TIC110, INVOLVED IN CHLOROPLAST PROTEIN TRANSLOCATION, CONTAINS HIGHLY DIVERGENT HEAT-LIKE REPEATED MOTIFS, AS REVEALED BY HYDROPHOBIC CLUSTER ANALYSIS

ADRIAN JOSE JAIMES BECERRA

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE CIENCIAS
ESCUELA DE BIOLOGIA
2009

TIC110, INVOLVED IN CHLOROPLAST PROTEIN TRANSLOCATION, CONTAINS HIGHLY DIVERGENT HEAT-LIKE REPEATED MOTIFS, AS REVEALED BY HYDROPHOBIC CLUSTER ANALYSIS

ADRIAN JOSE JAIMES BECERRA

Trabajo de investigación para optar por el titulo de biólogo

DIRECTOR:
Jorge Hernández Torres
PhD Ciencias

CODIRECTOR:
Jacques Chomilier
PhD Física

UNIVERSIDAD INDUSTRIAL DE SANTANDER
FACULTAD DE CIENCIAS
ESCUELA DE BIOLOGIA
2009

CONTENIDO

Pág.

LISTA DE FIGURAS

Pág.

RESUMEN

TITULO: TIC110, INVOLUCRADA EN EL PROCESO DE TRANSPORTE DE CLOROPLASTICO CONTIENE REPETICIONES SEMEJANTES A HEAT ALTAMENTE DIVERGENTES REVELADO POR HCA[*].

AUTOR: ADRIAN JOSE JAIMES BECERRA[**].

PALABRAS CLAVE: Tic110, HCA, Repetición HEAT, Translocon del cloroplasto, Transporte de membrana.

RESUMEN: La teoría endosimbiotica admite que los plastidios se originaron a partir de una bacteria fotosintética tragada por una célula eucariótica primitiva. En consecuencia, el genoma del cloroplasto permanece como una reminiscencia de este evento ancestral, aunque reducido en tamaño y número de genes. La mayoría de los genes en el plastidio fueron transferidos al genoma nuclear de la célula huésped y desde entonces, estos son codificados en el núcleo. Así, las proteínas del cloroplasto son sintetizadas en citósol con extensiones N-terminales llamadas péptidos transito y la maquinaria de importe fue requerida que apareciera en la evolución para transferir estas proteínas al estroma. Hasta ahora, dos complejos de proteínas han sido encontrado que median el proceso de importe: los translocones de la envoltura de la membrana Toc (externa) y Tic (interna). El origen evolutivo de muchas de las proteínas de los complejos Toc y tic han sido establecidos, pero esto no es aún el caso para la subunidad Tic110. Tic110 enlaza péptidos señal y sirve como un andamio molecular para el reclutamiento de componentes estromales. En este estudio después de un análisis de agregados hidrofóbicos (HCA) y reconocimiento de pliegue de las proteínas, nosotros concluimos que Tic110 esta compuesto de un seria de *motifs* repetidos relacionados a las repeticiones HEAT.

La explicación para la presencia de tales repeticiones en Tic110 y su función más probable en el proceso de importe del cloroplasto es discutida.

---

[*] Trabajo de Grado

[**] Facultad de Ciencias. Escuela de Biología. Director: Jorge Hernández Torres PhD. Ciencias. Co-Director: Jacques Chomilier PhD Física

ABSTRACT

TITLE: Tic110, involved in chloroplast protein translocation, contains highly divergent HEAT-like repeated motifs, as revealed by Hydrophobic Cluster Analysis.*

AUTOR: ADRIAN JOSE JAIMES BECERRA**.

KEYWORDS: Tic110; HCA; HEAT-repeat; chloroplast translocon; membrane transport.

ABSTRACT: The endosymbiotic theory admits that plastids originated from a photosynthetic bacterium engulfed by a primitive eukaryotic cell. In consequence, the chloroplast genome remains as reminiscences of this ancestral event, although reduced in size and number of genes. Most part of the plastid genes were transferred to the host cell nuclear genome and since then, they are nuclear-encoded. Thus, chloroplast proteins are synthesized in the cytosol as precursors with N-terminal extensions called transit peptides and import machinery was required to appear in evolution to transfer those proteins to the stroma. Until the present time, two protein complexes have been found to mediate the import process: the Toc (outer) and Tic (inner) envelope membrane translocons. The evolutionary origin of many of Tic and Toc proteins have been established, but this is not yet the case for the Tic110 subunit. Tic110 binds signal peptides and serves as a scaffold for the recruitment of stromal components. In this study, after a keen analysis of hydrophobic clusters (HCA) and protein fold recognition, we concluded that Tic110 is composed of a series of repeated motifs related to HEATrepeats.
The explanation for the presence of such repeats in Tic110 and its probable function in the chloroplast import process is discussed.

---

* Degree Work

** Science Faculty. School of Biology. Director: Jorge Hernández Torres PhD. Ciencias. Co-Director: Jacques Chomilier PhD Física

INTRODUCTION

To date the endosymbiotic theory widely accepted assumes that chloroplasts have originated from a photosynthetic bacterium taken up by a primitive eukaryotic cell. As a consequence, most of the original prokaryotic genome was lost or transferred to the host cell nuclear genome [1, 2]. From this event, most proteins in chloroplasts are encoded by the nuclear genome and synthesized in the cytosol as precursors with N-terminal targeting signals called transit peptides. These proteins are translocated into the chloroplast by the so-called general import pathway, mediated by the Toc (outer) and Tic (inner) envelope membrane translocons [3, 4, 5].

In contrast to the Toc complex in which the activities, functions and evolutionary origins of the largest part of components of the machinery have been established and well defined [6,7,8], the study of Tic complex has been hard to undertake by the fact that assembly of functional complexes is dynamic and occurs in response to preprotein translocation. Besides, from the point of view of sequence analysis, Tic proteins do not show homology to any known membrane transport system [9,10].

Among the various members of the Tic machinery, Tic110 (translocon at the inner envelope protein of 110 kDa) was the first to be identified as a true subunit [11, 12]. Moreover, it was proposed to be one of the essential components for protein translocation into plastids in association with at least two other Tic subunits, namely Tic20 and Tic40 [13, 14]. The membrane arrangement of Tic110 is still in debate, although it most likely projects a large domain into the stroma [13, 15]. According to previous studies on Tic110 stromal domain, it binds transit peptides and serves as a docking module for the recruitment of stromal components involved in late stages of protein import [11, 13, 14].

The evolutionary origin of Tic subunits has been established for at least four components, Tic20, Tic22, Tic55 and Tic62, which seem to derive from a cyanobacterial ancestor, since they show significant similarities to bacterial proteins. No homologous sequence has been found for Tic110 neither in bacteria nor in eukaryotic organisms, suggesting that this protein might have been newly developed in the arising plant cell, in concert with chloroplast development, or lost from ancestral cyanobacteria [10, 16].

Internal repetitive sequences within proteins have been a successful strategy throughout evolution. Protein repeats have regular secondary structures and form 3D multirepeat assemblies of diverse sizes and functions. Detection of protein repeats can be a particularly arduous task on the basis of the primary structure, because of significant sequence divergence or the short length of sequence repeats [17, 18, 19].

The aim of this paper is to provide evidence of the evolutionary origin of the chloroplast inner envelope translocon Tic110 subunit by means of thorough sequence analysis, including HCA (Hydrophobic Cluster Analysis) in combination with standard predictive methods. Herein, we propose that the largest part of Tic110 seems to be composed of at least 8 HEAT-like repeat motifs (37-50 amino acids), supplemented with N-terminal transmembrane and transit peptide domains. The HEAT repeat is a tandemly repeated, 37-47 amino acid module occurring in a number of cytoplasmic proteins, including the four name-giving proteins huntingtin, elongation factor 3 (EF-3), the 65 kDa α- regulatory subunit of protein phosphatase 2A (PP2A) and the yeast PI3-kinase TOR1.

It has been noted that many HEAT repeat containing proteins are involved intracellular transport processes. The canonical HEAT repeat consists of two helices, A and B, which form a helical hairpin; arrays of HEAT repeats consist of 3

to 36 units forming a rod-like structure and appear to function as protein-protein interactions surfaces [20].

1. MATERIALS AND METHODS

PSI-BLAST searches were performed using *Arabidopsis thaliana* atTic110 as a query, against the non redundant database at NCBI (http://www.ncbi.nlm.nih.gov/blast) [21] using the default parameters.

The 2D Hydrophobic Cluster Analysis (HCA) was performed as previously published [22]. Briefly, HCA designs a sequence on the surface of a cylinder with the connectivity of an alpha helix. The 2D planar surface is then duplicated in order to keep local environment for each amino acid, and hydrophobic residues (VILFMWY) in this plot are then clustered, provided they are first neighbors in this pattern. The shapes of the clusters are keen indication of the nature of the secondary structure. Besides, the hydrophobic clusters observed in an HCA plot are not distributed at random. Instead of this, it has been statistically demonstrated that centers of the hydrophobic clusters correspond to the centers of regular secondary structures [22].

HCA identity is calculated by the number of identical aligned hydrophobic and non hydrophobic amino acids in both sequences to the number of amino acids of the longest sequence. For each sequence, the HCA score is the ratio between the number of topologically conserved residues between sequences 1 and 2, to the total number of hydrophobic residues in both segments. A HCA score ≥60% is an indicator of high sequence identity [22].

Repeated motifs were detected with MEME [23] and REP [24] programs, which are homology based methods that use iterative algorithms to estimate the significance of possible repeats in a sequence. The SBASE program [25] has been employed to predict known domains from the single information of the sequence. The PSIPRED protein structure prediction server [26] was used to predict the

secondary structure [27] and fold recognition was performed using mGenTHREADER [28, 29]. Multiple fold recognition methods were done using the CBS Meta-Server [30].

2. RESULTS

2.1 HYDROPHOBIC CLUSTER ANALYSIS REVEALS DUPLICATION EVENTS WITHIN TIC110

HCA is a fine method of 2D structural analysis that allows alignments between very distantly related proteins, with as low as 10% sequence identity (the 'twilight zone', [22]). HCA analysis does not pay any attention to the strict conservation of the residues inside the clusters but rather to the conserved shapes of the clusters, keeping in mind the underlying idea that shape is a testimony of the secondary structure. Thus, HCA is a powerful tool to detect among others, internal repeated motifs or domains of high level of divergence.

By performing a detailed HCA analysis of the Tic110 protein, we suspected the existence of domain duplications along the protein, as it has been proposed for other proteins of Tic and Toc [8,31]. As shown in Fig. 1, at least 3 copies of an unknown domain can be detected by HCA into Tic110. However, it was intriguing to us that individual motifs within one domain could be aligned with several motifs of the other domains. It resulted that a short motif (37-50 amino acids) with high level of leucine is repeated tandemly throughout the protein. In Fig. 2a we aligned 8 of such motifs one can find in the *Arabidopsis thaliana* Tic110 protein, with 6 other plant species. An outstanding feature inside each motif is the regular arrangement of two blocks enriched in hydrophobic residues (vertical lines ending with a *) separated in two sections by charged or polar amino acids. We were able to align the putative motifs of *Arabidopsis thaliana* Tic110 protein by HCA, and we can observe in Fig. 2b that while they exhibit a relative level of degeneracy, the number and shape of hydrophobic clusters are conserved, as well as non hydrophobic residues. We concluded that Tic110 could be originated by duplications in tandem of this unknown motif, which covers more than 70% of the

protein. The HCA score, which is the ratio of the number of hydrophobic residues occupying equivalent positions in both sequences over the total number of hydrophobic residues, was in the range of 60 to 90 %.

2.2 RECOGNITION OF HEAT-LIKE MOTIFS BY COMPUTATIONAL METHODS

MEME (Multiple EM for Motif Elicitation) is a tool for discovering motifs in a group of related DNA or protein sequences [23]. Sequence analysis of Tic110 by the MEME algorithm identified four repetitions composed of ≈ 50 amino acid, predicted to contain two α helix with an E-value of 8.2e-280. Subsequently, we submitted the entire Tic110 sequence and the putative motifs to the REP servers focusing on all major classes of protein repeats [24] but no sequence similarity was found to known repeats.

Nevertheless, the pattern search service of the SBASE protein domain library [25] yielded more suitable results and detected the presence of one Helix hairpin Helix motif with 99% confidence.

Since this attempt to predict motifs and domains was not entirely concluding, we changed our approach by performing fold recognition and threading methods that can be used to assign tertiary structures to protein sequences, even in the absence of clear homology [28]. The PSIRED protein structure prediction server [26] revealed that some fragments of the Tic110 protein could share structural similarity with HEAT repeats with a confidence of $p < 0.0001$ in the range of 70%-80%. This prediction was confirmed by the multiple fold recognition methods of the CBS Meta-Server [30]. A striking observation is that the same motifs we discovered by HCA alignments are the same that were detected by fold recognition servers.

Taken together these results, we concluded that Tic110 proteins contain at least 8 putative HEAT-like sequences (37-50 amino acids). The consensus for the

secondary structure of each HEAT-like motif is a pair of helical domains (Fig. 2a, upper cylinders), separated by a non-helical spacer. Besides, 7 conserved hydrophobic residues are located at positions ca. 10, 13, 17, 24, 28, 32 and 35 (Fig. 2a, vertical lines).

The consensus residue for these positions is leucine, excepting positions 24 and 28, where valine and alanine are more frequent, respectively. The repetition motifs found within Tic110 by HCA analysis and fold predictors fit well with the basic pattern of the HEAT repeats: two α-helices separated by a variable region. We show in Fig. 3, HCA alignments between two HEAT-like and already characterized HEAT motifs of human protein phosphatase 2A, a 65 kDa regulatory subunit (Fig. 3a, 15 tandemly repeated HEAT motifs) and Cand1, a 120 kDa HEAT repeat protein (Fig. 3b, 27 HEAT repeats).

It is noteworthy the strong conservation of hydrophobic clusters and non hydrophobic amino acids. Besides, the shape of the clusters is fully compatible with α-helices [22], indicating unequivocally the existence of two α-helices separated by a loop of polar or charged residues.

We propose in Fig. 3c a model of the *Arabidopsis thaliana* Tic110 protein, where at least 8 HEAT-like repeats are arranged in groups of 3 to 4. Furthermore, recognized domains according to biochemical analyses are also underlined.

3. DISCUSSION

Although previous studies based on *in vitro* analyses have given some insights into the function and topology of the Tic110 protein in the inner membrane of chloroplasts [13, 14], its evolutionary origin has been impossible to determine by traditional 1D sequence alignments or bank screening methods.

The HCA methodology in combination with standard prediction methods enabled us to evidence highly divergent HEAT-like motifs that have been duplicated across the protein for at least 8 times. This task would be unattainable by other alignment methods, particularly because the absence of a Tic110 3D structure and the fact that HEAT-like repeats are quite degenerated throughout evolution.

Since internal repetition affords an enlargement of its available interacting surface, it is also clear that the most common function of repeat ensembles is protein binding [20].

Such a global structure provides opportunities to the new proteins to expand its repertoire of cellular functions, such as protein-protein interaction, transport across membranes and protein-complex assembly [33]. Our results could explain most of the *in vitro* data published so far, since Tic110 has revealed potential roles in key elements of protein import into plastids, including Tic complex assembly, preprotein binding and the recruitment of molecular chaperones to import sites [14]. Although HEAT-repeat proteins are involved in a great diversity of cellular processes, a common function is to mediate multiple protein-protein interactions like a scaffold, for translocation processes of diverse macromolecules [34, 35].

One can advance the following arguments to support the presence of HEAT repeats within Tic110 protein: i) Predicted Tic110 HEAT-like repeats accomplish

the structural requirements to configure two helical patches with appropriate spacing, ii) each predicted repeat contain sequence elements that match the consesus sequence for HEAT repeats, mainly on the hydrophobic core and iii) Tic110 and HEAT-repeat proteins involve protein-protein interactions in their activities, such as assembly and membrane transport [35]. These critical actions would require multiple HEAT-repeats to collect an assortment of subunits to assemble active complexes. Because other HEAT-repeat proteins play key roles in transport processes as scaffolding, this finding provides a plausible mechanistic explanation for the origin of the Tic110 proteins.

Figura 1. HCA plot of Arabidopsis thaliana Tic110 protein showing putative repeated domains. Because of the duplication (see methods), sequence is read vertically, one line over two, and the secondary structure is read horizontally, a cluster corresponding statistically to a regular secondary structure. (The onset of Fig. 2 helps interpreting the HCA plots). Vertical lines connect the occurrences of analogous clusters. Conserved hydrophobic clusters are shaded in grey. Strict identities are indicated by white letters on a black background. H1 and H2 represent two predicted α-helices (See the text)**.**

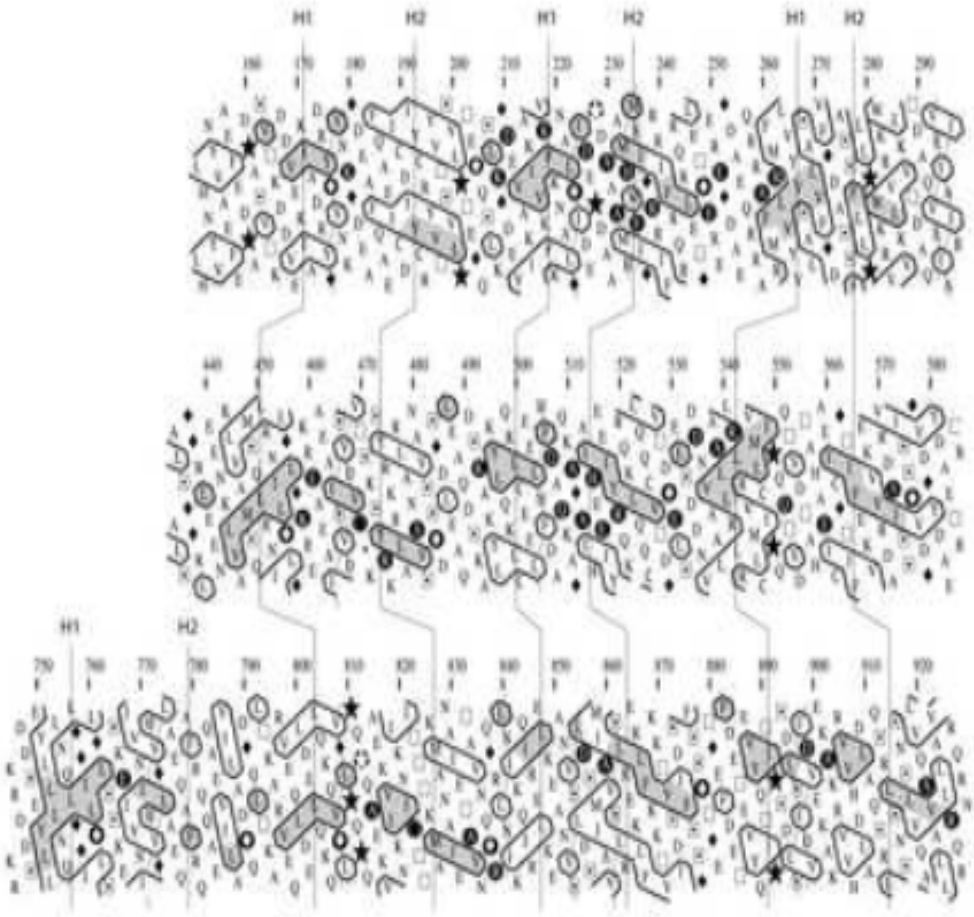


Figura 2. a) 1D alignment of repeated motifs within Tic110, from green plants (land plants and green algae). Above the alignment, predicted secondary structures are displayed, where α-helices are represented by cylinders. Vertical lines ending with * indicate conserved hydrophobic positions. Abbreviations are: ARATH, Arabidopsis thaliana (NP_172176.1); MEDTR, Medicago truncatula (ABE84639.1); OSTLU, Ostreococcus lucimarinus (XP_001418787.1); ORYSA, Oryza sativa (AAP54402.2); PHYPA, Physcomitrella patens (XP_001785732.1); PISSA, Pisum sativum (CAA92823.1); POPTR, Populus trichocarpa (XP_002326080.1). b) Multiple HCA alignment of Tic110 HEAT-like repeat motifs. Vertical lines indicate correspondence between putative α-helices.

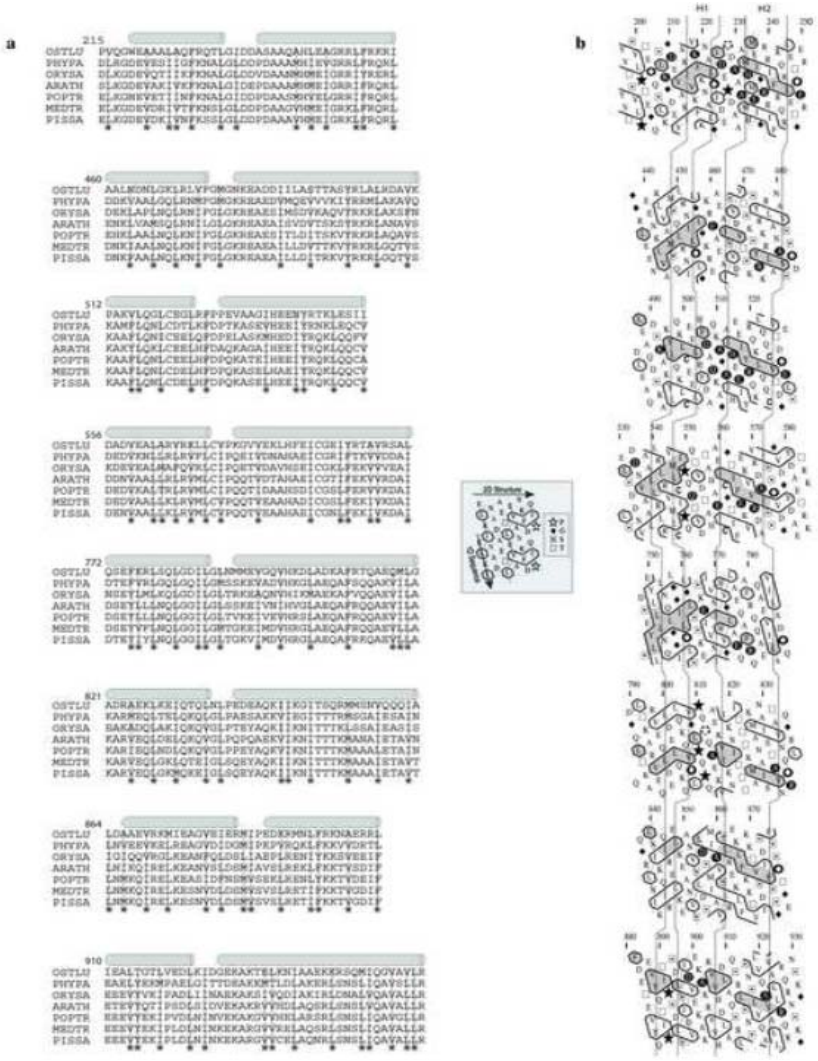
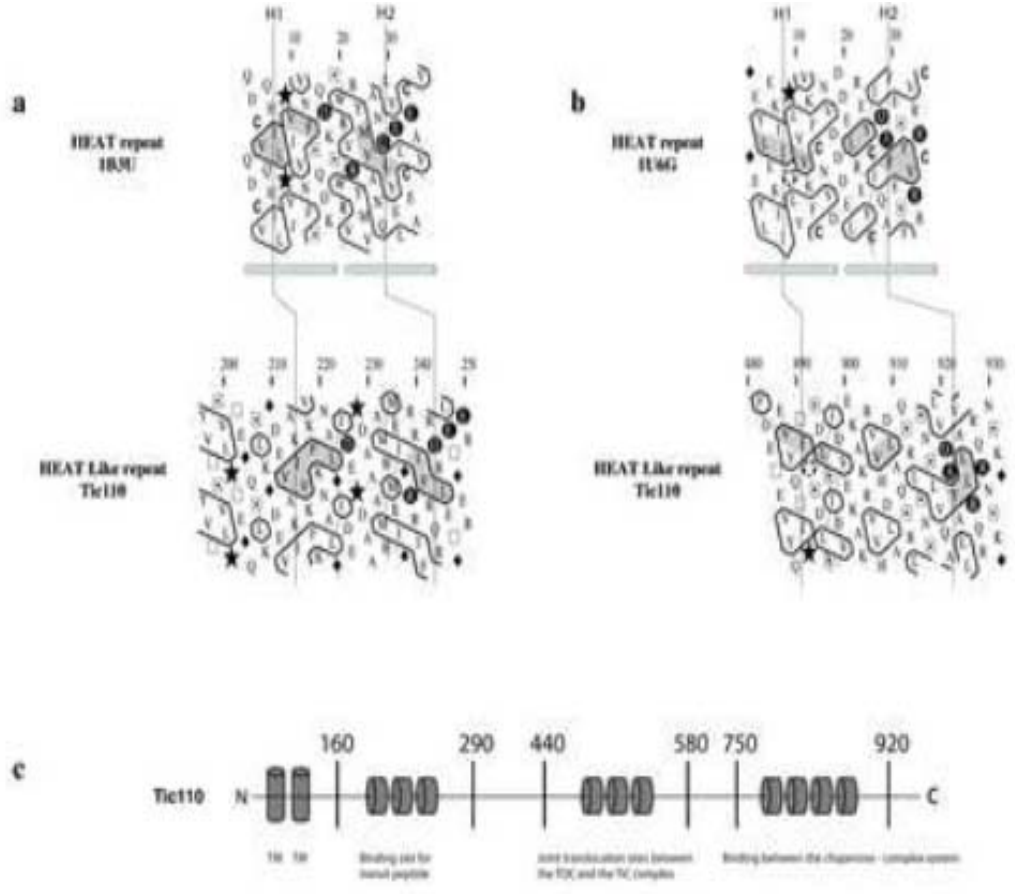


Figura 3. HCA plots of pairs of aligned sequences. Tic110 HEAT-like repeated motifs are aligned with well known HEAT-repeats a) from human protein phosphatase 2A, regulatory subunit (1B3U) and b) from Cand1 (1U6G). c) Proposed model of Tic110 showing tandems of predicted HEAT-like repeats. Functional domains are underlined according to [14].



BIBLIOGRAFIA

[1] T. Cavalier-Smith, Membrane heredity and early chloroplast evolution, Trends Plant Sci. 5 (2000) 174-182.

[2] G.I. McFadden, Primary and secondary endosymbiosis and the origin of plastids, J.Phycol. 37 (2001) 951–59

[3] A. Stengel, J. Soll, B. Bolter, Protein import into chloroplast: new aspects of a well known topic, Biol. Chem. 388 (2007) 765-772.

[4] P. Jarvis, Targeting of nucleus encoded proteins to chloroplast in plants, New Phytol. 179 (2008) 257-85.

[5] J. Benz, J. Soll, B.Bolter, Protein transport in organelles: The composition function and regulation of the Tic complex in chloroplast protein import, FEBS journal 276 (2009) 1166-1176.

[6] A. Hiltbrunner, J. Bauer, PA. Vidi, S. Infanger, P. Weibel, M. Hohwy, F. Kessler, Targeting of an abundant cytosolic form of the protein import receptor at Toc159 to the outer chloroplast membrane, J. Cell Biol. 154 (2001) 309–316.

[7] F. Kessler, D.J. Schnell, The function and diversity of plastid protein import pathways: a multilane GTPase highway into plastids, Traffic 7 (2006) 248–257.

[8] J. Hernandez, M. Arias, J. Chomilier, Tandem duplications of a degenerated GTPbinding domain at the origin of GTPase receptors Toc159 and thylakoidal SRP. Biochem. Biophys. Res. Commun, 364 (2007) 325–331.

[9] A. Kouranov, X. Chen, B. Fuks, D..J. Schnell, Tic20 and Tic22 are new components of the protein import apparatus at the chloroplast inner envelope membrane, J. Cell Biol. 143 (1998) 991–1002.

[10] S. Reumann, K. Inoue, K. Keegstra, Evolution of the general protein import pathway of plastids, Mol. Membr. Biol. 22 (2005) 73–86.

[11] F. Kessler, G. Blobel, Interaction of the protein import and folding machineries of the chloroplast, Proc. Natl. Acad. Sci. 93 (1996) 7684–7689.

[12] J. Lubeck, J. Soll, M. Akita, E. Nielsen, K. Keegstra, Topology of IEP110, acomponent of the chloroplastic protein import machinery present in the inner envelope membrane, EMBO J. 15 (1996) 4230–4238.

[13] T. Inaba, M. Li, M. Alvarez, F. Kessler, D.J. Schnell, atTic110 functions as a scaffold for coordinating the stromal events of protein import into chloroplasts, J. Biol. Chem. 278 (2003) 38617–38627.

[14] T. Inaba, M. Alvarez, M. Li, J. Bauer, C. Ewers, F. Kessler, D.J. Schnell, *Arabidopsis* tic110 is essential for the assembly and function of the protein import machinery of plastids, Plant Cell 17 (2005) 1482–1496.

[15] D. Jackson, J. Froehlich, K. Keegstra, The hydrophilic domain of Tic110, an inner envelope membrane component of the chloroplastic protein translocation apparatus, faces the stromal compartment, J. Biol. Chem. 273 (1998) 16583–16588.

[16] S. Reumann, K. Keegstra, The endosymbiotic origin of the protein import machinery of chloroplastic envelope membranes. Trends Plant Sci. 4 (1999) 302–307.

[17] E. Marcotte, M. Pellegrini, T. Yates, D. Eisenberg, A census of protein repeats, J. Mol. Biol. 293 (1999) 151-160.

[18] M. Andrade, C. Perez, C. Ponting, Protein repeats: Structures, Functions, and Evolution, J. Struct. Biol. 134 (2001) 117–131.

[19] A.Bjorklund, D. Ekman, A. Elofsson, Expansion of Protein Domain Repeats, PLoS Comput. Biol. 2(8) (2006) 114. DOI: 10.1371/journal.pcbi.0020114.

[20] M. Andrade, C. Petosa, S. O´Donoghue, C. Muller, P. Bork, Comparison of ARM and HEAT Protein Repeats, J. Mol. Biol. 309 (2001) 1-18.

[21] S. Altschul, T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. Lipman, Gapped-BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25 (1997) 3389-3402.

[22] I. Callebaut, G. Labesse, P. Durand, A. Poupon, Canard L, J. Chomilier, B. Henrissat, J. Mornon, Deciphering protein sequence information though hydrophobic cluster analysis. Current status and perspectives, Cell. Mol. Life Sci. 53 (1997) 621-645.

[23] T. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *In: Second International Conference on Intelligent Systems for Molecular Biology; Menlo Park, California.* (1994) 28-36.

[24] M. Andrade, C. Ponting , T. Gibson, , P. Bork, Homology-based method for identification of protein repeats using statistical significance estimates, J. Mol. Biol. 298 (2000) 521-537.

[25] K. Vlahovicek, L. Kajan, V. Agoston and S. Pongor, The SBASE domain sequence resource, release 12: prediction of protein domain-architecture using support vector machines, Nucleic Acids Research 33 (2005) 223-225.

[26] K. Bryson, L. McGuffin, R. Marsden, J. Ward, J. Sodhi, D. Jones, Protein structure prediction servers at University College London. Nucl. Acids Res. 33 (2005) 36-38.

[27] D. Jones, Protein secondary structure prediction based on position-specific scoring matrices, J. Mol. Biol. 292 (1999) 195-202.

[28] D. Jones, GenTHREADER: an efficient and reliable protein folds recognition method for genomic sequences, J. Mol. Biol. 287 (1999) 797-815.

[29] L. McGuffin, D. Jones, Improvement of the GenTHREADER method for genomic fold recognition, Bioinformatics, 19 (2003) 874-881.

[30] D. Douguet, G. Labesse, Easier threading through web-based comparisons and cross-validations, Bioinformatics, 17(8) (2001) 752-3.

[31] M. Kuchler, S. Decker, F. Hormann, J. Soll, L. Heins, Protein import into chloroplasts involves redox regulated proteins. EMBO J. 21(2002) 6136–6145.

[32] A. Lupas, C. Ponting, R. Russell, On the Evolution of Protein Folds: Are Similar Motifs in different Protein Folds the Result of Convergence, Insertion or Relics o fan Ancient Peptide World?, J. Struct. Biol. 134 (2001) 191-203.

[33] T. Street, G. Rose, D. Barrick, The role of introns in repeat protein gene formation, J. Mol. Biol. 360 (2006) 258–266.

[34] M. Groves, D. Barford, Topological characteristics of helical repeat proteins. Curr. Opin. Struct. Biol. 9 (1999) 383-389.

[35] G. Cingolani, H. Lashuel, L. Gerace, C. Muller, Nuclear import factors importin alpha and importin beta undergo mutually induced conformational changes upon association, FEBS Lett. 484 (2000) 291-298