doi:10.1016/j.jmb.2004.12.057 *J. Mol. Biol.* (2005) **347**, 81–93

JMB

Available online at www.sciencedirect.com
SCIENCE DIRECT®

ELSEVIER

Patterns that Define the Four Domains Conserved in Known and Novel Isoforms of the Protein Import Receptor Tom20

Vladimir A. Likić[1], Andrew Perry[1], Joanne Hulett[1], Merran Derby[1] Ana Traven[2], Ross F. Waller[3], Patrick J. Keeling[3], Carla M. Koehler[4] Sean P. Curran[4], Paul R. Gooley[1]* and Trevor Lithgow[1]*

[1]*Russell Grimwade School of Biochemistry and Molecular Biology, University of Melbourne, Melbourne 3010 Australia*

[2]*St. Vincent's Institute of Medical Research, Fitzroy 3065 Australia*

[3]*Botany Department University of British Columbia Vancouver BC, Canada V6T 1Z4*

[4]*Department of Chemistry and Biochemistry, UCLA, Los Angeles, CA 90095-1569 USA*

**Corresponding authors*

Tom20 is the master receptor for protein import into mitochondria. Analysis of motifs present in Tom20 sequences from fungi and animals found several highly conserved regions, including features of the transmembrane segment, the ligand-binding domain and functionally important flexible segments at the N terminus and the C terminus of the protein. Hidden Markov model searches of genome sequence data revealed novel isoforms of Tom20 in vertebrate and invertebrate animals. A three-dimensional comparative model of the novel type I Tom20, based on the structurally characterized type II isoform, shows important differences in the amino acid residues lining the ligand-binding groove, where the type I protein from animals is more similar to the fungal form of Tom20. Given that the two receptor types from mouse interact with the same set of precursor protein substrates, comparative analysis of the substrate-binding site provides unique insight into the mechanism of substrate recognition. No Tom20-related protein was found in genome sequence data from plants or protozoans, suggesting the receptor Tom20 evolved after the split of animals and fungi from the main lineage of eukaryotes.

Keywords: mitochondria; protein import; protein targeting sequences; import receptors; hidden Markov models

Introduction

Genomic and proteomic studies have revealed that 5–10% of the proteins made in a eukaryotic cell are targeted to mitochondria.[1–7] Specific mitochondrial targeting sequences distinguish these proteins from those that will stay in the cytosol. In some mitochondrial proteins, especially those destined for the mitochondrial matrix, the targeting sequences are cleavable basic, amphipathic helices.[8,9] However, for many mitochondrial membrane proteins, amphipathic segments serve as targeting signals, and these same amphipathic segments become the transmembrane segment or segments once the protein is embedded in the mitochondrial outer or inner membrane.[10,11]

All these mitochondrial targeting sequences have in common properties of positive charge and amphipathicity but there is no consensus in primary structure. Specialized protein import receptors recognize structural aspects common to all mitochondrial targeting sequences. The import receptors are mitochondrial outer membrane proteins that can bind each of the diverse protein substrates and transfer them efficiently to the protein translocation channel, formed from the essential protein Tom40, which is the central component of the translocase in the outer mitochondrial membrane (the TOM complex).[12–15] While it has been shown that the import receptors dock transiently to the core TOM complex to deliver their substrate protein cargo,[16–20] the structurally important features for interactions between the receptors and Tom40 or its attendant subunits Tom6 and Tom7 are not known.

At least three receptors, Tom20, Tom22 and Tom70, mediate protein import into mitochondria.[14] The master receptor is Tom20, which

Abbreviations used: TPR, tetratricopeptide repeat; HMM, hidden Markov model; RNAi, RNA interference; ML, maximum-likelihood; AU, approximately unbiased.

E-mail addresses of the corresponding authors: prg@unimelb.edu.au; t.lithgow@unimelb.edu.au

0022-2836/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.

binds mitochondrial targeting sequences directly, as shown with the Tom20 from the fungi *Neurospora crassa* and *Saccharomyces cerevisiae*, and from rats and humans.[14,21–24] Structural analysis of the central core domain of the rat protein has shown that it includes a tetratricopeptide repeat (TPR) fold and a distal helical segment, which together form a small, globular domain with a shallow groove. The surface of the groove, which represents the substrate-binding surface in which targeting sequences sit, is formed from hydrophobic side-chains to accommodate the hydrophobic surface of the targeting sequence ligands.[25] In addition to contributing to the ligand-binding groove, the TPR segment of Tom20 is needed for a productive interaction with the receptor Tom70, which facilitates recognition and binding of large hydrophobic precursor proteins.[26] Other regions of Tom20 have been shown to be functionally important, but have proved intransigent to direct structural analysis.[23,25,27,28]

In order to further understand its structure and function, we designed hidden Markov models (HMMs) to describe the Tom20 receptor, and uncovered conserved structural features in Tom20 proteins from animals and fungi. Those conserved motifs allowed us to search genome sequence data and identify new isoforms of the Tom20 receptor in animals. No Tom20-like sequence was found in plants or protozoans. Structural analysis of the novel form of Tom20 from mouse by comparative modeling, using the known structure of the classical Tom20 as a template, revealed that the two paralogs have distinguishing features in the ligand-binding groove. Furthermore, analyses in mice and worms showed that the variant Tom20 isoforms are functional, and likely provide for optimal expression of import receptors in specific cell types in metazoans.

Results

Hidden Markov models define conserved sequence characteristics of Tom20 receptors from animals and fungi

HMMs can be used to describe conserved features of a family of proteins with a view to defining domain structure and for searching for proteins from distantly related species.[29] To provide the sequences from which to build HMMs for Tom20, a BLAST search of GenBank was initiated with sequences of the functionally defined Tom20 from *N. crassa*, *S. cerevisiae* and *Homo sapiens* (see Materials and Methods). The initial set of Tom20 sequences consisted of 12 originating from animal species and six originating from fungi. HMMs were then constructed to represent the animal and fungal full-length sequences. The models were detected using three different programs, MEME,[30] ITERALIGN[31] and PROBE,[32] and HMMs were used to scan the UniProt database. A total of 1614 sequences matched HMMs with *E*-value <0.01 (790 matched the animal model and 824 matched the fungal model). Sequences longer than 90 residues and shorter than 200 residues were retained, giving a final set of 88 sequences. These were examined visually; an alignment of the most diverse sequences is shown in Supplementary Figure 1.

Two motifs were found in the animal set of sequences (A1 and A2) and three motifs were found in fungal sequences (F1, F2, and F3). When the combined sequences were analyzed, a single common motif emerged (AF1), which partially overlapped with A1 and F1. In each case a motif consensus was derived from an ungapped alignment and compiled with the program CONS from the EMBOSS suite.[33] Figure 1 shows the motifs mapped against other structural features of the animal (Figure 1(a)) and fungal (Figure 1(b)) Tom20 sequences.

The motif AF1 describes an N-terminal region that includes the transmembrane segment of Tom20. The residues in the transmembrane segment that are invariant in this AF1 motif are curious: the bulky aromatic residues: Y^{13}, Y^{16}, F^{17}, and highly conserved small residues, glycine or alanine, at G^4 and G^{12}. We suggest the conserved placement of these residues facilitates protein–protein contacts within the plane of the membrane.

In the Tom20 sequences from animal species, motif A2 (amino acid residues 63–137 of the rat protein) overlaps the region of Tom20 from rat, for which the structure has been solved by NMR (residues 57–124 of the 145 residue protein). The structure of this protein fragment corresponds to a set of five helices that form a compact globular domain.[25] Since the A2 motif is conserved in animal Tom20 sequences, the fold determined by NMR is likely conserved for all animal Tom20s. The first two helices in the region described by the A2 motif, $\alpha1$ and $\alpha2$, form a TPR fold.[25] The TPR represents an ancient protein fold, and key residues within the helix-turn-helix motif of a TPR (asterisked, Figure 1) are conserved.[34]

While the core of the animal Tom20 sequences is described by the single A2 motif, the corresponding region of Tom20 from fungi is described by two consecutive motifs, F2 and F3 (Figure 1(b)). These two motifs are broken only by a short, variable sequence that aligns with the "turn" sequence between helices A and B ($\alpha1$ and $\alpha2$) of the TPR segment (Figure 1(b)). The fact that F2 and F3 together describe the equivalent region of Tom20 as the A2 motif for animals suggests that this "core domain" of Tom20 is highly conserved in a structural sense, and that the variability between the animal and fungal groups of sequences is likely a consequence of evolutionary distance.

Our analysis also revealed two segments of Tom20, each labeled D in Figure 1. Variability in sequence precludes any motif being calculated for either of these two segments. However, analysis of each Tom20 sequence with the predictor DisEMBL[35] showed that each region is likely to be

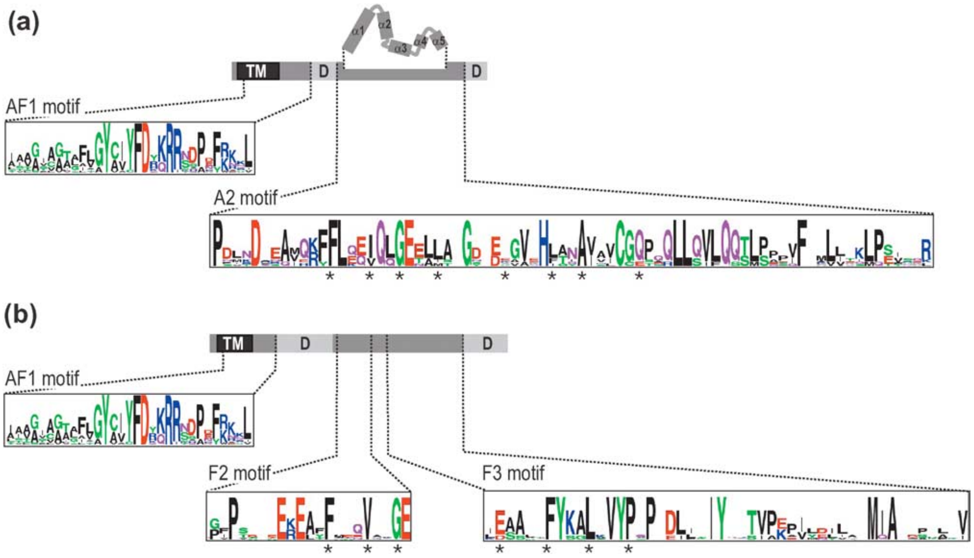

Figure 1. Motif analysis and domain structure of Tom20. (a) The motifs AF1 and A2, derived from eight sequences of Tom20 and modified after analysis of all Tom20 sequences currently available. A detailed sequence alignment in CLUSTALW is available as Supplementary Figure 1. Tom20 sequences from animals are shown mapped against a representation of Tom20 from rat. The transmembrane segment (TM) and regions predicting as disordered (D) are indicated. Also shown are the position for the five helices (α1–α5) that form the ligand-binding core of the receptor.[25] Asterisks (*) indicate the positions of residues corresponding to those in the TPR-like consensus. (b) The motifs AF1, F2 and F3, derived from six sequences of Tom20 and modified after ten Tom20 sequences from fungi are shown mapped against a representation of Tom20 from yeast.

disordered (data not shown). Therefore, despite the lack of primary structure definition, the need for flexibility in these positions seems to be a conserved feature for Tom20 receptors from all fungi and animals. In conclusion, four conserved regions were identified in Tom20 orthologs from all species of animals and fungi: the transmembrane segment, a disordered region rich in charged amino acid residues, the ligand-binding domain centered around a TPR segment and a disordered tail punctuated in acidic residues.

Two types of Tom20 exist in animal genomes

Phylogenetic analysis of the sequences retrieved by the HMM search showed that the Tom20 sequences from various fungi formed a well supported group and that this group is distinct from the animal sequences (Figure 2).

The search revealed an unexpected result from the animal genomes analyzed: variant forms of Tom20 seem to be encoded in animals. This characteristic is not shared with fungi, which have only one *TOM20* gene. The version of Tom20 from humans and rats that has been characterized structurally and functionally[23–25,36] is referred to as the type II Tom20. Figure 2 shows that type II sequences group with support with other vertebrate sequences. Humans and mice have a divergent Tom20 paralog encoded in their genome, which groups strongly on the phylogenetic tree with similar Tom20 paralogs represented as partial ESTs from rat and pig. We predict this paralogue, designated the type I Tom20, will be represented widely in mammals. The novel type I Tom20s form a group divergent from their type II counterparts, as evidenced by the long branch length and strong bootstrap support. However, the phylogenetic analysis is unable to resolve whether the type I gene developed early in animal evolution, or later during radiation of the vertebrates (Figure 2).

At least some fish (zebra fish, trout and fugu) have duplicate Tom20s; however, pair-wise identity of these duplicates is always very high (∼70–80%) within each species, indicating they represent a more recent gene duplication of the type II form. Amongst invertebrates, multiple forms of Tom20 are seen. In *Drosophila melanogaster*, there is one Tom20 encoded on the left arm of chromosome 3 (CG7654 76E1) and a distinct gene on the right arm of the same chromosome (CG14690 86C5). *Caenorhabditis elegans* has a gene encoding a conserved form of Tom20 (F23H12.2) and another, divergent form of Tom20 encoded by F32B4.2. Further, EST data suggest two splice variants of the transcript from this F32B4.2 gene (see Supplementary Figure 1). Without structural analysis it is premature to classify these invertebrate Tom20s as either type I or type II; however, topology testing rejected the grouping of the vertebrate type I with either of the *Drosophila* Tom20s, indicating that duplication of Tom20s in animals has happened independently multiple times.

The type I and type II forms of Tom20 in vertebrates

Pairwise alignment of mouse Tom20 type I and

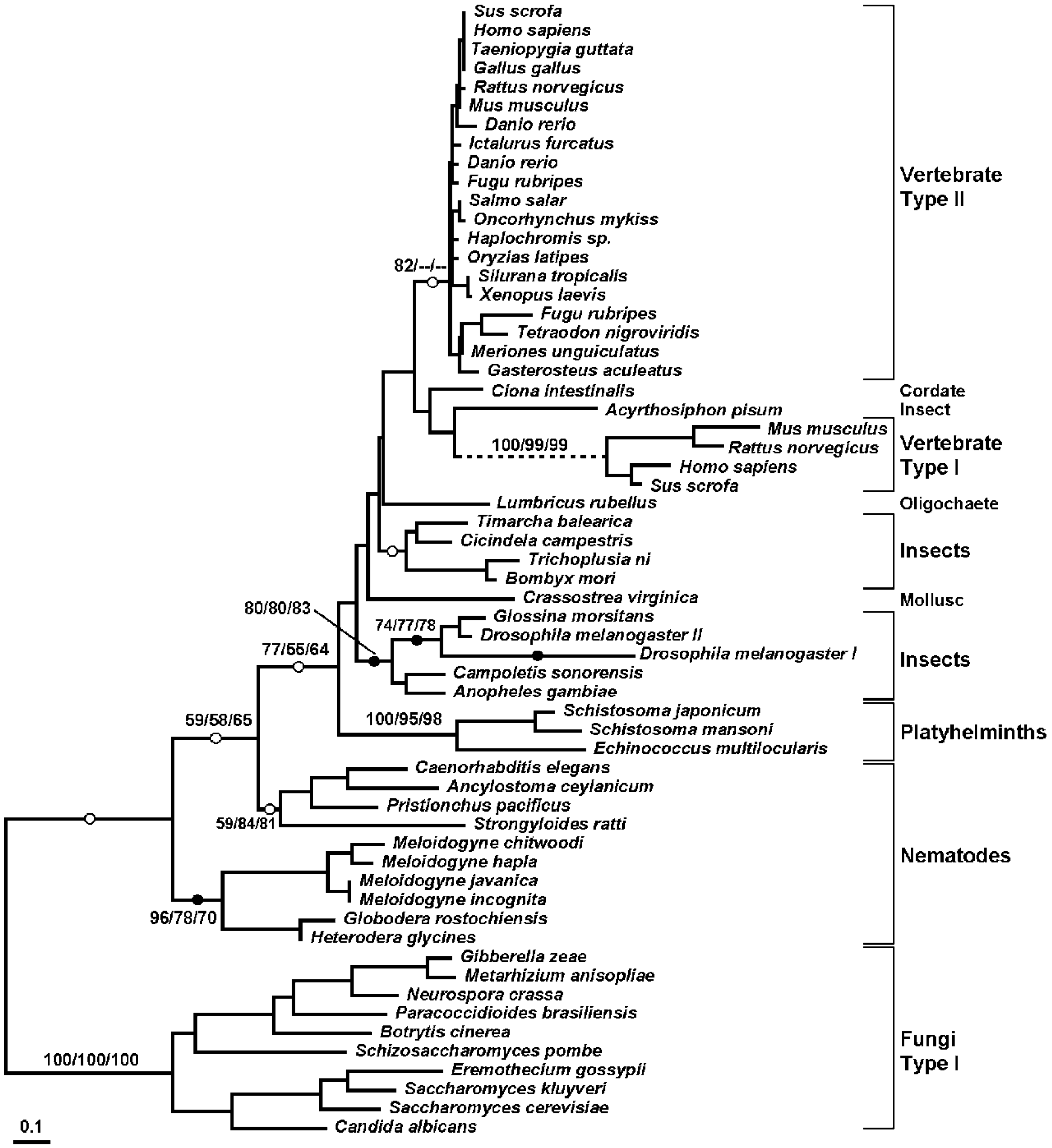

Figure 2. Tom20 protein maximum likelihood phylogeny (PhyML). Bootstrap values are shown for major nodes (left to right) using PhyML, and the distance analysis weighted neighbor-joining and Fitch–Margoliash (left to right) (dashes represent support lower than 50%). Alternate positions for the vertebrate type I Tom20 clade (broken line node) were tested by placing the clades at nodes placed at positions marked with a dot. Filled dots indicate topologies that were rejected as significantly worse than the best topology ($p < 0.05$), and open dots indicate where this topology was not rejected.

type II is shown in Figure 3(a). The level of sequence identity between the two is 34.9%, with many of the substitutions conservative. The predicted transmembrane segments are shaded and the predicted difference in length within the membrane bilayer is curious. Since this N-terminal region includes the targeting information for Tom20,[23,28] we constructed yellow fluorescent fusion proteins for each form of Tom20 and expressed these in cultured mammalian cells (Figure 3(b)). Both forms of Tom20 are mitochondrial proteins as judged by co-staining of Tom20-YFP protein and the mitochondria-specific dye MitoTracker.

A further clear distinction in the primary structure of the type I *versus* the type II protein is shown boxed in Figure 3(a). Motif analysis highlights replacements in the type I proteins at positions corresponding to five glutamine residues conserved

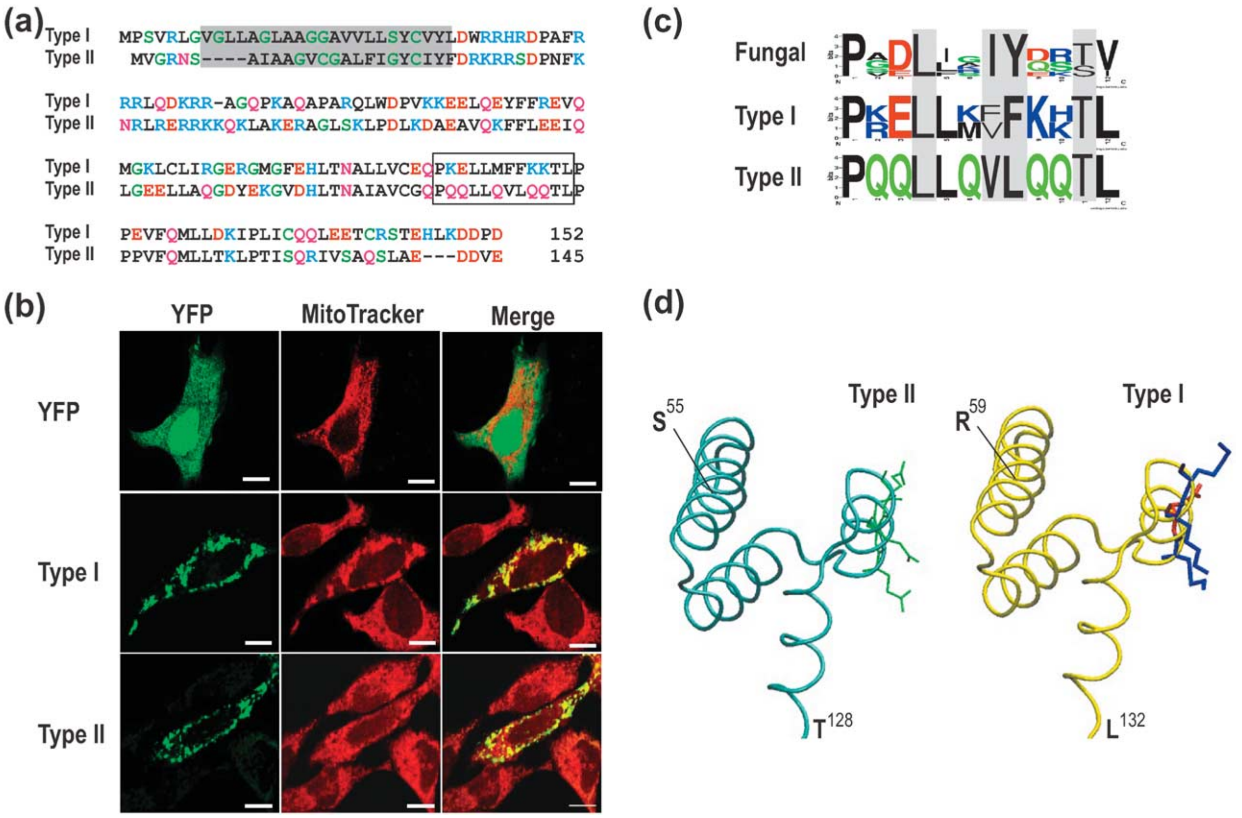
(a)
Type I MPSVRLGVGLLAGLAAGGAVVLLSYCVYLDWRRHRDPAFR
Type II MVGRNS----AIAAGVCGALFIGYCIYFDRKRRSDPNFK
Type I RRLQDKRR-AGQPKAQAPARQLWDPVKKEELQEYFFREVQ
Type II NRLRERRKKQKLAKERAGLSKLPDLKDAEAVQKFFLEEIQ
Type I MGKLCLIRGERGMGFEHLTNALLVCEQPKELLMFFKKTLP
Type II LGEELLAQGDYEKGVDHLTNAIAVCGQPQQLLQVLQQTLP
Type I PEVFQMLLDKIPLICQQLEETCRSTEHLKDDPD 152
Type II PPVFQMLLTKLPTISQRIVSAQSLAE---DDVE 145
(b)
YFP
MitoTracker
Merge
YFP
Type I
Type II
(c)
Fungal
Type I
Type II
PQQLLQVLQQTL
(d)
Type II
Type I
S55
R59
T128
L132

Figure 3. Structural definition of type I and type II Tom20 proteins. (a) Sequence alignment of the mouse sequences for Type I Tom20 (GenBank accession BY715706) and type II Tom20 (accession NM_024214). The transmembrane segment is shaded grey and the distinguishing "lip" helix is boxed. Residue colors are according to the MOTIF parameters (relatively hydrophobic, black; hydrophilic, green; glutamine, asparagine, magenta; acidic, red; relatively basic, blue). (b) HeLa cells transfected to express either type I Tom20-YFP or type II Tom20-YFP were co-stained with the mitochondria-specific dye Mitotracker red and examined by confocal microscopy. Filters selective for the fluorescence of yellow fluorescent protein and fluorescence of MitoTracker red were used. Merged green and red fluorescence images show no overlap between YFP and mitochondria, while Tom20-YFP fusions show distinct overlap with mitochondria. (c) MOTIF plots of the lip helix segment from the human and mouse type I sequences and various type II Tom20 sequences. These are shown compared to the equivalent segment of polypeptide from ten fungal Tom20 sequences. Relative height of the single letter representing each amino acid shows its relative abundance at each position. (d) Rat Tom20 type II structure (backbone trace in cyan) and mouse Tom20 type I model (backbone trace in yellow) showing the position of glutamine side-chains (green) in the Q-rich region. Shown highlighted in the type I Tom20 structure are the side-chain orientation for acidic (red) and basic (blue) residues that substitute for glutamine.

in all type II sequences. In this region, the primary structure of type I Tom20 is reminiscent of the Tom20 from fungal species (Figure 3(c)).

Three-dimensional modeling was employed to make a more detailed comparative analysis of the type I and type II Tom20s. The type II protein from rat has been characterized structurally,[25] and is 98.6% identical in sequence with the mouse type II protein. The use of the type II structure to model the type I protein is reasonable, with sequence identity and similarity (60.3%) after global alignment of template and model sequences justifying the use of comparative modeling.[37] A model structure calculated for type I Tom20 is shown in Figure 3(d). The program Modeller (6v2) was used to generate mouse Tom20 type I coordinates, using the rat Tom20 type II NMR structure[25] as the template. A total of 64 model structures of Tom20 type I were prepared and compared to the 20 structures of the rat Tom20 type II as determined by NMR.[25] After building the model of Tom20 type I we: (a) evaluated the model structures using ProsaII[38] and investigated the changes in the protein surface which resulted from sequence divergence between the two Tom20 isoforms (see Materials and Methods); and (b) investigated changes in the Tom20 presequence peptide binding site.

Analysis of the presequence peptide-binding groove

To determine which residues might make contact with the presequence peptide, we re-built the model of mouse type I isoform with the bound peptide, based on the complex of rat Tom20 with the presequence peptide pALDH(12–22).[25] We then evaluated the protein accessible surface shielded by the bound peptide, and filtered this surface to include only protein side-chain atoms (including C^{α}). The surface shielded by the presequence

Table 1. Presequence binding surfaces of Type I and Type II Tom20

Tom20 (Type I)	Tom20 (Type II)
Q^{71}	Q^{67}
F^{74}	F^{70}
F^{75}	L^{71}
V^{78}	I^{74}
Q^{79}	Q^{75}
K^{82}	E^{78}
L^{83}	E^{79}
I^{86}	A^{82}
E^{109}	Q^{105}
F^{113}	V^{109}
F^{114}	L^{110}
T^{117}	T^{113}
L^{118}	L^{114}

Side-chain solvent accessible surface shielded by the bound presequence peptide was calculated from the structure of the Type II Tom20 from rat with the pALDH(12–22) presequence peptide and calculated from the model of the mouse Tom20 Type I (see Methods). Residues shown are those with side-chains exhibiting shielded solvent accessible surface of >10 Å^2. Residues distinct between Type I and Type II Tom20 are shaded, heavy shading indicates non-conservative substitutions.

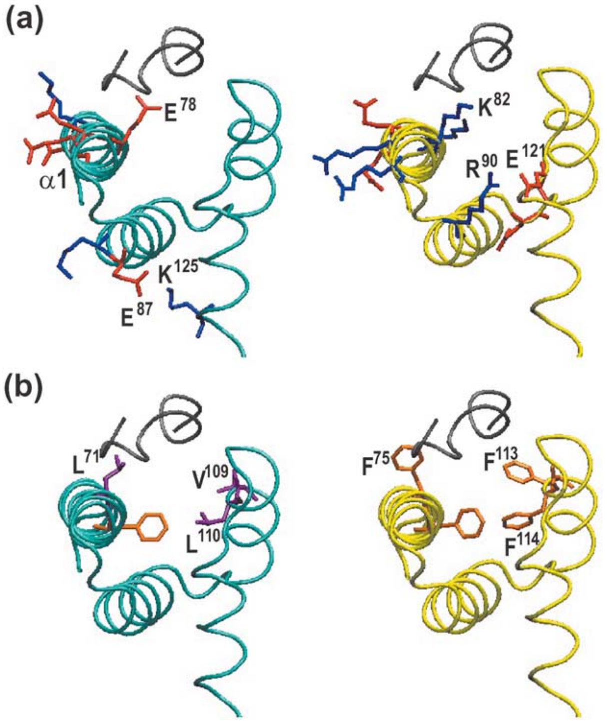

Figure 4. The presequence binding site on Tom20 type I. (a) Side-chains of charged residues that have been substituted between type II (rat) and type I (mouse) in helix α1 are shown (red for acidic, blue for basic residues). Most striking is the E^{78}K substitution, which protrudes into the peptide-binding groove. A potentially stabilizing electrostatic interaction between E^{87} and K^{125} (in the type I Tom20) would be conserved, but in type II Tom20 the interaction is *via* charge reversal a few residues further along the helix (R^{90} to E^{121}). (b) Rat Tom20 type II structure (backbone trace in cyan, left) and mouse Tom20 type I model (backbone trace in yellow, right) showing the position of the type II aliphatic side-chains in the binding groove (purple) substituted by phenylalanine (orange) in the type I isoform (L^{71}/F^{75}, V^{109}/F^{113}, L^{110}/F^{114}), as well as the single conserved phenylalanine (F^{70}/F^{74}). The presequence peptide backbone is shown in grey.

peptide was evaluated in both the modeled structure of mouse Tom20 type I and the structure of the rat Tom20 type II as determined by NMR.

The 13 side-chains of Tom20 listed in Table 1 form the presequence peptide-binding groove. Eight out of 13 residues that would form close contact with a presequence peptide are distinct in the type I and type II isoforms of Tom20, but five of these substitutions preserve hydrophobicity and would appear unlikely to perturb presequence binding. Strikingly, two substitutions either eliminate or introduce a charged residue ($E^{79}\rightarrow L$ and $Q^{105}\rightarrow E$), and one reverses the charge ($E^{78}\rightarrow K$). The mutations of E^{78} and E^{79} eliminate two negatively charged residues located near the N terminus of the presequence peptide in the structure of the complex: the lysine residue in the position corresponding to E^{78} seemed to be particularly unfavorable (Figure 4(a)). However, the peptide used in the NMR experiments represents residues 12–22 of the native substrate; in an intact mitochondrial precursor protein this region would be located distal from the amino terminus of the ligand and there would be no free amine group at position 12. Thus, the fact that a lysine residue can be accommodated in place of E^{78} provides independent support to the suggestion[39] that Tom20 receptors tend to bind to internal segments of presequences rather than close to the amino terminus.

Another striking change in the presequence binding groove is the replacement of three hydrophobic residues (two leucine and one valine) with phenylalanine (Figure 4(b)). While all of these residues are strongly hydrophobic,[40,41] and leucine and phenylalanine are relatively bulky, valine is both a shorter and less bulky residue. The additional aromatic residues convert what appeared as a hydrophobic groove in the type II form, to a raised hydrophobic platform in the type I Tom20. While this might have little effect on substrate binding, the aromatic residues thereby increase the bulk of the hydrophobic core and could play a role in increasing receptor stability.

Functional analyses of the variant Tom20 receptor in animals

There are several explanations for the existence of more than one gene encoding Tom20 in animals: (i) the two proteins might perform redundant roles; (ii) they could have distinct biological functions, perhaps being expressed in different tissues or

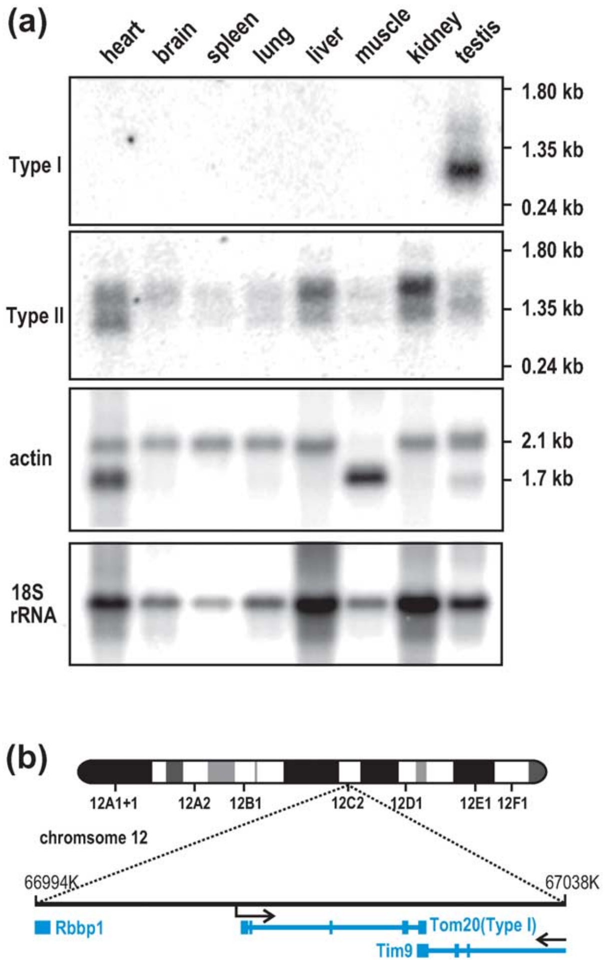

Figure 5. Mouse genes encoding type I and type II Tom20. (a) A filter carrying mRNA (20 μg) isolated from the indicated tissues was probed for the presence of messages encoding type I Tom20, type II Tom20 and actin and for the presence of 18 S rRNA. Migration of the 0.24 kb, 1.35 kb and 1.8 kb markers is shown for the top panels. The positions of the two, tissue-specific actin species at 1.7 kb and 2.1 kb are indicated. (b) The structure of the gene encoding Tom20 type I in mice.

during different stages of animal development; or (iii) the variant *TOM20* gene might not be expressed and therefore be without a meaningful cellular role. To address these possibilities, we used gene expression analysis for the two mouse *TOM20* genes using RNAs derived from a collection of tissues, analysis of available EST data for *D. melanogaster*, and RNA interference (RNAi) experiments in *C. elegans* to look at mutant phenotypes.

Northern analysis was used to look at expression of the two genes encoding the mouse type II and type I Tom20 proteins (Figure 5(a)), with RNAs isolated from heart, brain, spleen, lung, liver, skeletal muscle, kidney and testis. The type II probe recognized a doublet of RNA species of ∼1.4 kb and 1.5 kb, while the type I probe hybridized to a message of ∼1 kb. We found that the type II gene was expressed ubiquitously and to similar levels in all tissues tested. However, the type I gene showed a clear testis-specific expression, with no detectable signal in any other tissue analyzed.

In mouse, the type I Tom20 sequence corresponds to a multi-exon gene located on chromosome 12. The gene sits tail-to-tail with the gene called *TIMM10* (Gene ID 30056; Figure 5(b)). *TIMM10* (called *TIMM9* in humans) encodes Tim9, the small subunit of the TIM22 complex that mediates a later stage of protein import into mitochondria.[42] The 3′-untranslated regions of the longest cDNA corresponding to the type I Tom20 and the longest cDNA corresponding to Tim9 overlap by 48 bp. The overlap in the 3′-UTR regions of the genes encoding Tom20 type I and Tim9 in mouse is preserved in the syntenic gene structure on human chromosome 14 and, while this is suggestive of a potential co-expression of the two genes, our results for the mouse genes suggests that is not the case. The *TOM20* type I gene is transcribed in a tissue-specific manner (Figure 5(a)), while multiple ESTs and cDNAs exist to suggest *TIMM10* is expressed ubiquitously.

Although our phylogenetic analysis leaves open the question of whether the variant form of Tom20 in invertebrates is ancestrally related to the type I form from mammals, the Tom20s from flies and mammals have similar tissue-specific expression. Analysis of data contained in Flybase revealed the two forms of Tom20 from flies are encoded from distinct loci on chromosome III. Expressed sequence tags for CG7654 76E1 were found in samples prepared from a range of tissues, embryo libraries and cultured cells (e.g. RE28313, RE58084, RE42569, RE58619, SD03966, SD03031, SD20760, LD34461, LD06132, LD05932, LD34461, RE62148, RE62124, RE68359), while ESTs corresponding to the variant Tom20 (e.g. AI946818, BE97782, BE978106) could be found only in adult testis tissue, and the gene was listed on a genome-wide search for testis-specific transcripts.[43]

We set up RNAi experiments to independently knock down the function of the genes, F23H12.2 (GenBank accession 5M822) and F32B4.2 (GenBank accession 1M104), encoding the two Tom20 paralogs in *C. elegans*. Animals fed double-stranded (dsRNA) against F23H12.2 exhibited larval arrest and embryonic lethality phenotypes consistent with an essential function of one version of Tom20 (Figure 6(a)); in contrast, gene depletion of F32B4.2 encoding the novel form of Tom20 was not lethal, but instead resulted in slow growth (Slo) and small body (Sma) phenotypes (Figure 6(b)). These results show that the two forms of the Tom20 receptor in worms do not perform redundant roles: the essential role of the Tom20 encoded by F23H12.2 cannot be complemented by the protein product of F32B4.2. The slow development to a reproductive adult, and the Sma phenotype, would be best explained by cell-specific expression of F32B4.2.

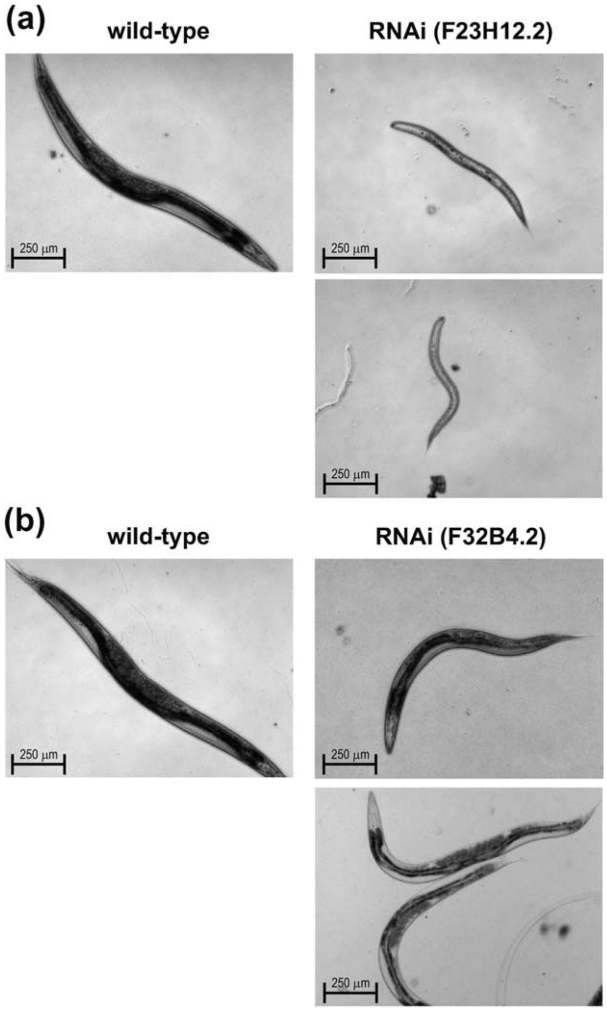

Figure 6. An essential and a non-essential gene encoding two types of Tom20 in *C. elegans*. (a) RNAi was used to knock-down the function of the gene F23H12.2, encoding Tom20 in *C. elegans.* The small worms shown represent the size of the very few worms that survived after feeding of RNA corresponding to F23H12.2. (b) RNAi knock-down of the gene F32B4.2, encoding the variant form of Tom20, limits both the rate and extent of growth. Similar numbers of worms hatched and lived in the absence or the presence of RNA corresponding to F32B4.2, indicating this gene is not essential. The scale bar represents 250 μm in all photographs.

Discussion

We collected and analyzed Tom20 sequences covering representative classes of animals and fungi for sequence motifs and used hidden Markov models to define consensus features of the import receptor and to comprehensively search the known data sets for Tom20 sequences. The search revealed a novel isoform of Tom20 in animals, and three-dimensional modeling allowed us to determine the extent of structural conservation in the Tom20 paralogs. In terms of the substrate-binding site, our analysis complements the experimental testing of a single species of protein by biophysical measurements, and revealed conserved structural features in the transmembrane tether and C-terminal tail that sit outside the core of the type II protein previously determined by NMR.[25]

The substrate-binding surfaces of Tom20

The ligand-binding surfaces differ between the type I and type II isoforms. In the type II Tom20, 13 residues were found to form the hydrophobic groove that cradles the presequence ligand.[25] Two non-conservative substitutions, $E^{78}\rightarrow K$ and $E^{79}\rightarrow L$, remove two negatively charged residues that were interacting with the terminal amine group of the synthetic peptide presequence complexed to the type II form of Tom20.[25] That these two glutamate residues are absent from the type I form is in keeping with the fact that Tom20 binds target sequences at regions situated further back from the N-terminal amino group.[39] Furthermore, since the two forms of Tom20 in mouse must interact with a similar set of protein substrates, the raised hydrophobic platform seen for the type I Tom20 leads us to suggest that in the hydrophobic groove noted by Abe *et al*.[25] it is the hydrophobic character and not the depth of the groove that is responsible for binding presequences. On this hydrophobic surface, presequence peptides can sit loosely, and thereby adopt helical character to display positively charged residues on the surface of the Tom20–substrate complex. Seen in this light, mitochondrial targeting information would be encoded in the structure of the peptide adopted in complex with Tom20, and the largely electrostatic characteristics of the peptide–receptor complex might then serve as a docking surface for other, negatively charged components of the import pathway.

A set of six glutamine residues on the lip of the ligand-binding groove were suggested to be a characteristic feature of Tom20[25], and most animals do have at least one Tom20 variant, referred to here as type II, with these glutamine residues conserved. However, analyses of the primary structure of the fungal Tom20 receptors and the primary and model tertiary structure of the novel mouse type I isoform of Tom20 suggest that charged residues can substitute for glutamine at these positions. This is true in the fungal proteins, in many of the invertebrate proteins and in the type I class of vertebrate Tom20. Three of these substitutions are likely to be near bound substrate peptides: K^{82} (E in type II Tom20), E^{109} and K^{116} (both glutamine in the Tom20 type II). Each of these residues is positioned in such a way that it may contribute to charge around the periphery of the binding groove, though they are unlikely to interact directly with the substrate. Instead, the basic and solvent-exposed residues of the bound substrate in combination with charged residues such as K^{82} and K^{116} on the receptor could provide an attractive surface for acidic partner proteins like Tom22 and Tom5.[19,44–46]

The C-terminal tail of Tom20 is rich in acidic residues, is strongly predicted to be natively disordered and is present in all Tom20 receptors that we analyzed. NMR experiments on the type II Tom20,[25] and our Modeller predictions for the type I Tom20 further suggest this disordered region is solvent exposed. Yeast mutants from which this acidic tail segment has been deleted suffer defects in protein import, due to either impaired ligand-binding or decreased interactions of Tom20 with partner subunits in the TOM complex.[27]

Tom20 is attached to the outer membrane by a flexible tether

Analysis of motifs in the various Tom20 sequences revealed conserved characteristics of the transmembrane segment and the presence of a flexible domain between the transmembrane part and the central, substrate-binding core domain. All Tom20 receptors have a single, amphipathic membrane-spanning region at the N terminus. This region has several highly-conserved glycine residues, several hydroxylated residues and absolutely conserved aromatic residues followed by an aspartate residue positioned at the cytosol–membrane interface. We suggest these conserved features of the transmembrane segment might contribute a surface used by Tom20 to interact with its integral membrane protein partners. Docking of Tom20 to the core TOM complex is required for transfer of substrates into the translocation channel.[19,20]

The evolution of the mitochondrial import machinery and novel isoforms of Tom20 in animals

Hidden Markov models are a sensitive means to retrieve Tom20 sequences from novel genome sequence data. We have sufficient sequence representation to propose that genes encoding Tom20 arose relatively late in the evolution of eukaryotes. Whereas the core TOM complex, comprising Tom40, Tom22 and Tom7, appears to have been operating in the earliest eukaryotes,[47] Tom20 was derived only after the split of animals and fungi from the ancestors of other eukaryotes (e.g. plants, and protistan groups). Purification and characterization of the TOM complex from higher plants showed it has protein import receptors, including a 20 kDa component that has, therefore, been called Tom20,[48,49] but it is unrelated in sequence to the Tom20 family of proteins defined here and might have evolved independently after an early split in eukaryote lineages.

The search for Tom20 homologs that we performed revealed that animals, both vertebrate and invertebrate, have more than one gene encoding Tom20. Our analysis of the variant forms of Tom20 in mice, *D. melanogaster* and *C. elegans* suggests tissue-specific expression as a driving force for duplication of the gene encoding Tom20 in metazoans. In worms, the novel Tom20 receptor is not essential. However, shut-down of the novel gene's expression gives rise to smaller animals that exhibit slow growth. That such a clear and distinct phenotype resulted from the decreased expression of this gene shows that the new Tom20 isoform is functional in worms and has a non-essential but specific role. That role could relate to a cell-specific expression pattern, consistent with the data obtained from flies and mice.

The type I gene in mouse is expressed only in testis tissue, while the type II gene shows ubiquitous expression in all of the other tissues tested. Data from *D. melanogaster*, including a very recent report confirms that the variant Tom20 is testis-specific, with *in situ* analysis revealing the transcript is restricted to primary spermatocytes and bundles of early spermatids.[50] Only recently have we started to understand how control over gene expression might be exerted during spermatogenesis.[51,52] It is known that regulation of gene expression, and the actual general transcription machinery, differ substantially between somatic and male germ cells.[53,54] One of the hallmarks of spermatogenesis is the massive mitochondrial biogenesis undertaken during the development of sperm cells.[53] The locus of the gene encoding Tom20 type II might be less active during spermatogenesis, requiring a second locus that can be activated during the germ cell developmental program. Why more complex organisms need more than one form of Tom20 protein invites further investigation, but our results in *C. elegans* and mice and the data from *Drosophila* suggest that variant forms of Tom20 are necessary in select animal cell types to allow for complex gene expression programs.

Materials and Methods

Tom20 sequences used in HMM search

The initial set of Tom20 sequences consisted of 12 animal sequences (from the insects *Bombyx mori* and *D. melanogaster*, the nematodes *C. elegans* and *Echinococcus multilocularis*, the flatworms *Schistosoma japonicum* and *Schistosoma mansoni*, the coelenterate *Ciona intestinalis*, the mollusk *Crassostrea virginica*, the fish *Danio rerio* and *Oryzias latipes*, the frog *Silurana tropicalis*, and from *H. sapiens*) and six fungal sequences (*Botrytis cinerea*, *Candida albicans*, *Gibberella zeae*, *N. crassa*, *S. cerevisiae* and *Schizosaccharomyces pombe*). The animal and fungal sequences were used to build HMMs to search for similar proteins. The building of HMM models and database search were performed with the HMMER package,[55] version 2.3.2. We searched the UniProt database version 1.9 (Swiss-Prot Release 43.3 and TrEMBL release 26.3 of 10 May, 2004). The results were processed with programs developed in-house.

Analysis of motifs

In order to eliminate incorrect/partial sequences from the set used to analyze the motifs, the animal and fungal sequences were initially examined for a single, best defined motif by MEME.[30] Four sequences from the initial 12 were excluded for these further analyses. In the animal set, *S. mansoni* and *C. virginicae* sequences appear to come from truncated cDNAs and do not contain the best defined motif detected in other sequences. When globally aligned, and examined for pairwise identity, the *H. sapiens* and *D. rerio* sequences showed identities around 90% with several other sequences and were excluded from further analysis to miminize bias. The final set of sequences selected for motif analysis consisted of eight animal sequences (*B. mori*, *C. elegans*, *C. intestinalis*, *D. melanogaster*, *E. multilocularis*, *O. latipes*, *S. japonicum* and *S. tropicalis*) and six fungal sequences (*S. cerevisiae*, *N. crassa*, *S. pombe*, *B. cinerea*, *G. zeae* and *C. albicans*).

For motif detection we used MEME,[30] ITERALIGN[31] and PROBE,[32] which employ different algorithms for motif detection. We applied all three programs to both animal and fungal sequences, as well as to their union, and the predicted motifs were compared. Only motifs found by all three programs were deemed to be genuine. In all instances, motifs predicted by these three programs were similar, with some variations in the exact start/end position of the reported motifs. This was resolved by comparing the exact positions of motifs predicted by MEME, ITERALIGN, and PROBE. Given the motif predicted by all three programs, the longest stretch of sequence common to all three predictions was deemed to represent a genuine motif.

Prediction of disordered regions

To determine regions likely to be disordered, individual analyses with each Tom20 sequence were undertaken with the predictor DisEMBL,[35] at the EMBL site†.

Homology modelling of the Tom20 type I isoform three-dimensional structure

The three-dimensional model of mouse type I Tom20 was generated using the MODELLER (6v2) program with default input parameters. Template coordinates were prepared from the NMR structure of rat Tom20 type II (PDB code 1om2, chain A), which corresponds to residues 51–143 of the full length protein. Rat Tom20 type II and mouse Tom20 type I, were globally aligned using CLUSTALW 1.81[56] and the default parameters. Residues 51–143 of the rat type II Tom20 align to residues 55–152 of the mouse type I Tom20. The regions of sequence corresponding to the ligand-binding domains (residues 55–128 in rat type II Tom20, 59–132 in the mouse type I Tom20) align without gaps. Weak alignment was observed in the C-terminal region of the two sequences, specifically after the residue V^{134} in the type II. This residue marks the end of the regular and well-defined structure in rat Tom20[25], and is not part of the ligand-binding region. ProsaII energies of the model structures are similar to the template structures, suggesting that the rat Tom20 type II provides a valid template for creating these models. Molecular images for Figures 3(c) and 4 were created using VMD‡.[57]

Calculation of solvent-accessible surfaces

The analysis of the solvent-accessible surfaces was confined to residues 59–132 of the model, which corresponds to residues 55–128 of the template Tom20 type II structure. This region forms a globular-like domain, and contains all well-defined secondary structure elements observed in the NMR structure of rat Tom20 type II.[25] It includes the presequence peptide-binding site, and the two sequences align in this region without any gaps.

The calculation of solvent-accessible surfaces was performed with the program CHARMM,[58] which implements the algorithm presented by Lee & Richards.[59] A probe radius of 1.6 Å was used. For the comparison of overall protein surfaces, the solvent-accessible surface for each side-chain was calculated as the sum of the contribution of individual atoms, with backbone atoms C′, N, and O excluded. For each model structure (20 NMR structures of rat Tom20 type I and 64 model structures of mouse Tom20 type II were analyzed) the accessible surface hydrophobicity per unit area was calculated, using hydrophobicity scales provided by Black & Mould,[60] Kyte & Doolittle[41] Eisenberg *et al*.[40] and Roseman.[61] For each hydrophobicity scale, the Shapiro–Wilk normality test was used to assess the normality of data, an *F*-test to test equality of variances, and Student's *t*-test was used to assess the equality of means.[62,63]

Analysis of the presequence peptide-binding groove

The model of the mouse Tom20 type I with bound presequence peptide was constructed by fitting the model coordinates into the position of the rat Tom20 type II with bound peptide. The C^{α} atoms positions of the globular domain (residues 59–132 of the full-length protein) were fit into the positions of the corresponding atoms of the type II structure. The fitted structure was minimized in CHARMM with fixed positions of backbone atoms to relieve bad side-chain contacts between the protein and the peptide with electrostatic interactions switched off. The solvent-accessible surface shielded by the peptide was calculated as the difference between the solvent-accessible surface of the protein with the presequence peptide deleted and the solvent-accessible surface of the protein–peptide complex.

Phylogenics

Maximum-likelihood (ML) and distance analyses were performed on an alignment of 60 Tom20 sequences from animals and fungi. Protein maximum-likelihood analyses were performed with 102 conserved characters using PhyML,[64] using an input tree generated by BIONJ, the JTT model of amino acid substitution, proportion of variable rates estimated from the data, and nine categories of substitution rates (eight variable and one invariable): 100 bootstrap trees were similarly calculated with PhyML. Topology tests were carried out by calculating site-likelihoods using PAML 3.12[65] with zero gamma categories for the ML tree, 100 ML bootstrap trees, and ten alternative topologies where the vertebrate type I clade was moved to ten different nodes. Approximately unbiased (AU) tests were then conducted on the site-likelihoods using CONSEL 0.1d.[66] For distance analyses, fewer characters (73) were used due to several incomplete sequences. Gamma corrected distances were calculated by TREE-PUZZLE 5.0,[67] using the WAG substitution matrix[68] with eight variable rate categories and invariable sites. Trees were inferred by Fitch–Margoliash using FITCH 3.6a§ and weighted neighbor joining using WEIGHBOR 1.0.1a.[69] Bootstrap resampling was

† http://dis.embl.de/

‡ http://www.ks.uiuc.edu/Research/vmd/

§ http://evolution.genetics.washington.edu/phylip.html performed using PUZZLEBOOT (shell script by Roger & Holder†) with rates and frequencies estimated using TREE-PUZZLE 5.0.

RNAi analysis in *C. elegans*

C. elegans strains were maintained at 20 °C unless specified otherwise, using standard techniques.[70,71] Wild-type genes were cloned from N2 Bristol and cloned into vector L4440 and the *Escherichia coli* host HT115 was used for feeding RNAi studies.[72] Live worms were mounted on a film of dried agarose in a small volume of M9 medium with 10 mM aldicarb. Worms were then visualized on a Zeiss Axiovert 200 microscope. OP50-fed wild-type adult worms were seeded onto NGM plates containing *E. coli* expressing dsRNA for F23H12.2 or F32B4.2. Adults were allowed to lay eggs for four hours and were then removed from the plate. The plates were incubated at the temperatures stated. The resulting progeny continued to eat the dsRNA food source and were further analyzed for body size, growth rate, and brood sizes.

Mouse gene expression and confocal microscopy

A mouse Multiple Tissue Northern (MTN™) Blot was purchased from CLONTECH Laboratories, Inc. Each lane contains a separation of 1 μg of poly(A)$^+$ RNA extracted from one of a selection of different tissue types. Probes complementary to the nucleotide sequence of *Mm*Tom20 type I (corresponding to amino acid residues 1–80) and *Mm*Tom20 type II (corresponding to amino acid residues 1–111) and β-*actin* were generated by PCR from cDNA clones. Random prime labeling with [α-^{32}P]dATP was carried out using Promega's Prime-a-Gene® system. Hybridization, autoradiography and membrane stripping were carried out according to the manufacturer's instructions.

HeLa cells were seeded on glass cover-slips and transiently transfected with the mammalian expression vector pEYFP-N1 encoding the mouse Tom20 proteins using FuGene 6 (Boehringer Mannheim). At 48 hours after transfection the cells were incubated in DMEM containing 150 nM Mitotracker® CMX-Ros (Molecular Probes) for 15 minutes at 37 °C. All cells were fixed in 4% (v/v) paraformaldehyde in PBS for 15 minutes and free aldehyde groups were quenched in 50 mM NH_4Cl/PBS. YFP and Mitotracker staining patterns were examined using a Bio-rad MRC-1024 confocal scanning laser microscope and images merged to examine co-localization.

Acknowledgements

We thank Daniel Bird for expert advice on Northern analysis, Paul Gleeson for expert advice and support with the mammalian cell culture experiments, and the European Bioinformatics Institute for access to CLUSTALW. This work was supported by a grant from the Australian Research Council (to T.L.), a Dora Lush Postgraduate Award (ID 310656) from the National Health and Medical Research Council (to J.H.) and an Australian Postgraduate Research Award (to A.P.).

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2004.12.057

References

1. Karlberg, O., Canback, B., Kurland, C. G. & Andersson, S. G. (2000). The dual origin of the yeast mitochondrial proteome. *Yeast*, **17**, 170–187.
2. Andersson, S. G., Karlberg, O., Canback, B. & Kurland, C. G. (2003). On the origin of mitochondria: a genomics perspective. *Phil. Trans. Roy. Soc. ser. B*, **358**, 165–179.
3. Taylor, S. W., Fahy, E., Zhang, B., Glenn, G. M., Warnock, D. E., Wiley, S. *et al*. (2003). Characterization of the human heart mitochondrial proteome. *Nature Biotechnol.* **21**, 281–286.
4. Mootha, V. K., Bunkenborg, J., Olsen, J. V., Hjerrild, M., Wisniewski, J. R., Stahl, E. *et al*. (2003). Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell*, **115**, 629–640.
5. McDonald, T. G. & Van Eyk, J. E. (2003). Mitochondrial proteomics. Undercover in the lipid bilayer. *Basic Res. Cardiol.* **98**, 219–227.
6. Sickmann, A., Reinders, J., Wagner, Y., Joppich, C., Zahedi, R., Meyer, H. E. *et al*. (2003). The proteome of *Saccharomyces cerevisiae* mitochondria. *Proc. Natl Acad. Sci. USA*, **100**, 13207–13212.
7. Guda, C., Fahy, E. & Subramaniam, S. (2004). MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics*, **20**, 1785–1794.
8. Roise, D., Horvath, S. J., Tomich, J. M., Richards, J. H. & Schatz, G. (1986). A chemically synthesized presequence of an imported mitochondrial protein can form an amphiphilic helix and perturb natural and artificial phospholipid bilayers. *EMBO J.* **5**, 1327–1334.
9. von Heijne, G. (1986). Mitochondrial targeting sequences may form amphiphilic helices. *EMBO J.* **5**, 1335–1342.
10. Smagula, C. & Douglas, M. G. (1988). Mitochondrial import of the ADP/ATP carrier protein in *Saccharomyces cerevisiae*. Sequences required for receptor binding and membrane translocation. *J. Biol. Chem.* **263**, 6783–6790.
11. Waizenegger, T., Stan, T., Neupert, W. & Rapaport, D. (2003). Signal-anchor domains of proteins of the outer membrane of mitochondria: structural and functional characteristics. *J. Biol. Chem.* **278**, 42064–42071.
12. Herrmann, J. M. & Neupert, W. (2000). Protein transport into mitochondria. *Curr. Opin. Microbiol.* **3**, 210–214.
13. Gabriel, K., Buchanan, S. K. & Lithgow, T. (2001). The alpha and the beta: protein translocation across mitochondrial and plastid outer membranes. *Trends Biochem. Sci.* **26**, 36–40.
14. Endo, T. & Kohda, D. (2002). Functions of outer membrane receptors in mitochondrial protein import. *Biochim. Biophys. Acta*, **1592**, 3–14.
15. Frazier, A. E., Chacinska, A., Truscott, K. N., Guiard,

† http:/www.tree-puzzle.de

B., Pfanner, N. & Rehling, P. (2003). Mitochondria use different mechanisms for transport of multispanning membrane proteins through the intermembrane space. *Mol. Cell. Biol.* **23**, 7818–7828.

16. Hachiya, N., Mihara, K., Suda, K., Horst, M., Schatz, G. & Lithgow, T. (1995). Reconstitution of the initial steps of mitochondrial protein import. *Nature*, **376**, 705–709.
17. Dekker, P. J., Ryan, M. T., Brix, J., Muller, H., Honlinger, A. & Pfanner, N. (1998). Preprotein translocase of the outer mitochondrial membrane: molecular dissection and assembly of the general import pore complex. *Mol. Cell. Biol.* **18**, 6515–6524.
18. Ahting, U., Thun, C., Hegerl, R., Typke, D., Nargang, F. E., Neupert, W. & Nussberger, S. (1999). The TOM core complex: the general protein import pore of the outer membrane of mitochondria. *J. Cell. Biol.* **147**, 959–968.
19. van Wilpe, S., Ryan, M. T., Hill, K., Maarse, A. C., Meisinger, C., Brix, J. *et al.* (1999). Tom22 is a multifunctional organizer of the mitochondrial preprotein translocase. *Nature*, **401**, 485–489.
20. Krimmer, T., Rapaport, D., Ryan, M. T., Meisinger, C., Kassenbrock, C. K., Blachly-Dyson, E. *et al.* (2001). Biogenesis of porin of the outer mitochondrial membrane involves an import pathway *via* receptors and the general import pore of the TOM complex. *J. Cell. Biol.* **152**, 289–300.
21. Sollner, T., Griffiths, G., Pfaller, R., Pfanner, N. & Neupert, W. (1989). MOM19, an import receptor for mitochondrial precursor proteins. *Cell*, **59**, 1061–1070.
22. Ramage, L., Junne, T., Hahne, K., Lithgow, T. & Schatz, G. (1993). Functional cooperation of mitochondrial protein import receptors in yeast. *EMBO J.* **12**, 4115–4123.
23. Iwahashi, J., Yamazaki, S., Komiya, T., Nomura, N., Nishikawa, S., Endo, T. & Mihara, K. (1997). Analysis of the functional domain of the rat liver mitochondrial import receptor Tom20. *J. Biol. Chem.* **272**, 18467–18472.
24. Yano, M., Kanazawa, M., Terada, K., Takeya, M., Hoogenraad, N. & Mori, M. (1998). Functional analysis of human mitochondrial receptor Tom20 for protein import into mitochondria. *J. Biol. Chem.* **273**, 26844–26851.
25. Abe, Y., Shodai, T., Muto, T., Mihara, K., Torii, H., Nishikawa, S. *et al.* (2000). Structural basis of presequence recognition by the mitochondrial protein import receptor Tom20. *Cell*, **100**, 551–560.
26. Haucke, V., Horst, M., Schatz, G. & Lithgow, T. (1996). The Mas20p and Mas70p subunits of the protein import receptor of yeast mitochondria interact *via* the tetratricopeptide repeat motif in Mas20p: evidence for a single hetero-oligomeric receptor. *EMBO J.* **15**, 1231–1237.
27. Bolliger, L., Junne, T., Schatz, G. & Lithgow, T. (1995). Acidic receptor domains on both sides of the outer membrane mediate translocation of precursor proteins into yeast mitochondria. *EMBO J.* **14**, 6318–6326.
28. Kanaji, S., Iwahashi, J., Kida, Y., Sakaguchi, M. & Mihara, K. (2000). Characterization of the signal that directs Tom20 to the mitochondrial outer membrane. *J. Cell. Biol.* **151**, 277–288.
29. Hofmann, K. (2000). Sensitive protein comparisons with profiles and hidden Markov models. *Brief Bioinform.* **1**, 167–178.
30. Bailey, T. L. & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36.
31. Brocchieri, L. & Karlin, S. (1998). A symmetric-iterated multiple alignment of protein sequences. *J. Mol. Biol.* **276**, 249–264.
32. Neuwald, A. F., Liu, J. S., Lipman, D. J. & Lawrence, C. E. (1997). Extracting protein alignment models from the sequence database. *Nucl. Acids Res.* **25**, 1665–1677.
33. Rice, P., Longden, I. & Bleasby, A. (2000). EMBOSS: the European molecular biology open software suite. *Trends Genet.* **16**, 276–277.
34. D'Andrea, L. D. & Regan, L. (2003). TPR proteins: the versatile helix. *Trends Biochem. Sci.* **28**, 655–662.
35. Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J. & Russell, R. B. (2003). Protein disorder prediction: implications for structural proteomics. *Structure (Camb)*, **11**, 1453–1459.
36. Hernandez, J. M., Giner, P. & Hernandez-Yago, J. (1999). Gene structure of the human mitochondrial outer membrane receptor Tom20 and evolutionary study of its family of processed pseudogenes. *Gene*, **239**, 283–291.
37. Marti-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F. & Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annu. Rev. Biophys. Biomol. Struct.* **29**, 291–325.
38. Sippl, M. J. (1993). Recognition of errors in three-dimensional structures of proteins. *Proteins: Struct. Funct. Genet.* **17**, 355–362.
39. Muto, T., Obita, T., Abe, Y., Shodai, T., Endo, T. & Kohda, D. (2001). NMR identification of the Tom20 binding segment in mitochondrial presequences. *J. Mol. Biol.* **306**, 137–143.
40. Eisenberg, D., Schwarz, E., Komaromy, M. & Wall, R. (1984). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. *J. Mol. Biol.* **179**, 125–142.
41. Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132.
42. Bauer, M. F., Rothbauer, U., Muhlenbein, N., Smith, R. J., Gerbitz, K., Neupert, W. *et al.* (1999). The mitochondrial TIM22 preprotein translocase is highly conserved throughout the eukaryotic kingdom. *FEBS Letters*, **464**, 41–47.
43. Andrews, J., Bouffard, G. G., Cheadle, C., Lu, J., Becker, K. G. & Oliver, B. (2000). Gene discovery using computational and microarray analysis of transcription in the *Drosophila melanogaster* testis. *Genome Res.* **10**, 2030–2043.
44. Schatz, G. (1997). Just follow the acid chain. *Nature*, **388**, 121–122.
45. Neupert, W. (1997). Protein import into mitochondria. *Annu. Rev. Biochem.* **66**, 863–917.
46. Chacinska, A., Pfanner, N. & Meisinger, C. (2002). How mitochondria import hydrophilic and hydrophobic proteins. *Trends Cell Biol.* **12**, 299–303.
47. Macasev, D., Whelan, J., Newbigin, E., Silva-Filho, M. C., Mulhern, T. D. & Lithgow, T. (2004). Tom22′, an 8-kDa trans-site receptor in plants and protozoans, is a conserved feature of the TOM complex that appeared early in the evolution of eukaryotes. *Mol. Biol. Evol.* **21**, 1557–1564.
48. Heins, L. & Schmitz, U. K. (1996). A receptor for protein import into potato mitochondria. *Plant J.* **9**, 829–839.
49. Werhahn, W., Niemeyer, A., Jansch, L., Kruft, V.,

Schmitz, U. K. & Braun, H. (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.

50. Hwa, J. J., Zhu, A. J., Hiller, M. A., Kon, C. Y., Fuller, M. T. & Santel, A. (2004). Germ-line specific variants of components of the mitochondrial outer membrane import machinery in *Drosophila*. *FEBS Letters*, **572**, 141–146.
51. Almstrup, K., Nielsen, J. E., Hansen, M. A., Tanaka, M., Skakkebaek, N. E. & Leffers, H. (2004). Analysis of cell-type-specific gene expression during mouse spermatogenesis. *Biol. Reprod.* **70**, 1751–1761.
52. Yu, Z., Guo, R., Ge, Y., Ma, J., Guan, J., Li, S. *et al.* (2003). Gene expression profiles in different stages of mouse spermatogenic cells during spermatogenesis. *Biol. Reprod.* **69**, 37–47.
53. Hecht, N. B. (1998). Molecular mechanisms of male germ cell differentiation. *Bioessays*, **20**, 555–561.
54. Sassone-Corsi, P. (2002). Unique chromatin remodeling and transcriptional regulation in spermatogenesis. *Science*, **296**, 2176–2178.
55. Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
56. Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* **22**, 4673–4680.
57. Humphrey, W., Dalke, A. & Schulten, K. (1996). VMD: visual molecular dynamics. *J. Mol. Graph.* **14**. 33-8, 27-8.
58. Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **4**, 187–217.
59. Lee, B. & Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **55**, 379–400.
60. Black, S. D. & Mould, D. R. (1991). Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications. *Anal. Biochem.* **193**, 72–82.
61. Roseman, M. A. (1988). Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds. *J. Mol. Biol.* **200**, 513–522.
62. Madansky, A. (1988). *Testing for normality Prescriptions for Working Statisticians*, Springer-Verlag, New York pp. 14–55.
63. Ott, R. L. & Longnecker, M. (2001). *An Introduction to Statistical Methods and Data Analysis*, Duxbury, Pacific Grove.
64. Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704.
65. Yang, Z. (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**, 555–556.
66. Shimodaira, H. & Hasegawa, M. (2001). CONSEL: for assessing the confidence of phylogenetic tree selection. *Bioinformatics*, **17**, 1246–1247.
67. Strimmer, K. & von Haeseler, A. (1996). Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.*, 964–969.
68. Goldman, N. & Whelan, S. (2000). Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Mol. Biol. Evol.* **17**, 975–978.
69. Bruno, W. J., Socci, N. D. & Halpern, A. L. (2000). Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol. Biol. Evol.* **17**, 189–197.
70. Sulston, J. E. & Hodgkin, J. (1988). The nematode *Caenorhabditis elegans*. In *Methods* (Wood, W.B., ed.), Cold Spring Harbor Laboratory Press, Plainview, NY, pp. 587–606.
71. Brenner, S. (1974). The genetics of *Caenorhabditis elegans*. *Genetics*, **77**, 71–94.
72. Timmons, L., Court, D. L. & Fire, A. (2001). Ingestion of bacterially expressed dsRNAs can produce specific and potent genetic interference in *Caenorhabditis elegans*. *Gene*, **263**, 103–112.

Edited by G. von Heijne

(Received 15 August 2004; received in revised form 17 December 2004; accepted 27 December 2004)