

# Extreme variations in the ratios of non-synonymous to synonymous nucleotide substitution rates in signal peptide evolution

Federico Garcia-Maroto\*, Atilio Castagnaro, Pilar Sanchez de la Hoz, Carmen Marañña\*\*, Pilar Carbonero and Francisco García-Olmedo

*Bioquímica y Biología Molecular, E.T.S. Ingenieros Agrónomos-UPM, E-28040 Madrid, Spain*

Received 17 April 1991; revised version received 21 May 1991

Nucleotide sequences encoding signal peptides from the precursors of  $\alpha$ -amylase/trypsin inhibitors from cereals are homologous to those corresponding to the precursors of thaumatin II and of plastocyanins. Non-synonymous ( $K_A$ ) and synonymous ( $K_S$ ) rates of nucleotide substitutions have been calculated for all possible binary combinations. Extreme variation in  $K_A/K_S$  ratios has been observed; from the 0.167 average found within the plastocyanin family to an average of 1.90 calculated for the inhibitors/thaumatin II transition. A similar calculation has been carried out for the signal peptide sequences of thionins, which are unrelated to those of the  $\alpha$ -amylase/trypsin inhibitor family, and an average  $K_A/K_S$  of 0.12 has been obtained. This variation can be largely explained in terms of an empirical index of stability related to amino acid composition and seems to be independent of functional constraints.

$\alpha$ -Amylase/trypsin inhibitor; Thaumatin II; Thionin; Plastocyanin; Signal peptide evolution

## 1. INTRODUCTION

It has been repeatedly observed that the rates of synonymous nucleotide substitutions ( $K_S$ ), which do not lead to amino acid changes, are greater than those of non-synonymous substitutions ( $K_A$ ) which do cause such changes, and that both rates are greater in pseudogenes (see [1,2]). These differences in substitution rates have been ascribed to selective constraints related to structure–function relationships. Thus, functionally constrained proteins or parts of proteins would evolve at a slower rate than less constrained ones, and residues at functionally important positions should be less variable than those at other positions (see [3]). A number of cases have been reported in which the spatial distribution and the rate of amino acid substitution do not seem to follow these rules [4–8]. Positive Darwinian selection has been involved to explain the accelerated changes taking place at functionally relevant sites in these cases, all of which involve families of proteinase inhibitors. However, an alternative explanation has been proposed that is consistent with the neutral theory of evolution [3]. The alternative explanation is in agreement with the contention that the mutability of an amino acid is determined not so much by its position,

but by the consequences of its being replaced on the overall chemical and spatial structure of the protein [9].

In the course of our study of two plant protein families, the  $\alpha$ -amylase/trypsin inhibitors and the thionins (for reviews, see [10,11]), we have observed extreme variation in the rates of change of their signal peptides. These peptides, which determine the co-translational translocation of the corresponding mature proteins across the endoplasmic reticulum membrane, have some common features, such as an N-terminal methionine, charged amino acids near the N-terminal, and a membrane-spanning hydrophobic sequence, but are otherwise highly variable among different genes and can be considered of polyphyletic origin [12]. However, different genes may code for homologous signal peptides that co-evolve with their neighbouring mature proteins. We report here an analysis of substitution rates in the signal peptide sequences of the two protein families which indicates that the observed variation can be explained to a great extent by compositional effects consistent with the neutral theory of evolution.

## 2. MATERIALS AND METHODS

Nucleotide sequences encoding signal peptides of the precursors of  $\alpha$ -amylase/trypsin inhibitors were determined by direct sequencing of the cloned cDNAs. Sequences, cloning and screening procedures corresponding to cDNAs not previously reported will be published elsewhere. Appropriate references for those already reported are given in the legend of Fig. 1.

A computer program kindly donated by W.H. Li (Houston, Texas) was used to estimate the rates of synonymous and non-synonymous nucleotide substitution [13]. The Microgenie Program (Beckman) was used to search for homologous sequences encoding signal peptides.

*Present address:* \*Max-Planck-Institut für Züchtungsforschung, Egelspfad, D-5000 Köln 30, Germany. \*\*Laboratorium voor Genetische Virologie, Vrije Universiteit Brussel, B-1640, Belgium.

*Correspondence address:* F. Garcia-Olmedo, Cátedra de Bioquímica y Biología Molecular, E.T.S. Ingenieros Agrónomos-UPM, E-28040 Madrid, Spain.

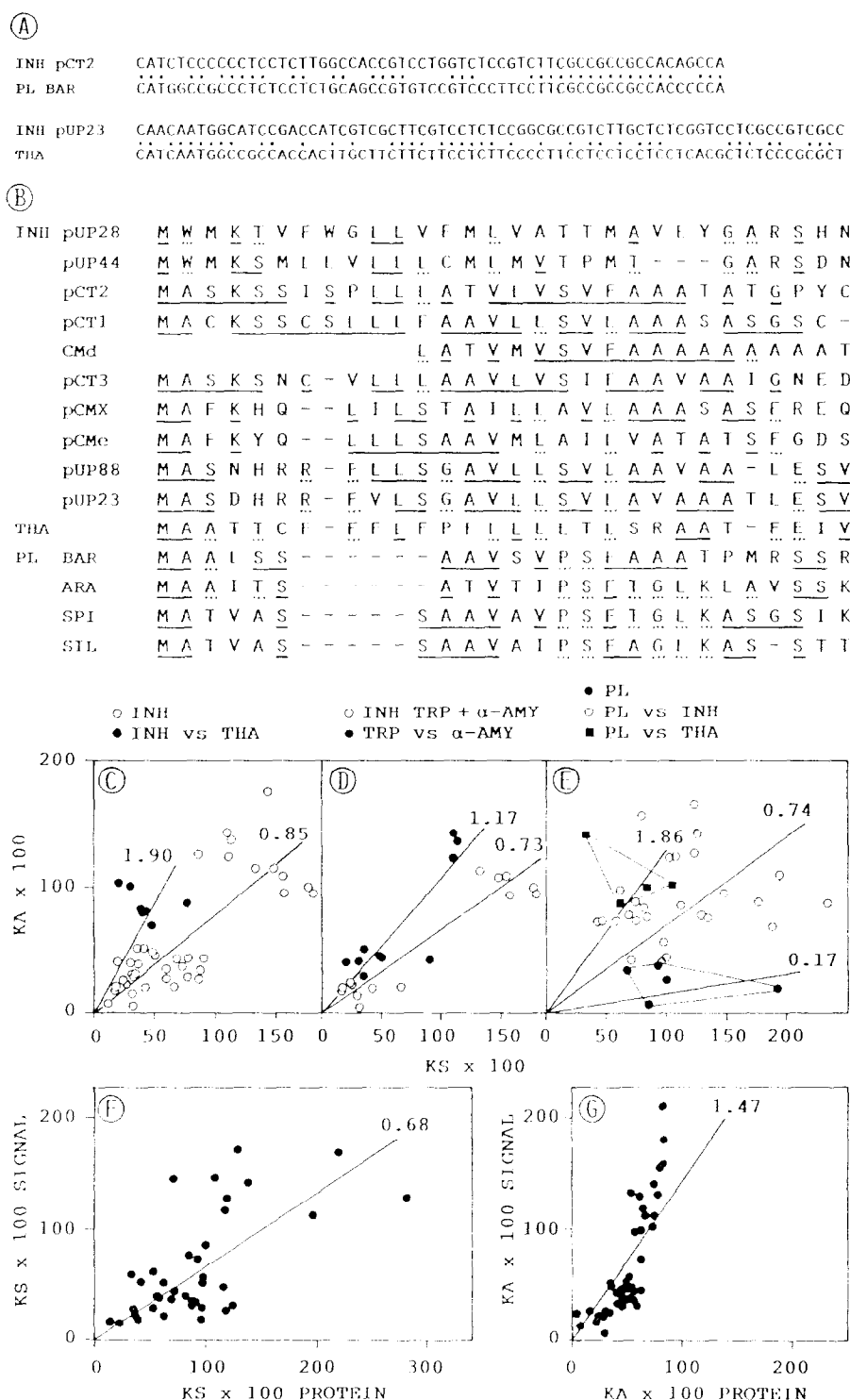


Fig. 1. Evolution of signal peptides from  $\alpha$ -amylase/trypsin inhibitors, thaumatin II and plastocyanins. A. Homology of nucleotide sequences encoding signal peptides as detected by the Microgenie (Beckman) computer program. B. Alignment of the amino acid sequences of the signal peptides used to calculate  $K_A$  and  $K_S$  with the computer programme of W.H. Li. The most frequent residues at a given position are underlined. (— and ...). C–E. Representation of  $K_A \times 100$  vs  $K_S \times 100$  values for the indicated binary combinations of signal peptides. The numbers indicate average  $K_A/K_S$  ratios and differed significantly ( $P < 0.05$ ) in the comparisons represented in each panel. F. Representation of  $K_S \times 100$  in signal peptide vs  $K_S \times 100$  in adjacent protein sequence. G. Representation of  $K_A \times 100$  of signal peptide vs  $K_A \times 100$  of adjacent protein. The inhibitor family (INH) includes inhibitors of trypsin (TRP) and  $\alpha$ -amylase ( $\alpha$ -AMY), as well as components whose activity has not been identified. INH sequences pUP28, pCT1, pCT2 and pCT2 correspond to unpublished subunits of  $\alpha$ -amylase inhibitors; pUP44 [16] and CMd [19] are also  $\alpha$ -amylase inhibitors; pCMX (unpublished) and pCMe [18] are trypsin inhibitors; pUP88 and pUP23 (unpublished) are components of unknown activity. THA corresponds to thaumatin II [17,21]. Plastocyanin (PL) sequences were from the Microgenie data base (PLA96 from barley, BAR; PLA30 from *Arabidopsis thaliana*, ARA; PLA869 from spinach, SPI; PLA702 from *Silene pratensis*, SIL).

Statistical significance of differences between average  $K_A/K_S$  ratios was calculated using the non-parametric Mann-Whitney test. The index of mutability derived by Graur [9] for the 7 relevant amino acids was calculated using the formula  $I_7 = 0.841 - 5.096 f \text{ Gly} + 24.145 f \text{ Asn} - 26.807 f \text{ Tyr} - 7.398 f \text{ Val} - 18.219 f \text{ Phe} - 8.263 f \text{ Asp} + 7.960 f \text{ Ile}$ .

### 3. RESULTS

The  $\alpha$ -amylase/trypsin inhibitor family in wheat and barley includes monomeric, dimeric, and tetrameric inhibitors of heterologous  $\alpha$ -amylases, as well as trypsin inhibitors, and is encoded by genes dispersed over several chromosomes [10]. These inhibitors are synthesized by membrane-bound polysomes as precursors in which the mature proteins are preceded by typical signal peptides, as shown by a study of their synthesis *in vivo* and *in vitro* [14] and by cloning and sequencing several cDNAs [15–20]. Homology was found between the signal peptides of the inhibitors and those of two other plant protein classes, thaumatin II [17,21] and plastocyanins [22–25] using the Microgenie Computer Programs, Beckman (Fig. 1A). In Fig. 1B, the signal peptides of ten members of the  $\alpha$ -amylase/trypsin inhibitor family have been aligned with those corresponding to thaumatin II [21] and to four plastocyanin clones [22–25]. Based on this alignment, both synonymous and non-synonymous nucleotide substitution rates have been estimated for all possible binary combinations, using the computer programme of Li [13]. The  $K_A$  and  $K_S$  values obtained within the inhibitor family and between members of this family and

thaumatin II are represented in Fig. 1C. The  $K_A/K_S$  ratio for each group was calculated as the slope of a least-squares-adjusted line intersecting the origin. This ratio was quite high for the signal peptide within the inhibitor family and significantly higher in the transition from the inhibitors to thaumatin II. Values obtained within the trypsin inhibitor and  $\alpha$ -amylase inhibitor subfamilies can be compared with those obtained in binary combinations involving one member of each subfamily in Fig. 1D. Similarly, in Fig. 1E, values obtained within the plastocyanin family are represented together with those corresponding to plastocyanin/inhibitor and plastocyanin/thaumatin II combinations. The signal peptide of the inhibitor is evolving faster than the adjacent mature protein sequence, i.e. lower  $K_S$  (Fig. 1F) and higher  $K_A$  (Fig. 1G).

In contrast with the above observations, the thionin signal peptide (Fig. 2A) is evolving at a slow rate (Fig. 2B), which is even slower than that of the adjacent mature protein, with lower  $K_S$  (Fig. 2C) and still lower  $K_A$  (Fig. 2D).

### 4. DISCUSSION

The data presented here indicate that amino acid sequences with a common functional constraint in their evolution, i.e. their function as signal peptides, show extreme variation in their  $K_A/K_S$  ratios. Graur [9] suggested that the propensity of an amino acid to remain conserved in the course of evolution depends not so much on its being in functional sites, but on an intrinsic

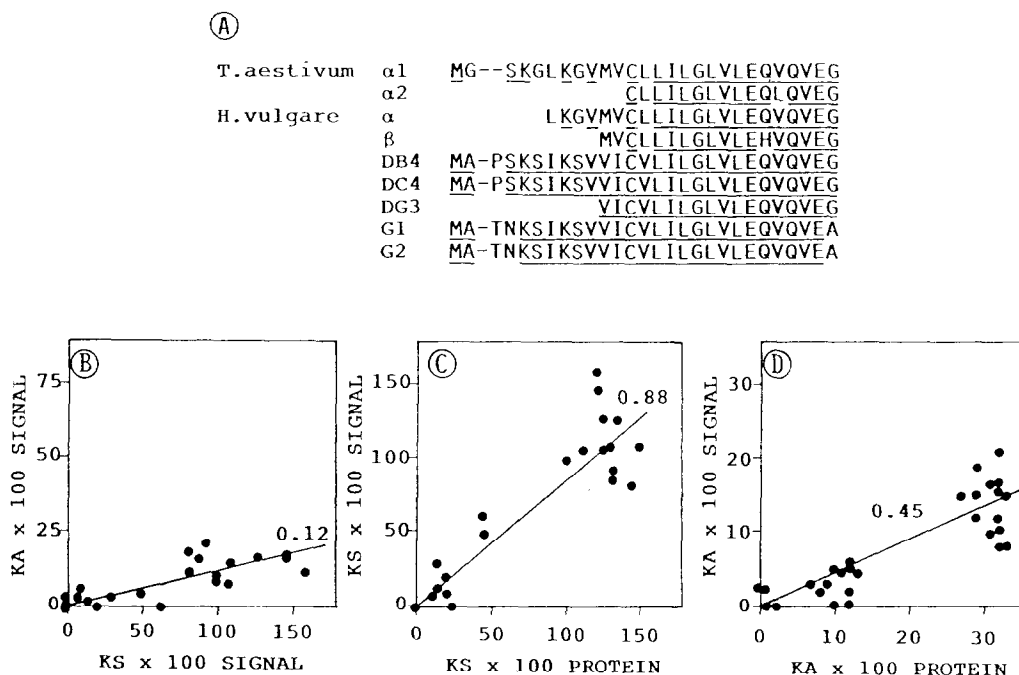


Fig. 2. Evolution of signal peptides from the thionin family. A. Alignment of known parts of signal peptides used to calculate  $K_A$  and  $K_S$  with the computer programme of W.H. Li. B. Representation of  $K_A \times 100$  vs  $K_S \times 100$  values of signal peptides. C. Representation of  $K_S \times 100$  of signal peptide vs  $K_S \times 100$  of adjacent protein. D. Representation of  $K_A \times 100$  of signal peptide vs  $K_A \times 100$  of protein.  $\alpha 1$  and  $\alpha 2$  from *Triticum aestivum* (unpublished), as well as  $\alpha$  and  $\beta$  from *Hordeum vulgare* [25,26] are endosperm thionins, while the others are from leaves [28,29].

Table I

Values of the mutability index  $I_m$ , taking into account the frequencies of the following 7 amino acids ( $I_7$ ): Gly, Asn, Tyr, Val, Phe, Asp, Ile

Sequence	Abbrev.	Value
Signal peptides		
Thaumatococcus	Tha-SP	4.637
Average inhibitors	Inh-SPav	1.455
Average plastocyanins	Pl-SPav	1.121
Average thionins	Thi-SPav	-0.825
Mature proteins		
Average inhibitors	Inhav	0.629
Average thionins	Thiav	1.036

stability index defined as the mean chemical distance between the amino acid and its mutational derivatives produced by a single nucleotide substitution. Reliable predictions of mutability could be obtained with an empirical index  $I_m$ , based on multiple linear regression equations involving the frequencies of particular amino acids in a given sequence [3,9]. The  $I_7$  values for the signal peptides studied here (Table I) rank in the order Tha-SP > Inh-SPav > Pl-SPav > Thi-SPav, which are in line with the calculated  $K_A/K_S$  values. Similarly, Inh-SPav > Inhav and Thi-SPav < Thiav, as expected from the  $K_A/K_S$  ratios. These results indicate that at least an important fraction of the observed variability in  $K_A/K_S$  ratios can be considered as consistent with the neutral theory.

In three situations, the non-synonymous rate ( $K_A$ ) significantly exceeds the synonymous one ( $K_S$ ): the inhibitor/thaumatococcus II transition (Fig. 1C), the trypsin inhibitor/ $\alpha$ -amylase inhibitor transition (Fig. 1D), and the plastocyanin/thaumatococcus II transition (Fig. 1E). In the first and third cases, the transition implies a change in the mature protein partner served by the signal peptide, while in the second case, a change of activity of the mature protein has occurred. A similar situation has been described by Hill and Hastie [4] in a comparison of the putative reactive center and the neighbouring sequences of a given serpin in rat, mouse, and man. Positive Darwinian selection was invoked by these authors to explain observed high rates of evolution, but Graur and Li [3] subsequently interpreted the same data in terms of compositional effects. In the present case, the data suggest a significant acceleration in the transitions.

**Acknowledgements:** Technical assistance from D. Lamonedá is gratefully acknowledged. Work was supported by Grant BI088-0216 from the Comisión Interministerial de Ciencia y Tecnología (Spain).

## REFERENCES

- [1] Li, W.-H., Gojobori, T. and Nei, M. (1981) *Nature* 292, 237-239.

- [2] Miyata, T. and Yasunaga, T. (1981) *Proc. Natl. Acad. Sci. USA* 78, 450-453.
- [3] Graur, D. and Li, W.-H. (1988) *J. Mol. Evol.* 28, 131-135.
- [4] Hill, R.E. and Hastie, N.D. (1987) *Nature* 326, 96-99.
- [5] Laskowski, M., Kato, I., Andelt, W., Cook, J., Denton, A., Empie, N.W., Kohr, W.J., Park, S.J., Parks, K., Shateley, B.L., Schoenberger, O.L., Tashiro, M., Vichot, G., Whatley, H.E., Wicczorek, A. and Wicczorek, M. (1987) *Biochemistry* 26, 202-221.
- [6] Laskowski, M., Kato, I., Kohr, W.J., Park, S.J., Tashiro, M. and Whatley, H.E. (1987) *Cold Spring Harbor Symp. Quant. Biol.* 52, 201-207.
- [7] Fioretti, E., Iacopino, G., Angeletti, M., Barra, D., Bossa, F. and Ascoli, F. (1985) *J. Biol. Chem.* 260, 11451-11455.
- [8] Creighton, T.E. and Charles, I.G. (1987) *J. Mol. Biol.* 194, 11-12.
- [9] Graur, D. (1985) *J. Mol. Evol.* 22, 53-62.
- [10] García-Olmedo, F., Salcedo, G., Sánchez-Monge, R., Gómez, L., Royo, J. and Carbonero, P. (1987) *Oxford Surveys Plant Mol. Cell Biol.* 4, 275-334.
- [11] García-Olmedo, F., Rodríguez-Palenzuela, P., Hernández-Lucas, C., Ponz, F., Marañón, C., Carmona, M.J., López-Fando, J., Fernández, J.A. and Carbonero, P. (1989) *Oxford Surveys Plant Mol. Cell Biol.* 6, 31-60.
- [12] Watson, M.E.E. (1984) *Nucleic Acids Res.* 12, 5145-5164.
- [13] Li, W.-H., Wu, C.I. and Luo, C.-C. (1984) *J. Mol. Evol.* 21, 58-71.
- [14] Paz-Ares, J., Ponz, F., Aragoncillo, C., Hernández-Lucas, C., Salcedo, G., Carbonero, P. and García-Olmedo, F. (1983) *Planta* 157, 74-80.
- [15] Paz-Ares, J., Ponz, F., Rodríguez-Palenzuela, P., Lázaro, A., Hernández-Lucas, C., García-Olmedo and Carbonero, P. (1986) *Theor. Appl. Genet.* 71, 842-846.
- [16] Lázaro, A., Sánchez-Monge, R., Salcedo, G., Paz-Ares, J., Carbonero, P. and García-Olmedo, F. (1988) *Eur. J. Biochem.* 172, 129-134.
- [17] Lázaro, A., Rodríguez-Palenzuela, P., Marañón, C., Carbonero, P. and García-Olmedo, F. (1988) *FEBS Lett.* 239, 147-150.
- [18] Rodríguez-Palenzuela, P., Royo, J., Gómez, L., Sánchez-Monge, R., Salcedo, G., Molina-Cano, J.L., García-Olmedo, F. and Carbonero, P. (1989) *Mol. Gene. Genet.* 219, 474-479.
- [19] Halford, N.G., Morris, N.A., Urwin, P., Williamson, M.S., Kasarda, D.D., Lew, W.J.-L., Kreos, M. and Shewry, P.R. (1988) *Biochim. Biophys. Acta* 950, 435-440.
- [20] García-Maroto, F., Marañón, C., Mena, M., García-Olmedo, F. and Carbonero, P. (1990) *Plant Mol. Biol.* 14, 845-853.
- [21] Edens, L., Heslinga, L., Klok, R., Ledebor, A.M., Maat, J., Toonen, M.Y., Visser, C. and Verrips, C.Th. (1982) *Gene* 18, 1-12.
- [22] Smekens, S., de Groot, M., van Binsbergen, J. and Weisbeek, P. (1985) *Nature* 317, 456-458.
- [23] Rother, C., Jansen, T., Tyagi, A., Tittgen, J. and Herrmann, R.G. (1986) *Curr. Genet.* 11, 171-176.
- [24] Vorst, O., Oosterhoff-Teertstra, R., Vankan, P., Smekens, S. and Weisbeek, P. (1988) *Gene* 65, 59-69.
- [25] Ponz, F., Paz-Ares, J., Hernández-Lucas, C., García-Olmedo, F. and Carbonero, P. (1986) *Eur. J. Biochem.* 156, 131-135.
- [26] Hernández-Lucas, C., Royo, J., Paz-Ares, J., Ponz, F., García-Olmedo, F. and Carbonero, P. (1986) *FEBS Lett.* 200, 103-105.
- [27] Rodríguez-Palenzuela, P., Pintor-Foro, J.A., Carbonero, P. and García-Olmedo, F. (1988) *Gene* 70, 271-281.
- [28] Gausing, K. (1987) *Planta* 171, 241-246.
- [29] Bohlmann, H. and Apel, K. (1987) *Mol. Gen. Genet.* 207, 446-454.