Evidence of Evolutionary Constraints That Influences the Sequence Composition and Diversity of Mitochondrial Matrix Targeting Signals

Stephen R. Doyle*, Naga R. P. Kasinadhuni, Chee Kai Chan, Warwick N. Grant

La Trobe Institute for Molecular Sciences, La Trobe University, Bundoora, Australia

Abstract

Mitochondrial targeting signals (MTSs) are responsible for trafficking nuclear encoded proteins to their final destination within mitochondria. These sequences are diverse, sharing little amino acid homology and vary significantly in length, and although the formation of a positively-charged amphiphilic alpha helix within the MTS is considered to be necessary and sufficient to mediate import, such a feature does not explain their diversity, nor how such diversity influences target sequence function, nor how such dissimilar signals interact with a single, evolutionarily conserved import mechanism. An *in silico* analysis of 296 N-terminal, matrix destined MTSs from *Homo sapiens*, *Mus musculus, Saccharomyces cerevisiae, Arabidopsis thaliana*, and *Oryza sativa* was undertaken to investigate relationships between MTSs, and/or, relationships between an individual targeting signal sequence and the protein that it imports. We present evidence that suggests MTS diversity is influenced in part by physiochemical and N-terminal characteristics of their mature sequences, and that some of these correlated characteristics are evolutionarily maintained across a number of taxa. Importantly, some of these associations begin to explain the variation in MTS length and composition.

Citation: Doyle SR, Kasinadhuni NRP, Chan CK, Grant WN (2013) Evidence of Evolutionary Constraints That Influences the Sequence Composition and Diversity of Mitochondrial Matrix Targeting Signals. PLoS ONE 8(6): e67938. doi:10.1371/journal.pone.0067938

Editor: Niall James Haslam, University College Dublin, Ireland

Received February 5, 2013; **Accepted** May 23, 2013; **Published** June 25, 2013

Funding: SRD was supported by an Early Career Development Fellowship from La Trobe University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: s.doyle@latrobe.edu.au

Introduction

It is generally accepted that mitochondria have evolved from an alpha-proteobacterium that was engulfed by an ancestral eukaryotic cell over one billion years ago. Over time, almost all of the original bacterial genes have been translocated, such that now the majority of mitochondrial proteins are transcribed in the nucleus, translated in the cytoplasm and are actively trafficked, potentially through multiple cellular compartments, to reach their final destination. The gradual loss of autonomy to the nucleus would have required a number of independent events to occur [1,2]: the transfer and integration of bacterial-derived genetic information into the nuclear genome, followed by genetic modifications to allow nuclear transcription and regulation, translation on cytoplasmic ribosomes and lastly, trafficking of the protein to its correct destination within the mitochondria. This trafficking process is mediated by multiple molecular interactions between the import apparatus and a mitochondrial targeting signal (MTS) sequence, a 'molecular address' that facilitates import and sorting to its correct destination (see Chacinska et al., [3], Mokranjac and Neupert [4] and Schleiff and Becker [5] for comprehensive descriptions of the import pathway).

Although it is nearly a quarter of a century since they were initially described [6–8], MTSs remain poorly characterised. MTSs from different proteins share virtually no sequence homology and vary extensively in length; proteins targeted to the mitochondrial matrix do however typically contain an N-terminal MTS and that these targeting sequences are loosely related by some characteristic physicochemical properties, including being rich in hydroxylated and basic residues, deficient in acidic resides [9], and most likely exhibiting a tendency to form an amphiphilic alpha helix (i.e. with opposing positively charged and hydrophobic faces) [6]. Some simple residue conservation sometimes applies, such as the proposed $\phi\chi\chi\phi\phi$ motif (where ϕ is hydrophobic and χ is any other residue) thought to mediate recognition of the presequence by Tom20 [10,11], and the -10/-3/-2 arginine motif [12], which has been proposed to be involved in cleavage site recognition by the mitochondrial processing peptidase (MPP). However, the determinants of cleavage recognition have not been fully elucidated as MTSs may contain either one or two cleavage sites (a second site recognised by the mitochondrial intermediate peptidase, MIP) or no cleavage site at all (i.e. the 'mature' protein retains the MTS in the matrix). These limited characteristics have been used in the development of software to predict signal sequences [13–16], however, the significant variation in both sequence composition and length of MTSs renders this task difficult, so that a predicted MTS must be verified experimentally. Furthermore, the majority of research dedicated to matrix protein import is focused primarily on understanding the molecular interactions between the components of the import translocases and the MTS is often overlooked, possibly due to the lack of obvious explanation for its diversity.

It is interesting to speculate on the nature of the selective pressures that may have been exerted on the development of any given matrix-targeted MTS. Matrix-destined MTSs have a functional role in multiple cellular compartments (cytoplasm, inner-membrane space and matrix space) as they interact with cytosolic chaperones and several different components of the membrane-bound mitochondrial import and processing machinery. Moreover, they direct a vast range of different proteins through a single import pathway to the matrix space. Although formation of an amphiphilic alpha helix within an N-terminal MTS is broadly described to be necessary and sufficient to direct import [4,17,18], it seems to be an overly simple explanation for a complex role, and is unlikely to be relevant in all stages of import [19]. Given the significant variation among MTSs and the complex processes they facilitate, it is more likely that there are additional functionally significant MTS characteristics that have not been described.

In this investigation, an *in silico* analysis of N-terminal, matrix-destined MTSs from five diverse taxa (*Homo sapiens*, *Mus musculus*, *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, and *Oryza sativa*) was undertaken to search for evolutionarily conserved characteristics that could shed light on the evolutionary relationships between diverse MTSs, and/or, between an individual MTS and the protein that it imports. We hypothesised that sequence characteristics of the mature portion of the protein to be imported are responsible for the extensive variation in MTS sequence composition and length. A number of physicochemical correlations were identified that support this hypothesis, and these results are discussed in reference to the evolutionary adaptation required of MTSs to enable efficient mitochondrial import.

Methods

Data mining

A sequence database of mitochondrial matrix-localised proteins was constructed in November 2012 from sequences retrieved from NCBI GenBank using Gene Ontology with the keywords 'mitochondrial matrix' and UniProt (http://www.uniprot.org/) databases, as well as from other sources including Appendix 4 of Methods in Cell Biology, Vol. 65, and websites such as the Human Mitochondrial Protein Database (http://bioinfo.nist.gov/hmpd/index.html) and MitoProteome Database (http://www.mitoproteome.org/). A large proportion of the *S. cerevisiae* [20], and all of the *A. thaliana* and *O. sativa* [21] sequences were obtained from previously published proteomics databases. Each candidate sequence was required to meet the following criteria: (i) the mature protein is imported and localised within the mitochondrial matrix, (ii) the mature protein is not part of the inner-mitochondrial membrane (iii), the sequence has a defined N-terminal targeting signal, and (iv) the sequence was catalogued within the NCBI protein reference sequence database, which was used to acquire accession numbers and to extract the amino acid sequence. Dataset S1 contains a complete list of the mitochondrial matrix proteins, gene IDs, accession numbers, MTS sequences and their mature primary amino acid sequences for all sequences used in this study.

Primary amino acid sequence analysis

Full length sequences were broken into two sub-sequences; the first containing the MTS and the second containing the remaining portion of the sequence found in the mature protein (designated 'mature sequence' or 'mtpt' throughout this study). Amino acid composition, length, and physicochemical characteristics were determined using a Perl program, *AAResidueFreqCalculator*, which enabled frequencies to be determined from a FASTA sequence input (see File S1 for Perl script and running details). Charge characteristics at pH 7.5 were calculated using CLC Genomics Workbench (5.5.1). Isoelectric point (pI) was calculated using the Compute pI/Mw tool (http://web.expasy.org/cgi-bin/compute_pi/pi_tool). Secondary structure analysis was undertaken using the tools PROFsec and MD (www.predictprotein.org, [22]).

A reduced amino acid alphabet [23] was developed to analyse physicochemical characteristics in their primary amino acid sequence context, which enabled the investigation of amino acid residue-type effects (as opposed to individual amino acid residue-specific effects). The 20-letter amino acid code was divided into nine groups; non-polar aliphatic (G, A, V, I, L), non-polar aromatic (F, W), non-polar cyclic (P), polar sulphur containing (C, M), polar hydroxyl (S, T), polar aromatic (Y), polar acidic-amide (N, Q), acidic (D, E), and basic (R, H, K), according to the biochemical properties explorer at NCBI (http://www.ncbi.nlm.nih.gov/Class/Structure/aa/aa_explorer.cgi). Conversion of each amino acid sequence from the standard alphabet to the reduced amino acid alphabet was performed using the Perl script, *ReducedAASequenceConverter* (see File S2 for Perl script and running details) prior to further analysis.

Statistical analyses

Data analysis, including multivariate principal component analysis (PCA), Spearman's rho correlation and regression analysis, Kruskal-Wallis test, and graph construction was performed using SPSS (version 17). All other data manipulation and analysis was performed in Microsoft Excel (2010).

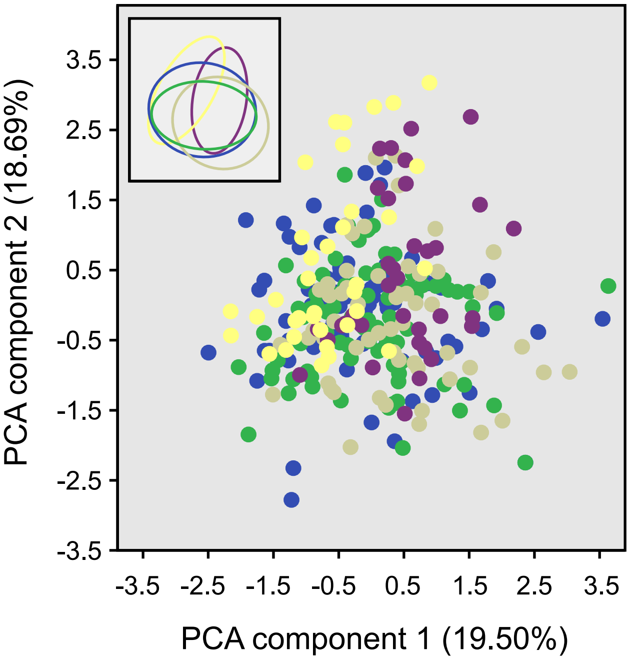
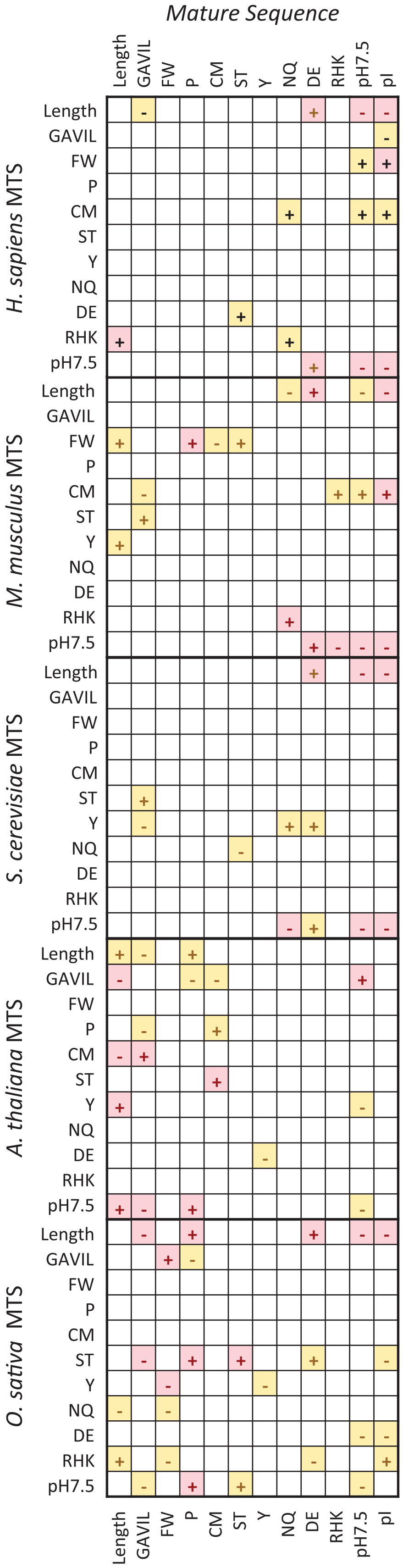


Figure 1. Comparison of MTS diversity among five taxa using multivariate PCA. The first two components are plotted, which represent 38.19% of variation among MTS sequences. Insert depicts simplified version of the scatterplot to emphasise boundaries of distribution. Blue – *H. sapiens*, Green – *M. musculus*, Beige – *S. cerevisiae*, Purple – *A. thaliana*, Yellow – *O. sativa*.
doi:10.1371/journal.pone.0067938.g001



Mature Sequence

		Length	GAVIL	FW	P	CM	ST	Y	NQ	DE	RHK	pH7.5	pI
H. sapiens MTS	Length		-							+		-	-
	GAVIL												-
	FW											+	+
	P												
	CM								+			+	+
	ST												
	Y												
	NQ												
	DE						+						
	RHK	+							+				
	pH7.5									+		-	-
M. musculus MTS	Length								-	+		-	-
	GAVIL												
	FW	+			+	-	+						
	P												
	CM		-								+	+	+
	ST		+										
	Y	+											
	NQ												
	DE												
	RHK								+				
	pH7.5									+	-	-	-
S. cerevisiae MTS	Length									+		-	-
	GAVIL												
	FW												
	P												
	CM												
	ST		+										
	Y		-						+	+			
	NQ						-						
	DE												
	RHK												
	pH7.5								-	+		-	-
A. thaliana MTS	Length	+	-		+								
	GAVIL	-			-	-						+	
	FW												
	P		-			+							
	CM	-	+										
	ST					+							
	Y	+										-	
	NQ												
	DE							-					
	RHK												
	pH7.5	+	-		+							-	
O. sativa MTS	Length		-		+					+		-	-
	GAVIL			+	-								
	FW												
	P												
	CM												
	ST		-		+		+			+			-
	Y			-				-					
	NQ	-		-									
	DE											-	-
	RHK	+		-						-			+
	pH7.5		-		+		+					-	

Figure 2. Correlation matrix to analyse pairwise characteristics between MTS and mature sequence. Spearman's rho correlation coefficients were calculated for each pair of variables between MTS (y-axis) and mature (x-axis) sequences for each species. Positive (+) and negative (-) correlations are indicated within each square. Yellow shading indicates p<0.05, red shading indicates p<0.01. Blank squares indicate no significant correlation was present between the pair of variables.

doi:10.1371/journal.pone.0067938.g002

Results and Discussion

Multivariate analysis of MTSs and associations with their mature sequence

A database and literature search yielded an extensive list of mitochondrial matrix proteins, of which a total of 296 candidates from five diverse taxa met the defined selection criteria. This list was comprised of two mammalian species, *H. sapiens* (85 sequences) and *M. musculus* (84), the bakers yeast *S. cerevisiae* (56), and two plant species, the monocot *O. sativa* (36) and the dicot *A. thaliana* (35). The primary amino acid sequence for each protein was divided into two sequences (MTS and the mature sequence), after which a range of sequence properties was determined for each MTS/mature protein pair. These properties included sequence length and charge (pI and charge at pH 7.5), as well as the proportion of nine groups of amino acids segregated based on shared physicochemical characteristics that may influence potential functional properties of the sequence (Table S1).

To analyse these characteristics, multivariate principle component analysis (PCA) was used to explore variation among the five species. Five components were found to explain 71.47% of the variance among MTS; sequence length and charge at pH 7.5 (19.5%), non-polar aliphatic and aromatic residues (18.69%), polar aromatic and acidic-amide residues (12.51%), polar hydroxyl and basic residues (11.51%), and finally acidic residues (9.61%). A pairwise comparison of the first two components demonstrated that all five species had diverse but somewhat overlapping distributions that reflect their phylogenetic history, with *H. sapiens* and *M. musculus* most similar, followed by *S. cerevisiae* and then both plant species (Figure 1). Some subtle differences in variance were evident, particularly between *O. sativa* and *A. thaliana*, which shared only a small proportion of their distribution. These subtle differences between the two plant species remained consistent across all pairwise PCA comparisons (Figure S1). Differences in MTS composition, particularly surrounding the cleavage site [21], and in Tom20 genomic copy number (*A. thaliana* contains four copies of the Tom20 gene whereas *O. sativa* only contains a single copy [24,25]) between both plant species have been described previously.

To investigate the hypothesis that features of a mature sequence influence the composition of the associated MTS, both MTS and mature sequence characteristics were analysed together using PCA. A significant positive association between both sequence length and charge at pH 7.5 of MTSs and the proportion of acidic residues in the mature sequence was found in the *H. sapiens*, *M. musculus* and *S. cerevisiae* datasets. A similar positive correlation between MTS length and charge and the mature sequence acidic residue content was seen in the *O. sativa* dataset, however, the charge association was due to a correlation with MTS basic residues rather than charge at pH 7.5 as seen in the *H. sapiens*, *M. musculus* and *S. cerevisiae* datasets. Acidic residues of the import machinery have been shown to be essential for import [26–31] and offer potential interaction sites for a positively charged MTS. These essential acidic regions, and the cation-selective nature of Tom40 [32], are likely responsible for the characteristic deficiency of acidic residues within the MTS due to negative selection from electrostatic repulsion, as import would be inhibited if the MTS

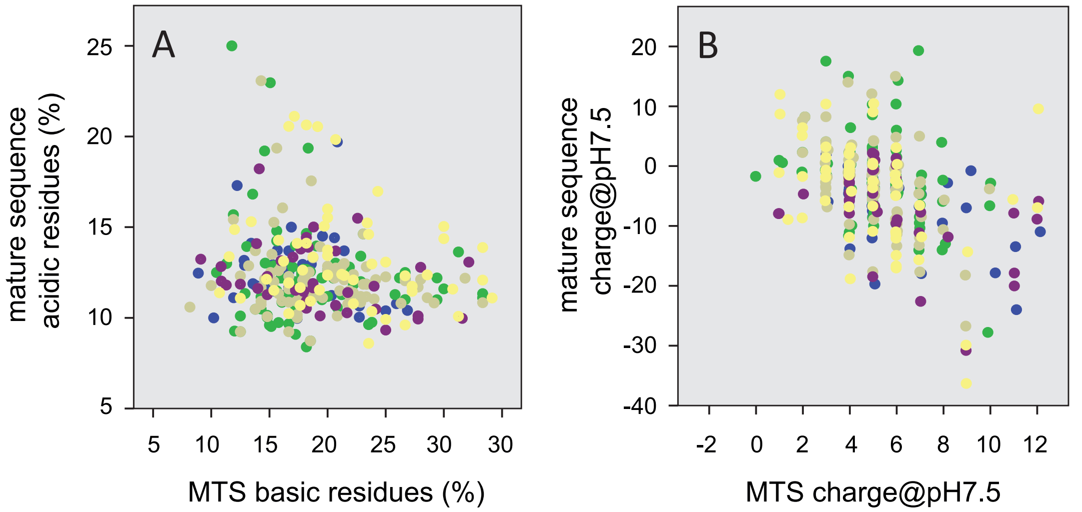


Figure 3. Correlation analysis of charge characteristics between MTS and mature sequences. Scatterplots were generated to investigate (A) basic residue composition (percentage of total) of the MTS against acidic residue composition of the mature sequence, and (B) global sequence charge characteristics by comparing charge at pH 7.5 of both MTS and mature sequences. Blue – *H. sapiens*, Green – *M. musculus*, Beige – *S. cerevisiae*, Purple – *A. thaliana*, Yellow – *O. sativa*.
doi:10.1371/journal.pone.0067938.g003

also contained a significant proportion of acidic residues. The residue composition of the mature sequences did not significantly differ from the global average amino acid composition of the proteome (data not shown), and therefore, mature sequences that contain and require acidic residues for normal structure and function would experience some electrostatic repulsion as they proceed though the import complexes following passage of the MTS. It is therefore possible that a longer MTS associated with more acidic mature sequence would traverse the outer membrane (via Tom40) and engage the membrane potential and mtHsp70 [33] at the inner membrane earlier relative to a shorter MTS, and thereby may permit the import machinery to overcome the potentially inhibitory interaction between acidic residues of mature proteins and the import complexes. Interestingly, the proportion of acidic residues found in the mature sequence did not correlate with any property of MTSs within the *A. thaliana* dataset. This may relate to the amino acid composition of the *A. thaliana* import complexes: TOM components from *A. thaliana* contain fewer acidic residues compared to the more acidic TOM components of yeast [34], and therefore *A. thaliana* mature sequences would not be inhibited by negative charge repulsion to the same extent as yeast mature sequences. The most prominent correlation within the *A. thaliana* group of sequences suggests that MTS length is positively associated with mature sequence length, a correlation that was not present in any of the other species (apart from a minor association in yeast). There was a smaller negative association between basic residues and pH 7.5 in the MTS and the charge at pH 7.5 in the mature sequence, which suggests that associations involving sequence charge may be present but are likely different than that observed in the other species.

Pairwise correlation analysis of MTS and its mature sequence

To investigate these observations further and to determine the significance of each association, a pairwise Spearman's rho correlation matrix was generated for each species to directly compare properties of both MTS and mature sequences (Figure 2). Similar patterns of association to the multivariate analysis were observed; MTS length was significantly correlated with charge characteristics of the mature sequence (negative correlations with mature sequence pH 7.5 and pI), and particularly acidic residue content in all species but *A. thaliana*. It has been speculated previously that charge characteristics influence mitochondrial import rate, whereby a mitochondrial isoform (pI 9–9.5) of aspartate aminotransferase was imported up to four times faster than its cytosolic isoform (pI 6.7) when fused to an identical MTS sequence [35,36]. Only a single correlation was shared among all five species; MTS charge at pH 7.5 was negatively correlated with the mature sequence charge at pH 7.5 (r_s range: -0.484>−0.372), which suggests that mature sequences that are more negatively charged tend to be associated with MTSs that are more positively charged. Basic residues have long been acknowledged to play an important role within the MTS, as they are thought to enable electrophoretic mobility of the precursor through the inner-membrane space in response to the membrane potential [37], which in turn increases the likelihood of MTS interaction with the import motor by initiating translocation through the Tim23 complex [38]. It is therefore surprising that there was little evidence from this analysis that positively charged basic content in the MTS was associated with the proportion of negatively charged acidic residues in the mature sequence ($r_s = -0.108$, $p = 0.062$; Figure 3A), leading us to conclude that the significant negative correlation between both MTS and mature sequence charge at pH 7.5 ($r_s = -0.437$, $p < 0.001$; Figure 3B) is likely indicative of the balance between positive- and negatively charged residues in each sequence rather than being significantly influenced by any one individual amino acid property (basic or acidic) alone.

The pairwise analysis revealed some correlations that were unique to plants, and so emphasised further the difference between *A. thaliana* and *H. sapiens*, *M. musculus* and *S. cerevisiae* seen in the multivariate analysis. Both plant species show positive associations between MTS length and mature sequence non-polar cyclic residues (P), and in addition, MTS pH 7.5 and mature sequence non-polar aliphatic (GAVIL, negative correlation) and cyclic residues (P), all of which are not correlated in the non-plant species. A number of differences have been previously described that differentiate plant and non-plant species, and considerable differences between *Arabidopsis* and yeast MTS length and

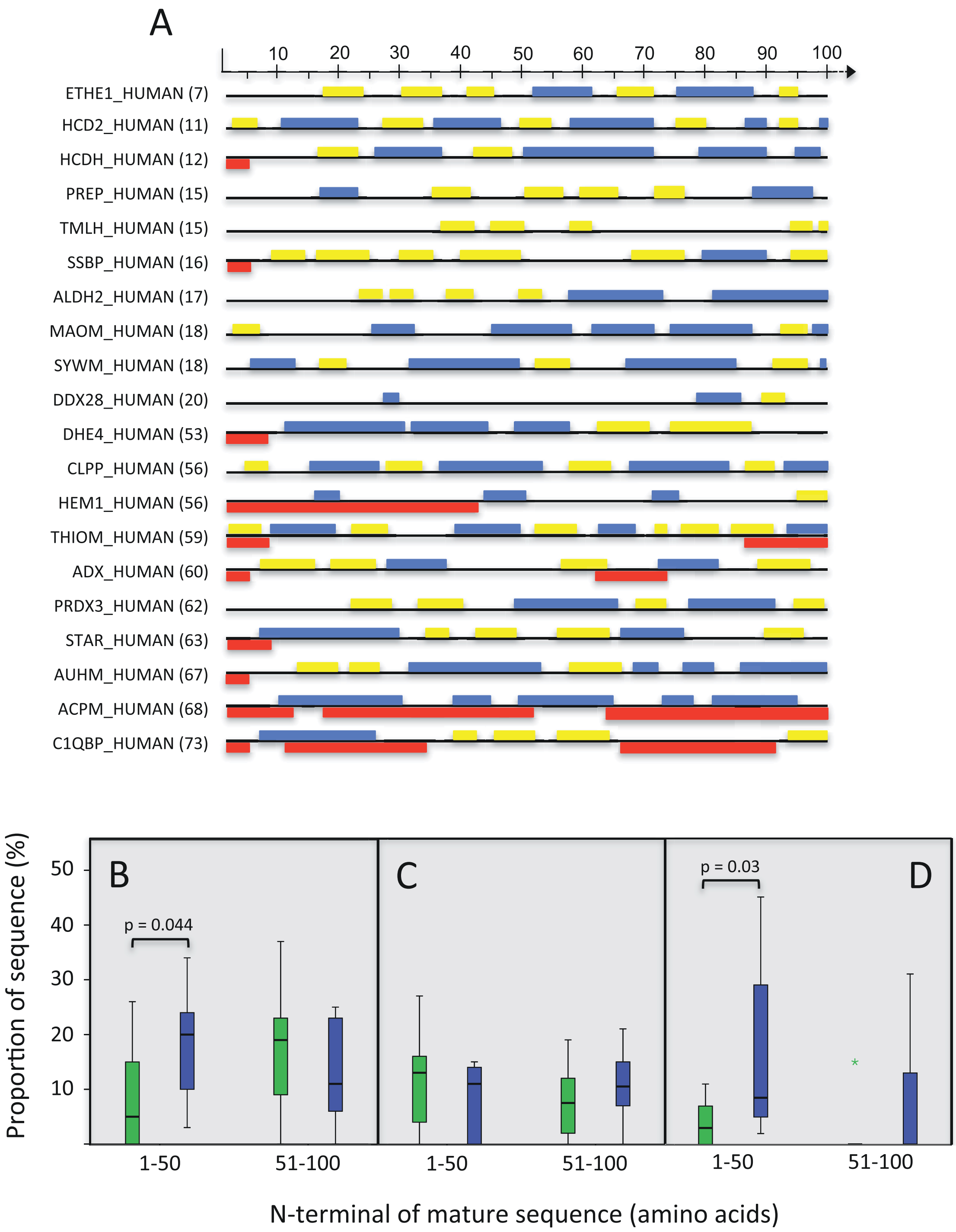


Figure 4. N-terminal secondary structure analysis of *H. sapiens* mature sequences associated with the ten shortest and ten longest MTSs. (A) Individual sequence IDs are indicated with MTS length in parentheses. Scale bar represents amino acid residue position after the MTS cleavage site. Yellow boxes – alpha helix, Blue boxes – beta sheet, Red boxes – predicted disordered sequence. Figure was adapted from secondary structure predictions of indicated mature sequences generated at www.predictprotein.org. Quantitative analysis of the proportion of predicted alpha helix (B), beta strand (C), and disordered (D) structures in the N-terminal mature sequences of short (green box plots) and long (blue box plots) MTSs for the first and second 50 amino acids of the mature sequence. The distribution of each structural group was analysed using a Kruskal-Wallis test for independent samples.
doi:10.1371/journal.pone.0067938.g004

composition [39] are consistent with the results presented here. Moreover, phylogenetic analysis of mitochondrial import components suggests that although they are functionally similar, plant and animal Tom20′ are evolutionarily distinct, having undergone convergent evolution following the divergence of the animal and plant lineages [40,41]. Some differences between plant and non-plant import mechanisms are likely due to the need to maintain specificity during the co-evolution of chloroplast and mitochon-

Table 1. Correlation analysis between MTS property and N-terminal mature sequence (MTPT) acidic residues relative to the cleavage site.

	MTS length vs MTPT acidic residues		**MTS charge vs MTPT acidic residues**	
Species	**1–40 aa**	**41–80 aa**	**1–40 aa**	**41–80 aa**
H. sapiens	0.360 (0.001)*	0.152 (0.166)	0.333 (0.002)	0.111 (0.311)
M. musculus	0.274 (0.012)	0.109 (0.323)	0.367 (0.001)	−0.004 (0.969)
S. cerevisiae	0.349 (0.014)	0.282 (0.05)	0.181 (0.213)	0.185 (0.203)
A. thaliana	0.064 (0.716)	0.423 (0.011)	0.102 (0.558)	0.494 (0.003)
O. sativa	0.372 (0.025)	0.208 (0.223)	0.029 (0.866)	0.443 (0.007)

*Data represented as Spearman's rho (non parametric) correlation coefficient with two-tailed p-values in parentheses.
doi:10.1371/journal.pone.0067938.t001

drial targeting in plant species [21]. However, the observation that *O. sativia* shares a number of correlated characteristics with both *A. thaliana* and non-plant species that were not otherwise shared (between *A. thaliana* and non-plant species; Figure 2) may be evidence that additional convergent evolutionary processes may have occurred in the development of the trafficking pathways between species.

Influence of the mature sequence N-terminal composition on MTS diversity

There have been a number of carefully controlled studies that suggest that the N-terminal of the mature sequence can influence the rate of precursor protein import [42–45]. The data presented in this investigation demonstrate that significant associations exist between MTS and the corresponding mature sequence. However, as the data presented here are derived from average composition values across the whole protein, the significance of the correlations described here is likely to be underestimated if local sequence features rather than average global sequence composition in the mature sequence plays a role in influencing MTS diversity. Therefore, we investigated whether two of the correlated properties described above were maintained between the immediate N-terminal 40 amino acids of the mature sequence and the second 40 amino acids (41–80 amino acids after the MTS)(Table 1). All species except for *A. thaliana* showed a significant correlation between MTS length and acidic residue content in the first 40 amino acids of the mature sequence that is then lost in the second 40 amino acids. *A. thaliana* shows an opposite effect; there is no association between MTS length and acidic residues in the first 40 amino acids, however, a significant correlation is evident in the second 40 amino acids. MTS charge was only associated with the acidic residues in the first 40 amino acids of *H. sapiens* and *M. musculus*; there was no association between either region in *S. cerevisiae*, and only the second 40 amino acids showed a correlation in both plant species. This analysis was expanded to determine if the predicted secondary structure in the N-terminal of the mature sequence was different between mature sequences with short or long MTSs (Figure 4; Figure S2). Experimental data does suggest that mature protein unfolding rate is inversely correlated with the stability of the N-terminal of the mature sequence [46] and that protein unfolding rate can be influenced by the secondary structure of N-terminal sequence immediately following the MTS [44]. Two observations were made: short MTSs were more likely to be associated with mature sequences that contain shorter alpha helix and beta strands, whereas mature sequences associated with longer MTS sequences tended to have a greater proportion of alpha helical regions (Figure 4B) and/or an increased occurrence of disordered or unstructured regions (Figure 4D), particularly in the first 50 amino acids of the mature sequence. These characteristics were most prevalent in the *H. sapiens* and *M. musculus* datasets, less prevalent in the *S. cerevisiae* and *O. sativa* datasets, and not obviously present in *A. thaliana*. Although it has been demonstrated in an *in vitro* model that mitochondrial import is more efficient when an alpha helix, rather than a beta strand, follows a targeting sequence in the mature sequence [44], this observation does not seem to be correlated with a change in MTS length and is not consistent across taxa. The prediction that disordered regions, which are characterised by a lack of bulky hydrophobic residues and rich in polar and charged residues, are associated with longer MTSs is somewhat counterintuitive, since such disordered regions are thought to be often unstructured [47] and therefore should be easier to import as they are likely loosely- or un-folded. These regions may however reflect more complex (and therefore difficult to predict bioinformatically) structural configurations that may be difficult to import.

Although the statistical mining of primary sequence data can only at best provide indirect evidence of a structure-function relationship, we believe that these observations suggest that negative charge interactions and structural features, particularly in the immediate N-terminal of the mature protein, reflect evolutionary constraints that likely influence the susceptibility or resistance of a protein to be imported. These limitations are consistent with a number of experimental observations, including that the mature portion of mitochondrial protein must be essentially unfolded to allow import [48], that some proteins require active unfolding to ensure they are import competent [49], and that longer MTSs unfold mature proteins faster than shorter variants [50]. It is therefore likely that a proportion of MTS diversity exists to mitigate these constraints in the N-terminal of the mature sequence and overcome potential resistance to unfolding prior to import.

Conclusions

In this investigation, we present evidence that suggests MTS diversity is influenced in part by physiochemical and structural characteristics of the mature protein that it imports, and that some of these correlated characteristics are evolutionarily maintained across a number of taxa. MTS diversity therefore likely reflects adaptation that balances the engagement of import-promoting mechanisms (such as the electrophoretic effect driven by the inner mitochondrial membrane potential and the matrix import motor, mtHsp70) and the likely import rate-limiting properties of the mature sequence (such as physicochemical interactions between the precursor protein and the import complex, and unfolding potential of the mature sequence prior to import). Charge characteristics of the MTS and mature sequence seem to be most influenced by these adaptations; however, they do not seem to influence plant MTS diversity to the extent seen in non-plant species, which may be due to comparatively longer MTSs in plants that allow greater interaction with mtHsp70 and in turn, possibly less dependency on electrophoretic effect across the inner membrane. Despite the significance of the data presented, a large proportion of the variation among these sequences remains unexplained. This is not necessarily surprising, given that predominantly primary amino acid sequence properties were investigated here. Further understanding of the folding variation among mature sequences, particularly in the N-terminus, will likely account for additional variation among MTS sequences.

Supporting Information

Figure S1 Pairwise factor analysis of multivariate PCA. Blue – *H. sapiens*, Green – *M. musculus*, Biege – *S. cerevisiae*, Purple – *A. thaliana*, Yellow – *O. sativa*.
(EPS)

Figure S2 N-terminal secondary structure analysis of all mature sequences (excluding *H. sapiens*) associated the five shortest and five longest MTSs. See Figure 4 for description.
(PDF)

Table S1 Quantitative comparison of sequence characteristics and amino acid groupings of MTS and mature protein sequences for each species used in this investigation.
(PDF)

Dataset S1 Mitochondrial matrix protein dataset. A complete list of mitochondrial matrix proteins, gene IDs, accession numbers and both MTS and mature sequences used in this study.
(TXT)

Dataset S2 Data used to generate Spearman's rho correlation matrix in Figure 2.
(TXT)

File S1 Amino acid frequency analyser (*AAResidueFreqCalculator*). Perl script that enables the calculation of both individual and grouped amino acid frequencies from a file containing multiple FASTA sequences. Open the text file for instructions on how to use the program. Convert attached text document file into Perl recognised file by changing the file extension from.txt to.pl (note: requires Perl installed on your system).
(TXT)

File S2 Reduced amino acid sequence converter (*ReducedAASequenceConverter*). Perl script that enables the conversion of amino acidic primary sequence into reduced amino acid sequence. Program will convert multiple FASTA sequences for a text file into reduced alphabet FASTA sequences in a new text file. Open the text file for instructions on how to use the program. Convert attached text document file into Perl recognised file by changing the file extension from.txt to.pl (note: you must have Perl installed on your system).
(TXT)

Author Contributions

Conceived and designed the experiments: SRD WNG. Performed the experiments: SRD. Analyzed the data: SRD. Contributed reagents/materials/analysis tools: NRPK. Wrote the paper: SRD WNG. Wrote the Perl scripts used to extract and analyse the initial sequence datasets: NPRK. Helped refine the project and study design, and in drafting the manuscript: CKC.

References

1. González-Halphen D, Funes S, Pérez-Martínez X, Reyes-Prieto A, Claros MG, et al. (2004) Genetic correction of mitochondrial diseases: using the natural migration of mitochondrial genes to the nucleus in chlorophyte algae as a model system. Ann N Y Acad Sci 1019: 232–239.
2. Brennicke A, Grohmann L, Hiesel R, Knoop V, Schuster W (1993) The mitochondrial genome on its way to the nucleus: different stages of gene transfer in higher plants. FEBS Lett 325: 140–145.
3. Chacinska A, Koehler CM, Milenkovic D, Lithgow T, Pfanner N (2009) Importing mitochondrial proteins: machineries and mechanisms. Cell 138: 628–644.
4. Mokranjac D, Neupert W (2009) Thirty years of protein translocation into mitochondria: unexpectedly complex and still puzzling. Biochim Biophys Acta 1793: 33–41.
5. Schleiff E, Becker T (2011) Common ground for protein translocation: access control for mitochondria and chloroplasts. Nat Rev Mol Cell Biol 12: 48–59.
6. von Heijne G (1986) Mitochondrial targeting sequences may form amphiphilic helices. EMBO J 5: 1335–1342.
7. Roise D, Horvath SJ, Tomich JM, Richards JH, Schatz G (1986) A chemically synthesized pre-sequence of an imported mitochondrial protein can form an amphiphilic helix and perturb natural and artificial phospholipid bilayers. EMBO J 5: 1327–1334.
8. Roise D, Theiler F, Horvath SJ, Tomich JM, Richards JH, et al. (1988) Amphiphilicity is essential for mitochondrial presequence function. EMBO J 7: 649–653.
9. von Heijne G, Steppuhn J, Herrmann RG (1989) Domain structure of mitochondrial and chloroplast targeting peptides. Eur J Biochem 180: 535–545.
10. Muto T, Obita T, Abe Y, Shodai T, Endo T, et al. (2001) NMR identification of the Tom20 binding segment in mitochondrial presequences. J Mol Biol 306: 137–143.
11. Saitoh T, Igura M, Obita T, Ose T, Kojima R, et al. (2007) Tom20 recognizes mitochondrial presequences through dynamic equilibrium among multiple bound states. EMBO J 26: 4777–4787.
12. Schneider G, Sjoling S, Wallin E, Wrede P, Glaser E, et al. (1998) Feature-extraction from endopeptidase cleavage sites in mitochondrial targeting peptides. Proteins 30: 49–60.
13. Höglund A, Dönnes P, Blum T, Adolph HW, Kohlbacher O (2006) MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. Bioinformatics 22: 1158–1165.
14. Small I, Peeters N, Legeai F, Lurin C (2004) Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics 4: 1581–1590.
15. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2: 953–971.
16. Emanuelsson O, von Heijne G, Schneider G (2001) Analysis and prediction of mitochondrial targeting peptides. Methods Cell Biol 65: 175–187.
17. Dudek J, Rehling P, van der Laan M (2013) Mitochondrial protein import: common principles and physiological networks. Biochim Biophys Acta 1833: 274–285.
18. Schmidt O, Pfanner N, Meisinger C (2010) Mitochondrial protein import: from proteomics to functional mechanisms. Nat Rev Mol Cell Biol 11: 655–667.
19. Taylor AB, Smith BS, Kitada S, Kojima K, Miyaura H, et al. (2001) Crystal structures of mitochondrial processing peptidase reveal the mode for specific cleavage of import signal sequences. Structure 9: 615–625.
20. Vögtle FN, Wortelkamp S, Zahedi RP, Becker D, Leidhold C, et al. (2009) Global analysis of the mitochondrial N-proteome identifies a processing peptidase critical for protein stability. Cell 139: 428–439.
21. Huang S, Taylor NL, Whelan J, Millar AH (2009) Refining the definition of plant mitochondrial presequences through analysis of sorting signals, N-terminal modifications, and cleavage motifs. Plant Physiol 150: 1272–1285.
22. Rost B, Yachdav G, Liu J (2004) The PredictProtein server. Nucleic Acids Res 32: W321–326.
23. Peterson EL, Kondev J, Theriot JA, Phillips R (2009) Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment. Bioinformatics 25: 1356–1362.

24. Lister R, Carrie C, Duncan O, Ho LH, Howell KA, et al. (2007) Functional definition of outer membrane proteins involved in preprotein import into mitochondria. Plant Cell 19: 3739–3759.
25. Werhahn W, Niemeyer A, Jänsch L, Kruft V, Schmitz UK, et al. (2001) Purification and characterization of the preprotein translocase of the outer mitochondrial membrane from Arabidopsis. Identification of multiple forms of TOM20. Plant Physiol 125: 943–954.
26. Schleiff E, Heard TS, Weiner H (1999) Positively charged residues, the helical conformation and the structural flexibility of the leader sequence of pALDH are important for recognition by hTom20. FEBS Lett 461: 9–12.
27. Yano M, Hoogenraad N, Terada K, Mori M (2000) Identification and functional analysis of human Tom22 for protein import into mitochondria. Mol Cell Biol 20: 7205–7213.
28. Hammen PK, Weiner H (2000) Structure of the cytosolic domain of TOM5, a mitochondrial import protein. FEBS Lett 468: 101–104.
29. Bauer MF, Sirrenberg C, Neupert W, Brunner M (1996) Role of Tim23 as voltage sensor and presequence receptor in protein import into mitochondria. Cell 87: 33–41.
30. Geissler A, Chacinska A, Truscott KN, Wiedemann N, Brandner K, et al. (2002) The mitochondrial presequence translocase: an essential role of Tim50 in directing preproteins to the import channel. Cell 111: 507–518.
31. Meier S, Neupert W, Herrmann JM (2005) Conserved N-terminal negative charges in the Tim17 subunit of the TIM23 translocase play a critical role in the import of preproteins into mitochondria. J Biol Chem 280: 7777–7785.
32. Hill K, Model K, Ryan MT, Dietmeier K, Martin F, et al. (1998) Tom40 forms the hydrophilic channel of the mitochondrial import pore for preproteins [see comment]. Nature 395: 516–521.
33. Geissler A, Rassow J, Pfanner N, Voos W (2001) Mitochondrial import driving forces: enhanced trapping by matrix Hsp70 stimulates translocation and reduces the membrane potential dependence of loosely folded preproteins. Mol Cell Biol 21: 7097–7104.
34. Werhahn W, Jänsch L, Braun HP (2003) Identification of novel subunits of the TOM complex from Arabidopsis thaliana. Plant Physiology and Biochemistry 41: 407–416.
35. Hartmann CM, Lindenmann JM, Christen P, Jaussi R (1991) The precursor of mitochondrial aspartate aminotransferase is imported into mitochondria faster than the homologous cytosolic isoenzyme with the same presequence attached. Biochem Biophys Res Commun 174: 1232–1238.
36. Hartmann C, Christen P, Jaussi R (1991) Mitochondrial protein charge. Nature 352: 762–763.
37. Martin J, Mahlke K, Pfanner N (1991) Role of an energized inner membrane in mitochondrial protein import. Delta psi drives the movement of presequences. J Biol Chem 266: 18051–18057.
38. Krayl M, Lim JH, Martin F, Guiard B, Voos W (2007) A cooperative action of the ATP-dependent import motor complex and the inner membrane potential drives mitochondrial preprotein import. Mol Cell Biol 27: 411–425.
39. Braun HP, Schmitz UK (1999) The protein-import apparatus of plant mitochondria. Planta 209: 267–274.
40. Perry AJ, Hulett JM, Likić VA, Lithgow T, Gooley PR (2006) Convergent evolution of receptors for protein import into mitochondria. Curr Biol 16: 221–229.
41. Lister R, Whelan J (2006) Mitochondrial protein import: convergent solutions for receptor structure. Curr Biol 16: R197–199.
42. Verner K, Lemire BD (1989) Tight folding of a passenger protein can interfere with the targeting function of a mitochondrial presequence. EMBO J 8: 1491–1495.
43. Matouschek A (2003) Protein unfolding - an important process in vivo? Curr Opin Struct Biol 13: 98–109.
44. Wilcox AJ, Choy J, Bustamante C, Matouschek A (2005) Effect of protein structure on mitochondrial import. Proc Natl Acad Sci U S A 102: 15435–15440.
45. Waltner M, Hammen PK, Weiner H (1996) Influence of the mature portion of a precursor protein on the mitochondrial signal sequence. J Biol Chem 271: 21226–21230.
46. Huang S, Ratliff KS, Schwartz MP, Spenner JM, Matouschek A (1999) Mitochondria unfold precursor proteins by unraveling them from their N-termini. Nat Struct Biol 6: 1132–1138.
47. Schlessinger A, Punta M, Yachdav G, Kajan L, Rost B (2009) Improved disorder prediction by combination of orthogonal approaches. PLoS One 4: e4433.
48. Schwartz MP, Huang S, Matouschek A (1999) The structure of precursor proteins during import into mitochondria. J Biol Chem 274: 12759–12764.
49. Lim JH, Martin F, Guiard B, Pfanner N, Voos W (2001) The mitochondrial Hsp70-dependent import system actively unfolds preproteins and shortens the lag phase of translocation. EMBO J 20: 941–950.
50. Matouschek A, Azem A, Ratliff K, Glick BS, Schmid K, et al. (1997) Active unfolding of precursor proteins during mitochondrial protein import. EMBO J 16: 6727–6736.