Volume 261, number 2, 455–458 FEBS 08193 February 1990

A conserved cleavage-site motif in chloroplast transit peptides

Ylva Gavel and Gunnar von Heijne*

Research Group for Theoretical Biophysics, Department of Theoretical Physics, Royal Institute of Technology, S-100 44 Stockholm, Sweden and **Department of Molecular Biology, Karolinska Institute Center for Biotechnology, NOVUM, S-141 52 Huddinge, Sweden*

Received 18 January 1990

A collection of 32 stroma-targeting chloroplast transit peptides with known cleavage sites have been analysed in terms of amino acid preferences in the vicinity of the processing site. A loosely conserved consensus motif (Val/Ile)-X-(Ala/Cys)↓Ala is found in the majority of the transit peptides. About 30% of the sequences have a perfect match to the consensus. When such a match is found, there is a 90% probability that it specifies the correct cleavage site.

Protein targeting; Protein import; Chloroplast; Transit peptide; Stromal protease

1. INTRODUCTION

Most chloroplast proteins are made as precursors in the cytoplasm and imported post-translationally into the organelle [1]. Targeting to the stromal compartment is effected by amino terminal chloroplast transit peptides (cTPs) that are subsequently removed by stromal proteases. In general, cTPs from higher plants have a high content of hydroxylated Ser and Thr residues but contain very few acidic amino acids. In addition, 3 structural domains can be discerned in higher plant cTPs: an uncharged amino-terminal domain, a central positively charged domain lacking acidic residues, and a carboxy-terminal domain with a high potential for forming an amphiphilic β-strand [2]. cTPs from the green alga *Chlamydomonas reinhardtii* have a domain structure that is distinct from the higher plant cTPs and closer to the structure of mitochondrial targeting peptides [3].

Although an early study reported that cTPs contain a conserved 'homology block' Gly-X-Gly-Arg-Val close to the processing site [4], it is now clear that most cTPs do not have such a well-conserved cleavage site. In this paper, we report that higher plant cTPs fit a more loosely defined cleavage site consensus sequence: (Val/Ile)-X-(Ala/Cys)↓Ala. In addition, arginines are often found in positions −6 to −10.

Correspondence address: G. von Heijne, Department of Molecular Biology, Karolinska Institute Center for Biotechnology, NOVUM, S-141 52 Huddinge, Sweden

2. MATERIALS AND METHODS

32 non-homologous cTPs with known cleavage sites were collected from the literature (table 1 [5–38]). Amino acid counts as a function of position relative to the cleavage site were obtained, and the statistical significance of prominent peaks in the amino acid distribution profiles was assessed by calculating the probability of obtaining a similar or larger deviation from the mean in a binomial distribution with the mean frequency of the amino acid in question put equal to its mean frequency in the sample. For peaks within the transit peptides, the overall frequencies in the whole sample of cTPs (with initial methionines removed) were used. Overall frequencies according to [2] were used for peaks situated in the mature parts.

3. RESULTS

3.1. *Ile and Val are enriched in position −3*

About 60% of the cTPs in our sample have a β-carbon branched hydrophobic residue (Ile or Val) in position −3 relative to the cleavage site (fig.1). Assuming a binomial distribution, the probability for a deviation of this magnitude in any one position is $P < 5 \times 10^{-6}$ for both Ile and Val, which is clearly significant even though we are making multiple tests (20 amino acids – i.e. 20 observations – for each position, and 10–15 independent positions). A fairly high count is found also for position −2, though statistically this is not significant ($P \approx 0.05$ for both Ile and Val).

3.2. *Ala and Cys are often found in positions −1 and +1*

A high significant enrichment for Cys ($f_{Cys} \approx 20\%$, $P < 1 \times 10^{-6}$) is seen in position −1 (fig.2A). Likewise, about 45% of the sequences have Ala in positions −1

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/90/$3.50

Table 1

Chloroplast transit peptides

Known cleavage sites are indicated by +. Perfect matches to the conserved motif are underlined

```
1)                  ....ELGERNLLPQKKPPKMACSLSFSSSVSTFHLPTTTQSRQAPPNNNATTLPTTNPIQC+ANLRELRDRIGSVKNTQKKI
2)                      MTTAVTAAVSFPSTKTTSLSARSSSVISPDKISYKKVPLYYRNVSATGKMGPIRA+QIASDVEAPPPAPAKVEKMS
3)           MAAATASLSSTLLAPCSSKQPQPQQQHQHQQLKCKSFSGLRPLKLNISSNNSSSSLSMSSARRSMTCRA+ELSPSLVISLSTGLSLFLGR
4)               MAAHTIFTTTSYYNSFLFPIASSNTNSAPSLSSSFHGVSLKVKSKTPQSLTLSSVTSPKPFIVFA+ATKKAVAVLKGTSMVEGVVT
5)         MAALATSQLVATRAGLGVPDASTFRRGAAQGLRGARASAAADTLSMRTSARAAPRHQQQARRGGRFPSLVVC+ASAGMNVVFV
6)           MAALQNPVALQSRTTTAVAALSTSSTTSPPKPFSLSFSSSTATFNPLRLKILTASKLTAKPRGGALGTRM+VDSTASRYASALADVADVTG
7)                                          MAASSSSSMALSSPTLAGKQLKLNPSSQELGAARFT+MRKSA
8)          MAASVSRAICVQKPGSKCTRDREATSFARRSVAAPRPPHAKAAGVIRSDSGAGRGQHCSPLRAVVDAAPIQ+TTKKRVFHFGKGKSEGNKTM
9)                        MAATTTMATLNLPSLTSHPNSSTFPKHPQPLQFPFRTTTNPISLSSTRTTRLRPIA+AVEAPEKIEQLGTQLSGLTL
10)                    MAHCLAAVSSFSPSAVRRRLSSQVANVVSSRSSVSFHSRQMSFVSISSRPSSLRFKICC+AAMGEAQAKKETVDKVCMIV
11)           MAIENCLQLSTSASVGTVAVKSHVHHLQPSSKVNPVPTFRGLKRSFPALSSSVSSSSPRQFRYSSVVC+KASEAVKEVQDVNDSSWKEF
12)      MALQLLPSTLSVPKKGSSMGAVAVKDTAAFLGVSSKAKKASLAVRTQVATAPSPVTTSPGSTASSPSGKKTLRQ+GVVVITGASSGLGLAAAKAL
13)                              MAMATQATLFSPSSLSSAKPIDTRLTTSFKQPSAVTFASKPASRHHSIRA+AAAAEGKRAAATETKEAPKG
14)        MAQINNMAQGIQTLNPNSNFHKPQVPKSSSFLVFGSKKLKNSANSMLVLKKDSIFMQKFCSFRISASVATAQ+KPSEIVLQPIKEISGTVKLP
15)           MASATFSVAKPAIKANGKGFSEFSGLRNSSRHLPFSRKSSDDFHSLVTFQTNAVGSSGGHKKSLVVEA+KQLKVAINGFGRIGRNFLRC
16)                                      MASIMMNKSVVLSKECAKPLATPKVTLNKRGFATTIATKNRE+MMVWQPFNNKMFETFSFLPP
17)                                 MASLATLAAVQPTTLKGLAGSSIAGTKLHIKPARQSFKLNNVRSGAIVA+KYGDKSVYFDLEDNANTTGQ
18)                                          MASLPVNKIIPSSTTLLSSSNNNRRRNNSSIR+CQKAVSPAAETAAVSPSVDA
19)                          MASLSATTTVRVQPSSSSLHKLSQGNGRCSSIVCLDWGKSSFPTLRTSRRRSFISA+AKKETIDKVCDIVKEKLALG
20)                                  MASQTLVSPSPLSSHSLLRTSFSGVSVKLAPQFSTLATSNFKPLTVVA+AAKKAVAVLKGTSAVEGVVT
21)                          MASSIVSSAAVATRSNVAQASMVAPFTGLKSAASFPVTKKNNNVDITSLASNGGRVRC+MQVWPPINMKKYETLSYLPD
22)                 MASSMLSATTVPLQQGGGLSEFSGLRSSASLPMRRNATSDDFMSAVSFRTHAVGTSGGPRRAPTEA+KLKVAINGFGRIGRNFLRCW
23)                                  MASTLSTLSVSASLLPKQQPMVASSLPTNMGQALFGLKAGSRGRVTAM+ATYKVTLITK
24)                                    MASTNMASATSRFMLAAGIPSGANGGVSSRVSFLPSNRLGLKLVAR+AEEPTAAAPAEPAPAADEKP
25)                                         MASTVMSSLSLKPTFTLEKTSVKGLPSLARSSSSFKVVA+SGVKKLKTDKPYGINGSMAL
26)                          MATAVSTVGAATRAPLNLNGSSAGASVPTSGFLGSSLKKHTNVRFPSSSRTTSMTVKA+AENEEKNTDKWAHLAKDFSD
27)                                          MATSVMSSLSLKPSSFGVDTKSAVKGLPSLSRSSASFTVRA+SGVKKIKVDKPLGIGGGMKL
28)                                 MATTFSASVSMQATSLATTTRISFQKPVLVSNHGRTNLSFNLSRTRLSISC+AAKQETVEKVSEIVKKQLSL
29)                                 MAVCTVYTIPTTTHLGSSFNQNNKQVFFNYKRSSSSNNTLFTTRPSYVITC+SQQQTIVIGLAADSGCGKST
30)                            MGLSTVYSPAGPRLVPAPLGRCRSAQPRRPRRAPLATVRCSVDATKQAQDGVATAVA+TEAPASRKECFGVFCTTYDL
31)               MIISIFNQLHLTENSSLMASFTLSSATPSQLCSSKNGMFAPSLALAKAGRVNVLISKERIRGMKLTCQ+ATSIPADNVPDMQKRETNLN
32)                                  MTANGAHLFNHYSSNSRFIHFTSRNTSSKLLTKTSHFRRPKRCFHVNN+TLSEKIHHPITQGGESDLSS
```

Protein names and references:

1) ATP synthase gamma subunit, spinach [5,6]
2) FNR, spinach [7]
3) PSI V, spinach [8]
4) SOD, petunia [9]
5) UDPGlc:starch glucosyl transferase, maize [10]
6) ATPase delta subunit, spinach [11]
7) LHCII, pea [12]
8) pyruvate, ortophosphate dikinase, maize [13]
9) ribosomal protein L12, spinach [14]
10) acyl carrier protein, barley [15]
11) thioredoxin m, spinach [16]
12) PCR, barley, stroma [17]
13) PS I 20kDa subunit, spinach [18]
14) ESPS, petunia [19]
15) GapA, pea [20]
16) SSU, Acetabularia mediterranea [21]
17) PSI VI, spinach [22]
18) nitrate reductase, spinach [23]
19) acyl carrier protein I, spinach [24]
20) Cu/Zn-superoxide dismutase, pea [25]
21) SSU, Solanum tuberosum, stroma [26]
22) glyceraldehyde-3-phosphate dehydrogenase, maize [27,28]
23) FD, ferredoxin, Silene, stroma [29]
24) PSI 10.8 kDa protein, barley [30]
25) PS II ST-LSI, potato [31]
26) ribulose-bisphosphate carboxylase/oxygenase activase, spinach [32]
27) 10 kd, spinach, thylakoid membrane [33]
28) acyl carrier protein, Brassica napus [34]
29) phosphoribulokinase, spinach [35]
30) NADP-malate dehydrogenase, maize [36]
31) Rieske FeS protein, spinach [37]
32) alpha-glycan phosphorylase, potato [38]

and/or +1 ($P < 1 \times 10^{-6}$) (fig.2B). All of the cTPs with Cys_{-1}, 65% with Ala_{-1}, and 70% with Ala_{+1} have Ile or Val in −3, suggesting that the $(Ile/Val)_{-3}$ and the $(Ala/Cys)_{-1}\downarrow Ala_{+1}$ patterns correlate.

3.3. *Arg is abundant in −10 to −6 and in −2*

Although not clearly concentrated in one or a couple of positions, Arg is quite abundant throughout the region −10 to −6, and also in −2 (fig.3).

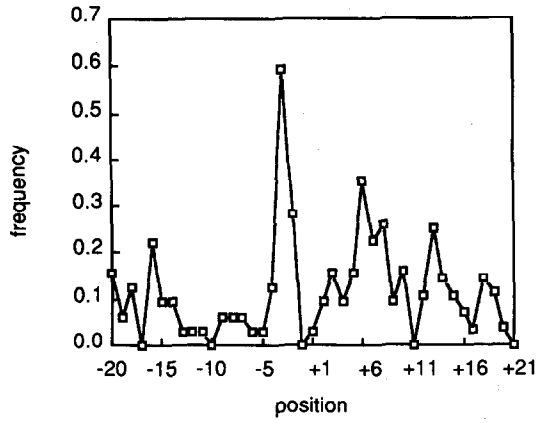

Fig.1. Positional distribution of Ile + Val residues in cTPs aligned according to the cleavage site (between positions −1 and +1).

4. DISCUSSION

The statistical data presented above suggests that a majority of higher plant cTPs are characterized by a reasonable fit to the cleavage site motif (Ile/Val)-X-(Ala/Cys)↓Ala. In addition, one often finds one or more arginines in the region −6 to −10. It is possible that sequences which do not fit this pattern may at least in some cases be cleaved in more than one step by different proteases, since reasonable matches are often found in internal positions.

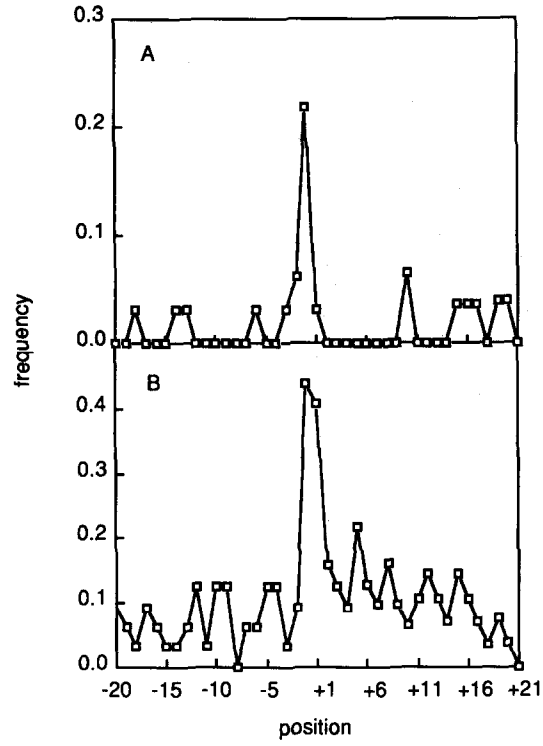

Fig.2. Positional distribution of Cys (panel A) and Ala (panel B) residues in cTPs.

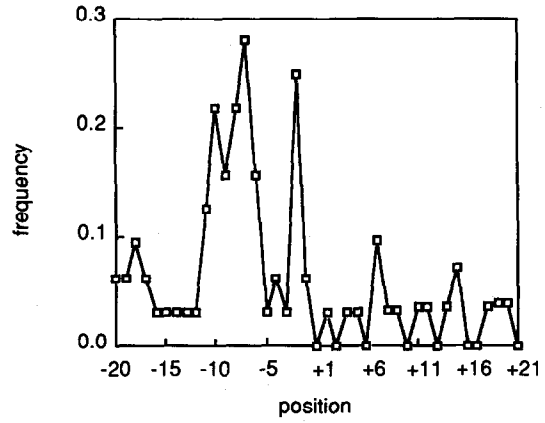

Fig.3. Positional distribution of Arg residues in cTPs.

Only a small number of mutations in the region of the cleavage site have been studied so far. In general, when more than a couple of residues are removed from the cleavage region, either no processing or aberrant cleavage products are observed [39–41]. The study by Smeekens et al. [41] is particularly interesting, since removal of residues −1 and −2 from the *Silene* ferredoxin cTP have no apparent effect on processing. The ferredoxin cTP has the cleavage site ...Arg-Val-Thr-Ala-Met↓Ala-Thr... with the consensus motif (underlined) displaced one step from its normal position. However, since N-terminal methionines on plastid-encoded proteins are normally removed [42–45], the Met may be cleaved off by a stromal methionyl-aminopeptidase after an initial cleavage at the consensus motif. In the deletion mutant, the sequence becomes ...Arg-Val-Thr-Ala↓Thr..., and cleavage after the consensus motif would produce a molecule only one residue shorter than the wild-type protein and most likely indistinguishable on SDS gels.

We have shown previously [2,3] that the cleavage region has a high potential for forming amphiphilic β-strands in many cTPs. In this regard, it is notable that both Ile and Val have a high potential for forming extended β-strand structures [46], whereas Leu, which is largely absent from the cleavage site region, is a strong helix-former. Arg is indifferent in terms of secondary structure propensities, but Lys, the other positively charged amino acid (which is not enriched in cTPs), is a strong helix-former. Ala and Cys tend to be found in helical regions of proteins. The Arg-...-(Ile/Val) part of the consensus cleavage site motif is thus consistent with the idea of an amphiphilic β-strand being an important element in the cleavage region; the (Ala/Cys)↓Ala part may signal a transition from β-strand to a possibly more helical structure.

Nine of the 32 cTP cleavage sites in the sample (30%) match the consensus motif perfectly. As shown in table 1, there are only 3 other perfect matches in the sample, two of which are very close to the amino-terminus of the respective cTP. The third is found in a cTP that

does not have a good match to the consensus at its reported cleavage site, suggesting that it may be processed in more than one step. At any rate, when a perfect match to the consensus motif is found at a reasonable distance from the amino terminus (30–75 residues, say) – this should happen in about one-third of all newly sequenced cTPs – one can be about 90% confident that it signals the correct cleavage site. When no perfect match is found, one can still guess at the most likely cleavage site by roughly locating it to a region where there is a drop in the frequency of serine and a corresponding increase in the frequency of acidic residues, and then looking for partial matches to the consensus within this region. We feel, however, that more sequence-data is needed before one can design an automatic routine along these lines.

Acknowledgement: This work was supported by a grant from the Swedish Natural Sciences Research Council to G.v.H.

REFERENCES

[1] Keegstra, K. (1989) Cell 56, 247–253.
[2] Von Heijne, G., Steppuhn, J. and Herrmann, R.G. (1989) Eur. J. Biochem. 180, 535–545.
[3] Franzén, L.G., Rochaix, J.D. and Von Heijne, G. (1990) FEBS Lett., in press.
[4] Karlin-Neumann, G.A. and Tobin, E.M. (1986) EMBO J. 5, 9–13.
[5] Miki, J., Maeda, M., Mukohata, Y. and Futai, M. (1988) FEBS Lett. 232, 221–226.
[6] Herrmann, R.G., personal communication.
[7] Jansen, T., Reiländer, H., Steppuhn, J. and Herrmann, R.G. (1988) Curr. Genet. 13, 517–522.
[8] Steppuhn, J., Hermans, J., Nechushtai, R., Ljungberg, U., Thümmler, F., Lottspeich, F. and Herrmann, R.G. (1988) FEBS Lett. 237, 218–224.
[9] Tepperman, J., Katayama, C. and Dunsmuir, P. (1988) Plant Mol. Biol. 11, 871–872.
[10] Klosgen, R.B., Gierl, A., Schwarz-Sommer, Z. and Saedler, H. (1986) Mol. Gen. Genet. 203, 237–244.
[11] Hermans, J., Rother, C., Bichler, J., Steppuhn, J. and Herrmann, R.G. (1988) Plant Mol. Biol. 10, 323–330.
[12] Cashmore, A.R. (1984) Proc. Natl. Acad. Sci. USA 81, 2960–2964.
[13] Matsuoka, M., Ozeki, Y., Yamamoto, N., Hirano, H., Kano-Murakami, Y. and Tanaka, Y. (1988) J. Biol. Chem. 263, 11080–11083.
[14] Giese, K. and Subramanian, A.R. (1989) Biochemistry 28, 3525–3529.
[15] Hansen, L. (1987) Carlsberg Res. Commun. 52, 381–392.
[16] Wedel, N., Clausmeyer, S., Gardet-Salvi, L., Herrmann, R.G. and Schürmann, P. (1990) submitted.
[17] Schulz, R., Steinmüller, K., Klaas, M., Forreiter, C., Rasmussen, S., Hiller, C. and Apel, K. (1989) Mol. Gen. Genet. 217, 355–361.
[18] Lagoutte, B. (1988) FEBS Lett. 232, 275–280.
[19] Klee, H.J., Muskopf, Y.M. and Gasser, C.S. (1987) Mol. Gen. Genet. 210, 437–442.
[20] Brinkmann, H., Cerff, R., Salomon, M. and Soll, J. (1989) Plant Mol. Biol. 13, 81–94.
[21] Schneider, S.U., Leible, M.B. and Yang, X.-P. (1989) Mol. Gen. Genet. 218, 445–452.
[22] Steppuhn, J., Hermans, J., Nechushtai, R., Herrmann, G.S. and Herrmann, R.G. (1989) Curr. Genet. 16, 99–108.
[23] Back, E., Burkhart, W., Moyer, M., Privalle, L. and Rothstein, S. (1988) Mol. Gen. Genet. 212, 20–26.
[24] Scherer, D.E. and Knauf, V.C. (1987) Plant Mol. Biol. 9, 127–134.
[25] Sciloli, J.R. and Zilinskas, B.A. (1988) Proc. Natl. Acad. Sci. USA 85, 7661–7665.
[26] Wolter, F.P., Fritz, C.C., Willmitzer, L., Schell, J. and Schreier, P.H. (1988) Proc. Natl. Acad. Sci. USA 85, 846–850.
[27] Quigley, F., Martin, W.F. and Cerff, R. (1987) Proc. Natl. Acad. Sci. USA 85, 2672–2676.
[28] Gowri, G. and Campbell, W.H. (1989) Plant Physiol. 90, 792–798.
[29] Smeekens, S., Van Binsbergen, J. and Weisbeek, P. (1985) Nucleic Acids Res. 13, 3179–3194.
[30] Okkels, J.S., Jepsen, L.B., Hønberg, L.S., Lehmbeck, J., Scheller, H.V., Brandt, P., Høyer-Hansen, G., Stummann, B., Henningsen, K.W., Von Wettstein, D. and Møller, B.L. (1988) FEBS Lett. 237, 108–112.
[31] Webber, A.N., Packman, L.C. and Gray, J.C. (1989) FEBS Lett. 242, 435–438.
[32] Werneke, J.M., Zielinski, R.E. and Ogren, W.L. (1988) Proc. Natl. Acad. Sci. USA 85, 787–791.
[33] Lautner, A., Klein, R., Ljungberg, U., Reiländer, H., Bartling, D., Andersson, B., Reinke, H., Beyreuther, K. and Hermann, R.G. (1988) J. Biol. Chem. 263, 10077–10081.
[34] Safford, R., Windust, J.H.C., Lucas, C., DeSilva, J., James, C.M., Hellyer, A., Smith, C.G., Slabas, A.R. and Hughes, S.G. (1988) Eur. J. Biochem. 174, 287–295.
[35] Roesler, K.R. and Ogren, W.L. (1988) Nucleic Acids Res. 16, 7192.
[36] Metzler, M.C., Rothermel, B.A. and Nelson, T. (1989) Plant Mol. Biol. 12, 713–722.
[37] Steppuhn, J., Rother, C., Hermans, J., Jansen, T., Salnikow, J., Hauska, G. and Herrmann, R.G. (1987) Mol. Gen. Genet. 210, 171–177.
[38] Nakano, K., Mori, H. and Fukui, T. (1989) J. Biochem. 106, 691–695.
[39] Reiss, B., Wasmann, C.C., Schell, J. and Bohnert, H.J. (1989) Proc. Natl. Acad. Sci. USA 86, 886–890.
[40] Ostrem, J.A., Ramage, R.T., Bohnert, H.J. and Wasmann, C.C. (1989) J. Biol. Chem. 264, 3662–3665.
[41] Smeekens, S., Geerts, D., Bauerle, C. and Weisbeek, P. (1989) Mol. Gen. Genet. 216, 178–182.
[42] Ikeuchi, M., Takio, K. and Inoue, Y. (1989) FEBS Lett. 242, 263–269.
[43] Webber, A.N., Packman, L., Chapman, D.J., Barber, J. and Gray, J.C. (1989) FEBS Lett. 242, 259–262.
[44] Michel, H., Hunt, D.F., Shabanowitz, J. and Bennett, J. (1988) J. Biol. Chem. 263, 1123–1130.
[45] Flinta, C., Persson, B., Jörnvall, H. and Von Heijne, G. (1986) Eur. J. Biochem. 154, 193–196.
[46] Levitt, M. (1978) Biochemistry 17, 4277–4285.