Title:

Ribosomal protein with conserved function has entirely different structures in different organisms

Authors:

Léon Schierholz[1,*], Charlotte R. Brown[2,*], Karla Helena-Bueno[2,*], Vladimir N. Uversky[3], Jonas Barandun[1], Sergey V. Melnikov[2]

Affiliations

1. Department of Molecular Biology, Laboratory for Molecular Infection Medicine Sweden, Umeå Centre for Microbial Research, Science for Life Laboratory, Umeå University, Umeå, 901 87, Sweden
2. Biosciences Institute, Newcastle University School of Medicine, Newcastle upon Tyne, NE2 4HH, UK
3. Department of Molecular Medicine and USF Health Byrd Alzheimer's Research Institute, Morsani College of Medicine, University of South Florida, Tampa, FL 33612, USA

* Contributed equally to this work

Corresponding authors:

jonas.barandun@umu.se and sergey.melnikov@ncl.ac.uk

Abstract

It is generally accepted that protein structures are more conserved than protein sequences. This notion is widely used to identify distant protein homologs and predict the structure and functions of uncharacterized proteins. However, this principle has been derived from studies of globular proteins, leaving it unclear whether the same principle is applied to non-globular (or intrinsically disordered) proteins. Here, to help answer this question, we describe the evolution of the ribosomal protein msL1/msL2 that was recently found in ribosomes from the parasitic microorganisms microsporidia. We first show that this protein has conserved function but entirely dissimilar structures in different organisms: in each of the analyzed species, msL1/msL2 exhibits an altered secondary structure, an inverted orientation of the N- and C-termini, and a completely transformed fold. We then show that this fold change is likely caused by the evolution of the msL1/msL2-binding site in the ribosome; specifically, by variations in microsporidian rRNA. These findings illustrate that structure may evolve faster than sequence in non-globular proteins. Hence, non-globular proteins can completely transform their fold without loss of function, challenging the current sequence-structure-function paradigm that is viewed as a universal principle of protein evolution.

Significance

Our current understanding of protein evolution is largely based on studies of stand-alone globular proteins. Hence, we know little about the evolution of other essential and abundant proteins, including non-globular proteins and proteins that comprise multi-subunit assemblies. Here, we describe a ribosomal protein that evolves in a strikingly dissimilar fashion to globular proteins. While this protein has a conserved function and similar sequences in two different organisms, it has entirely dissimilar folds. This finding may help to better understand the sequence-structure-function interplay during the evolution of non-globular proteins, thereby enhancing our ability to predict the structure of newly discovered proteins, annotate genomes, and peer into the origin and evolutionary history of life on Earth.

INTRODUCTION

The classical view of protein evolution, mostly derived from studies of globular enzymes, postulates that protein sequences evolve faster than protein structures (1-8). This model assumes that when two initially identical proteins accumulate mutations relative to each other, their structures undergo slow transformation that occurs at about a three- to ten-times slower rate than the rate of sequence evolution (7, 8). Hence, although evolving sequences may become too dissimilar to be seen as homologs, their structures may remain similar, making it possible to identify homologous proteins that are overlooked by the sequence analyses (9, 10).

This view of protein evolution is widely used in current biology to predict the functions of uncharacterized proteins (11-13), shed light on the origin of life (14-18), and engineer minimal enzymes to help create artificial life (19-21). However, in addition to globular proteins, nature harbors an abundance of non-globular (or intrinsically disordered) proteins and proteins with long disordered segments (22-27). Based on recent estimates, these intrinsically disordered proteins include more than 1,150 protein families, and more than 50% of eukaryotic proteins bear at least one long disordered segment (28-31). These proteins are typically multifunctional and represent an illustration of the "protein structure-function continuum" model, where a protein structure is viewed as a dynamic ensemble of multiple conformations with various structural features and various functionalities (32, 33). The abundance of non-globular proteins, along with their frequent omission from structural studies, raises the question: do they follow the same principle of molecular evolution as their globular counterparts?

Recently, an unexpected opportunity to answer this question arose from structural studies of ribosomes from the group of fungal parasites known as microsporidia (34-36). These studies revealed that microsporidian ribosomes possess an additional non-globular protein, msL1/msL2. This protein was first discovered in ribosomes of the parasite *Vairimorpha necatrix*, where it was annotated as msL1-*Vn*; its homologs were subsequently found in 10 microsporidian species, including the parasite *Encephalitozoon cuniculi* (35, 36). Later, however, an investigation of the structure of *E. cuniculi* ribosomes revealed that the msL1-*Vn* binding site was occupied by a structurally dissimilar protein, which was annotated as msL2-*Ec* (36). The study of *E. cuniculi* ribosomes, however, has overlooked that the proteins msL1-*Vn* and msL2-*Ec* were previously described as distant homologs following an iterative search of msL1-*Vn* homologs (35). This homology was easy to overlook because msL1-*Vn* and msL2-*Ec* have highly dissimilar structures, with a root-mean-square deviation (RMSD) of approximately 11.4 Å between their backbone residues, compared with an RMSD of less than 2.4 Å for close homologs and approximately 4.5 Å for highly distant homologs.

Here, we describe this previously overlooked structural anomaly, in which a seemingly conserved ribosomal protein completely transforms its three-dimensional structure in one species relative to another. We then provide an evolutionary insight into the mechanism of this fold change. We show that the fold change of msL1/msL2 was likely driven by its adaptation to rRNA reduction in microsporidian parasites. Furthermore, we show that while msL1-*Vn* and msL2-*Ec* share approximately 41% sequence similarity, their conserved residues form dissimilar contacts in the from *E. cuniculi* and *V. necatrix*. Thus, while the proteins msL1-*Vn* and msL2-*Ec* are classified as structural homologs—based on canonical sequence analyses—these proteins have in fact entirely dissimilar three-dimensional structures with dissimilar function of their conserved residues and different mechanisms of ribosome attachment.

This example illustrates that, in non-globular proteins, their structure may evolve faster than their sequence. This finding suggests that proteins, whose structure is shaped primarily by molecular contacts with other molecules, may evolve in a drastically different manner compared with globular

proteins, and that widely accepted concepts derived from studies of well-folded proteins may not represent universal principles of protein evolution.

RESULTS

The msL1/msL2 fold change in microsporidian species

To better understand how the proteins msL1-*Vn* and msL2-*Ec* could have changed their fold during their evolution, we first compared their structures and ribosome attachment mechanisms. Both msL1-*Vn* and msL2-*Ec* were clearly visualized in the cryo-EM structures of their respective ribosomes and display unique and distinct folds compared to each other and to AlphaFold predictions (**Fig. 1a**). Although the two proteins are of identical length (72 amino acids), share 41% sequence similarity, and some secondary structure similarity, namely two short α-helical segments, between residues 23-40 and 58-68, the overall folds are dissimilar with the backbone RMSD value of 11.4Å.

We next compared the ribosomal binding of msL1-*Vn* and msL2-*Ec* and found that these proteins also differ in their orientation in the ribosome. While msL1-*Vc* and msL2-*Ec* occupy the same site in the large ribosomal subunit, our analysis showed that in *V. necatrix* ribosomes, the N-terminus of msL1-*Vn* is directed toward the solvent side of the ribosome, whereas in *E. cuniculi* ribosomes, the N-terminus of msL2-*Ec* is oriented in the opposite direction, toward the ribosomal core. Thus, the proteins msL1-*Vn* and msL2-*Ec* are inverted relative to each other in the ribosome structure, pointing their N- and C-termini in opposite directions (**Fig. 1b**).

Given the extent of sequence similarity between msL1-*Vn* and msL2-*Ec*, which share approximately 41% similar residues, we also compared the molecular contacts of these conserved residues with the ribosome. In so doing, we wished to understand whether the conserved residues of these proteins could make conserved contacts with the ribosome. In total, we analyzed the contacts of 21 identical and 8 similar residues of msL1-*Vn* and msL2-*Ec*. We found that most of these residues mediated msL1-*Vn*/msL2-*Ec* attachment to the ribosome; however, none bound the ribosome in a similar manner in msL1-*Vn* compared with msL2-*Ec*. For example, in *V. necatrix* ribosomes, the conserved motif ^{43}KIKxLKxKK^{51} of msL1-*Vn* bound the bundle of rRNA helices H5/H8/H10/H52/H53, whereas in *E. cuniculi* ribosomes the same motif was moved approximately 22 Å away, binding a different bundle, H55/H56/H66 (**Fig. 1b,c**).

Overall, our analysis revealed a vivid contrast between a relatively high conservation of msL1-*Vn*/msL2-*Ec* sequences and complete change of fold. Thus, msL1-*Vn* and msL2-*Ec* represent a new example of proteins classified as sequence homologs that adopt dissimilar folds and completely repurpose their conserved residues in the ribosome structure to retain their biological function.

Each microsporidian lineage bears a dissimilar ortholog of msl1/msL2

To better understand the evolution of msL1/msL2 folds, we next compared msL1-*Vn* and msL2-*Ec* with their structurally uncharacterized homologs from other microsporidian species. In doing so, we sought to understand whether these uncharacterized homologs resemble msL1-*Vn* or msL2-*Ec* in ribosome attachment and possibly in fold. To address this, we first searched for msL1/msL2 homologs in the NCBI genome database using a sensitive homolog search based on hidden Markov models (**Materials and Methods**, **Supplementary Data S1**). This search revealed detectable msL1/msL2 homologs in 16 of the 24 characterized microsporidian species, indicating that msL1/msL2 is a widely occurring protein among microsporidians (**Fig. 2**).

We then assessed the sequence similarity between these msL1/msL2 homologs and msL1-*Vn* and msL2-*Ec* (**Fig. 2b**, Supplementary **Figs. 1 and 2**). We found that among *Encephalitozoon* species, msL1/msL2 homologs shared more than 95% of sequence conservation in those residues that mediate the ribosome attachment of msL2-*Ec*, suggesting that in *Encephalitozoon* species these homologs share the same ribosome attachment strategy and possibly the same fold as in msL2-*Ec*. We also found that msL1-*Vn* shared less than 48% sequence conservation in those residues that mediate ribosome attachment in msL2-*Ec*, consistent with the fact that msL1-*Vn* has a dissimilar ribosome attachment strategy and fold compared with those of msL2-*Ec*.

Remarkably, we found that the remaining microsporidian species carried mutations in most of the residues that are responsible for ribosome attachment of msL1-*Vn* or msL2-*Ec* (**Fig. 2a,b**). For example, the *Trachipleistophora hominis* homolog of msL1/msL2 carries mutations in approximately 66% of the ribosome-binding residues of msL1-*Vn* and 68% of the ribosome-binding residues of msL2-*Ec*. Hence, if the msL1/msL2 homolog from *T. hominis* has the same fold as msL1-*Vn* or msL2-*Ec*, it would carry mutations in most of its ribosome-binding residues. This high variability of the ribosome-binding interface suggests that each lineage of microsporidian parasites (such as *Encephalitozoon*, *Nosematidae*, *Enterospora*, and *Trachipleistophora*) encodes structurally dissimilar variants of msL1/msL2.

The msL1/msL2 fold change is likely caused by rRNA truncation

Finally, to understand what may have caused the fold change of msL1-*Vn* and msL2-*Ec* relative to each other, we compared the molecular surroundings of these proteins in the ribosome. In doing so, we wished to test our hypothesis that the fold change is induced by ribosomal surrounding of msL1-*Vn* and msL2-*Ec*. To test this idea, we first superposed the structures of *V. necatrix* and *E. cuniculi* ribosomes and found that the msL1/msL2-binding pockets are not identical, with the most prominent difference present in helix H52 of the 25S rRNA (**Fig. 3a**). This helix is three bases shorter in *E. cuniculi* compared with the helix in *V. necatrix*. Furthermore, if this helix in *V. necatrix* ribosomes was the same size as that in *E. cuniculi*, it would eliminate 24% of the ribosome-binding surface for msL1-*Vn*, potentially explaining why msL1-*Vn*/msL2-*Ec* cannot adopt the same fold in *E. cuniculi* and *V. necatrix* ribosomes.

Because this finding suggested that the H52 truncation was the major driver of the msL1-*Vn*/msL2-*Ec* fold change, we next assessed the length of H52 in microsporidian species (**Fig. 3b**, **Supplementary Data 2**). We found that in typical eukaryotes, such as *Saccharomyces cerevisiae*, H52 comprises 61 bases, whereas in microsporidian species H52 is reduced to 12-33 bases. In microsporidian lineages with short H52s, msL1/msL2 homologs are present in every species. However, in lineages with longer H52s, msL1/msL2 homologs were only occasionally observed. For one of these species with longer H52, *Paranosema locustae*, the structure of the ribosome was previously determined, and the absence of msL1/msL2 was confirmed experimentally (34).

To better understand the evolutionary interplay between the length of H52 and the occurrence of msL1/msL2 homologs in microsporidia, we finally docked msL1-*Vn* and msL2-*Ec* into their ribosomal binding sites in *P. locustae* and *S. cerevisiae* ribosomes (**Fig. 3c**). We found that the longer H52 would not allow msL1-*Vn* and msL2-*Ec* to bind the ribosome, suggesting that these proteins would require a different structure to bind the ribosome. Collectively, our findings suggest that variations in rRNA size can indeed be a driver of the msL1-*Vn*/msL2-*Ec* fold change.

Discussion

A ribosomal protein with conserved function has entirely dissimilar folds in different organisms

In this study, we capitalized on the recent advances in structural biology of microsporidian ribosomes to establish a new approach to explore the evolution of non-globular proteins and proteins that evolve in the context of complex molecular machines. Here, instead of studying ribosomes from the most evolutionary distant species, we compared ribosomes from the closest evolutionary branches of organisms. This approach enabled us to identify an evolutionary transition in the ribosomal protein msL1/msL2 that would be impossible to observe using computational predictions or sequence analyses. We found that this ribosomal protein exhibits a change in the overall fold when one organism is compared with another, and that this fold change occurs without loss of sequence similarity or loss of function. Therefore, unlike the case of globular proteins where sequence is thought to carry all the information needed for folding (known as the Anfinsen principle), proteins msL1/msL2 illustrate a different evolutionary principle in which the structure of a protein depends on other factors than solely on its sequence.

In non-globular proteins, structure can evolve faster than sequence

What makes the msL1/msL2 structure evolve faster than its sequence, thus defying the principle that protein structures evolve three- to ten-times slower than protein sequences? Our analysis suggests that this evolutionary transformation is primarily caused by changes in the msL1/msL2 ribosomal binding site. This finding is consistent with previous studies showing that microsporidian rRNA appears to evolve faster than microsporidian ribosomal proteins (34-43). It is, therefore, possible that microsporidian ribosomal proteins are frequently required to adapt to rRNA variations to compensate for rRNA variations.

This finding also illustrates that most of what we know about protein evolution is based on studies of globular proteins, yet msL1/msL2 belongs to the class of relatively understudied non-globular proteins, also known as intrinsically disordered proteins. In fact, our bioinformatics analysis revealed that all msL1/msL2 studied here are characterized by high levels of predicted intrinsic disorder (see **Supplementary Figures S1** and **S2**). Furthermore, based on the structural analysis presented here, we documented that not only msL1/msL2 lacks a hydrophobic globule, but its three-dimensional structure appears to depend on contacts with the helix H52. This helix neutralizes the positive charge of the msL1/msL2 molecule, promoting the msL1/msL2 folding. Our analysis revealed that this rRNA helix varies in size and sequence among microsporidian species. Hence, the binding site of msL1/msL2 has a variable structure that is most compatible only with a certain fold of the msL1/msL2 molecule. The transition of the msL1/msL2 structure enables it to adapt to rRNA truncations and remain bound to the ribosome despite structural variations in its ribosomal binding site. Thus, non-globular proteins such as msL1/msL2 evolve in a conceptually different way compared with the evolution of globular proteins: instead of their structure gradually changing in response to sequence mutations, their structure changes in response to structural changes in their binding partners. This could explain how non-globular proteins such as msL1/msL2 evolve dissimilar structures following relatively small changes in their sequence.

Non-globular proteins such as msL1/msL2 require a different strategy of homology search

Our study also shows potential pitfalls of structure-based searches of homologous proteins that might be ineffective when applied to non-globular proteins. This finding is important because structure-based detection of homologous proteins is rapidly becoming the major tool to predict the biological functions of uncharacterized proteins: with nearly 200,000 experimentally determined structures deposited in the Protein Data Bank, and even more structures predicted by AlphaFold, we now have structural information about almost every protein known to science (44). Hence, it is

becoming more and more tempting to directly compare models of protein structures, bypassing the need to compare protein sequences (45-48).

However, our efforts to model msL1/msL2 structure using AlphaFold illustrated on potential issue of this approach, as the program generated highly similar structures (**Fig. 1a**). This was unsurprising given the high sequence similarity; however, ", both predicted structures deviate significantly from the experimentally determined and the predicted free form (**Fig. 1a**). What's more, our analysis shows that difference in structure does not necessarily translate into difference in function, which needs to be taken into account when conducting functional annotation of uncharacterized proteins based on structural similarity.

The example of msL1/msL2 also highlights that lack of structural similarity does not necessarily mean independent or very distant evolutionary origin. Instead, we show that a single polypeptide could transform its fold and alter 100% of its functional interface without any loss of biological function. Hence, for proteins such as msL1/msL2, experimental structural information is still the gold standard which cannot be replaced by today's methods of bioinformatics and structural simulations.

Notably, the example of msL1/msL2 shows that the most conserved feature of this protein is neither its sequence nor its structure, but its position in the ribosome. It is therefore possible that for proteins that act as part of complex molecular assemblies, their position in these assemblies can be used as an independent indicator of the common evolutionary origin of seemingly unrelated proteins.

Taken together, we urge the community to exercise caution when comparing protein structures for homology searching, functional annotation or modeling of proteins that are non-globular, include disordered regions, or have evolved as a part of macromolecular assemblies. Based on our results, those are the cases where the principles that drive folding and molecular evolution may deviate significantly from our current understanding, exposing the need to revise the current sequence-structure-function paradigm of protein evolution.

MATERIALS AND METHODS

Comparative analysis of microsporidian ribosome structures. The msL1/msL2 structures in *E. cuniculi* and *V. necatrix* ribosomes were compared using Coot (0.9.4.1) to superpose structures, measure inter- and intramolecular distances, and calculate the RMSD between protein backbone residues (49). The contact area between msL1/msL2 and rRNA was calculated using the CCP4 package AreaIMol (8.0.001), with the solvent sphere diameter set to 1.7Å (50). The figures were prepared using ChimeraX (51). Intrinsic disorder propensities of the ribosomal proteins from the msL1/msL2 family were evaluated by the PONDR® VLXT computational tool (52). To evaluate global disorder predisposition of a query protein, we calculated the percent of predicted disordered residues (i.e., residues with the disorder scores above 0.5) and calculated the average disorder score as a protein length-normalized sum of all the per-residue disorder scores in a query protein. For further classification of these proteins, we used CH-CDF analysis (53-55).

Searching for msL1/msL2 homologs. Putative msL1/msL2 homologs were identified using two iterations of HMMER (56), with sequences of msL1/msL2 from either *E. cuniculi* or *V. necatrix* as the initial input. To maximize the success rate, the search was repeated using the following databases of protein sequences: Reference Proteomes, UniProtKB, SwissProt, and Ensemble. The obtained results from each of these searches were combined, reduced to remove redundant findings, and listed (**Supplementary Data 1**).

Analysis of variations in microsporidian rRNA. To analyze variations in microsporidian rRNA, microsporidian 25S rRNA sequences were consolidated using the SILVA ribosomal RNA gene database (57), the RNAcentral database (58), and local genome databases (*V. necatrix* and *P. locustae*). These sequences were then aligned using Clustal Omega with default parameters, the resulting alignment is shown in (**Supplementary Data 2**). Where available, ribosomal structures were used to manually model the 2D structures of helix H52. These structures included the models of *S. cerevisiae* (PDB 4V88), *V. necatrix* (PDB 6RM3), *P. locustae* (PDB 6ZU5) and *E. cuniculi* (PDB 7QEP) ribosomes. For the remaining species, respective segments of the 25S rRNA alignment were used to predict 2D structures with Mfold using the RNA Folding Form feature (59). The sequences corresponding to the 2D structures of Helix H52 were then assessed to calculate the total length of the rRNA segment in each microsporidian species.

Assembling a microsporidian phylogenetic tree. The microsporidian -phylogeny shown in (**Fig. 2b**) and (**Fig. 3b**) was assembled based on ribosomal protein sequences. Ribosomal protein sequences were retrieved by using the translated nucleotide blast (tblastn) with an E-value cutoff of 0.05 based on sequences of the *S. cerevisiae* or *V. necatrix* hits as query and the NCBI NIH and the MicrosporidiaDB (60) as the search databases. Protein sequences for *V. necatrix* and *P. locustae* were procured from local databases (34, 35). The protein sequences were aligned using MUSCLE (61), followed by trimming with trimAl (62) using the – gappyout option. Trimmed alignments were subsequently concatenated using FASconcat (63). The resulting phylogenetic tree was assembled with IQ-TREE and visualized using FigTree.v1.4.4 using *S. cerevisiae* as the root.

ACKNOWLEDGEMENTS AND FUNDING SOURCES

We thank members of the Barandun and Melnikov laboratories for help with the manuscript preparation, Robert Hirt and Johan Panek (both at Newcastle University, UK) for their critical comments. This work was funded by the BBSRC UK Studentship BB/T008695/1 (to C.R.B.), the NUORS 2021 award (to K.H-B.), the Swedish Research Council (2019-02011), the European Research Council (ERC Starting Grant PolTube 948655), the SciLifeLab National Fellows program, and MIMS (all to J.B.) and the European Union's Horizon 2020 research and innovation programme (the Grant Agreements No. 895166 to S.M).

References

1. S. Bershtein, D. Kleiner, D. Mishmar, Predicting 3D protein structures in light of evolution. *Nat Ecol Evol* **5**, 1195-1198 (2021).
2. T. Bepler, B. Berger, Learning the protein language: evolution, structure, and function. *Cell Syst* **12**, 654-669 e653 (2021).
3. N. Bordin, I. Sillitoe, J. G. Lees, C. Orengo, Tracing evolution through protein structures: nature captured in a few thousand folds. *Front Mol Biosci* **8**, 668184 (2021).
4. A. S. Konagurthu *et al.*, Universal architectural concepts underlying protein folding patterns. *Front Mol Biosci* **7**, 612920 (2021).
5. G. Postic, Y. Ghouzam, R. Chebrek, J. C. Gelly, An ambiguity principle for assigning protein structural domains. *Sci Adv* **3**, e1600552 (2017).
6. L. N. Kinch, N. V. Grishin, Evolution of protein structures and functions. *Curr Opin Struct Biol* **12**, 400-408 (2002).
7. K. Illergard, D. H. Ardell, A. Elofsson, Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. *Proteins* **77**, 499-508 (2009).
8. M. I. Sadowski, D. T. Jones, The sequence-structure relationship and protein function prediction. *Curr Opin Struct Biol* **19**, 357-362 (2009).
9. E. V. Koonin, K. S. Makarova, Y. I. Wolf, Evolution of microbial genomics: conceptual shifts over a quarter century. *Trends Microbiol* **29**, 582-592 (2021).
10. C. Zhang, P. L. Freddolino, Y. Zhang, COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information. *Nucleic Acids Res* **45**, W291-W299 (2017).
11. A. Rives *et al.*, Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* **118** (2021).
12. J. C. Whisstock, A. M. Lesk, Prediction of protein function from protein sequence and structure. *Q Rev Biophys* **36**, 307-340 (2003).
13. V. Gligorijevic *et al.*, Structure-based protein function prediction using graph convolutional networks. *Nat Commun* **12**, 3168 (2021).
14. C. Alvarez-Carreno, P. I. Penev, A. S. Petrov, L. D. Williams, Fold evolution before LUCA: common ancestry of SH3 domains and OB domains. *Mol Biol Evol* **38**, 5134-5143 (2021).
15. V. Alva, J. Soding, A. N. Lupas, A vocabulary of ancient peptides at the origin of folded proteins. *Elife* **4**, e09410 (2015).
16. J. C. Bowman, A. S. Petrov, M. Frenkel-Pinter, P. I. Penev, L. D. Williams, Root of the Tree: the significance, evolution, and origins of the ribosome. *Chem Rev* **120**, 4848-4878 (2020).
17. M. Krupovic, V. V. Dolja, E. V. Koonin, Origin of viruses: primordial replicators recruiting capsids from hosts. *Nat Rev Microbiol* **17**, 449-458 (2019).
18. N. A. Kovacs, A. S. Petrov, K. A. Lanier, L. D. Williams, Frozen in time: the history of proteins. *Mol Biol Evol* **34**, 1252-1260 (2017).
19. J. Rees-Garbutt, J. Rightmyer, O. Chalkley, L. Marucci, C. Grierson, Testing theoretical minimal genomes using whole-cell models. *ACS Synth Biol* **10**, 1598-1604 (2021).
20. C. A. Hutchison, 3rd *et al.*, Design and synthesis of a minimal bacterial genome. *Science* **351**, aad6253 (2016).
21. J. C. Lachance, S. Rodrigue, B. O. Palsson, Minimal cells, maximal knowledge. *Elife* **8** (2019).
22. A. Toto, P. Sormanni, C. Paissoni, V. N. Uversky, Intrinsically disordered proteins and regions: the challenge to the structure-function relationship. *Front Mol Biosci* **9**, 962643 (2022).
23. P. E. Wright, H. J. Dyson, Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol* **16**, 18-29 (2015).
24. R. van der Lee *et al.*, Classification of intrinsically disordered regions and proteins. *Chem Rev* **114**, 6589-6631 (2014).
25. C. J. Oldfield, A. K. Dunker, Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu Rev Biochem* **83**, 553-584 (2014).
26. P. Tompa, Intrinsically disordered proteins: a 10-year recap. *Trends Biochem Sci* **37**, 509-516 (2012).
27. V. N. Uversky, A. K. Dunker, Understanding protein non-folding. *Biochim Biophys Acta* **1804**, 1231-1264 (2010).
28. P. Kulkarni, V. N. Uversky, Intrinsically disordered proteins: the dark horse of the dark proteome. *Proteomics* **18**, e1800061 (2018).
29. S. Contreras-Martos *et al.*, Quantification of intrinsically disordered proteins: a problem not fully appreciated. *Front Mol Biosci* **5**, 83 (2018).
30. B. Xue, A. K. Dunker, V. N. Uversky, Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J Biomol Struct Dyn* **30**, 137-149 (2012).
31. Z. Peng *et al.*, Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life. *Cell Mol Life Sci* **72**, 137-151 (2015).
32. V. N. Uversky, Dancing protein clouds: the strange biology and chaotic physics of intrinsically disordered proteins. *J Biol Chem* **291**, 6681-6688 (2016).
33. V. N. Uversky, Protein intrinsic disorder and structure-function continuum. *Prog Mol Biol Transl Sci* **166**, 1-17 (2019).
34. K. Ehrenbolger *et al.*, Differences in structure and hibernation mechanism highlight diversification of the microsporidian ribosome. *PLoS Biol* **18**, e3000958 (2020).
35. J. Barandun, M. Hunziker, C. R. Vossbrinck, S. Klinge, Evolutionary compaction and adaptation visualized by the structure of the dormant microsporidian ribosome. *Nat Microbiol* **4**, 1798-1804 (2019).
36. D. Nicholson *et al.*, Adaptation to genome decay in the structure of the smallest eukaryotic ribosome. *Nat Commun* **13**, 591 (2022).

37. E. Peyretaillade *et al.*, Microsporidian Encephalitozoon cuniculi, a unicellular eukaryote with an unusual chromosomal dispersion of ribosomal genes and a LSU rRNA reduced to the universal core. *Nucleic Acids Res* **26**, 3513-3520 (1998).
38. O. Lecompte, R. Ripp, J. C. Thierry, D. Moras, O. Poch, Comparative analysis of ribosomal proteins in complete genomes: an example of reductive evolution at the domain scale. *Nucleic Acids Res* **30**, 5382-5390 (2002).
39. S. Melnikov, K. Manakongtreecheep, D. Soll, Revising the structural diversity of ribosomal proteins across the three domains of life. *Mol Biol Evol* **35**, 1588-1598 (2018).
40. E. M. O'Mahony, W. T. Tay, R. J. Paxton, Multiple rRNA variants in a single spore of the microsporidian Nosema bombi. *J Eukaryot Microbiol* **54**, 103-109 (2007).
41. L. Wadi, A. W. Reinke, Evolution of microsporidia: an extremely successful group of eukaryotic intracellular parasites. *PLoS Pathog* **16**, e1008276 (2020).
42. N. Jespersen, L. Monrroy, J. Barandun, Impact of genome reduction in microsporidia. *Exp Suppl* **114**, 1-42 (2022).
43. S. V. Melnikov *et al.*, Muller's ratchet and ribosome degeneration in the obligate intracellular parasites microsporidia. *Int J Mol Sci* **19** (2018).
44. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).
45. T. Aderinwale *et al.*, Real-time structure search and structure classification for AlphaFold protein models. *Commun Biol* **5**, 316 (2022).
46. S. Ovchinnikov *et al.*, Protein structure determination using metagenome sequence data. *Science* **355**, 294-298 (2017).
47. L. Deng, G. Zhong, C. Liu, J. Luo, H. Liu, MADOKA: an ultra-fast approach for large-scale protein structure similarity searching. *BMC Bioinformatics* **20**, 662 (2019).
48. J. Pereira *et al.*, High-accuracy protein structure prediction in CASP14. *Proteins* **89**, 1687-1699 (2021).
49. P. Emsley, B. Lohkamp, W. G. Scott, K. Cowtan, Features and development of Coot. *Acta Crystallogr D Biol Crystallogr* **66**, 486-501 (2010).
50. M. A. Hough, K. S. Wilson, From crystal to structure with CCP4. *Acta Crystallogr D Struct Biol* **74**, 67 (2018).
51. E. F. Pettersen *et al.*, UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* **30**, 70-82 (2021).
52. P. Romero *et al.*, Sequence complexity of disordered protein. *Proteins* **42**, 38-48 (2001).
53. F. Huang *et al.*, Improving protein order-disorder classification using charge-hydropathy plots. *BMC Bioinformatics* **15 Suppl 17**, S4 (2014).
54. A. Mohan, W. J. Sullivan, Jr., P. Radivojac, A. K. Dunker, V. N. Uversky, Intrinsic disorder in pathogenic and non-pathogenic microbes: discovering and analyzing the unfoldomes of early-branching eukaryotes. *Mol Biosyst* **4**, 328-340 (2008).
55. F. Huang *et al.*, Subclassifying disordered proteins by the CH-CDF plot method. *Pac Symp Biocomput*, 128-139 (2012).
56. R. D. Finn, J. Clements, S. R. Eddy, HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* **39**, W29-37 (2011).
57. C. Quast *et al.*, The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**, D590-596 (2013).
58. R. C. The, RNAcentral: a hub of information for non-coding RNA sequences. *Nucleic Acids Res* **47**, D1250-D1251 (2019).
59. M. Zuker, Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31**, 3406-3415 (2003).
60. S. Warrenfeltz *et al.*, EuPathDB: the eukaryotic pathogen genomics database resource. *Methods Mol Biol* **1757**, 69-113 (2018).
61. R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797 (2004).
62. S. Capella-Gutierrez, J. M. Silla-Martinez, T. Gabaldon, trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972-1973 (2009).
63. P. Kuck, K. Meusemann, FASconCAT: Convenient handling of data matrices. *Mol Phylogenet Evol* **56**, 1115-1118 (2010).

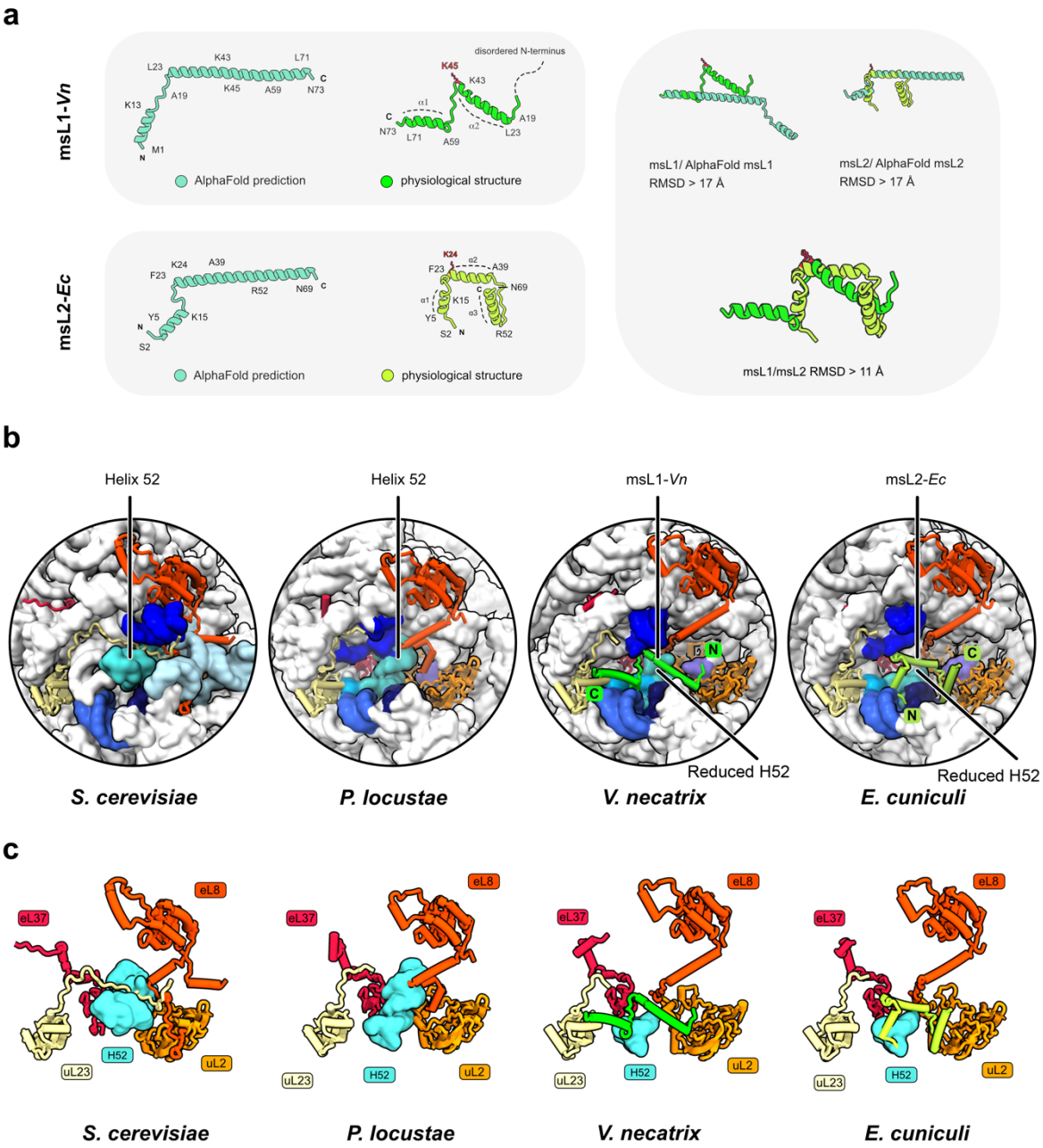

Figure 1. The msL1/msL2 fold change in microsporidian ribosomes. a) Structures of two homologous ribosomal proteins, msL1-*Vn* and msL2-*Ec*, as predicted by AlphaFold or observed in the structures of *V. necatrix* and *E. cuniculi* ribosomes, respectively. **b)** Zoom-in views of ribosome structures compare the binding sites for msL1-*Vc* and msL2-*Ec* and corresponding ribosomal sites in *S. cerevisiae* and *P. locustae*. In microsporidian ribosomes compared to ribosomes from other eukaryotes, the helix H52 is truncated. In *V. necatrix*, the void of the truncated H52 is filled with msL1-*Vn*, and in *E. cuniculi* ribosomes this void is filled with msL1-*Ec*. **c)** Side-by-side comparison of the molecular surrounding of msL1-*Vn*, msL2-*Ec* and H52 in the ribosome from different species.

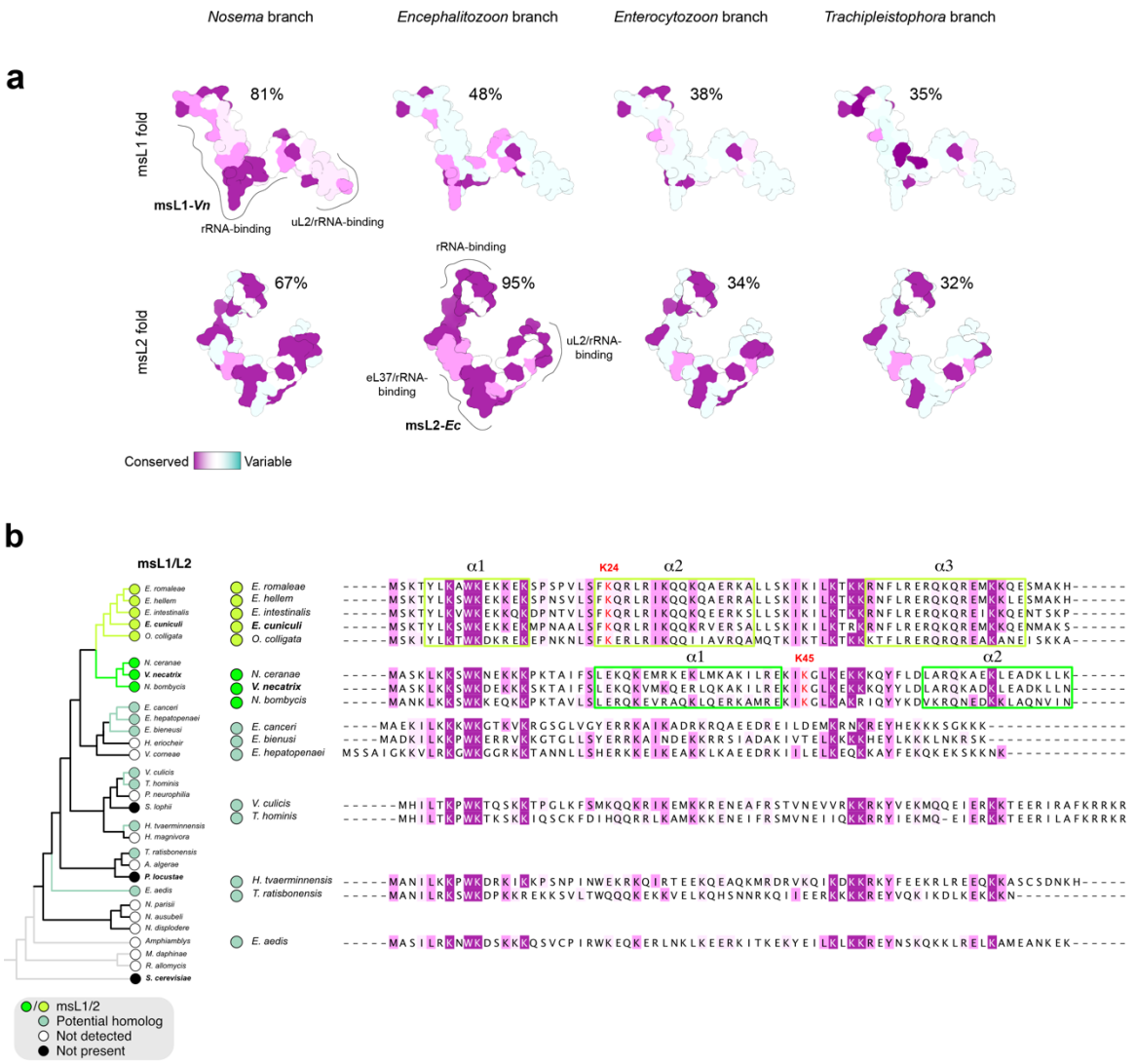

Figure 2. Each lineage of microsporidians encodes a dissimilar ortholog of msl1/msL2. a) Structures of msL1-*Vn* and msL2-*Ec* are colored by conservation of their ribosome-binding residues in each microsporidian lineage. The panel shows that if msL1/msL2 homologs from other microsporidian branches would have the same fold as either msL1-*Vn* or msL2-*Ec,* then these proteins would carry mutations in most of their ribosome-binding residues. **b)** A ribosomal protein-based phylogenetic tree highlights microsporidian species with detectable msL1/msL2 homologs. A multiple sequence alignment illustrates the conservation msL1/msL2 homologs from different microsporidian species.

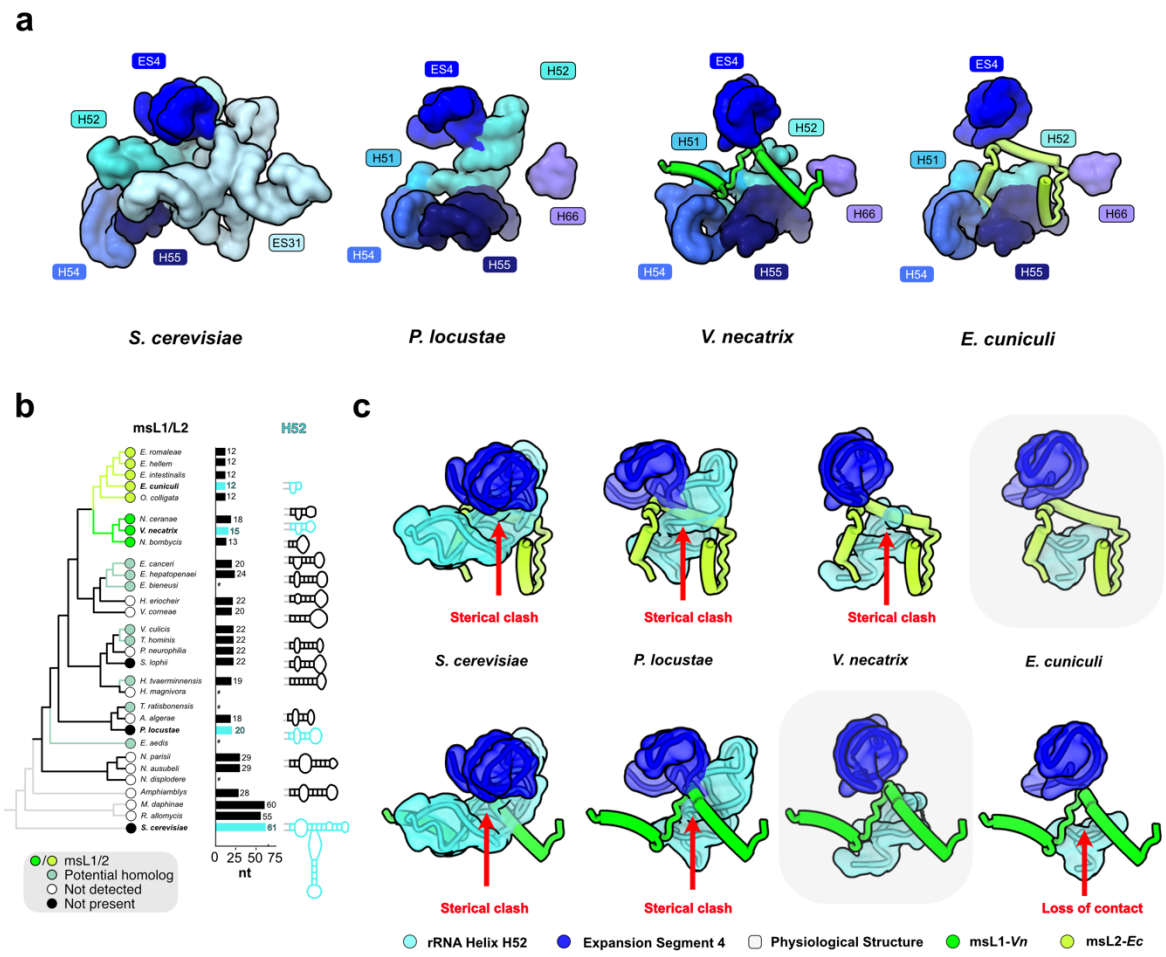

Figure 3. The msL1/msL2 fold change is likely caused by rRNA truncations. a) The structures compare ribosomal binding pockets of msL1-*Vn* and msL2-*Ec* and their corresponding ribosome segments in *S. cerevisiae* and *P. locustae*. **b)** A ribosomal protein-based phylogenetic tree is shown with detectable msL1/msL2 homologs indicated. For species with available 5.8S-25S rRNA sequences, the length and predicted secondary structure of the helix H52 are shown next to their corresponding species. Species with missing or incomplete sequences for 5.8S-25S rRNA are highlighted by asterisks, and species with experimentally defined ribosome structures are highlighted in cyan. **c)** The structures of msL1-*Vn* and msL2-*Ec* docked into the ribosomal environment of *S. cerevisae, P. locustae, V. necatrix and E. cuniculi* ribosomes in order to illustrate potential steric clashes or loss of ribosome-binding surface in one species relative to each other.