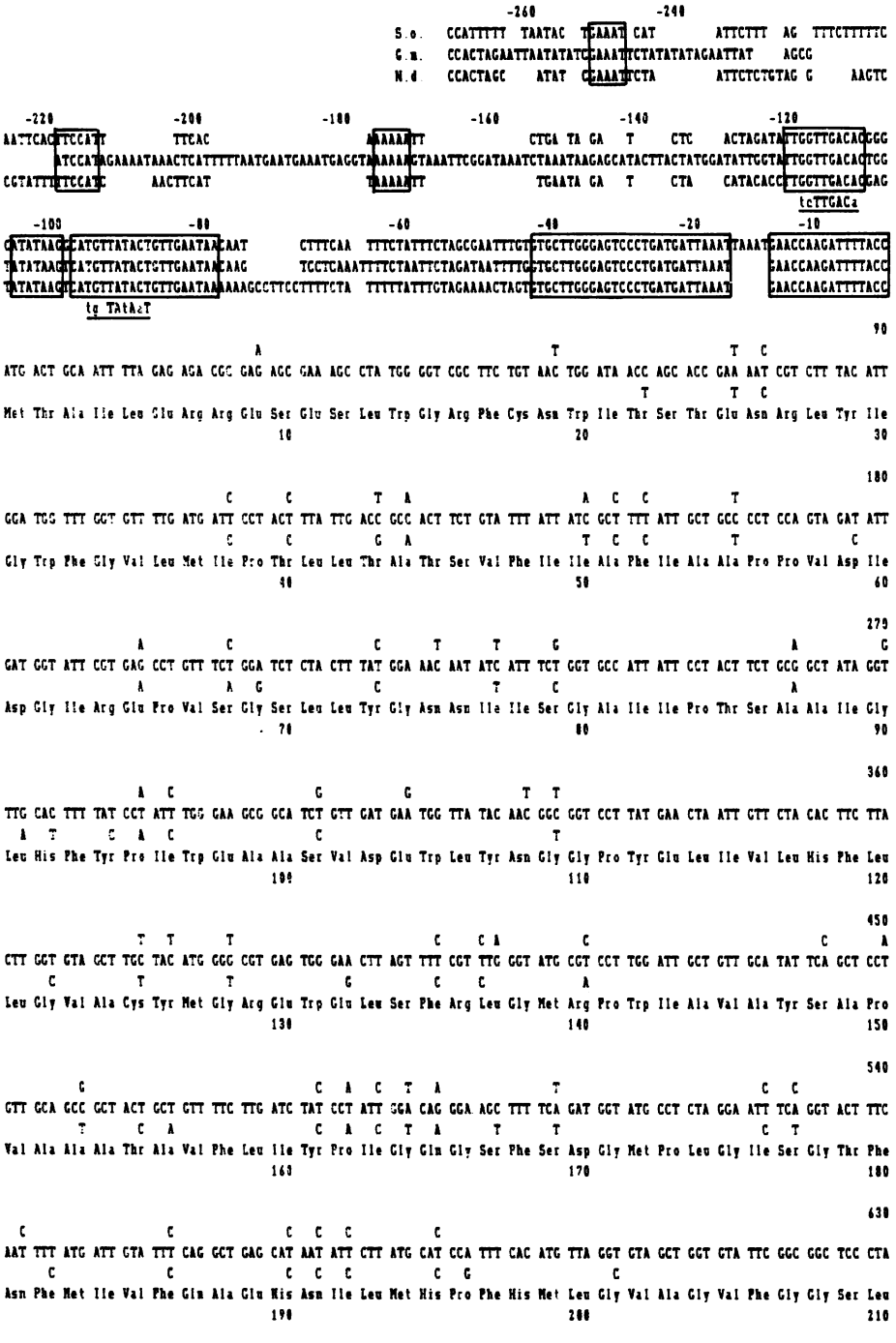
Fig. 1. Restriction sites map and strategy used to sequence the psb A and trn H gene region of the soybean chloroplast DNA. A. Restriction sites of SmaI-H fragment (1). B. Arrows represent portions of the 5' end labeled fragments from which unambiguous sequences could be established (13); ⟶ RNA-like strand, ⟵ coding strand; arrows with a short vertical line indicate regions sequenced according to (14). C. Location of structural parts of the psb A gene, trn H gene and parts of the inverted repeat [IR1]. D. Restriction sites map of SmaI-F fragment and position of parts of the inverted repeat [IR2]. HindIII, SmaI, Sau3A, HinfI, HaeIII, PstI. Numbers on scales in A and D refer to nucleotide sequence position given in Figs 2 and 3.

England Biolabs and used following the instructions of the supplier. ^{32}P-ATP was from Amersham.

RESULTS

1. Sequence analysis of the psb A gene

In Fig. 1 we show a stretch of the soybean circular chloroplast DNA which carries the psb A gene. In Fig. 2 we give the nucleotide sequence of the



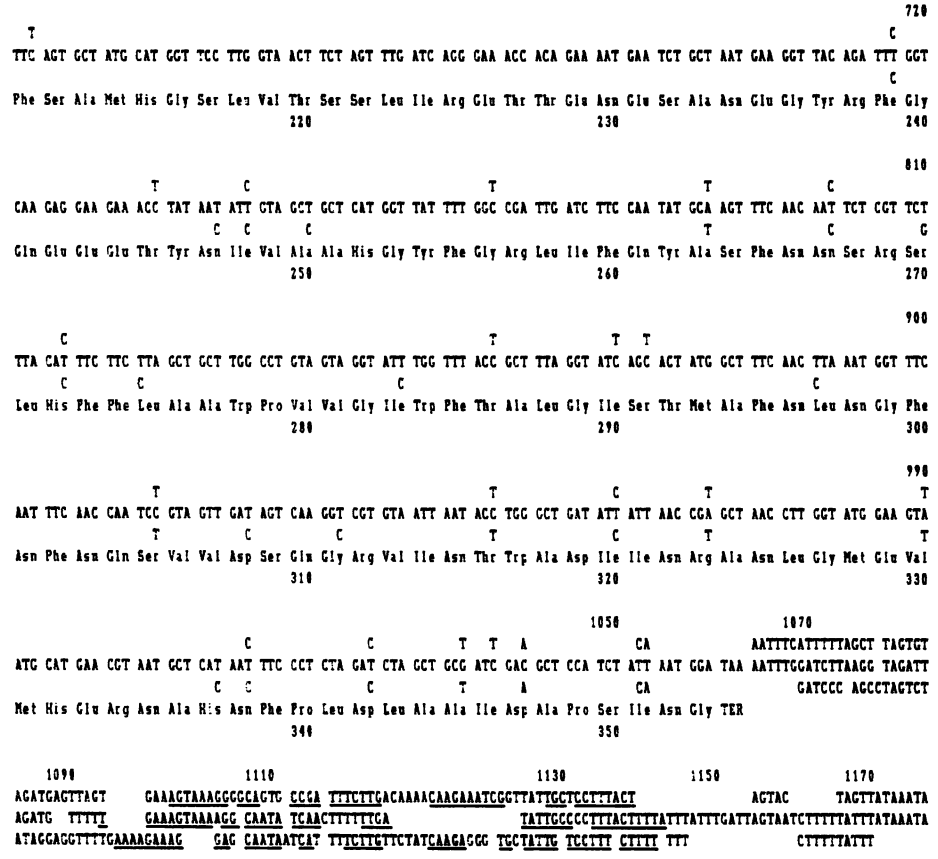


Fig. 2. Nucleotide sequence of the psb A gene and flanking regions. Only the RNA like strand is given starting with the 5' position. Aligned with the soybean chloroplast psb A gene region (G.m.) are the corresponding sequences of Spinacia oleracea (S.o.) and Nicotiana debneyi (N.d.); within the structural part only deviations from the soybean sequence are given; within the flanking parts the entire sequences are shown and aligned such as to maximize sequence homology. Homologous regions are boxed. The deduced aminoacid sequence is given; the first methionine of the open reading frame is taken as position 1. Potential regulatory sites in the 5' flanking and a potential stem-loop structure in the 3' flanking part are underlined. Prokaryotic promoter consensus sequences for the -35 and -10 region (15) are given.

psb A gene and its flanking regions along with the deduced aminoacid sequence for the structural part of the gene. For comparative reasons we add the nucleotide sequences of the corresponding parts of the S. oleracea and N. debneyi chloroplast DNA. The results of this comparative study can be summarized as follows : 1) The coding part of the soybean psb A gene is

identical in length to that of S. oleracea and N. debneyi, i.e., the reading frame is also open for a maximum of 353 aminoacids. 2) The soybean psb A gene diverges in its coding part at 73 (78) nucleotide positions from the S. oleracea (N. debneyi) gene. 3) With only three exceptions, the nucleotide differences occur in the wobble position, and only two mutations lead to a change in the aminoacid composition : aspartic acid replaces glutamic acid (position 347) and isoleucine replaces threonine (position 351). 4) Codons for lysine are absent in both cases. 5) About 130 nucleotide positions of the 5' flanking part are highly conserved, the sequence homology being in the range of 80 to 85%. This region contains potential promotor sites. Within this region the soybean sequences shares once a pentanucleotide gap with the N. debneyi sequence and once a heptanucleotide gap with the S. oleracea sequence. 6) Within the positions -130 to -271 sequence homology is very low especially due to multiple short insertion/deletions. 7) The 115 positions of the 3' flanking part are homologous, respectively, to 63% and 57% with the S. oleracea and N. debneyi counterparts. This includes gaps required for maximal sequence alignement. 8) A stem and loop structure with 21 basepairs can be formed [positions 1094 - 1115 ∩ 1121 - 1143] similar to but not identical with the stem and loop structures proposed for the corresponding sequences of S. oleracea and N. debneyi (6). They may serve as transcription termination signals.

2. Sequence analysis of the trn H gene and of a segment overlapping the large single copy region and the inverted repeats

The circular soybean chloroplast genome contains two inverted repeats which separate a large (84 kb) from a small (24 kb) single copy region (1,8). The psb A gene maps close to one of the inverted repeats (1) the identified trn H gene (9) was not located yet. We anticipated, however, that the trn H gene would map between the psb A gene and the IR1, as seen in other angiosperm chloroplast genomes (10). In order to exactly locate the relative map positions of these genes and study the fine anatomy we sequenced the gap between the psb A gene and the beginning of the IR1, including a small part of the IR2. To identify the beginning of the IR1 it was necessary to sequence the corresponding DNA segment on the other side of the large single copy region which was mapped as shown in Fig. 1D. The sequence results are given in Fig. 3. The gap between the terminator codon of the psb A coding region and the first nucleotide of the inverted repeat is 319 positions (nucleotide positions

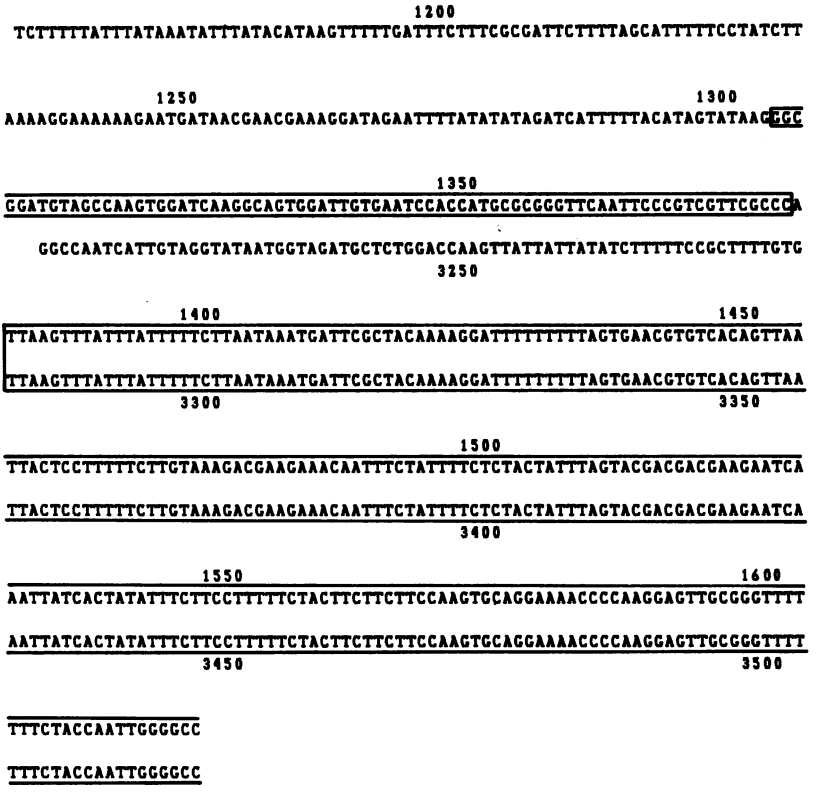


Fig. 3. Nucleotide sequence of the trn H gene region, its flanking parts and parts of the inverted repeats IR1 and IR2. Only the RNA like strand is shown. The structural part of the trn H gene and IR1 and IR2 are framed. Note that the psb A and trn H gene have the same polarity. Counting of nucleotide positions is continuous from Fig. 2, note the overlap. For IR1 the strand with the polarity of the trn H gene is given. For IR2 the sequence of the opposite strand (by definition) is given including 70 positions of the large single copy region adjacent to IR2.

in Fig. 3 are counted as in Fig. 2, note the overlap). Within this segment we found the gene for $tRNA^{His}_{GUG}$ (trn H). The structural part and therefore the secondary structure of the transcript (Fig. 4) are to 100% identical with the recently sequenced trn H gene of Zea mays (16). Upstream of the structural part at positions 1193 to 1198 and 1215 to 1219 we recognize sequences which may qualify as promoter sequences (15). Downstream and already within the IR1 we recognize inverted repeats (9-mer) which could form a stem and loop structure and qualify as gene terminators.

Seventy positions of both ends of the large single copy region and 250

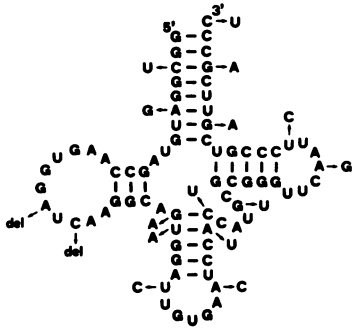


Fig. 4. Cloverleaf structure (unmodified) of soybean chloroplast $tRNA^{His}_{GUG}$ as deduced from the trn H sequence. The 5' terminal G is taken as position 1. Arrows point towards nucleotides present at that position in Euglena gracilis $tRNA^{His}_{GUG}$ (22); del : deletion.

positions of the IR1 and IR2 are aligned and compared in Fig. 3. IR1 and IR2 show perfect sequence homology for the entire analysed segment (250 positions). The two ends of the single copy region have nothing in common within the analysed stretch (70 positions), i.e., the structural part of the trn H gene is certainly not duplicated in this genome contrary to the situation in Zea mays (16). The size and function of an open reading frame which starts at position 3348 within IR2 (Fig. 3) and continues in the large single copy region (coding strand) is presently under investigation.

DISCUSSION

The psb A gene

The soybean chloroplast psb A gene is in its structural properties essentially identical to that of S. oleracea and N. debneyi (6). The length of the transcribed region (distance between the most likely transcription initiation and termination) is well within 1.2 kb, what agrees with the size of the major RNA which interacted with a psb A gene probe (17). These authors compared the psb A gene transcripts of several angiosperm chloroplasts including soybean and spinach. They observed for soybean, but not for spinach, that two minor transcription products of about 1.0 and 0.2 kb also interacted with the psb A DNA probe. They thought, however, that these minor RNAs were specific degradation products of the 1.2 kb RNA, the specific cleavage site being on the 3' terminal part of the coding region. A sequence comparison between the soybean and spinach psb A gene around nucleotide position 900

(Fig. 2), where preferential cleavage would occur in case of soybean, reveals no particular differences, i.e., the soybean chloroplast must contain a ribonuclease(s) which differ in specificity from that of spinach chloroplasts.

Hoffman-Falk et al. (18) studied the Mr 32 000 thylakoid membrane protein from several angiosperms and the alga Chlamydomonas rheinhardii. They found extensive similarities at levels of precursor maturation, membrane orientation and primary structure. This is in accordance with the psb A sequence data published sofar. There is no doubt that constraints on the primary structure of the entire translated region are very rigid, permitting only point mutations (with few exceptions) in the wobble position as shown now in chloroplast genomes of representatives from three distant plant families (Chenopodiaceae, Solanaceae, Leguminosae). Of particular importance is the observation that also in case of soybean the psb A gene codes for 353 aminoacids, the N-terminal part being to 100% homologous with that of S. oleracea and N. debneyi. This strongly suggests that the first 36 codons of the open reading frame are translated and essential in the primary translation product. However, McIntosh (19) reported very recently that the psb A genes from Zea mays and Amaranthus hybridus encode a protein of 317 aminoacids only. According to this preliminary report, it seems possible that the first 36 aminoacids are not required for a functional 32 kd thylakoid membrane protein. More analytical data, however, are necessary to obtain a clear picture concerning the size difference between the coding region and the gene product. Zurawsky et al. (6) discussed several possibilities to explain this discrepancy.

The trn H gene and the terminal part of the inverted repeats

A trn H gene was mapped adjacent to one of the inverted repeats on the chloroplast genomes of Spinacia oleracea (10) and Phaseolus vulgaris (20). For Zea mays the trn H gene was mapped within the inverted repeats and it was shown that the trn H gene slightly overlaps with a gene of opposite polarity which is transcribed into a RNA of 1.6 kb (16). The soybean trn H gene sits with its structural part right at one end of the large single copy region and a potential transcription termination site is located on the inverted repeat. The ribonucleotide sequence deduced from the trn H gene is to 100% identical with that of Zea mays chloroplasts. All highly conserved positions (21) are maintained in the trn H sequence. In Fig. 4 we compare the soybean with the Euglena gracilis chloroplast $tRNA^{His}_{GUG}$ (22) as deduced from the corresponding

structural genes. There is about 80% sequence homology. In both cases the D-stem is shortened and the extra loops are of identical length. The 3' terminal CCA are not encoded in the genes, a property chloroplast tRNA genes seem to share with eukaryotic tRNA genes (21). All three chloroplasts trn H genes sequenced so far code for $tRNA^{His}_{GUG}$. According to extensive hybridization experiments using a variety of chloroplast genomes (for references, see 10) no second trn H gene was identified, i.e., a possible gene for $tRNA^{His}_{AUG}$ is still undetected. Nevertheless, we can see (Fig. 2) that the codon CAT (CAU) is frequently used in translating both the spinach and soybean Mr 32 000 thylakoid membrane protein, this in agreement with the wobble hypothesis.

The three chloroplasts trn H genes sequenced so far have a different genetic environment. The soybean trn H gene is at one end of the large single copy region proximate to and with the same polarity as the psb A gene. Its transcription termination site is most likely part of the inverted repeat. The maize trn H gene is integral part of the inverted repeat, occurs therefore twice per circular genome and it overlaps slightly with a protein coding gene of opposite polarity (16). Finally the Euglena gracilis trn H gene is the second gene in a cluster of five tandemly arranged tRNA genes (22), which most likely are co-transcribed and under the control of one promoter. trn H gene transcription regulation must be different in the three types of chloroplasts, a point of considerable interest.

It is known from cross hybridization experiments that the inverted repeats of chloroplast genomes are structurally related, displaying considerable sequence homology (7). Our sequence results define for the first time the exact beginning of IR1 and IR2 on a angiosperm chloroplast genome. A comparison with the Zea mays chloroplast IR (16) reveals that a short stretch of 32 positions is to 91% homologous with a soybean chloroplast DNA segment (position 1510-1542). Most likely, sequence homology between the two IRs continues beyond the sequenced positions (16). It's noteworthy that the short homologous parts map at different places within the two kinds of inverted repeats. In case of soybean this short segment starts 138 positions inside of the beginning of the IR, for maize it is most likely about 0.5 kb away from the IR start. This and the different location of the trn H gene shows that during evolution not only the single copy region underwent DNA rearrangments but also the rather conservative inverted repeats region.

ACKNOWLEDGEMENTS

We received financial support from the Swiss National Science Foundation (3.183.82 to E.S.) and from Nestlé Products Technical Assistance Co., Ltd. We are grateful to Ch. Bachmann for secretarial and B. Schlunegger and I. Howald for technical assistance. E. Roux gave us helpful advices for DNA sequencing and computer analysis. This report represents part of a Ph.D. thesis (A.S.) submitted to the Science Department of the University of Neuchâtel.

*To whom reprint requests should be addressed.

REFERENCES

1. Spielmann, A., Ortiz, W. and Stutz, E. (1983) Mol. Gen. Genet. 190, 5-12.
2. Bedbrook, J.R., Link, G., Coen, D.M., Bogorad, L. and Rich, A. (1978) Proc. Natl. Acad. Sci. USA 75, 3060-3064.
3. Driesel, A.J., Speirs, J. and Bohnert, H.J. (1980) Biochim. Biophys. Acta 610, 297-310.
4. Mattoo, A.K., Pick, U., Hoffman-Falk, H. and Edelman, M. (1981) Proc. Natl. Acad. Sci. USA 78, 1572-1576.
5. Steinback, K.E., McIntosh, L., Bogorad, L. and Arntzen, C.J. (1981) Proc. Natl. Acad. Sci. USA 78, 7463-7467.
6. Zurawski, G., Bohnert, H.J., Whitfeld, P.R. and Bottomley, W. (1982) Proc. Natl. Acad. Sci. USA 79, 7699-7703.
7. Palmer, J.D. and Thompson, W.F. (1982) Cell 29, 537-550.
8. Palmer, J.D., Singh, G.P. and Pillay, D.T.N. (1983) Mol. Gen. Genet. 190, 13-19.
9. Swamy, G.S. and Pillay, D.T.N. (1982) Plant Sci. Lett. 25, 73-84.
10. Bohnert, H.J., Crouse, E.J. and Schmitt, J.M. (1982) Encyclopedia of Plant Physiology, New Series, vol. 14B, Parthier, B. and Boulter, D., eds, Springer-Verlag Berlin-Heidelberg, pp. 475-530.
11. Rigby, P.W.J., Dieckman, H., Rhodes, C. and Berg, P. (1977) J. Mol. Biol. 113, 237-251.
12. Graf, L., Roux, E., Stutz, E. and Kössel, H. (1982) Nucleic Acids Res. 10, 6369-6381.
13. Maxam, A.M. and Gilbert, H. (1980) Methods in Enzymology, vol. 65, pp. 499-560, Grossman, L. and Moldave, K., eds, Academic Press, New York.
14. Sanger, F., Coulson, A.R., Barrel, B.G., Smith, A.J.H. and Roe, B.A. (1980) J. Mol. Biol. 143, 161-178.
15. Hawley, D.K. and McClure, W.R. (1983) Nucleic Acids Res. 11, 2237-2255.
16. Schwarz, Z., Jolly, S.O., Steinmetz, A.A. and Bogorad, L. (1981) Proc. Natl. Acad. Sci. USA 78, 3423-3427.
17. Palmer, J.D., Edwards, H., Jorgensen, R.A. and Thompson, W.F. (1982) Nucleic Acids Res. 10, 6819-6832.
18. Hoffman-Falk, H., Mattoo, A.K., Marder, J.B. and Edelman, M. (1982) J. Biol. Chem. 257, 4583-4587.
19. McIntosh, L. (1983) J. Cell. Biochem. s 7B, 295.
20. Mubumbila, M., Gordon, K.H.J., Crouse, E.J., Burkard, G. and Weil, J.H. (1983) Gene 21, 257-266.
21. Sprinzl, M. and Gauss, D.H. (1982) Nucleic Acids Res. 10, r1-r56.
22. Hollingsworth, M.J. and Hallick, R.B. (1982) J. Biol. Chem. 257, 12795-12799.