

8 Insights into the evolution of the nucleolus by an analysis of its protein domain repertoire

Die folgende Erklärung legt die individuellen Beiträge der Co-Autoren zum nachfolgenden Manuskript dar. Von den Co-Autoren unterschriebene Fassungen dieser Erklärung liegen dieser Dissertation bei. Die Erklärungen sollen belegen, dass die erzielten Resultate im wesentlichen auf meiner Arbeit beruhen und ihre Verwendung innerhalb dieser Dissertation gerechtfertigt ist.

The NUCLEOLUS manuscript: declaration about contributions of authors

Hereby I, co-author of the manuscript „*Insights into the evolution of the nucleolus by an analysis of its protein domain repertoire*“ which is accepted for publication in *BioEssays*, declare my contributions to the results and to the preparation of the manuscript. Furthermore, I agree that the statements below describe the contributions of the other co-authors correctly.

1) Eike Staub

- conducted the whole sequence analysis process,
- determined the strategy for the discovery of new domains,
- characterised the new domains by exhaustive literature searches,
- determined the distributions of domains across phyla,
- noted the link between biochemical function of the domains and their differing patterns of occurrence in species from different phyla,
- outlined the strategy how to represent the complete data in the WWW,
- interpreted the results with regard to nucleus and nucleolus evolution,
- wrote the text and prepared the figures for the manuscript,
- serves as the corresponding author during the review process.

2) Petko Fiziev

- was responsible for the visualisation of the complete data set on nucleolar protein domains and nucleolar protein domain architectures on the WWW.

2) André Rosenthal

- contributed to the final stages of manuscript preparation by helpful comments on the style of the manuscript.

3) Bernd Hinzmann

- was the supervisor of the project,

- served as a partner in discussions about the meaning of the results for the reconstruction of early eukaryotic evolution,
- contributed to the final stages of manuscript preparation by helpful comments on the style of the manuscript.

Insights into the evolution of the nucleolus by an analysis of its protein domain repertoire

Eike Staub^{*,1,2}, Petko Fiziev^{1,2}, André Rosenthal¹ and Bernd Hinemann¹

¹ metaGen Pharmaceuticals GmbH, Oudenarder Str. 16, D-13347 Berlin, Germany

² Max-Planck Institute for Molecular Genetics, Dept. of Computational Molecular Biology, Ihnestr. 73,
D-14195 Berlin, Germany

*author to whom correspondence should be addressed

tel +49-(0)30-8413 1157

fax +49-(0)30 8413 1152

eike.staub@molgen.mpg.de

keywords: nucleolus, protein domains, proteomics repeats, ribosome, RNA metabolism, nucleolar organiser region

Abstract

Recently, the first investigation of nucleoli using mass spectrometry led to the identification of 271 proteins. This represents a rich resource for a comprehensive investigation of nucleolus evolution. We applied a protocol for the identification of known and novel conserved protein domains of the nucleolus, resulting in the identification of 115 known and 91 novel domain profiles. The phyletic distribution of nucleolar protein domains in a collection of complete proteomes of selected organisms from all domains of life confirms the archaeobacterial origin of the core machinery for ribosome maturation and assembly, but also reveals substantial eubacterial and eukaryotic contributions to nucleolus evolution. We predict that in different phases of nucleolus evolution, protein domains with different biochemical functions were recruited to the nucleolus. We suggest a model for the late and continuous evolution of the nucleolus in early eukaryotes and argue against an endosymbiotic origin of the nucleolus and the nucleus.

Supplementary information

We present alignments of the novel motifs and sketches of domain compositions of hundreds of already known or novel homologues of nucleolar proteins on our website (<http://www.nucleolus.net/nucleolus/>).

Introduction

Nucleoli are membrane-less dense compartments in the nuclei of eukaryotic cells ⁽¹⁾. They are associated with regions on chromosomes that comprise arrays of ribosomal RNA (rRNA) genes, so called nucleolar organiser regions (NOR). Nucleoli are thought to be the ribosome factories of the cell. Consequently, numerous building blocks of ribosomes can be found in nucleoli, both rRNAs and proteins. However, because many steps are required to build a ribosome, all those proteins are present in nucleoli which contribute to its biogenesis. In recent years, evidence emerged that nucleoli also have other functions than the assembly of ribosomes. They were proposed to function in the assembly of the signal recognition particle, in the processing of certain mRNAs, tRNAs and small nuclear RNAs, in the maturation of telomerase, nuclear export, sequestering of gene silencers, and in the regulation of the cell cycle ⁽²⁻⁷⁾. This illustrates that the knowledge about the biological function of nucleoli is still fragmentary, despite the fact that the first nucleoli were already purified 40 years ago.

A breakthrough in nucleolus research was recently presented by Andersen and co-workers ⁽⁸⁾. They presented the first proteomic analysis of purified human nucleoli using mass spectrometry. One half of the 271 identified peptides were known proteins. Only ten percent were known to be nucleolar before. For the other known proteins their association with the nucleolus was shown for the first time. The study also revealed the nucleolar localisation of many previously uncharacterised hypothetical proteins. Some of these were confirmed to be part of the nucleolus by fluorescence microscopy after expression with a YFP tag. Given that the dynamic change in the localisation of some nucleolar proteins depends on the status of a cell, the authors did not claim that they captured all nucleolar proteins by their approach. Nevertheless, their study is a big step towards a complete inventory of the human nucleolar proteome. It provides the seeds for the identification of homologous proteins in other species and in the human proteome itself, possibly leading to the discovery of additional building blocks of the nucleolus. Moreover, it facilitates the comprehensive investigation of the evolutionary past of the nucleolus by the analysis of nucleolar sequences.

In this manuscript, we describe the results of a search for known and novel conserved motifs of nucleolar proteins using sensitive sequence analysis techniques. After the identification and analysis of known protein domains and sequence features, we isolated novel repeats and domains in previously

uncharacterised sequence fragments of nucleolar proteins. We identified homologous protein domains in the proteomes of human, mouse, fly, worm, yeast, ear cress and of diverse eubacteria and archaeobacteria which allowed us to determine the distribution of a comprehensive set of conserved nucleolar protein domains across phyla. The implications of these results for the evolution of the nucleolus and the nucleus are discussed.

Results and Discussion

Identification of 115 known and 91 novel protein domains in nucleolar proteins

On the basis of the results of Andersen et al. ⁽⁸⁾ we extracted a set of 235 nucleolar protein sequences from public databases. Our approach for the discovery of new motifs is similar to that successfully applied by Doerks et al. to identify novel protein domains in nuclear proteins ⁽⁹⁾. Our set of proteins was searched for low-complexity regions, transmembrane helices and coiled-coil regions to exclude these regions from the subsequent analysis. We localized 115 different known protein domains from the Pfam database (version 7.3) in 177 (75%) proteins of our set (see figure 1). Subsequences of known domains were cut out in the nucleolar sequences. We identified intra-molecular repeats in 21 masked protein sequences which were also excluded, but were kept for manual evaluation of the repeats. Finally, approximately 55% of the original sequence remained unmasked. Because fragments of less than 30 amino acids length are not suitable for the detection of novel domains, we discarded another 7% of the sequence. 513 protein fragments remained, representing 48% of the sequence. To reduce the redundancy in this set of sequence fragments we performed pair-wise sequence similarity searches using *BLASTP* of all fragments versus each other. Detected similarities were used as relations in a single linkage clustering of the fragments. As only one fragment per cluster was selected for the subsequent analysis, a set of 488 fragments remained which had the potential to comprise novel conserved domains.

For the detection of sequences which are homologous to our set of 488 sequence fragments we performed iterative *PSIBLAST* searches in the NCBI non-redundant protein database (nr) using an expectation (E) value of 0.001 as a threshold to include a detected database sequence into the sequence profile of the next iteration. The number of maximum iterations was restricted to eight. Further iterations are unlikely to provide new information as usually a search either converges or

becomes unspecific due to an incorporation of false-positive sequences or sequences of low complexity into the profile. Therefore, we discarded *PSIBLAST* results which did not converge within 8 rounds. Using *PSIBLAST* profiles of each fragment we searched a database containing all copies of the 177 Pfam domains in the pfamseq database. When a profile detected one of these known domain copies, we also excluded the fragment from the analysis because it is likely to be distantly related to a known domain. We automatically built alignments from successful *PSIBLAST* results. To correct misalignments, each alignment was trimmed manually and reduced to regions of sufficient sequence conservation. Alignments that only presented trivial sequence similarities were discarded. For each of the remaining 213 alignments we built profile Hidden Markov Models (HMM) which allowed us to search for the conserved domains with high sensitivity in the nrdb90 database. The visualisation of known and new domains in all identified proteins in nrdb90 facilitated the exclusion of less interesting alignments from the analysis. We excluded those motifs that exclusively occurred in direct proximity to known domains. These can simply be regarded as domain extensions. Conflicts between overlapping novel domains were resolved. By picking only those domains that were characteristic of a set of sequences, we ended up with a set of 91 new domain signatures and repeats from 89 of the 235 proteins in the original set. Using our HMMs together in combination with Pfam HMMs we redetected 210 out of the 235 original proteins compared to 177 of 235 proteins using only the Pfam HMMs. The coverage of the total sequence space of the 235 proteins with domains was raised by 10.5%. We conclude that our set of HMMs is a tool which will enhance the detection of nucleolar protein domains in uncharacterised protein sequences. For all novel domains and repeats we provided a basic annotation based on the available annotations of single family members that were already characterised (see Table 2).

The distribution of nucleolar protein domains across the kingdoms of life

To elucidate the evolutionary history of the nucleolus, we decided to search several protein sets from completely sequenced genomes for occurrences of all known and novel conserved domains that are present in the investigated nucleolar proteins. We analysed species from different branches of the tree of life: human and mouse as mammals, the worm *Caenorhabditis elegans*, the insect *Drosophila melanogaster*, the baker's yeast *Saccharomyces cerevisiae* as a unicellular eukaryote, as well as multiple archaea and eubacteria representing the major bacterial lineages. Of the

conserved protein domains that can be found in human nucleolar proteins, the largest fraction of 59 domains (58 known and one novel) can be found in a minimum of one protein in all three domains of life. 25 domains were found in archaea and eukaryotes to the exclusion of eubacteria. 13 protein domains are present in eubacteria and eukaryotes to the exclusion of archaea. The vast majority, here 109, can be detected only in eukaryotes.

These results are only meaningful if we can exclude extensive lateral gene transfer (LGT) between phyla as an explanation for the distribution patterns of these domains. A hint for such a late spread of the protein domains across phyla would be the occurrence of a single domain in only a small subset of organisms from one phylum, either eubacteria or archaea. Therefore, we applied a more stringent rule to conclude that a distinct protein domain is present in a certain domain of life. In the following, only those protein domains were discussed which allowed a clearer statement about their presence or absence in each of the phyla: archaeobacteria, eubacteria and eukaryotes. To be considered, a protein domain had to be present in a minimum of 4 different proteins from one phylum.

Each proteins domain was classified according to its phyletic distribution, thus providing information about its putative evolutionary age. Subsequently, the cellular functions of the domains that fit a particular phyletic pattern were analysed. The interpretation of the evolutionary age of the domains in combination with their cellular function allowed conclusions about the timely order in which the pre-nucleolar protein machinery could, at the earliest, have acquired certain domains and their associated cellular functions (see also box 1).

Ancient nucleolar protein domains mainly stem from ribosomal proteins or ribosome maturation factors

The fact that 59 human nucleolar protein domains can be found in all kingdoms of life indicates that a large fraction of the building blocks for nucleolar proteins was already present in the last universal common ancestor (LUCA). Ignoring the domains with less than four hits in a distinct kingdom, we yielded a set of 54 domains which can be regarded as ancient nucleolar domains. The proteins comprising these domains form the ancient core of the nucleolar protein machinery. They include the large group of protein domains from ribosomal proteins, reflecting the well recognised role of the nucleolus as the ribosome assembly factory. DEAD/DEAH box helicases are among the most abundant nucleolar proteins.

These RNA helicases are thought to unwind RNA during the assembly of diverse nucleoprotein complexes ⁽¹⁰⁾.

Diverse ancient RNA binding protein domains can be found in nucleoli. Many of these occur in ribosomal proteins, but are also present in non-ribosomal proteins. The S1 domain, named after its occurrence in the ribosomal protein S1, is a widespread RNA binding domain that can be found in a large number of other RNA-associated proteins ⁽¹¹⁾. The S4 domain is a putative RNA binding domain of diverse bacterial and eukaryotic ribosomal proteins and of RNA modifying enzymes like pseudouridine synthases and deaminases, RNA methylases, and tyrosyl-tRNA synthetases ⁽¹²⁾. The transcription antitermination protein NusG of bacteria and various ribosomal proteins like L24 have a common RNA associated domain which is named KOW after its discoverers (Kyprides, Ouzounis, Woese) ⁽¹³⁾. The PUA domain is a further putative RNA binding domain. Its name reflects its occurrence in pseudouridine synthase and archaeosine transglycosylase, but it is also present in other RNA modifying proteins like archaeosine synthases, rRNA methylases, and other families related to RNA metabolism ⁽¹²⁾. The K homology domain (KH) is defined by its similarity to the human heterogeneous nuclear ribonucleoprotein (hnRNP) K. It is an RNA binding module that is present in a wide variety of quite diverse nucleic acid-binding proteins, e.g. the prokaryotic ribosomal protein S3 ⁽¹⁴⁾.

Apart from RNA binding domains, there are other ubiquitous protein domains which are diagnostic of RNA modification functions in proteins from the nucleolus. The RTC domain is named after its presence in RNA 3'-terminal phosphate cyclases which catalyse the ATP-dependent conversion of the 3'-phosphate to the 2',3'-cyclic phosphodiester in RNA ⁽¹⁵⁾. The ribonuclease PH family signature is specific for 3'-5' exoribonucleases. Among these are ribonuclease PH which removes nucleotides from the CCA terminus of tRNA, polyribonucleotide nucleotidyltransferase (PNPase) that degrades messenger RNA starting from the 3' end, and diverse proteins of the exosome which is responsible for 3' processing of the 5.8S rRNA ^(16,17). The detection of the TruB domain reveals the base modification function of pseudouridylate synthases in nucleoli. Named after the prototype TruB which converts uracil to pseudouridine in many tRNAs, this family also comprises Cbf5p that modifies uracil in rRNA ⁽¹⁸⁾. It is reasonable to assume that these RNA binding domains, which are either enzymatic themselves or associated with other catalytic RNA modifying domains, are relicts from an ancient RNA world and constitute the oldest part of the nucleolus.

Other ancient protein domains in the nucleolus are involved in the folding of proteins, they are so called chaperones. DnaJ domains (J-domains) are associated with the hsp70 heat-shock system, a ubiquitous protein folding system ⁽¹⁹⁾. The cpn60-like proteins are proteins with homology to components of the bacterial GroEL protein folding system and which is essential for the correct folding and assembly of polypeptides into oligomeric structures ^(20,21). The presence of DnaJ-like and TCP-1/cpn60-like proteins in the nucleolus and all domains of life suggests that the original function of these chaperones was ribosome assembly.

Several other ancient protein families with diverse functions can be found in the nucleolus. A few are related to the modification of DNA structure. We detected domain signatures of subunits of topoisomerase II, including those of DNA gyrase A and of DNA gyrase B ^(22,23). Another ancient domain is the forkhead-associated domain (FHA), a phosphopeptide recognition domain found in many regulatory proteins like kinases, phosphatases, kinesins, transcription factors, RNA-binding proteins and metabolic enzymes ⁽²⁴⁾. To our knowledge, the emergence of FHA domains in genomes of archaea has not been described before. The GTP binding domain of elongation factor Tu is an ancient part of the translation machinery which is functionally linked to the nucleolus via its ribosomal association ⁽²⁵⁾. A second type of GTPase domain with prototypes in mouse MMR1 and human HSR1 is also found in the nucleolus ⁽²⁶⁾. The Metallophosphoesterase family comprises enzymes of different substrate specificity like nucleases (yeast MRE11, bacterial SbcD), phosphoserine phosphatases, nucleotidases, sphingomyelin phosphodiesterases and 2'-3' cAMP phosphodiesterases ⁽²⁷⁾. The Nol1_Nop2_Sun domain is the characteristic central domain of the proliferating cell nuclear antigen (PCNA) p120, which is encoded by the NOL1 gene and is thought to function as a RNA methylase in the nucleolus. The structural maintenance of chromosomes (SMC) proteins have recently been shown to act in processes like DNA repair, epigenetic silencing, and sister chromatid cohesion where they form ring-like structures around the chromatids. Their N- and C-terminal domains are ATPase domains which are linked by two coiled-coil hinge regions and a central globular domain. The central globular domain is the only ancient domain of this screen that has not been integrated into Pfam before ⁽²⁸⁾. The function of SMC proteins in the nucleolus is yet unknown. It is reasonable to assume that either they regulate DNA structure or transcription in the nucleolar organiser region (NOR). A similar function can be anticipated for SNF2_N domains which occur in proteins involved in transcription regulation (e.g., SNF2, STH1, brahma or MOT1) and chromatin

unwinding (e.g. ISWI) and other processes related to DNA structure regulation ⁽²⁹⁾. Thioredoxin domains are ancient domains that catalyse the oxidation and reduction of disulfide bonds, thereby facilitating protein folding ⁽³⁰⁾. The A1pp domain is a modular domain that is present in proteins like the rat macro-H2A histone protein, proteins from single strand RNA viruses and a third largely uncharacterised protein family with members in all kingdoms of life. A function in an ubiquitous cellular process was proposed. The detection of the A1pp domain in our analysis suggests that its role is associated with the nucleolus ⁽³¹⁾. ABC transporters are responsible for the active transport of small molecules across cellular membranes. Their two ATP binding subunits can either be joined to the two transmembrane domains in one protein or exist as a separate protein. We also found ABC transporter ATP binding domains ⁽³²⁾ and a Band 7 domain ⁽³³⁾ in nucleolar proteins. Their occurrences among nucleolar proteins are hard to explain. Band 7 proteins are integral membrane proteins which should not co-purify with nucleoli, suggesting that their identification during mass spectrometry is possibly an artefact.

The distribution of functional classes of ancient nucleolar protein domains shows a strong bias towards ribosomal domains and domains acting in RNA modification and binding. Recently, Anantharaman et al. provided an excellent analytical review about the enzymes of RNA metabolism, many of which can be found in the set of ancient nucleolar domains presented here ⁽³⁴⁾. The various other ancient nucleolar protein domains mostly function in the regulation of DNA structure or in protein folding, probably regulating the accessibility and transcription of ribosomal genes in nucleolar organiser regions or supporting ribosome assembly. The structural core of the ribosome, the enzymes modifying the rRNA, and those supporting the correct assembly of the ribosome apparently represent the oldest part of the human nucleolus.

Nucleolar protein domains of archaeobacterial origin function in ribosome maturation and translation

Another large fraction of nucleolar protein domains, 25 in this study, can be detected only in archaeobacteria, but not in eubacteria. In the light of the previously proposed chimeric origin of the eukaryotic genome this finding is not surprising ⁽³⁵⁾. As the nucleolus is spatially linked to the rRNA genes and therefore adapted to them, it is reasonable to assume that a considerable fraction of the ribosome factory has the same archaeobacterial origin as the eukaryotic rRNA genes. Evidence that core parts of the required RNA modification machinery were derived from an

archaeobacterial ancestor comes from several earlier studies. Recently, Omer et al. have shown that small RNAs (sRNAs) exist in archaea and that they are homologous to eukaryotic small nucleolar RNAs (snoRNAs) ⁽³⁶⁾. Single examples of archaeal homologues of nucleolar proteins have also been noted before and were confirmed by this study. Those of fibrillarin (yeast Nop1p), NOP56/NOP58 or Imp4 ^(36,37) are already known and were proposed as indicators of an archaeal origin of eukaryotic RNA processing, and even as indicators of the archaeal origin of the nucleus ⁽³⁷⁾. We found that several other protein families or domains of the nucleolus are only present in archaea to the exclusion of eubacteria. Among these are four ribosomal protein domains, characteristic extensions of the small subunit proteins S3A and S4 and the large subunit proteins L15 and L31 (Ribosomal_L15e, Ribosomal_L31e, Ribosomal_S4e, Ribosomal_S3Ae) ^(38,39). Several motifs of proteins which function in the process of translation in eukaryotes have also been found in archaea, but not in eubacteria (eIF-5a, EIF-5a_N, eIF6, eRF1_1, eRF1_2, eRF1_3) ⁽⁴⁰⁻⁴³⁾. The eIF-5a proteins are linked to the nucleolus via their functional relation to translation. The roles of these proteins in the nucleolus are not clear yet.

We are aware of the fact that archaeal and eubacterial eIF-5a proteins are likely to be homologous to eubacterial EFP proteins (alignment not shown). The rather close relationship of archaeobacterial and eukaryotic eIF-5a proteins and the distant relationship of both subfamilies to eubacterial EFPs has prevented the detection of this homology. This case illustrates the limited sensitivity of sequence searches, even when using HMMs. However, it also shows that limited sensitivity is not rendering our evolutionary interpretation invalid: Classifying the common domain of the eIF-5a and EFP families as “ancient” would have made these families uninformative with regard to the question about the origin of the eukaryotic eIF-5a proteins. However, the closer relationship of eukaryotic eIF-5a proteins with the archaeal ones can clearly be deduced from the sequences. Therefore we expect that also other hypothetical cases of undetected homology will not change the general tendency of our results.

In combination with the results for ancient protein domains, these findings on archaeobacterial sequence families support the hypothesis that the ribosome itself, many domains from the functionally related translation machinery, and the core human nucleolar machinery which includes RNA modification enzymes, stem from an archaeobacterial ancestor.

The Cbfd_nfyb_hmf family of proteins is characterised by a common domain between mammalian transcription factors of the CCAAT-binding factor (CBF) family

and archaeal histone proteins. It is probably involved in the regulation of ribosomal gene regulation in the nucleolar organiser regions ⁽⁴⁴⁾. Sm proteins are found in small nuclear ribonucleoprotein particles (snRNPs) like the spliceosomal U1, U2, U4/U6 and U5 and are also found in archaebacteria which do not have a splicing apparatus. The detection of a human Sm protein in the nucleolus points to an original role in ribosome maturation for Sm proteins ⁽⁴⁵⁾. Homologues of the archaebacterial subunit H of DNA-dependent RNA polymerases can be found in all eukaryotic RNA polymerases ⁽⁴⁶⁾. Their appearance in this analysis simply documents the transcriptional activity of ribosomal genes in NORs. This finally stresses the attractiveness of a model for the evolution of the nucleolus, in which a continuity of all aspects of ribosome generation is proposed; namely that ribosomal genes, the transcription machinery of ribosomal genes, and the machinery for maturation and assembly of the ribosome all stem from the genome of a single archaebacterial ancestor.

Nucleolar domains of eubacterial origin fulfil rather modern cellular functions

A considerable number of nucleolar protein domains are found in eubacteria and eukaryotes but not in archaea. However, they are much less abundant than the archaea-only protein domains: only 8 out of 13 of these domains fulfil our stringent criteria. Among these are again several proven or hypothetical RNA binding domains like the widespread RNA recognition motif (RRM) ⁽⁴⁷⁾, the helicase and RNase D carboxy-terminal domain (HRDC) ⁽⁴⁸⁾, the double stranded RNA binding (DsRBD/DSRM) domain ⁽¹⁴⁾ and the R3H domain, named after its conserved arginine and histidines ⁽⁴⁹⁾. In contrast to the archaeal RNA binding domains, these eubacterial RNA binding domains can not be found in enzymes which modify the bases of rRNA. Instead, they seem to be involved in more modern cellular functions related to RNA, e.g. like the regulation of splicing, the regulation of translocation of mRNAs or the control of the cell cycle. The functions of most of these eubacterial RNA binding domains in the nucleolus are not completely understood.

The 3'-5' exonuclease domain is the only eubacterial domain which is known to be catalytic. Prototypes of this domain are defined by the proofreading domain of E.coli DNA polymerase I, RNase D and Werner syndrome helicase ^(50,51). RNase D is involved in the processing of tRNA, suggesting a similar function for its nucleolar counterpart. WD40 domains are β -propeller-like protein-protein interaction domains that are present in a wide range of proteins with various roles and diverse

cellular functions ⁽⁵²⁾. The frequent occurrence of WD40 domains among eukaryotic nucleolar proteins might be a sign of an increasing tendency towards compaction that is mediated or facilitated by protein-protein interaction domains. The BRCT domain is characteristic of proteins with functional relations to eukaryotic cell cycle control ⁽⁵³⁾. They might provide a link to the timely regulation of nucleolus disassembly and reassembly during the cell cycle.

Based on the function of the mentioned eubacterial domains as interaction-mediating and regulatory components and based on the fact that they rarely are present in the key rRNA mediating enzymes, we hypothesise that these domains did not take part in the key function of the early nucleolus. Because of the limited sensitivity of sequence searches it could be possible that some of these domains are actually hidden ancient domains. Nevertheless, the eubacterial sequences would then be more closely related to their eukaryotic homologs than to their undetected archaeobacterial counterparts. We conclude that these protein domains are eubacterial contributions to nucleolus evolution that were acquired relatively late. We think that these domains have been recruited to a kind of ancestral nucleolar structure, probably of lower density than today's nucleoli, after the core rRNA modification enzymes and the core ribosome assembly machinery had evolved.

A large fraction of nucleolar protein domains evolved in eukaryotes

Most of the domains which were newly characterised in this study, precisely 80, are specific for eukaryotes. The new domains do not necessarily represent new protein folds: they are rather lineage-specific conserved sequence regions of unknown structure. Their co-occurrence with other well-characterised domains, which in many cases define large protein families, means that most of them are subfamily-specific extensions. This can be observed for the ancient family of DEAD box RNA helicases which have various different C-terminal extensions and are the most widespread class of proteins in the nucleolus. In this type of families, it is unlikely that all the different extensions represent new domain folds. The homology between different types of extensions may simply remain undetected because of a too high degree of sequence divergence. Thus, different types of extensions may well have a common structural fold or function. However, they certainly can be regarded as specific adaptations to the developing structure of the nucleolus, probably playing novel roles that had to be fulfilled during the formation of the nucleolus as a compartment and during its subsequent evolution.

Among the 115 known protein domains that occurred in the set of 235 nucleolar proteins, 29 domains were only found in eukaryotes (Tab 2). As these domains were not found in prokaryotes, and some even not in yeast, they probably represent those types of nucleolar protein domains that have evolved most recently and that are the latest acquisitions of the nucleolus. Limited sensitivity in sequence searches could have prevented the detection of some of these domains in bacteria. However, for these cases the degree of sequence divergence of eukaryotic and prokaryotic relatives must have been so high that it is disputable whether structure and function of the yet undetected relatives are still similar. With regard to the question whether the evolution of the nucleolus was dominated by archaebacterial or eubacterial influences the known and novel eukaryote-specific domains are not informative. However, the abundance of eukaryote-specific domains that occur in all eukaryotic phyla considered here suggests that large sequence parts of today's nucleoli evolved early or at least changed fast during early eukaryotic evolution.

Some of the eukaryotic domain families have undergone a dramatic increase in the number of copies per genome. For example the exclusively eukaryotic high mobility group (HMG) box ⁽⁵⁴⁾ can be found in seven yeast proteins, whereas the human genome already encodes 124 proteins with this motif.

Only four of the eukaryotic-only domains stem from the ribosome (Ribosomal_L6e, Ribosomal_L14e, Ribosomal_L22e, Ribosomal_L27e) ^(55,56), thus reflecting the ancient origin of the ribosomal proteins. There are two other eukaryotic domains which are thought to function in RNA binding. The SRP14 protein is a part of the signal recognition particle (SRP) which targets secretory proteins to the membrane of the rough endoplasmic reticulum. SRP14 is essential for RNA binding in the SRP ⁽⁵⁷⁾. Recently, the assembly of the SRP has been linked to the nucleolus ⁽⁴⁾. The D111/G-patch domain occurs in diverse eukaryotic proteins related to RNA processing. Based on associated sequence features the G-patch domain was predicted to function in mRNA splicing or polyadenylation ⁽⁵⁸⁾.

A well represented class of eukaryotic-only nucleolar domains is involved in the regulation of the compactness of DNA and in the assembly of complexes of nucleic acids and protein. Among them are the HMG box domains which are typically found in proteins that preferentially bind to distorted DNA structures ^(54,59). They function in diverse eukaryotic nucleoprotein assemblies like the signal recognition particle, the nucleolus, or the transcription initiation complex ⁽⁵⁹⁾. Poly-ADP ribose polymerases and their PARP domains, PARP-like zinc fingers and PARP regulatory regions are not present in the yeast proteome, but are abundant in multicellular

eukaryotes with ten PARP family members in humans. PARPs catalyse the DNA-dependent transfer of ADP-ribose to some DNA-binding proteins, thereby decreasing their affinity to DNA, e.g. in response to DNA damage ⁽⁶⁰⁾. In the nucleolus, PARP activity might regulate the condensation of nucleolar matter and the accessibility of nucleosomal DNA. The CHROMO (CHRromatin Organization MODifier) domain ⁽⁶¹⁾ and the CHROMO shadow domain ⁽⁶²⁾ are present in proteins that function in the regulation of chromatin condensation and gene silencing. Another identified putative chromatin regulating domain is the AT-rich interaction domain (ARID) ⁽⁶³⁾. The SAP motif (after SAF-A/B, Acinus and PIAS) motif is a possible DNA binding domain which also seems to be implicated in the organisation of chromatin ⁽⁶⁴⁾. The histone superfamily comprises the proteins of the nucleosomal core, the histones, as well as other DNA binding proteins ⁽⁶⁵⁾. The histone fold seems to be a general motif regulating the compactness of complexes between DNA and proteins. Proteins of the nucleoplasmin family are chromatin decondensation proteins and directly interact with histones, thereby regulating the structure of nucleosomes ⁽⁶⁶⁾.

The domains involved in chromatin organisation are the most abundant functional class within the group of eukaryotes-only nucleolar domains. How can this be explained? It is obvious that the emergence of chromatin during eukaryote evolution must have been a challenge for the correct assembly of ribosomes. Parallel to the evolution of gene-deactivating chromatin, the accessibility of rRNA and protein genes must have been maintained. Only then, the assembly of ribosomes, and thus protein synthesis in general, could be ensured. We hypothesise that the same machinery which regulated DNA structure in early eukaryotes, also was required for the evolution of compactness of the nucleolus. This assumption would also explain where the nucleolus life cycle has its origin and how the nucleolar structure depends on the cell cycle. A consequence of this hypothesis is that the compactness of present day nucleoli is made possible by proteins with chromatin-related functions.

Several eukaryotic protein domains with other functions can be found in the nucleolus. For some domains a role in nucleolus biology can be assumed, for others a function in the nucleolus is hard to imagine. The zinc knuckle is a zinc binding motif of the CCHC type. Besides its frequent occurrence in retroviral nucleocapsid proteins and plant transposases, it is found in a family 5'-3'-exoribonucleases of which some act as DNA strand transferases and others in nucleocytoplasmic transport of RNA ⁽⁶⁷⁻⁷⁰⁾. Based on its ssDNA and RNA-related function, a role for this domain in the nucleolus seems to be plausible. We also detected a signature of the

N-terminal domain of DNA Topoisomerase I, an enzyme that relaxes positive and negative supercoils ⁽⁷¹⁾. It is generally necessary for replication, recombination, and for transcription, here probably of ribosomal genes in the NOR. The armadillo repeat mediates protein-protein interactions and was first discovered in the *Drosophila* segment polarity gene *armadillo*, a homologue of the human nucleocytoplasmic signalling protein β -catenin. Both regulate transcription and cell division via HMG box transcription factors of the TCF/LEF family ⁽⁷²⁾. Armadillo repeats are as well present in the yeast nucleolar protein Srp1p, which is essential for the crescent shape of yeast nucleoli ⁽⁷³⁾. The IBB domain mediates the assembly of the importin complex which is required for the nuclear localisation signal-dependent import of proteins into the nucleus ⁽⁷⁴⁾. The detection of proteins acting in nuclear import is not surprising when one considers the enormous amount of rRNA and protein that has to be imported into the eukaryotic nucleus ⁽⁷⁵⁾. The C2 domain is thought to be involved in calcium-dependent phospholipid binding of protein kinase C (PKC) ⁽⁷⁶⁾. The FAT and FATC domains were first characterised by their presence in a family of large proteins with partial similarity to phosphatidylinositol kinases (PIK). Although they were called PIK-related kinases, none of them was shown to possess PIK activity, but some were shown to function as Ser/Thr kinases. Members of the FAT/FATC family include such prominent members as the Ataxia telangiectasia mutant (ATM) protein or the RAD3 protein, regulators of DNA damage response and the cell cycle ⁽⁷⁷⁾. The annexins are a protein family which is involved in cytoskeletal interactions and in the inhibition of phospholipases. They bind to phospholipids in a calcium-dependent manner ⁽⁷⁸⁾. Given the identification of different signatures of phospholipid signalling-related proteins among nucleolar proteins, it is reasonable to assume a special function of these modules in the regulation of nucleolar function or structure. The cellular function of another interesting protein family, the translationally controlled tumor proteins (TCTP), is largely unknown, although it was shown to bind tubulin and calcium. The TCTP is expressed in normal mammalian cells, but preferably in growing tumours ^(79,80) and its 3D structure shows similarity to the human chaperone protein Mss4 ⁽⁸¹⁾. Finally, we detected the C subunit of V-type ATP synthases. For their occurrence among nucleolar proteins there is no plausible explanation. An artefact in the purification process of nucleoli for mass spectrometry can not be excluded.

Conclusions

The core proteins of the eukaryotic nucleolus stem from an archaeobacterial ancestor

Nucleoli can be observed in eukaryotes but not in bacteria. On the other hand, the key function of nucleoli, ribosome biogenesis, is crucial for all living species. Their importance is stressed by the estimation that 60% of transcription in a rapidly growing yeast cell is devoted to rRNA synthesis. Generally, the process of ribosome maturation involves molecules which are not parts of the ribosome itself, as for example rRNA base modification enzymes or small guide RNAs. Because ribosome maturation seems to be essential, the core parts of the eukaryotic nucleolar machinery already must have been present in the first eukaryote and also in the last universal common ancestor (LUCA) of all presently living organisms. This requirement is reflected by the huge number of ancient protein domains in nucleolar proteins which function in the ribosome itself, in ribosome assembly or in ribosome maturation.

Some younger RNA-associated protein domains seem to have evolved after the split of archaea and eubacteria in an archaeobacterial ancestor of contemporary eukaryotes. It is widely accepted that this ancestor carried rRNA genes of an archaeobacterial type in its genome. Also the presence of homologous small nucleolar RNAs (snoRNAs) in archaeal and eukaryotic genomes has been reported⁽³⁶⁾. In this study, we found that far more homologues of human nucleolar protein domains occur in archaea and not in eubacteria than vice versa. This supports a theory which proposes an archaeobacterial origin of the nucleolus. In this theory, the archaeobacterial domains were already present in the pre-nucleolar proteins of the first eukaryote, whereas the eubacterial domains were added subsequently. The cellular functions of most archaeal domains are directly related to the ribosome or to protein translation, others to gene regulation and transcription. This suggests a common archaeal origin of the ribosomal genes, their transcription machinery, and the apparatus for maturation as well as assembly of the ribosome.

Eubacterial nucleolar protein domains were added lately in nucleolus evolution

In later phases of nucleolus evolution, some eubacterial protein domains with other RNA-related functions or with capabilities to mediate protein-protein interactions

appeared. We assume that their genes have been transferred to the nucleolus from a eubacterial genome and that they have contributed new functions in the early evolution of eukaryotes. Furthermore, the requirement to keep the ribosome assembly process efficient in a large eukaryotic cell must have been important, finally leading to a dense sub-nuclear organelle without membranous borders. In a large eukaryotic cell, all components of the ribosome assembly process had to be brought or kept in close proximity to each other. A dilution of the key components of ribosome biogenesis would have meant to generate ribosomes less efficiently. Having in mind that eukaryotic cells have become much larger than their prokaryotic ancestors, we believe that this anti-dilution effect was the major driving force in the evolution of the nucleolar machinery towards a dense sub-nuclear compartment. The nucleolar machinery had to develop the capability to retain their function in a densely-packed environment of DNA, RNA and proteins. To achieve this goal, certainly many novel functions had to be invented to fine-tune the nucleolar system. This is reflected in our results by the huge amount of known and novel eukaryotic protein domains which mediate protein-protein interactions (e.g. WD40 or armadillo repeats) or function in the packing of nucleic acids and proteins in chromatin. In parallel to the evolution of a nuclear membrane, an efficient transport system had to be invented to transport ribosomal proteins in and fabricated subunits out of the nucleus. In a growing yeast cell, each minute ~1000 ribosomal proteins have to be imported and ~25 subunits have to be exported through nuclear pores ⁽⁷⁵⁾. This would explain the detection of protein domains among nucleolar proteins that are related to the transport of proteins and RNA through nuclear pore.

The chimeric nature of the nucleolar protein domain repertoire does not support an endosymbiotic origin of the nucleus

It is currently under debate whether the nucleus has an endosymbiotic origin or has evolved gradually around the genomic DNA of an archaeal precursor cell ⁽⁸²⁻⁸⁶⁾. Our findings show the chimeric nature of an essential part of the nucleus, the nucleolus. It also revealed that not only the ribosome itself, but also the core nucleolar components involved in ribosomal RNA maturation and ribosome assembly are of archaeobacterial origin. These findings support and extend the view that those parts of the first eukaryote which relate to the processing of genomic information stem from an archaeobacterial ancestor of early eukaryotes ⁽⁸⁷⁾.

What does this mean for a hypothetical scenario in which a eubacterial endosymbiont becomes the nucleus of the first eukaryote? According to such a model, an archaeal origin of the nucleolar ribosome biogenesis machinery would mean that the ribosomal and nucleolar genes were transferred from an archaeal host genome to the eubacterial symbiont nucleus to replace the endogenous genes. Given the importance of a durable integrity of the ribosome synthesis machinery to maintain effective protein synthesis which is reflected by the enormous energy cost of ribosome synthesis ⁽⁷⁵⁾, we consider such a scenario to be highly unlikely.

Other models for nucleus evolution aim to explain the chimeric nature of the eubacterial nucleus ^(84,88). Using endosymbiosis as an explanation, either an archaeobacterial symbiont could have invaded a eubacterial host or an archaeobacterium could have invaded another archaeobacterium. Alternatively, a fusion event between an archaeobacterium and a eubacterium could have led to the chimeric nucleus. In all these models a subsequent step has to be integrated in which endosymbiosis of another eubacterium finally lead to the evolution of mitochondria. Although such models can not be fully excluded by the data of this study, several points argue against them. These models predict the existence, or eventually co-existence, of three different genomes and protein synthesis machineries in the early eukaryotes, a redundancy which hardly is an effective evolutionary strategy. Probably, successful endosymbiosis between prokaryotes depends on a favourable energy constitution of the resulting cell-hybrid, e.g. the exchange and use of each others waste metabolites to produce energy. Energetic advantages are not explained by theories that propose fusion or endosymbiosis as mechanisms leading to chimeric eukaryotic nuclear genomes. In addition, an endosymbiotic origin of the nucleus fails to explain other features of nucleus biology, e.g. the nature of the nucleus membrane (no free-living prokaryote is separated from the environment in the same manner in which the nucleus is separated from the cytoplasm) or the mode of nucleus replication (no organism is known which disintegrates its cell membrane during cell division) ^(86,89).

Recently, Martin and Müller proposed the 'hydrogen hypothesis', a more parsimonious model of early eukaryotic evolution regarding events like endosymbiosis or fusion ⁽⁹⁰⁾ (see figure 1). According to these authors, mitochondria evolved by endosymbiosis of an anaerobic hydrogen-producing heterotrophic α -proteobacterium in an autotrophic hydrogen-dependent archaeobacterium. The chimeric origin of nuclear genes could be explained by stepwise gene transfer from the symbiont to the host genome. The nuclear membrane and nucleus

substructures like nucleoli could have evolved slowly: the origin of intracellular membrane systems like the endoplasmatic reticulum and the nucleus could have been a result of an excess of membrane synthesis enzymes ⁽⁸²⁾. With regard to nucleolus evolution, the hydrogen hypothesis is consistent with an archaeal origin of the ribosome as well as an archaeal origin of the core nucleolar machinery. It can explain subsequent eubacterial contributions of nucleolar protein domains to the nucleus by gene transfer from the hydrogenosome (=mitochondrial) genome. It is compatible with the findings, that a substantial amount of nucleolus protein domains were invented after the common ancestor of eukaryotes emerged. It is not in conflict with the structure of the nuclear membrane or its disintegration during mitosis. It avoids critical steps that are energetically not favourable in a theory proposing nucleus endosymbiosis, like the maintenance of three genomes and protein synthesis machineries without a compensating advantage in energy metabolism for each cell. Thus, it seems to be a parsimonious and elegant model that is able to explain the chimeric nature of the nucleolus proposed in this study.

Text Box: remarks on the interpretation of phylogenetic profiles of protein domains with regard to cellular evolution

Consider a contemporary protein that comprises a certain domain. It is clear that the emergence of this domain in an ancestral organism is a prerequisite for the emergence of the protein during evolution: the time-point of the emergence of the domain during evolution must have preceded, or at least coincided with the time-point of the emergence of the protein. Often the protein domain is older than the protein architecture in which it is used today. The reason for this is the frequent reuse of protein domains as functional modules during evolution. Additionally, the protein domain could in principle work in completely unrelated functional contexts in those proteins where it is detected.

So which conclusions can be drawn from a single phylogenetic domain profile? One can conclude from a phylogenetic domain profile that the protein domain was already available as a potential building block of cellular structure in those ancestral organism whose descendants have the domain. One can not conclude that this protein domain was actually already used in the context of a particular cellular structure or function in that ancestral organism.

The same rules hold for the interpretation of a whole collection of protein domain profiles. Therefore, domain profiles are valuable tools to exclude that a certain cellular structure (like a biochemical pathway or the nucleolus) could have already existed at a certain timepoint during evolution. As such, they are helpful to deduce the earliest possible timepoints at which particular modules of a cellular unit, here the nucleolus, could have evolved.

We point out that similar guidelines also apply to the interpretation of phylogenetic profiles determined by other measures than protein domain absence or presence. One example is the interpretation of the presence/absence patterns of orthologous proteins from signal transduction pathways in metazoan species. Here, the uncertainty of the assignment of a clear function of a particular domain is analogous to the uncertainty about the functional meaning of the detection of an ortholog. The reasons are a) that orthologous proteins are reused during various developmental stages in a single organism and b) that different organism use the same sets of orthologous genes for the control of different developmental programs. This is illustrated by the diverse functions of the *wingless* gene of the fruit fly and its numerous metazoan orthologs: nobody would expect that the common ancestor of all organisms which have *wingless* orthologs had wings.

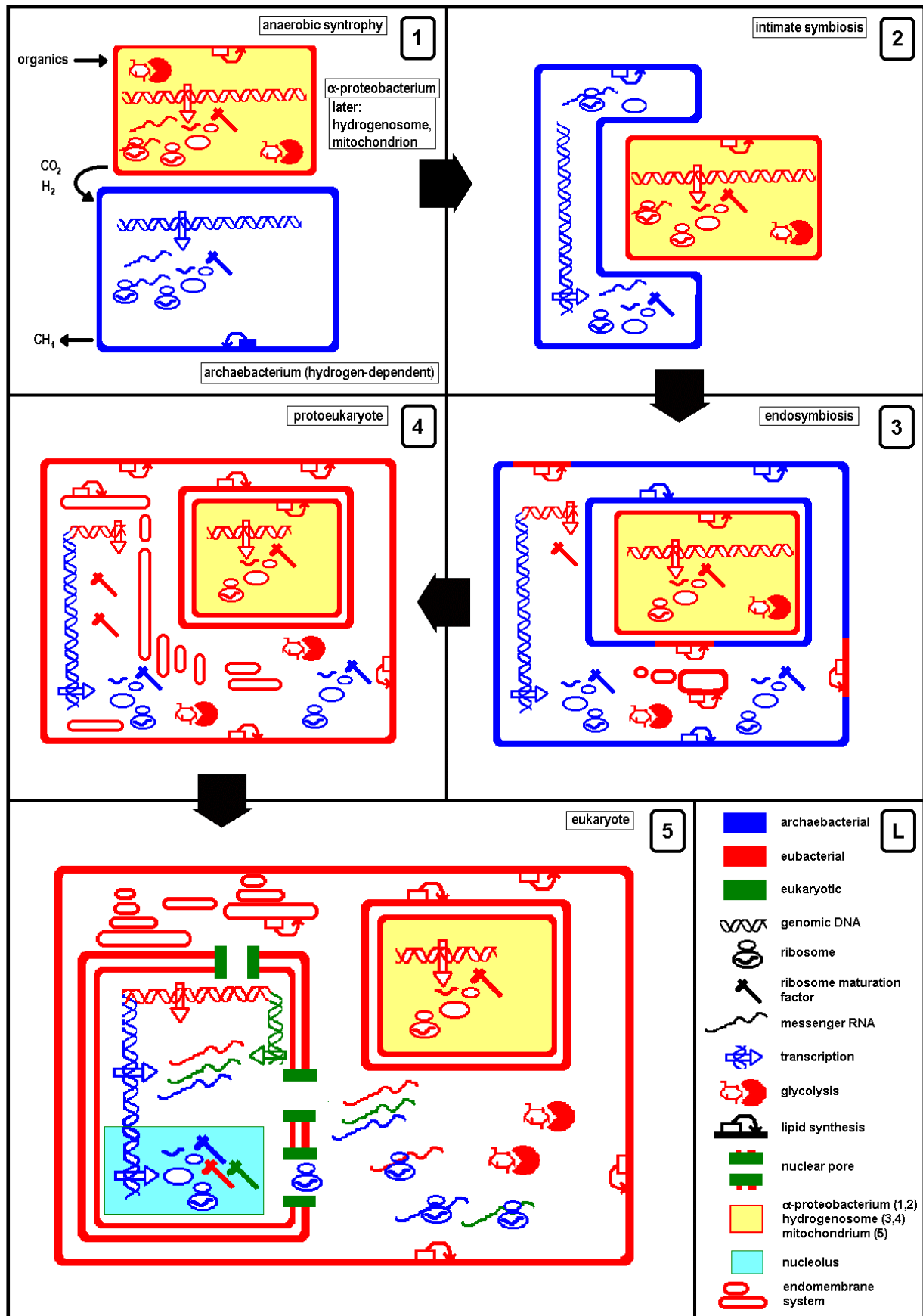


Figure 1. A possible scenario of nucleolus evolution according to the “hydrogen hypothesis for the first eukaryote”.

Figure 1.

Here we illustrate how the evolution of the nucleolus can be incorporated into the “hydrogen hypothesis for the first eukaryote”⁽⁹⁰⁾. This figure is adapted from Martin and Russel⁽⁹¹⁾. Our version of the figure accomodates our view of a late and continuous evolution of the chimeric eukaryotic nucleolus. The “hydrogen hypothesis” predicts a late emergence of the nucleus (subsequent to the emergence of the mitochondrial precursor) and is therefore well suited to explain our results. We describe the evolution of nucleolar components in five key phases leading to the evolution of the first eukaryotic cell according to Martin and Russel⁽⁹¹⁾. For completeness substantial parts of their argumentations are repeated here. (1) Anaerobic syntrophy. Because it is energetically favourable, an α -proteobacterium and an archaeobacterium share the same anaerobic environment: The (possibly facultative) anaerobic eubacterium is chemoheterotoph, can use organic molecules as energy and carbon sources and generates hydrogen as a waste product. This is willingly used by an obligate anaerobic hydrogen-dependent archaeobacterium, possibly a methanogen. At this stage both prokaryotes have their own types of ribosomes and ribosome maturation factors. No characteristic compartments for ribosome maturation in these cells exist. (2) Intimate and stable symbiosis. The hydrogen-dependent archaeobacterium tries to maximise its hydrogen consumption. Therefore it maximises its interacting surface while at the same time it has to ensure that the flow of carbon sources to the eubacterium continues, not to let the eubacterial fermentative production of hydrogen run dry. (3) Endosymbiosis. As soon as the archaeobacterial host has found a way to feed the α -proteobacterium with carbohydrates (e.g. by symbiont-to-host lateral transfer of carbohydrate transporter genes), endosymbiosis can complete. The former external symbiont becomes a hydrogenosome. It is possible that also copies of other eubacterial genes, e.g. for glycolytic enzymes, membrane synthesis or RNA metabolism were already transferred to the archaeobacterial host at this timepoint. Glycolysis probably has worked in both, the host and the symbiont cytoplasm. The enzymatic production of lipids of the eubacterial type in the archaeobacterial cytoplasm could have resulted in the production of host-incompatible lipid vesicles: the beginning of an endomembrane system that will later evolve into the endoplasmatic reticulum and the nuclear membrane. The eubacterial contributions to the future eukaryotic nucleolus could have entered the host in this phase of evolution, although it is not clear whether they were used in the context of ribosome maturation so soon after

the symbiont-to-host transfer. (4) The protoeukaryote. Symbiont-to-host gene transfer has continued. Proteins synthesised in the host cytosol can now be transferred back to the hydrogenosome, allowing for a reduction of the symbiont genome. The hydrogenosome's ability to metabolise sugars is lost on its way to become a specialised organelle. The endomembrane system has extended and lipids of the eubacterial type have replaced their archaeobacterial counterparts in all cellular membranes. The protoeukaryote cell is already substantially larger than its precursors. It still lacks a nuclear membrane and a nucleolus. For ribosome assembly the situation is suboptimal, because the components are diluted in the cytoplasm of the large protoeukaryotic cell. Therefore, the protoeukaryote is under pressure to form a "genome compartment" which serves to concentrate components that act in the assembly and regulation of large information-processing machineries (like the nucleolus or the transcription initiation complex). (5) The eukaryote. A facultative anaerobic heterotrophic cell with a mitochondrial precursor of endosymbiotic origin. Now a nuclear membrane is established, separating genome information management from the cytosol. Transcription and translation are uncoupled. Many new eukaryotic genes were already invented to regulate nuclear structure, e.g. proteins for nuclear import/export, the nuclear matrix and the reorganisation of the nuclear membrane during cell division. The genomic site of ribosomal gene transcription has now evolved into a dense subnuclear compartment by the reuse of eubacterial protein domains, the invention of new eukaryotic proteins and many new eukaryotic extensions of old proteins. It is not clear whether a dense pre-nucleolar structure evolved before or after the nuclear membrane. We suggest that the main driving force for the evolution of a densely-packed pre-nucleolar compartment was the compensation for the dilution of nucleolar components in a cell of larger volume. This dilution-effect would have been even larger in cells lacking nuclei. Thus, the start of the evolution of the ribosome assembly machinery towards a densely-packed compartment could have coincided with or even preceeded the start of nucleus evolution.

Materials and Methods

Sequence databases

During this study we used the following databases: the non-redundant protein database (nr) at the NCBI, the pfamseq database version 7, the nrdb90 database, the NCBI pdbaa database of protein sequences with solved 3D structures, the International Protein Index (IPI) databases of *Homo sapiens* and *Mus musculus* proteins, the wormpep database version 79 of *Caenorhabditis elegans* proteins, the NCBI databases yeast.aa and drosoph.aa of *Saccharomyces cerevisiae* and *Drosophila melanogaster*, the *Arabidopsis thaliana* protein set from the EBI, and protein sets from completely sequenced bacterial genomes provided by the EBI, namely those of the eubacteria *Bacillus subtilis*, *Borrelia burgdorferi*, *Brucella melitensis*, *Campylobacter jejuni*, *Caulobacter crescentus*, *Chlamydia trachomatis*, *Clostridium acetobutylicum*, *Deinococcus radiodurans*, *Escherichia coli* K12, *Haemophilus influenzae*, *Lactococcus lactis*, *Pseudomonas aeruginosa*, *Rhizobium meliloti*, *Rickettsia prowazekii*, *Salmonella typhimurium*, *Synechocystis* sp. PCC6803, *Thermotoga maritima*, *Treponema pallidum* and of the archaeobacteria *Archaeoglobus fulgidus*, *Halobacterium* sp. strain NRC-1, *Methanobacterium thermoautotrophicum*, *Methanococcus jannaschii*, *Pyrobaculum aerophilum*, *Pyrococcus abyssi*, *Pyrococcus horikoshi*, *Sulfolobus solfataricus*, *Sulfolobus tokodaii*, *Thermoplasma acidophilum*, *Thermoplasma volcanicum*.

Detection of known protein domains and other sequence features

Protein sequences were scanned for known domains and repeats using the Pfam database (version 7.3) ⁽⁹²⁾. Transmembrane helices were predicted using *TMHMM* version 2.0 ⁽⁹³⁾. For the prediction of signal peptides we used *SIGNALP V2.0* ⁽⁹⁴⁾. Sequences were investigated for the presence of coiled coils using the *COILS* algorithm ⁽⁹⁵⁾. Low-complexity regions were detected using the *SEG* program ⁽⁹⁶⁾.

Repeat analysis

The program *DOTTER* ⁽⁹⁷⁾ was used to visualise local sequence similarity when we compared sequences with themselves in order to examine them for repeats. Additionally, we refined the borders of repeat regions prior to their selection for the alignment with the help of *DOTTER*. The programs *PROSPERO* ⁽⁹⁸⁾ and *PRSS* ⁽⁹⁹⁾ from the *FASTA* program package were used to evaluate the significance of the repeats.

Sequence similarity searches, multiple alignments and phylogenetic trees

Pairwise sequence similarity searches were carried out using the gapped versions of the programs of the *BLAST* program package version 2.1.2 with default scoring schemes ⁽¹⁰⁰⁾. The *PSIBLAST* program was used to identify profiles and alignments based on single sequence queries. *PSIBLAST* profiles were stored using the -C option and applied using the -R option. Alignments were generated using *CLUSTALX* ⁽¹⁰¹⁾ and edited using *JALVIEW* by written by M. Clamp. The *hmmbuild* and *hmmcalibrate* programs of the *HMMER* package were used to construct HMMs from alignments with default options for model building with *hmmbuild* (hmmls/domain alignment) and calibration (sampled sequences: 5000; mean length 350) ⁽¹⁰²⁾. Database searches using these HMMs were carried out using the *hmmsearch* program of the same package.

Table 1**Distribution of known protein domains of the nucleolus across phyla.**

The abbreviations in columns stand for *Homo sapiens* (hs), *Mus musculus* (mm), *Caenorhabditis elegans* (ce), *Drosophila melanogaster* (dm), *Arabidopsis thaliana* (at), *Saccharomyces cerevisiae* (sc), Archaeobacterial species (ar), Eubacterial species (eu). For each domain, the number of domain copies per eukaryotic genome or per bacterial lineage is given. The domains are ordered according to their distribution in eukaryotic, archaeal and eubacterial lineages. In summary lines for each classification (e.g. “eukaryotic only” or “archaeobacterial plus eukaryotic”), the numbers of domains per class are given: the left number considers all domains for which a minimum of one copy has been found in a bacterial lineage or eukaryotic genome; the right number counts only those domains which fulfil a more stringent threshold for the conclusion that a domain occurs in a distinct lineage (see results section). These latter domains can be recognised by the use of brackets which mark distinct counts. Those counts are considered less significant for the conclusion whether a domain is present in a certain lineage.

Pfam Name	hs	mm	ce	dm	at	sc	ar	eu
Domains detected in eubacteria and eukaryotes (10/7)								
3_5_exonuclease	6	4	8	5	11	1	0	30
BRCT	39	27	29	12	14	10	0	21
GTP_CDC ^s	33	18	2	7	2	7	0	(1)
HEAT ^s	16	11	6	2	7	3	0	(1)
HRDC	6	3	3	2	3	2	0	19
Topoisomerase_I ^s	2	2	2	1	2	1	0	(3)
rrm	475	320	128	141	237	55	0	12
WD40	424	291	130	161	221	87	0	34
dsrm	43	23	14	16	18	2	0	21
R3H	10	11	3	5	2	2	0	11
Domains detected in archaea and eukaryotes (18/16)								
CBFD_NFYB_HMF	32	22	38	6	32	8	15	0
Fibrillarin	2	4	1	2	3	1	12	0
IF_tail ^s	7	10	12	2	0	0	(1)	0
IMP4	2	4	2	2	2	2	8	0
LIM ^s	107	95	40	37	11	4	(1)	0
Sm	31	26	17	16	25	16	21	0
eIF-5a	5	1	2	1	3	2	11	0
EIF-5a_N	7	2	2	1	3	2	12	0
eIF6	2	1	1	1	2	1	11	0
eRF1_1	6	2	3	3	5	2	23	0
eRF1_2	4	2	3	3	5	2	22	0
eRF1_3	4	2	3	3	5	2	23	0
RNA_pol_H	3	1	1	1	6	1	12	0
Nop	6	3	3	3	7	3	12	0
Ribosomal_L15e	14	5	1	1	2	2	12	0

Pfam Name	hs	mm	ce	dm	at	sc	ar	eu
Ribosomal_L31e	27	13	2	1	3	2	12	0
Ribosomal_S4e	16	4	1	2	3	2	12	0
Ribosomal_S3Ae	49	5	1	1	2	2	12	0
Domains detected only in eukaryotes (29)								
ARID	30	15	4	6	8	3	0	0
Armadillo_seg	79	49	9	15	70	4	0	0
C2	211	188	58	42	105	10	0	0
Chromo_shadow	12	7	4	4	0	0	0	0
FAT	11	8	4	4	4	5	0	0
FATC	13	11	7	6	4	5	0	0
G-patch	37	31	17	17	14	4	0	0
HMG_box	124	82	17	23	15	7	0	0
IBB	14	11	3	4	7	1	0	0
Nucleoplasmin	28	9	0	2	1	0	0	0
PARP	10	6	4	2	3	0	0	0
PARP_reg	5	3	4	1	3	0	0	0
PI3_PI4_kinase	35	29	13	11	9	8	0	0
Ribosomal_L6e	14	3	1	2	3	2	0	0
Ribosomal_L14e	3	3	1	1	2	2	0	0
Ribosomal_L22e	10	3	1	2	2	2	0	0
Ribosomal_L27e	8	4	1	1	3	2	0	0
SAP	32	24	7	8	8	5	0	0
SRP14	4	1	1	1	1	1	0	0
TCTP	11	2	1	1	2	1	0	0
Topoisomer_I_N	3	1	2	1	2	1	0	0
V-ATPase_C	2	4	1	3	1	1	0	0
actin	64	35	12	14	19	8	0	0
annexin	31	19	4	4	7	0	0	0
chromo	39	29	19	15	13	2	0	0
histone	73	50	74	5	46	10	0	0
ubiquitin	91	75	27	26	68	12	0	0
zf-CCHC	53	54	34	22	173	11	0	0
zf-PARP	6	2	3	1	2	0	0	0
Ancient domains detected in eu- and archaeobacteria and in eukaryotes (58/53 ^s)								
Pfam Name	hs	mm	ce	dm	at	sc	ar	eu
A1pp	9	7	1	1	4	0	9	7
ABC_tran	141	122	70	67	132	36	386	1229
ATP-synt_ab	12	7	5	10	8	4	24	70
ATP-synt_ab_C	10	7	5	10	8	4	24	38
ATP-synt_ab_N	10	7	5	10	8	4	24	52
Band_7	26	17	13	13	15	2	19	57
DEAD	147	123	78	70	116	80	126	207
DNA_gyraseB	4	2	4	1	4	1	4	32
DNA_topoisoIV	3	2	4	1	3	1	4	34
DnaJ	74	62	36	37	98	21	5	74
Exonuclease ^s	25	13	15	7	13	6	(2)	46
FHA	41	27	10	18	15	15	4	32
GTP_EFTU	75	53	29	32	37	27	85	182
GTP_EFTU_D2	50	26	19	20	26	15	57	130
GTP_EFTU_D3	53	10	8	10	8	5	17	21
HATPase_c	33	26	11	7	32	8	58	613
KH-domain	74	47	30	22	25	7	59	69

Pfam Name	hs	mm	ce	dm	at	sc	ar	eu
KOW	39	20	10	9	19	12	49	38
MMR_HSR1	18	13	10	8	28	12	29	67
Metallophos	43	36	65	34	67	22	94	145
Mov34 ^s	24	15	8	10	14	4	6	(1)
Nol1_Nop2_Sun	10	6	5	6	7	3	26	21
PHD ^s	166	135	63	60	212	17	(2)	(1)
PUA	3	3	2	2	2	4	49	14
RNA_pol_A	5	6	3	3	7	3	13	21
RNA_pol_A2	4	6	3	3	5	3	12	18
RNase_PH	8	7	7	6	9	6	19	29
RTC	4	3	1	2	0	1	10	4
Ribosomal_L10	11	3	2	2	6	3	12	19
Ribosomal_L13	19	3	2	2	6	3	12	18
Ribosomal_L2	3	2	2	3	6	3	12	19
Ribosomal_L22	29	24	3	2	5	3	12	18
Ribosomal_L3	13	6	2	5	4	2	12	19
Ribosomal_L30	33	7	1	2	5	4	11	13
Ribosomal_L4	16	4	2	2	4	3	12	18
Ribosomal_L5	2	5	2	1	6	3	12	19
Ribosomal_L5_C	2	7	2	1	4	3	12	19
Ribosomal_L6	11	7	1	2	5	3	12	19
Ribosomal_L7Ae	40	33	6	6	11	6	20	6
Ribosomal_S13	7	6	1	1	5	3	12	18
Ribosomal_S15	3	3	2	2	5	2	12	19
Ribosomal_S17	7	5	1	1	6	3	12	19
Ribosomal_S2	55	7	2	2	5	3	12	19
Ribosomal_S4	7	2	1	3	3	3	4	21
Ribosomal_S7	8	2	2	3	5	2	12	19
Ribosomal_S9	8	4	2	1	5	3	12	19
S1	8	9	5	6	14	4	35	119
S4	12	3	3	4	13	6	23	125
SMC_C	13	7	9	6	8	6	20	18
SMC_N	10	5	10	6	10	7	29	52
SNF2_N	48	42	29	19	40	21	19	25
TruB_N	3	4	1	1	2	2	12	19
cpn60_TCP1	34	18	11	14	19	11	23	26
helicase_C	181	165	93	75	144	77	107	197
ku ^s	4	3	2	2	3	2	(1)	6
pro_isomerase ^s	96	30	20	19	30	8	(1)	28
thioredo	45	32	32	30	66	10	19	57
tubulin	66	21	17	14	17	4	21	19

Table 2**Phyletic distribution and descriptions for novel protein domains of the nucleolus.**

The basic organisation of the table and the abbreviations in column headings are the same as table 1. Additionally, we proposed names for each novel domain. We also provided accession numbers (ACC) which can be used to retrieve information about the alignment of a domain and the domain architecture of all proteins of a domain family from our website (see supplement). Short descriptions of each domain are given as an initial annotation for each domain.

Acc	Proposed Name	Comment	size (aa)	representative	Pfam-seq	hs	mm	ce	dm	at	sc	ar	eu
Domains detected in archaeobacteria and eukaryotes (7/7)													
NUC001	NOSIC	central domain in Nop56/SIK1-like proteins	53	SIK1_YEAST	49	5	3	3	3	8	3	8	0
NUC002	GAR1L	characteristic domain in GAR1-like snoRNPs	61	GAR1_YEAST	20	4	3	2	2	3	2	7	0
NUC011	DKCLD	TruB_N/PUA domain associated; N-terminal domain of Dyskerin-like proteins	59	DKC1_RAT	27	1	1	1	1	1	1	10	0
NUC020	RS11NT	N-terminal domain of ribosomal S11/S17 proteins	39	RS11_MAIZE	44	3	4	1	1	3	2	12	0
NUC021	RS13NT	N-terminal domain of ribosomal S13/S15 proteins	60	RS13_MAIZE	37	5	2	1	1	2	1	12	0
NUC023	RS4NT	N-terminal domain of Ribosomal S4 / S4e proteins; associated with KOW domains	41	RS4_DROME	45	8	3	1	2	3	2	9	0
NUC168	MRACN	central domain in nucleolar proteins of the multi-copy repressor of ras (Mra) family	79	MRA1_SCHPO	16	1	1	1	1	1	1	9	0
Domains detected in eubacteria and eukaryotes (3/1)													
NUC009	PADR2 ^s	domain in poly(ADP-ribose) polymerases; associated with zf-PARP, BRCT, SAM, PARP and ankyrin repeats/domains	76	PPO2_HUMAN	35	5	3	3	1	3	0	0	(2)
NUC108	MLECT	C-terminal domain of maleless-like RNA helicase family	41	MLE_DROME	45	18	27	9	9	18	6	0	4
NUC185	DSHCT ^s	characteristic C-terminal domain of DOB1/SKI2/helY-like DEAD box helicases	202	DOB1	24	4	2	2	1	5	2	0	(2)
Ancient domains detected in eubacteria, archaeobacteria and in eukaryotes (1/1)													
NUC060	SMChinge	SMC hinge region	153	XCPC_XENLA	86	9	7	5	4	6	4	7	6

Acc	Proposed Name	Comment	size (aa)	representative	Pfam-seq	hs	mm	ce	dm	at	sc	ar	eu
Domains detected only in eukaryotes (80)													
NUC003	TCOFD	Treacher Collins-Franceschetti syndrome 1 protein tandem repeat	65	TCOF_HUMAN	3	3	1	0	0	0	1	0	0
NUC004	P68HR	characteristic repeat of p68-like RNA helicases	35	DDX5_MOUSE	7	1	1	0	0	0	0	0	0
NUC006	P120R	characteristic repeat of proliferating cell nuclear antigen P120	23	Q922K7	3	2	2	0	0	0	0	0	0
NUC007	KI67R	KI67/Chmadrin-repeat	113	KI67_HUMAN	3	3	1	0	0	0	0	0	0
NUC008	PADR1	novel domain in poly(ADP-ribose)-synthetases; located between zf-PARP domains and BRCT repeats	57	PPOL_DROME	16	3	1	1	0	2	0	0	0
NUC010	UME	characteristic domain in UVSB PI-3 kinase, MEI-41 and ESR1; associated with FAT, FATC, PI3_PI4_kinase modules	110	ESR1_YEAST	11	1	0	0	1	1	1	0	0
NUC014	ROKNT	N-terminal domain in RNP K-like proteins with KH-domains	45	ROK_MOUSE	4	4	3	0	0	0	0	0	0
NUC016	PMC2NT	N-terminal domain in 3'-5'-exonucleases with HRDC domain; putative exosome components; Polymyositis autoantigen 2	98	PMC2_HUMAN	7	2	1	1	1	0	1	0	0
NUC017	RL6NT	N-terminal domain of ribosomal L6 proteins	57	Q9HBB3	8	11	3	0	1	0	0	0	0
NUC018	RL30NT	N-terminal domain of ribosomal L30 proteins	71	RL7_MOUSE	21	18	3	1	1	4	2	0	0
NUC029	DTHCT	C-terminal domain of DNA gyrases B / topoisomerase IV / HATPase proteins	110	TP2B_HUMAN	15	3	2	0	0	0	0	0	0
NUC031	BDHCT	C-terminal domain in Bloom's syndrome DEAD helicase subfamily	41	BLM_HUMAN	4	3	1	0	0	0	0	0	0
NUC034	CHDNT	N-terminal domain in PHD/RING finger and chromo domain-associated helicases	55	CHD4_HUMAN	7	4	2	2	1	0	0	0	0
NUC036	CHDCT1	C-terminal domain A in PHD/RING finger and chromo domain-associated CHD-like helicases	120	CHD4_HUMAN	14	6	3	2	0	1	0	0	0
NUC038	CHDCT2	C-terminal domain B in PHD/RING finger and chromo domain-associated CHD-like helicases	180	CHD4_HUMAN	11	6	3	2	0	0	0	0	0
NUC045	CAFNT	N-terminal domain in family of CCR4-associated factor-like proteins with zf-CCCH and R3H domains; part of the CCR4/NOT transcription complex	136	CNO7_MOUSE	34	9	6	3	1	14	1	0	0
NUC046	PARNUCT	C-terminal domain in Poly(A)-specific ribonucleases	46	O95453	9	2	3	2	0	2	0	0	0
NUC056	IPN	domain in ILF3/p122/NF45 transcription factors; associated with dsrm repeats	154	ILF3_HUMAN	43	12	7	2	2	0	0	0	0
NUC059	NOPS	C-terminal domain of NONA and PSP1 proteins	53	SFPQ_HUMAN	20	9	4	2	1	0	0	0	0

Acc	Proposed Name	Comment	size (aa)	representative	Pfam-seq	hs	mm	ce	dm	at	sc	ar	eu
NUC062	CBFMK21	characteristic domain of CCAAT-box binding transcription factors and MAK21-like proteins; implications in ribosome biogenesis and transcription regulation	40	CBF_HUMAN	19	3	4	3	4	2	3	0	0
NUC063	zf-RNPHF	novel putative zinc-binding domain (CHHC motif) in RNP H and F; rrm repeat-associated	36	ROH1_HUMAN	6	3	4	0	0	0	0	0	0
NUC064	RBM1CTR	C-terminal repeat in RBM1-like RNA binding hnRNPs; associated with rrm repeats in the N-terminus	46	O75526	15	21	3	0	0	0	0	0	0
NUC068	PrCBPCN	central domain in Poly(rC)-binding proteins; associated with KH domain	132	PCB2_HUMAN	14	12	5	0	1	0	0	0	0
NUC069	PRO8NT	N-terminal domain in pre-mRNA splicing factors of PRO8 family	155	YLJ6_CAEEL	13	6	1	1	1	2	1	0	0
NUC071	PROCN	central domain in pre-mRNA splicing factors of PRO8 family	426	YLJ6_CAEEL	13	5	1	1	1	2	1	0	0
NUC072	PRO8CT	C-terminal domain in pre-mRNA splicing factors of PRO8 family	129	YLJ6_CAEEL	13	5	1	1	1	2	1	0	0
NUC083	DIP2CT	novel domain C-terminal to WD40 repeats in Dom34p-interacting protein 2 from yeast; role in regulation of translation	103	DIP2_YEAST	8	2	1	1	1	1	1	0	0
NUC086	BysCR	conserved region in proteins of the Bystin family; interaction with trophinin, tasin and cytokeatin; unusual occurrence in nucleolar protein	256	BYST_HUMAN	9	2	1	1	1	1	1	0	0
NUC087	NOGCT	C-terminal domain characteristic of NOG subfamily of nucleolar GTP-binding proteins	134	NOG1_TRYBB	15	3	1	1	1	2	1	0	0
NUC091	NGP1NT	N-terminal domain characteristic for subfamily of hypothetical nucleolar GTP-binding proteins similar to human NGP1	134	NGP1_HUMAN	14	2	1	2	1	1	1	0	0
NUC094	FerI	present in proteins of the Ferlin family; often central between two C2 domains	72	DYSF_HUMAN	10	9	7	1	1	0	0	0	0
NUC095	FerA	central domain A in proteins of the Ferlin family	67	DYSF_HUMAN	18	8	6	2	0	0	0	0	0
NUC096	FerB	central domain B in proteins of the Ferlin family	79	DYSF_HUMAN	18	11	7	2	0	0	0	0	0
NUC098	FerC	central domain C in proteins of the Ferlin family	120	DYSF_HUMAN	18	12	8	1	1	0	0	0	0
NUC102	TRAUB	C-terminal conserved domain of traube proteins	87	Q9JKX4	11	2	1	1	1	1	1	0	0
NUC103	NUC103/4	central domain hypothetical nucleolar proteins of novel family defined by alignments NUC103/104	156	YIJ1_YEAST	8	1	1	1	1	1	1	0	0
NUC104	NUC103/4	C-terminal domain hypothetical nucleolar proteins of novel family defined by alignments NUC103/104	149	YIJ1_YEAST	8	1	1	1	1	1	1	0	0
NUC105	MLENT	N-terminal domain of maleless-like RNA helicase family	128	MLE_DROME	6	2	0	1	1	1	1	0	0

Acc	Proposed Name	Comment	size (aa)	representative	Pfam-seq	hs	mm	ce	dm	at	sc	ar	eu
NUC109	DPOCT	central domain of proteins from DNA polymerase type V subfamily	71	DPO5_YEAST	7	2	1	0	0	0	1	0	0
NUC110	CDC5PAD	central domain between Ubox (RING-finger like) domain and WD40 repeats in spliceosome/cdc5p-associated proteins; possibly degenerated WD40 repeats	117	CWF8_SCHPO	10	2	1	1	1	2	0	0	0
NUC111	PESCNT	N-terminal domain in pescadillo-like proteins with BRCA1 C-terminus domain	139	YG2S_YEAST	10	3	3	1	1	1	1	0	0
NUC114	NUBF	N-terminal domain in UBF transcription factors; possibly degenerated HMG box	100	UBF1_MOUSE	9	4	2	0	0	0	0	0	0
NUC119	CPL	C-terminal domain in Penguin-like proteins associated with Pumilio repeats	159	PEN_DROME	3	1	1	1	1	0	1	0	0
NUC121	AARP2CN	AARP2 central domain; weakly similar to GTP-binding domain of elongation factor TU	91	Q94649	18	6	6	2	2	2	2	0	0
NUC123	AARP2CT	AARP2 family C-terminal domain	208	Q94649	19	11	6	2	2	2	2	0	0
NUC125	NUC125	central conserved domain in novel family of hypothetical proteins defined by NUC125	73	Q9Y3B9	9	1	1	1	1	1	1	0	0
NUC126	NUC126	novel family of hypothetical nucleolar proteins defined by NUC126	194	YQ52_CAEEL	12	1	1	1	1	1	1	0	0
NUC127	NOP5NT	N-terminal domain in RNA-binding proteins of the NOP5 family	68	NOP5_RAT	27	2	1	2	2	4	2	0	0
NUC129	NUC129	C-terminal domain in novel family of hypothetical nucleolar proteins defined by NUC129	63	Q9UMY1	4	1	2	0	0	0	0	0	0
NUC130	NUC130/3NT	N-terminal domain of novel nucleolar protein family defined by NUC130/133; weakly similar to TFIIF beta subunit	52	YBLE_SCHPO	8	3	1	1	1	2	1	0	0
NUC133	NUC130/3CT	C-terminal domain of novel family of nucleolar proteins defined by NUC130/133	114	YBLE_SCHPO	11	2	1	1	1	2	1	0	0
NUC135	NLE	redefined Nle domain of a family of proteins founded by fly notchless protein and yeast microtubule-associated protein YTM1; located N-terminal to WD40 repeats	71	YTM1_YEAST	13	4	2	1	2	1	1	0	0
NUC136	zf-LYAR	novel C2HC-type zinc finger in LYAR-like cell growth-regulating proteins; associated with rrm domains; present in one or two copies per protein	62	LYAR_MOUSE	8	3	12	2	7	2	1	0	0
NUC141	BING4CT	C-terminal domain in BING4 family of nucleolar WD40 repeat proteins	80	BIN4_HUMAN	12	2	1	1	1	1	1	0	0

Acc	Proposed Name	Comment	size (aa)	representative	Pfam-seq	hs	mm	ce	dm	at	sc	ar	eu
NUC142	KRSL	characteristic KR-rich domain for novel family of nucleolar proteins; SAS10 is a derepressor of silencing; LCP5 is a U3 snRNP component; domain is combined with a basic leucine zipper in one protein	69	LCP5_YEAST	12	2	2	1	2	1	1	0	0
NUC145	NNRR	central domain in NNP1/RRP1-like proteins	144	NNP1_HUMAN	10	2	3	1	1	0	1	0	0
NUC152	GUCT	C-terminal domain characteristic for RNA helicase II / Gu protein family	108	DD21_HUMAN	14	2	3	0	0	1	0	0	0
NUC153	NUC153	small domain in novel nucleolar protein family defined by NUC153	30	YG3J_YEAST	8	3	4	1	1	1	2	0	0
NUC156	NUC156	C-terminal domain in nucleolar proteins of family NUC156	151		6	1	1	1	1	1	0	0	0
NUC160	DBP10CT	characteristic C-terminal domain for Dbp10p subfamily of hypothetical RNA helicases	68	DBPA_YEAST	8	2	1	2	1	1	1	0	0
NUC161	CBFNT	N-terminal domain of CARG-binding factor A-like proteins; combined with rrm domains	76	Q98UD3	12	3	1	0	0	0	0	0	0
NUC162	RBB1NT	characteristic domain N-terminal to ARID/BRIGHT domain in DNA binding proteins of Retinoblastoma-binding protein 1 family	100	RBB1_HUMAN	4	6	2	0	1	0	0	0	0
NUC164	MAK16NT	N-terminal domain in MAK16-like proteins	139	MK16_YEAST	12	2	1	1	1	1	1	0	0
NUC167	Y112CN	central domain in nucleolar proteins of family NUC167	50	Y112_HUMAN	10	2	2	1	0	1	1	0	0
NUC169	BOP1NT	N-terminal domain in BOP1-like WD40 proteins	286	P97452	9	1	1	1	1	1	1	0	0
NUC173	NUC173	central domain of novel family of hypothetical nucleolar proteins defined by NUC173	203	Q9VYA7	8	2	2	1	1	2	1	0	0
NUC176	NOPP140CT	C-terminal domain in Nopp140-like proteins	72	Q91803	9	2	1	1	1	1	1	0	0
NUC177	TAHNT	N-terminal domain defining a novel family of nucleolar translational activator proteins with HEAT repeats	66	YAQ5_SCHPO	5	3	3	1	0	1	1	0	0
NUC188	POPLD	novel domain in family POP1-like nucleolar proteins	108	POP1_HUMAN	6	1	1	1	1	0	1	0	0
NUC189	NUC189	characteristic domain in NUC189 family of nucleolar proteins	90	Q9LFN2	9	1	2	1	1	2	2	0	0
NUC191	NUC191	domain A in the catalytic subunit of DNA-dependent protein kinases	515	PRKD_HUMAN	4	1	1	0	0	0	0	0	0
NUC194	NUC194	domain B in the catalytic subunit of DNA-dependent protein kinases	399	PRKD_HUMAN	4	1	1	0	0	0	0	0	0
NUC200	MPP10	characteristic domain in U3 snRNP mpp10-like proteins	88	MP10_YEAST	7	1	1	1	1	1	1	0	0
NUC201	NUC201	N-terminal domain in hypothetical nucleolar proteins with NUC202 tandem repeat	86	Q9DBD5	4	3	2	0	0	0	0	0	0
NUC202	NUC202	NUC202 repeat; characteristic for a novel family of nucleolar proteins	76	Q9DBD5	4	3	1	0	0	0	0	0	0

Acc	Proposed Name	Comment	size (aa)	representative	Pfam-seq	hs	mm	ce	dm	at	sc	ar	eu
NUC203	NUC203	C-terminal domain in novel family of hypothetical nucleolar WD40 repeat proteins	87	YC47_SCHPO	5	3	1	1	1	1	1	0	0
NUC205	NUC205	characteristic domain for novel family NUC205 of nucleolar proteins	44	Q9VW10	3	1	1	0	1	0	0	0	0
NUC209	BP28NT	N-terminal domain of BAP28-like nucleolar proteins	286	BP28_DROME	6	1	2	1	1	1	1	0	0
NUC211	BP28CT	C-terminal domain of BAP28-like nucleolar proteins	171	BP28_DROME	6	1	1	1	1	1	1	0	0
NUC213	NUC213	N-terminal domain in hypothetical nucleolar proteins of novel NUC213 family	36	YEV6_YEAST	14	2	4	1	1	1	1	0	0

References

1. Melese T, Xue Z. The nucleolus: an organelle formed by the act of building a ribosome. *Curr Opin Cell Biol* 1995;7(3):319-324.
2. Olson MO, Dundr M, Szebeni A. The nucleolus: an old factory with unexpected capabilities. *Trends Cell Biol* 2000;10(5):189-196.
3. Schneider R, Kadowaki T, Tartakoff AM. mRNA transport in yeast: time to reinvestigate the functions of the nucleolus. *Mol Biol Cell* 1995;6(4):357-370.
4. Politz JC, Yarovoi S, Kilroy SM, Gowda K, Zwieb C, Pederson T. Signal recognition particle components in the nucleolus. *Proc Natl Acad Sci U S A* 2000;97(1):55-60.
5. Gerbi SA, Lange TS. All small nuclear RNAs (snRNAs) of the [U4/U6.U5] Tri-snRNP localize to nucleoli; Identification of the nucleolar localization element of U6 snRNA. *Mol Biol Cell* 2002;13(9):3123-3137.
6. Mitchell JR, Wood E, Collins K. A telomerase component is defective in the human disease dyskeratosis congenita. *Nature* 1999;402(6761):551-555.
7. Bertrand E, Houser-Scott F, Kendall A, Singer RH, Engelke DR. Nucleolar localization of early tRNA processing. *Genes Dev* 1998;12(16):2463-2468.
8. Andersen JS, Lyon CE, Fox AH, Leung AK, Lam YW, Steen H, Mann M, Lamond AI. Directed proteomic analysis of the human nucleolus. *Curr Biol* 2002;12(1):1-11.
9. Doerks T, Copley RR, Schultz J, Ponting CP, Bork P. Systematic identification of novel protein domain families associated with nuclear functions. *Genome Res* 2002;12(1):47-56.
10. de la Cruz J, Kressler D, Linder P. Unwinding RNA in *Saccharomyces cerevisiae*: DEAD-box proteins and related families. *Trends Biochem Sci* 1999;24(5):192-198.
11. Bycroft M, Hubbard TJ, Proctor M, Freund SM, Murzin AG. The solution structure of the S1 RNA binding domain: a member of an ancient nucleic acid-binding fold. *Cell* 1997;88(2):235-242.
12. Aravind L, Koonin EV. Novel predicted RNA-binding domains associated with the translation machinery. *J Mol Evol* 1999;48(3):291-302.
13. Kyrpides NC, Woese CR, Ouzounis CA. KOW: a novel motif linking a bacterial transcription factor with ribosomal proteins. *Trends Biochem Sci* 1996;21(11):425-426.
14. Burd CG, Dreyfuss G. Conserved structures and diversity of functions of RNA-binding proteins. *Science* 1994;265(5172):615-621.
15. Palm GJ, Billy E, Filipowicz W, Wlodawer A. Crystal structure of RNA 3'-terminal phosphate cyclase, a ubiquitous enzyme with unusual topology. *Structure Fold Des* 2000;8(1):13-23.
16. Mitchell P, Petfalski E, Shevchenko A, Mann M, Tollervey D. The exosome: a conserved eukaryotic RNA processing complex containing multiple 3'→5' exoribonucleases. *Cell* 1997;91(4):457-466.
17. Allmang C, Mitchell P, Petfalski E, Tollervey D. Degradation of ribosomal RNA precursors by the exosome. *Nucleic Acids Res* 2000;28(8):1684-1691.

18. Lafontaine DL, Bousquet-Antonelli C, Henry Y, Caizergues-Ferrer M, Tollervey D. The box H + ACA snoRNAs carry Cbf5p, the putative rRNA pseudouridine synthase. *Genes Dev* 1998;12(4):527-537.
19. Kelley WL. The J-domain family and the recruitment of chaperone power. *Trends Biochem Sci* 1998;23(6):222-227.
20. Prasad TK, Stewart CR. cDNA clones encoding *Arabidopsis thaliana* and *Zea mays* mitochondrial chaperonin HSP60 and gene expression during seed germination and heat shock. *Plant Mol Biol* 1992;18(5):873-885.
21. Hemmingsen SM, Woolford C, van der Vies SM, Tilly K, Dennis DT, Georgopoulos CP, Hendrix RW, Ellis RJ. Homologous plant and bacterial proteins chaperone oligomeric protein assembly. *Nature* 1988;333(6171):330-334.
22. Qi Y, Pei J, Grishin NV. C-terminal domain of gyrase A is predicted to have a beta-propeller structure. *Proteins* 2002;47(3):258-264.
23. Wigley DB, Davies GJ, Dodson EJ, Maxwell A, Dodson G. Crystal structure of an N-terminal fragment of the DNA gyrase B protein. *Nature* 1991;351(6328):624-629.
24. Durocher D, Henckel J, Fersht AR, Jackson SP. The FHA domain is a modular phosphopeptide recognition motif. *Mol Cell* 1999;4(3):387-394.
25. Stark H, Rodnina MV, Rinke-Appel J, Brimacombe R, Wintermeyer W, van Heel M. Visualization of elongation factor Tu on the *Escherichia coli* ribosome. *Nature* 1997;389(6649):403-406.
26. Vernet C, Ribouchon MT, Chimini G, Pontarotti P. Structure and evolution of a member of a new subfamily of GTP-binding proteins mapping to the human MHC class I region. *Mamm Genome* 1994;5(2):100-105.
27. Aravind L, Koonin EV. Phosphoesterase domains associated with DNA polymerases of diverse origins. *Nucleic Acids Res* 1998;26(16):3746-3752.
28. Harvey SH, Krien MJ, O'Connell MJ. Structural maintenance of chromosomes (SMC) proteins, a family of conserved ATPases. *Genome Biol* 2002;3(2).
29. Goodwin GH. Isolation of cDNAs encoding chicken homologues of the yeast SNF2 and *Drosophila* Brahma proteins. *Gene* 1997;184(1):27-32.
30. Martin JL. Thioredoxin--a fold for all reasons. *Structure* 1995;3(3):245-250.
31. Martzen MR, McCraith SM, Spinelli SL, Torres FM, Fields S, Grayhack EJ, Phizicky EM. A biochemical genomics approach for identifying genes by the activity of their products. *Science* 1999;286(5442):1153-1155.
32. Hung LW, Wang IX, Nikaido K, Liu PQ, Ames GF, Kim SH. Crystal structure of the ATP-binding subunit of an ABC transporter. *Nature* 1998;396(6712):703-707.
33. Tavernarakis N, Driscoll M, Kyrpides NC. The SPFH domain: implicated in regulating targeted protein turnover in stomatins and other membrane-associated proteins. *Trends Biochem Sci* 1999;24(11):425-427.
34. Anantharaman V, Koonin EV, Aravind L. Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res* 2002;30(7):1427-1464.
35. Vellai T, Takacs K, Vida G. A new aspect to the origin and evolution of eukaryotes. *J Mol Evol* 1998;46(5):499-507.

36. Omer AD, Lowe TM, Russell AG, Ebhardt H, Eddy SR, Dennis PP. Homologs of small nucleolar RNAs in Archaea. *Science* 2000;288(5465):517-522.
37. Mayer C, Suck D, Poch O. The archaeal homolog of the Imp4 protein, a eukaryotic U3 snoRNP component. *Trends Biochem Sci* 2001;26(3):143-144.
38. Davies C, Gerstner RB, Draper DE, Ramakrishnan V, White SW. The crystal structure of ribosomal protein S4 reveals a two-domain molecule with an extensive RNA-binding surface: one domain shows structural homology to the ETS DNA-binding motif. *Embo J* 1998;17(16):4545-4558.
39. Zwickl P, Lupas A, Baumeister W. The *Thermoplasma acidophilum* rpl15 gene encodes a homologue of eukaryotic ribosomal proteins L15/YL10. *Biochem Biophys Res Commun* 1995;209(2):684-688.
40. Koonin EV. Multidomain organization of eukaryotic guanine nucleotide exchange translation initiation factor eIF-2B subunits revealed by analysis of conserved sequence motifs. *Protein Sci* 1995;4(8):1608-1617.
41. Peat TS, Newman J, Waldo GS, Berendzen J, Terwilliger TC. Structure of translation initiation factor 5A from *Pyrobaculum aerophilum* at 1.75 Å resolution. *Structure* 1998;6(9):1207-1214.
42. Si K, Maitra U. The *Saccharomyces cerevisiae* homologue of mammalian translation initiation factor 6 does not function as a translation initiation factor. *Mol Cell Biol* 1999;19(2):1416-1426.
43. Song H, Mugnier P, Das AK, Webb HM, Evans DR, Tuite MF, Hemmings BA, Barford D. The crystal structure of human eukaryotic release factor eRF1-- mechanism of stop codon recognition and peptidyl-tRNA hydrolysis. *Cell* 2000;100(3):311-321.
44. Burley SK, Xie X, Clark KL, Shu F. Histone-like transcription factors in eukaryotes. *Curr Opin Struct Biol* 1997;7(1):94-102.
45. Hermann H, Fabrizio P, Raker VA, Foulaki K, Hornig H, Brahms H, Luhrmann R. snRNP Sm proteins share two evolutionarily conserved sequence motifs which are involved in Sm protein-protein interactions. *Embo J* 1995;14(9):2076-2088.
46. Cramer P, Bushnell DA, Fu J, Gnatt AL, Maier-Davis B, Thompson NE, Burgess RR, Edwards AM, David PR, Kornberg RD. Architecture of RNA polymerase II and implications for the transcription mechanism. *Science* 2000;288(5466):640-649.
47. Birney E, Kumar S, Krainer AR. Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors. *Nucleic Acids Res* 1993;21(25):5803-5816.
48. Morozov V, Mushegian AR, Koonin EV, Bork P. A putative nucleic acid-binding domain in Bloom's and Werner's syndrome helicases. *Trends Biochem Sci* 1997;22(11):417-418.
49. Grishin NV. The R3H motif: a domain that binds single-stranded nucleic acids. *Trends Biochem Sci* 1998;23(9):329-330.
50. Moser MJ, Holley WR, Chatterjee A, Mian IS. The proofreading domain of *Escherichia coli* DNA polymerase I and other DNA and/or RNA exonuclease domains. *Nucleic Acids Res* 1997;25(24):5110-5118.
51. Gray MD, Shen JC, Kamath-Loeb AS, Blank A, Sopher BL, Martin GM, Oshima J, Loeb LA. The Werner syndrome protein is a DNA helicase. *Nat Genet* 1997;17(1):100-103.

52. Smith TF, Gaitatzes C, Saxena K, Neer EJ. The WD repeat: a common architecture for diverse functions. *Trends Biochem Sci* 1999;24(5):181-185.
53. Bork P, Hofmann K, Bucher P, Neuwald AF, Altschul SF, Koonin EV. A superfamily of conserved domains in DNA damage-responsive cell cycle checkpoint proteins. *Faseb J* 1997;11(1):68-76.
54. Bustin M, Lehn DA, Landsman D. Structural features of the HMG chromosomal proteins and their genes. *Biochim Biophys Acta* 1990;1049(3):231-243.
55. Kuwano Y, Olvera J, Wool IG. The primary structure of rat ribosomal protein L38. *Biochem Biophys Res Commun* 1991;175(2):551-555.
56. Gallagher RA, McClean PM, Malik AN. Cloning and nucleotide sequence of a full length cDNA encoding ribosomal protein L27 from human fetal kidney. *Biochim Biophys Acta* 1994;1217(3):329-332.
57. Birse DE, Kapp U, Strub K, Cusack S, Aberg A. The crystal structure of the signal recognition particle Alu RNA binding heterodimer, SRP9/14. *Embo J* 1997;16(13):3757-3766.
58. Aravind L, Koonin EV. G-patch: a new conserved domain in eukaryotic RNA-processing proteins and type D retroviral polyproteins. *Trends Biochem Sci* 1999;24(9):342-344.
59. Thomas JO, Travers AA. HMG1 and 2, and related 'architectural' DNA-binding proteins. *Trends Biochem Sci* 2001;26(3):167-174.
60. Smith S. The world according to PARP. *Trends Biochem Sci* 2001;26(3):174-179.
61. Koonin EV, Zhou S, Lucchesi JC. The chromo superfamily: new members, duplication of the chromo domain and possible role in delivering transcription regulators to chromatin. *Nucleic Acids Res* 1995;23(21):4229-4233.
62. Aasland R, Stewart AF. The chromo shadow domain, a second chromo domain in heterochromatin-binding protein 1, HP1. *Nucleic Acids Res* 1995;23(16):3168-3174.
63. Gregory SL, Kortschak RD, Kalionis B, Saint R. Characterization of the dead ringer gene identifies a novel, highly conserved family of sequence-specific DNA-binding proteins. *Mol Cell Biol* 1996;16(3):792-799.
64. Aravind L, Koonin EV. SAP - a putative DNA-binding motif involved in chromosomal organization. *Trends Biochem Sci* 2000;25(3):112-114.
65. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 1997;389(6648):251-260.
66. Ito T, Tyler JK, Bulger M, Kobayashi R, Kadonaga JT. ATP-facilitated chromatin assembly with a nucleoplasmin-like protein from *Drosophila melanogaster*. *J Biol Chem* 1996;271(40):25041-25048.
67. Zhang M, Yu L, Xin Y, Hu P, Fu Q, Yu C, Zhao S. Cloning and mapping of the XRN2 gene to human chromosome 20p11.1-p11.2. *Genomics* 1999;59(2):252-254.
68. Till DD, Linz B, Seago JE, Elgar SJ, Marujo PE, Elias ML, Arraiano CM, McClellan JA, McCarthy JE, Newbury SF. Identification and developmental expression of a 5'-3' exoribonuclease from *Drosophila melanogaster*. *Mech Dev* 1998;79(1-2):51-55.
69. Dykstra CC, Kitada K, Clark AB, Hamatake RK, Sugino A. Cloning and characterization of DST2, the gene for DNA strand transfer protein beta from *Saccharomyces cerevisiae*. *Mol Cell Biol* 1991;11(5):2583-2592.

70. Amberg DC, Goldstein AL, Cole CN. Isolation and characterization of RAT1: an essential gene of *Saccharomyces cerevisiae* required for the efficient nucleocytoplasmic trafficking of mRNA. *Genes Dev* 1992;6(7):1173-1189.
71. Redinbo MR, Stewart L, Kuhn P, Champoux JJ, Hol WG. Crystal structures of human topoisomerase I in covalent and noncovalent complexes with DNA. *Science* 1998;279(5356):1504-1513.
72. Cavallo R, Rubenstein D, Peifer M. Armadillo and dTCF: a marriage made in the nucleus. *Curr Opin Genet Dev* 1997;7(4):459-466.
73. Yano R, Oakes ML, Tabb MM, Nomura M. Yeast Srp1p has homology to armadillo/plakoglobin/beta-catenin and participates in apparently multiple nuclear functions including the maintenance of the nucleolar structure. *Proc Natl Acad Sci U S A* 1994;91(15):6880-6884.
74. Moroianu J, Blobel G, Radu A. The binding site of karyopherin alpha for karyopherin beta overlaps with a nuclear localization sequence. *Proc Natl Acad Sci U S A* 1996;93(13):6572-6576.
75. Warner JR. The economics of ribosome biosynthesis in yeast. *Trends Biochem Sci* 1999;24(11):437-440.
76. Ponting CP, Parker PJ. Extending the C2 domain family: C2s in PKCs delta, epsilon, eta, theta, phospholipases, GAPs, and perforin. *Protein Sci* 1996;5(1):162-166.
77. Bosotti R, Isacchi A, Sonnhammer EL. FAT: a novel domain in PIK-related kinases. *Trends Biochem Sci* 2000;25(5):225-227.
78. Barton GJ, Newman RH, Freemont PS, Crumpton MJ. Amino acid sequence analysis of the annexin super-gene family of proteins. *Eur J Biochem* 1991;198(3):749-760.
79. Bohm H, Benndorf R, Gaestel M, Gross B, Nurnberg P, Kraft R, Otto A, Bielka H. The growth-related protein P23 of the Ehrlich ascites tumor: translational control, cloning and primary structure. *Biochem Int* 1989;19(2):277-286.
80. Chitpatima ST, Makrides S, Bandyopadhyay R, Brawerman G. Nucleotide sequence of a major messenger RNA for a 21 kilodalton polypeptide that is under translational control in mouse tumor cells. *Nucleic Acids Res* 1988;16(5):2350.
81. Thaw P, Baxter NJ, Hounslow AM, Price C, Waltho JP, Craven CJ. Structure of TCTP reveals unexpected relationship with guanine nucleotide-free chaperones. *Nat Struct Biol* 2001;8(8):701-704.
82. Martin W. A briefly argued case that mitochondria and plastids are descendants of endosymbionts, but that the nuclear compartment is not. *Proc R Soc Lond* 1999;266:1387-1395.
83. Margulis L. Archaeal-eubacterial mergers in the origin of Eukarya: phylogenetic classification of life. *Proc Natl Acad Sci U S A* 1996;93(3):1071-1076.
84. Margulis L, Dolan MF, Guerrero R. The chimeric eukaryote: origin of the nucleus from the karyomastigont in amitochondriate protists. *Proc Natl Acad Sci U S A* 2000;97(13):6954-6959.
85. Horiike T, Hamada K, Kanaya S, Shinozawa T. Origin of eukaryotic cell nuclei by symbiosis of Archaea in Bacteria is revealed by homology-hit analysis. *Nat Cell Biol* 2001;3(2):210-214.

86. Poole A, Penny D. Does endo-symbiosis explain the origin of the nucleus? *Nat Cell Biol* 2001;3(8):E173-174.
87. Rivera MC, Jain R, Moore JE, Lake JA. Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci U S A* 1998;95(11):6239-6244.
88. Lake JA, Rivera MC. Was the nucleus the first endosymbiont? *Proc Natl Acad Sci U S A* 1994;91(8):2880-2881.
89. Rotte C, Martin W. Does endo-symbiosis explain the origin of the nucleus? *Nat Cell Biol* 2001;3(8):E173-174.
90. Martin W, Muller M. The hydrogen hypothesis for the first eukaryote. *Nature* 1998;392(6671):37-41.
91. Martin W, Russell MJ. On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. *Philos Trans R Soc Lond B Biol Sci* 2003;358(1429):59-83; discussion 83-55.
92. Bateman A, Birney E, Cerruti L, Durbin R, Ewinger L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL. The Pfam protein families database. *Nucleic Acids Res* 2002;30(1):276-280.
93. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 2001;305(3):567-580.
94. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int J Neural Syst* 1997;8(5-6):581-599.
95. Lupas A, Van Dyke M, Stock J. Predicting coiled coils from protein sequences. *Science* 1991;252(5010):1162-1164.
96. Wootton JC. Non-globular domains in protein sequences: automated segmentation using complexity measures. *Comput Chem* 1994;18(3):269-285.
97. Sonnhammer EL, Durbin R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* 1995;167(1-2):GC1-10.
98. Mott R. Accurate formula for P-values of gapped local sequence and profile alignments. *J Mol Biol* 2000;300(3):649-659.
99. Pearson WR. Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol* 2000;132:185-219.
100. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;25(17):3389-3402.
101. Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ. Multiple sequence alignment with Clustal X. *Trends Biochem Sci* 1998;23(10):403-405.
102. Eddy SR. Profile hidden Markov models. *Bioinformatics* 1998;14(9):755-763.