Novel Insights into the Evolution and Structural Characterization of Dyskerin Using Comprehensive Bioinformatics Analysis

Carolina Susana Cerrudo,*† Diego Luis Mengual Gómez,‡ Daniel Eduardo Gómez,‡§ and Pablo Daniel Ghiringhelli†§

†Laboratory of Genetic Engineering and Cellular and Molecular Biology, Department of Science and Technology, Quilmes National University, Roque Saenz Peña 352, 1876 Bernal, Buenos Aires, Argentine
‡Laboratory of Molecular Oncology, Department of Science and Technology, Quilmes National University, Roque Saenz Peña 352, 1876 Bernal, Buenos Aires, Argentine
§Supporting Information

ABSTRACT: Dyskerin is a conserved nucleolar protein. Several related genetic diseases are caused by defects in dyskerin. We hypothesized that having a comprehensive bioinformatic analysis of dyskerin will help to develop new drugs for these diseases. We predicted protein domains and compared sequences and structures to detect the universe of dyskerin-like proteins. We identified conserved features of shared domains in the three superkingdoms. We analyzed the phylogenetic diversity, confirming that there is a strong structural conservation. Also, we studied the relationship of dyskerin-like proteins with other proteins through an integrative protein–protein interaction approach. Most of them are conserved among homologous eukaryotic and archaeal proteins. Our results highlighted the preservation of proteins interacting with dyskerin. We identified conserved dyskerin interactor proteins between the different eukaryotes organisms. Furthermore, we studied the existence of dyskerin-like proteins in different species. Also, we compared and analyzed the secondary structure with the hydrophobic profile, confirming that all have hydrophilic properties highly conserved among proteins. The greatest difference was observed in the NTE and CTE regions. Another aspect studied was the comparison and analysis of tertiary structures. In our knowledge, this is the first time that these analyses were performed in such a comprehensive manner.

KEYWORDS: dyskerin-like protein, DKCLD, TruB_N, PUA, dyskerin phylogeny

INTRODUCTION

Dyskerin is a vastly conserved nucleolar protein encoded in the human genome by the DKC1 gene at Xq28 and is required for the biogenesis of ribonucleoprotein particles (RNPs) that incorporate H/ACA RNAs. The DKC1 gene is a member of the H/ACA small nucleolar ribonucleoprotein (snoRNP) gene family. Other proteins, as GAR1, NHP2, and NOP10, assist dyskerin in the interaction and processing of H/ACA-containing RNAs. The main function of Box H/ACA RNPs is to convert uridine (U) into pseudouridine (Ψ) post-transcriptionally at specific sites of rRNAs and snRNAs. Dyskerin is also needed for normal function of telomerase RNP and for telomere maintenance. Furthermore, some reports suggest a role for dyskerin in the regulation of microRNAs and in translational control of mRNA, mainly those containing internal ribosome entry site elements such as those encoding some antiapoptotic factors and tumor suppressors. Dyskerin is evolutionarily conserved in different species from the three superkingdoms. All dyskerin-like proteins present two well-characterized domains: TruB_N and PUA. Eukaryotes and archaeal dyskerin-like proteins also present another conserved domain that is not detected in bacteria: DKCLD (dyskerin-like domain). Additionally, only eukaryotic dyskerin-like proteins have N and C terminal extension (NTE and CTE). In the PDB database structures of dyskerin-like proteins corresponding to bacteria, archaea, and yeast are filed, while other eukaryotic sequences have not been solved yet. Rocchi et al. have recently created a homology model of human dyskerin by using a template crystal structure from Saccharomyces cerevisiae (PDB ID: 3U28), which has a sequence identity of 73% with the human dyskerin sequence. However, the structures of eukaryotic dyskerins remain poorly understood.

A number of related genetic diseases are caused by defects in the telomere maintenance machinery. These disorders, often called telomeropathies, share symptoms and molecular mechanisms, and mounting evidence indicates they are points along a spectrum of disease. Defects in genes involved in telomere maintenance result in a large overlapping spectrum of symptoms. The best known of these diseases is dyskeratosis...
congenital disorder (DKC), aplastic anemia, Hoyeraal–Hreidarsson syndrome, idiopathic pulmonary fibrosis, and cancer. The symptoms of these disorders are extensive, and the age of onset is highly variable; however, the disorders share similar underlying molecular mechanisms and have overlapping, incompletely penetrant phenotypes. Recently, an increasing number of reports have identified syndromes that do not include the classic symptoms but are still caused by mutations in core telomere maintenance proteins such as Revesz syndrome and Coats plus syndrome. Despite this heterogeneity it is important to go a bit deeper in the role of dyskerin in cancer due to the incidence and mortality of this disease and the fact that dyskerin has been explored for the development of drugs to selectively or preferentially kill cancer cells. A pioneer study has reported dyskerin expression to be increased in several human cancer types, especially in breast cancers. In accord with the known biological functions of the protein, breast cancers with low dyskerin expression contained lower levels of pseudouridine and telomerase RNA than those with high expression. Moreover, cancers with high expression generally exhibited worse histopathological features and prognosis. In line with these findings, DKC1 overexpression is associated with prostate cancer progression. Also, DKC1 is downregulated in sporadic chronic lymphocytic leukemia.

Eighty-five percent of the malignant tumors are immortal due to the fact that they possess an enzyme called telomerase that maintains the telomeric length. The human holoenzyme telomerase is a RNP composed of a catalytic subunit, hTERT and a RNA component (hRT), which acts as a template for the addition of a short repetitive sequence d(TTAGGG)_n in the 3' end of the telomeric DNA, and species-specific accessory proteins. These accessory proteins regulate telomerase biogenesis, subcellular localization, and function in vivo. For instance, analysis of affinity-purified telomerase from HeLa cells has identified integral protein components of human telomerase: dyskerin, NHP2, NOP10, RUVBL1(pontin)/ RUVBL2(reptin), GAR1, and TCAB1. Dyskerin, NHP2, and NOP10 are required for the stability and accumulation of hTR in vivo. The heterotrimer of dyskerin, NOP10, and NHP2 is deposited onto each hairpin unit of the H/ACA motif in a highly chaperoned biogenesis process. Cotranscriptional association of the heterotrimer is followed by an exchange of regulatory region.

A serial of causal mutations have been identified in different genes encoding telomerase components. Although not as frequent, DKC1 has a direct correlation with a great number of mutations in the DKC1 gene (Table 1). These mutations are mainly single-amino missense mutations, and many are present at the NTE and CTE of dyskerin, which are conserved between yeast and human but are missing in archaea and bacteria. Even yet, DKC1 mutations are not randomly distributed in the other conserved domains. Only two mutations have been found in the TruB N catalytic domain, one of which is located three residues from the essential aspartate required for pseudouridylation and is associated with a severe form of DKC. Mostly, DKC1 mutations are concentrated in or near to the PUA domain, and these mutations decrease the affinity for RNA binding. Therefore, we hypothesized that it is very important to have a comprehensive bioinformatics characterization of dyskerin, which will allow us to expand the knowledge of its function, evolution, and secondary and tertiary structure, opening the way to the development of new drugs to be employed in the treatment of telomeropathies. Hence, our present study focuses on function, evolution, and structure of dyskerin-like proteins.

MATERIAL AND METHODS

Sequence Search

NCBI Server (National Center for Biotechnology Information, NIH, Bethesda) was used to recover sequences corresponding to dyskerin-like proteins using the human dyskerin amino acid sequence, BLASTP program, and a profile inclusion expectation (E) value threshold of 0.01. The sequences belonging to the following species were only recovered: Homo sapiens, Rattus norvegicus, Mus musculus, Bos taurus, Gallus gallus, Danio rerio, Drosophila melanogaster, Arabidopsis thaliana, Anopheles gambiae, Aedes aegypti, Caenorhabditis elegans, Schizosaccharomyces pombe, Saccharomyces cerevisiae, Kluyveromyces lactis, Escherichia coli, Salmonella enterica, Shigella dysenteriae, Pseudomonas psychrotolerans, Pyrococcus furiosus, Pyrococcus abyssi, and Methanocaldococcus jannaschii (GenBank accession nos. NP_001354, NP_596910, NP_001025478, NP_001098865, NP_001026286, NP_001028279, NP_525120, NP_191274, XP_318082.4, XP_001656912, NP_499370, NP_594878, NP_013276, XP_453273, NP_289742, WP_000089686, WP_000089712, WP_007162646, NP_579514, WP_010867646, and AAB98132).

<table>
<thead>
<tr>
<th>gene (protein)</th>
<th>chromosomal location</th>
<th>number of mutations</th>
<th>additional references</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Telomerase</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DKC1</strong> (dyskerin)</td>
<td>Xq28</td>
<td>42&lt;sup&gt;b&lt;/sup&gt;</td>
<td>19, 20</td>
</tr>
<tr>
<td><strong>dykerin domains</strong></td>
<td>NTE</td>
<td>12</td>
<td>21–24</td>
</tr>
<tr>
<td></td>
<td>DKCLD</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>TruB</td>
<td>2</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>PUA</td>
<td>8</td>
<td>20, 21, 27</td>
</tr>
<tr>
<td></td>
<td>CTE</td>
<td>11</td>
<td>19–21, 26–29</td>
</tr>
<tr>
<td><strong>TERC</strong> (telomerase RNA component)</td>
<td>3q21-q28</td>
<td>33</td>
<td>30</td>
</tr>
<tr>
<td><strong>TERT</strong> (telomerase reverse transcriptase)</td>
<td>5p15.33</td>
<td>34</td>
<td>30</td>
</tr>
<tr>
<td><strong>NHP2</strong> (nonhistone chromosome protein 2)</td>
<td>Chr 5</td>
<td>3</td>
<td>31</td>
</tr>
<tr>
<td><strong>NOP10</strong> (nucleolar protein 10)</td>
<td>15q14-q15</td>
<td>1</td>
<td>32</td>
</tr>
<tr>
<td><strong>WRD79</strong> (TCAB1)</td>
<td>17p13.1</td>
<td>4</td>
<td>33</td>
</tr>
<tr>
<td><strong>Shelterin</strong></td>
<td>TINF2 (TIN2)</td>
<td>14q12</td>
<td>18</td>
</tr>
</tbody>
</table>

<sup>a</sup>Modified from the telomerase database [http://telomerase.asu.edu/diseases.html](http://telomerase.asu.edu/diseases.html).<sup>b</sup>Two of these mutations occur in the upstream regulatory region.

DOI: 10.1021/pr500956k

*J. Proteome Res.* 2015, 14, 874–887
Table 2. Identification of Protein Domains in Dyskerin-Like Proteins

<table>
<thead>
<tr>
<th>protein name</th>
<th>accession numbers</th>
<th>organism</th>
<th>PFAM domains</th>
<th>PROSITE domains</th>
<th>CDD domains</th>
</tr>
</thead>
<tbody>
<tr>
<td>dyskerin</td>
<td>NP_001354.1</td>
<td>Homo sapiens</td>
<td>DKCLD (PF08068.7)</td>
<td>PUA (PS0890)</td>
<td>PseudoU synth superfamily (cl0010)</td>
</tr>
<tr>
<td>H/ACA ribonucleoprotein complex subunit 4</td>
<td>NP_001025478.1</td>
<td>Mus musculus</td>
<td>PUA (PF01509.13)</td>
<td>LYS_RICH (PS0318)</td>
<td>PseudoU synth_hDyskerin (cl02572)</td>
</tr>
<tr>
<td>H/ACA ribonucleoprotein complex subunit 4</td>
<td>NP_001098865.1</td>
<td>Bos taurus</td>
<td>A7YWH5</td>
<td>PUA (PF01472.15)</td>
<td>TruB_N superfamily (cl17104)</td>
</tr>
<tr>
<td>H/ACA ribonucleoprotein complex subunit 4</td>
<td>NP_001026286.1</td>
<td>Gallus gallus</td>
<td>Q5JZK9</td>
<td>PUA (PF01472.15)</td>
<td>DKCLD superfamily (cl006987)</td>
</tr>
<tr>
<td>dyskerin</td>
<td>NP_001028279.1</td>
<td>Danio rerio</td>
<td>O4R281</td>
<td>PUA (PF01472.15)</td>
<td>PUA superfamily (cl00607)</td>
</tr>
<tr>
<td>Nop60B</td>
<td>NP_596910.1</td>
<td>Rattus norvegicus</td>
<td>PUA (PF01472.15)</td>
<td>Idem H. sapiens</td>
<td>H. sapiens</td>
</tr>
<tr>
<td>pseudouridylate synthase</td>
<td>XP_318082.4</td>
<td>Anopheles gambiae</td>
<td>PUA (PF01472.15)</td>
<td>Idem H. sapiens</td>
<td>H. sapiens</td>
</tr>
<tr>
<td>centromere/microtubule binding protein cbf5</td>
<td>XP_001656912.1</td>
<td>Aedes aegypti</td>
<td>PUA (PF01472.15)</td>
<td>Idem H. sapiens</td>
<td>H. sapiens</td>
</tr>
<tr>
<td>protein K01G5.5</td>
<td>NP_499370.1</td>
<td>Caenorhabditis elegans</td>
<td>PUA (PF01472.15)</td>
<td>Idem H. sapiens</td>
<td>H. sapiens</td>
</tr>
<tr>
<td>hypothetical protein</td>
<td>NP_594878.1</td>
<td>Schizosaccharomyces pombe</td>
<td>PUA (PF01472.15)</td>
<td>Idem H. sapiens</td>
<td>H. sapiens</td>
</tr>
<tr>
<td>Cbf5p</td>
<td>NP_013276.1</td>
<td>Saccharomyces cerevisiae</td>
<td>PUA (PF01472.15)</td>
<td>Idem H. sapiens</td>
<td>H. sapiens</td>
</tr>
<tr>
<td>hypothetical protein</td>
<td>XP_453273.1</td>
<td>Kluyveromyces lactis</td>
<td>PUA (PF01472.15)</td>
<td>Idem H. sapiens</td>
<td>H. sapiens</td>
</tr>
<tr>
<td>H/ACA ribonucleoprotein complex subunit 4</td>
<td>NP_191274.1</td>
<td>Arabidopsis thaliana</td>
<td>PUA (PF01472.15)</td>
<td>Idem H. sapiens</td>
<td>H. sapiens</td>
</tr>
<tr>
<td>tRNA pseudouridine synthase B</td>
<td>NP_289742.1</td>
<td>Escherichia coli</td>
<td>TruB_N</td>
<td>PUA (PF01509.13)</td>
<td>TruB superfamily (cl0010)</td>
</tr>
<tr>
<td>TruB</td>
<td>WP_000089686.1</td>
<td>Salmonella enterica</td>
<td>PUA (PF01509.13)</td>
<td>Idem E. coli</td>
<td>E. coli</td>
</tr>
<tr>
<td>TruB</td>
<td>WP_000089712.1</td>
<td>Shigella dysenteriae</td>
<td>PUA (PF01509.13)</td>
<td>Idem E. coli</td>
<td>E. coli</td>
</tr>
<tr>
<td>TruB</td>
<td>WP_007162646.1</td>
<td>Pseudomonas psychrotemonas</td>
<td>PUA (PF01509.13)</td>
<td>Idem E. coli</td>
<td>E. coli</td>
</tr>
<tr>
<td>Cbf5p</td>
<td>NP_579514.1</td>
<td>Pyrococcus furiosus</td>
<td>DKCLD</td>
<td>PUA (PF01509.13)</td>
<td>PUA (PF01509.13)</td>
</tr>
<tr>
<td>Cbf5p</td>
<td>WP_010867646.1</td>
<td>Pyrococcus abyssi</td>
<td>Cbf5p</td>
<td>PUA (PF01509.13)</td>
<td>PUA (PF01509.13)</td>
</tr>
</tbody>
</table>
Protein Characterization

To obtain all updated information about recovered proteins, we searched amino acid sequences against the UniProtKB/Swiss-Prot database. In addition, neXProt was also used for Homo sapiens dyskerin.

The identification of motifs, patterns, and domains in dyskerin-like proteins was performed using three known servers: Pfam, PROSITE, and CDD. Sequence Logos were performed using the WebLogo server version 2.8.2. Secondary structure predictions of regions not included in 3-D models were performed with JPRED 3 server. All servers are utilized with default parameters. Consensus hydrophobicity profiles were generated using a program designed ad hoc (Ghiringhelli, P.D.; unpublished). In brief, individual hydrophobicity profiles were obtained for each protein using a sliding windows strategy and standard hydrophobicity table. After this, gaps present in a multiple alignment of all considered proteins

Figure 1. Sequence conservation in dyskerin-like protein domains. (A) Schematic representation of dyskerin-like proteins architecture showing different protein domains present in species of the three superkingdoms. Sequence logos showing amino acid variability in the three domains: (B) DKCLD (archaea and eukaryotes), (C) TruB_N (bacteria, archaea, and eukaryotes), arrow indicate the active site, and (D) PUA (bacteria, archaea, and eukaryotes). Logos were constructed starting with alignments of each domain from different species. The height of the whole letter stack indicates the information content at that position, and the height of each letter is proportional to the relative frequency of the amino acid. Hydrophobic residues (L, V, I, W, M, F, P) are green, basic residues (R, K) are red, acidic residues (D, E) are blue, and all other residues are black. Asterisks and unfilled letters indicate the positions where human dyskerin mutations occur.
were automatically translated to each individual numeric profile. Average hydrophobicity and standard deviation values were calculated column by column. Average represents the consensus hydrophobicity profile and standard deviation represents the dispersion of hydrophobicity values in each column.

**Phylogeny**

Phylogenetic and molecular evolutionary analyses were conducted using MEGA version 5.48 with the complete dyskerin. The following parameters were used: UPGMA; Bootstrap with 1000 replicates; gap/Missing data = Pairwise deletion; Model = Amino (Dayhoff Matrix); patterns among sites = Same (Homogeneous); rates among sites = Different (Gamma Distributed); gamma parameter = 2.25.

**Ancestral Sequences Inference**

Ancestral states for eukaryotic dyskerin-like proteins were inferred using MEGA version 5.48. The following parameters were used: Method = Maximum Likelihood; Model = Dayhoff matrix based; Initial tree = precomputed tree file; Rate heterogeneity among sites = Gamma distribution using 5 gamma categories; Gaps missing/data treatment = use all sites; Branch swap filter = moderate. Pairwise alignments were performed with the ClustalW2 software using default parameters.

**Homology Modeling and Structural Alignment**

The BLAST program and *H. sapiens* dyskerin sequence (NP_001354) were used to identify putative dyskerin-like structures from Protein Data Bank (PDB). The top ranked record identified was the unique eukaryotic sequence, which has experimentally determined structure: the yeast Cbf5 structure (PDB ID: 3U28A; resolution of 1.90 Å), and this was selected for the following analysis. Thus, seven homology models were built for other selected eukaryotic proteins. Three-dimensional structures of dyskerin-like eukaryotic proteins from *Homo sapiens*, *Mus musculus*, *Gallus gallus*, *Danio rerio*, *Drosophila melanogaster*, *Anopheles gambiae*, *Aedes aegypti*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae*, *Kluyveromyces lactis*, *Escherichia coli*, *Salmonella enterica*, *Shigella dysenteriae*, *Pseudomonas psychrotolerans*, *Pyrococcus furiosus*, *Pyrococcus abyssi*, and *Methanocaldococcus jannaschii*. By comparing these recovered sequences with the UniProtKB/Swiss-Prot database, we were able to carry out a first general protein characterization. For human dyskerin we also used neXProt. Accession numbers of proteins in these databases are shown in Table 2.

Eukaryotic dyskerin have three well-characterized domains (DKCLD, TruB_N, and PUA) besides nuclear and nucleolar localization signals (Figure 1). Archaeal and bacterial pseudouridylate synthase obviously lack nuclear and nucleolar localization signals but keep the TruB_N and PUA domains. Finally, archaea pseudouridylate synthas also retain DKCLD domain.

**RESULTS**

**Domains in Dyskerin-Like Proteins**

As previously mentioned, dyskerin-like proteins are found in different species from the three superkingdoms. From this universe, we selected a set of 21 representative species to analyze the distribution of characteristic domains: *Homo sapiens*, *Rattus norvegicus*, *Mus musculus*, *Bos taurus*, *Gallus gallus*, *Danio rerio*, *Drosophila melanogaster*, *Anopheles gambiae*, *Aedes aegypti*, *Caenorhabditis elegans*, *Arabidopsis thaliana*, *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae*, *Kluyveromyces lactis*, *Escherichia coli*, *Salmonella enterica*, *Shigella dysenteriae*, *Pseudomonas psychrotolerans*, *Pyrococcus furiosus*, *Pyrococcus abyssi*, and *Methanocaldococcus jannaschii*. By comparing these recovered sequences with the UniProtKB/Swiss-Prot database, we were able to carry out a first general protein characterization. For human dyskerin we also used neXProt. Accession numbers of proteins in these databases are shown in Table 2.

Eukaryotic dyskerin have three well-characterized domains (DKCLD, TruB_N, and PUA) besides nuclear and nucleolar localization signals (Figure 1). Archaeal and bacterial pseudouridylate synthase obviously lack nuclear and nucleolar localization signals but keep the TruB_N and PUA domains. Finally, archaea pseudouridylate synthase also retain DKCLD domain.

**DKCLD** is an N-terminal domain present in 459 species, and 504 sequences were found deposited in Pfam under this name. This domain showed interactions with other three domains stored in the Pfam (Nop10p, TruB_N, and PUA). A sequence logo corresponding to this domain showed wide sequence conservation (Figure 1), especially at residues 23, 28, 31, 34, 43, 48–50, 53, and 55–58. This domain is detected by Pfam and CDD in all eukaryotic and archael studied proteins (Table 2). Six of the mutations described in human dyskerin are found in this domain.

**TruB_N family** (N terminal domain of TruB family pseudouridylate synthase) includes all eukaryotic, bacterial, and archael pseudouridylate synthases similar to human dyskerin, minify protein of Drosophila melanogaster, Trub of bacterial, and Cbf5 of archaeas and eukaryotes. These members are involved in modifying bases in RNA molecules. They carry
out the conversion of uracil bases to pseudouridine most in tRNAs and rRNA through TruB_N catalytic domain that contains the active site (Asp 29 in Figure 1C). The TruB_N domain is present in 4803 species, and 5359 sequences were found deposited under this name. This domain showed interactions with other five domains stored in the Pfam (Nop10p, PUA, TruB-C_2, TruB-C, and DKCLD). The analysis of the sequence logo corresponding to TruB_N domain (Figure 1) shows that even though this domain is present in all proteins the sequence conservation is very poor. In particular, because bacterial proteins have insertions in this domain they do not appear in the other sequences. Furthermore, this domain contains only two of the mutations previously described in human dyskerin.25 (Table 1). TruB_N domain is detected by Pfam and CDD in all analyzed proteins (Table 2).

PUA (pseudouridine synthase and archaeosine transglycosylase) domain is a highly conserved RNA-binding domain contained for many archaeal, bacterial, and eukaryotic proteins. These proteins include those involved in ribosome biogenesis and translation, enzymes that catalyze tRNA and rRNA post-transcriptional modifications, as well as in enzymes involved in proline biosynthesis, translation factors, and yeast glutamate kinases.59−61 Accordingly, PUA domain is detected in three databases (Pfam, Prosite, and CDD) in all studied sequences (Table 2). The PUA domain is also present in 4603 species, and 3369 sequences were found deposited under this name. This domain shows interactions with the other five domains stored in the Pfam (Nop10p, PUA, TruB_C_2, TruB_C, and DKCLD). The PUA domain is a small, compact domain that consists of 64−96 amino acids and presents a common RNA recognition surface, revealed by the structures of several PUA−RNA complexes.59,60 As can be seen in the sequence logo, this domain presented residues that are even conserved in bacterial sequences (Figure 1); among them we can mention the hydrophobic residues and two highly conserved positions occupied by glycines (23 and 66). As previously mentioned, PUA domain accumulates the greatest amount of mutations described in human dyskerin (Table 1).20,21,27 Of all of the proteins analyzed in this work, the dyskerin-like bacterial proteins are the ones with the lowest sequence conservation in the PUA domain.

Finally, eukaryotic dyskerin-like proteins also contain two intracellular localization signals: nuclear localization sequence (NLS) and nucleolar localization sequences (NoLSs). NLSs and NoLSs have a very similar amino acid composition with a high prevalence of basic residues in both cases. However, Scott et al. demonstrated that these two types of signals are recognized as different by the cell molecular machinery.62 These authors also classified NoLSs and NLSs signals in three
different types: NLS-only, NoLS-only, and joint NoLS–NLS. Proteins containing the joint NoLS–NLS usually are reported to have overlapping NLS and NoLS near its C-terminus, but other ones are reported to contain one of these in the N-terminus and other in the C-terminus. Eukaryotic dyskerin-like proteins belong to the latter type. These motifs are also identified as lysine-rich sequences (LYS_ RICH) by some servers (Table 2).

Secondary Structure and Hydrophobicity Conservation

As seen in Figure 2, all dyskerin-like proteins are highly hydrophilic proteins and exhibit quite similar hydrophobicity profiles (represented by the low dispersion in the graphs) when analyzed by superkingdoms. In Figure 2 it can also be seen that there is a strong conservation of secondary structure (derived from PDB structures) for each of the domains among proteins of different groups (eukaryotic, archaea, and bacteria). As previously mentioned, the eukaryotic dyskerin-like proteins have the three characteristic domains (DKCLD, TruB_N, and PUA) and also have the NTE and CTE domains. Although the last two domains are present in all eukaryotic species, these are the most variable ones. They have great length differences (NTE length varies from 16 to 56 residues and the CTE between residues 87–204), and residue conservation is not as marked as in the other domains. The secondary structure of NTE domain was predicted with the Jpred3 server because it is not in the structures stored in the PDB. Jpred3 predicts that NTE is a low-structured region, mainly loop regions, except in some proteins where α helix or a β sheet (of 4 or 3 residues) are predicted. CTE domain also has high variability in length and sequence, but the nearest region to the PUA domain has a secondary structure highly conserved in all analyzed proteins (according to the PDB and Jpred3): two α helices of 9 and 5 residues were detected.

A remarkable feature of the secondary structure of the analyzed domains is that both DKCLD and TruB_N domains share an α-helix. The terminal region of the DKCLD domain begins with one α-helix of 11 or 12 residues (depending on the species), while TruB_N domain begins having the last four residues of the same α-helix. This could be the reason why some servers predict a longest TruB_N domain (which includes the complete α helix).

Phylogeny

Dyskerin belongs to a family highly conserved from archaea to eukaryotes. Members of this family are named Cbf5p in yeast, trypanosomes, and archaea; Minily (MFL)/Nop60B in Drosophila; NAP57 in rat; and dyskerin in other mammals. In eukaryotes, proteins of this family accumulate mainly into nucleoli and are involved in several essential processes. Their high level of evolutionary conservation further testifies the importance of proteins belonging to the dyskerin family. For instance, human dyskerin and its homologue (the MFL protein) share 64.5% of identity and 85.2% of similarity. With the exception of NAF1, homologues of dyskerin (αCbf5), NOP10 (αNop10), NHP2 (L7Ae), and GAR1 (αGar1) are found in archaea, where they assemble with sRNAs that are homologous to eukaryotic H/ACA snoRNAs. Archaeal sRNAs assemble directly into active pseudouridylating sRNPs, reflecting a simpler cellular organization.

Previously, only two works reported partial phylogenetic analysis inside the dyskerin-like protein collection. Zucchini et al. analyzed the human TruB_N domain and Yokobori et al. studied only the archaean dyskerin-like proteins (Figure 3A). The tree topology obtained with complete dyskerin-like proteins matches species phylogeny. As can be seen in Figure 3, when we considered human dyskerin the proteins that are more distant phylogenetically are bacterial TruB (dyskerin-like proteins from E. coli, S. enterica, S. dysenteriae, and P. psychrotolerans). This is in agreement with previous works because although both proteins exhibit high
Structural similarity they have low sequence similarity. Moreover, bacterial proteins have TruB_N domains longer than archaea or eukaryotes, and their PUA domain is shorter with a smaller amount of residues involving loss of a-helix, which is conserved in archaea and eukaryotes. These differences could be related to the fact that PUA domain of bacterial dyskerin-like proteins makes non sequence-specific contacts with the acceptor stem of tRNA, in contrast with the PUA domain of dyskerin-like proteins in archaea and eukaryotes.

Inference of Ancestral Eukaryotic Sequences

Evolutionary relationships between eukaryotic dyskerin-like proteins can be analyzed considering the three monophyletic groups that they constitute (Figure 3A). Vertebrata, invertebrata, and plant and fungi each have a last common ancestor that we called Va, Ia, and PFa, respectively (Figure 3B). To investigate the sequence variability throughout the evolution of each monophyletic group, we carried out a reconstruction of ancestral sequences starting with extant sequences and using as a guide tree the fractions corresponding to the tree previously shown (Figure 3A).

Later, a pairwise alignment between each extant sequence and the corresponding last common ancestor was performed, and conservative changes, nonconservative changes, insertions, and deletions were analyzed. As can be seen in Figure 3B, in the three monophyletic groups, the DKCLD, TruB_N, and PUA domains have lesser variability than NTE and CTE domains. Insertions and deletions occur only in the NTE and CTE domains but ever preserving the existence of NLSs and NoLSs. Vertebrata clade globally exhibits a less divergent evolution, ranging between 76.3 and 87.7% in amino acid identity and 90.0 and 94.6% in amino acid similarity. Plant and fungi clades occupy intermediate positions having a range of 67.0 to 77.7% in amino acid identity and 82.6 to 89.7% in amino acid similarity with respect to Ia. In this clade, the dyskerin-like protein of Caenorhabditis elegans is the most divergent eukaryotic sequence, ranging between 49.0 and 57.7% in amino acid identity and 69.7 and 79.4% in amino acid similarity, whereas dyskerin-like proteins of Drosophila melanogaster, Anopheles gambiae, and Aedes aegypti have a range of 54.5 to 80.5% in amino acid identity and 76.2 to 92.5% in amino acid similarity with respect to other extant sequences analyzed.

Structural Analysis of Eukaryotic Proteins

Three-dimensional structures of dyskerin-like proteins from H. sapiens, M. musculus, G. gallus, D. rerio, D melanogaster, A. thaliana, and C. elegans were built by homology modeling using SWISS-MODEL with the yeast Cbf5 structure (PDB ID: 3U28A) as a template (Figure 4A). The predicted models quality was evaluated by the PROCHECK and PROSA-web programs (Figure 4B). PROCHECK program predicts the percentage of residues present in core, allowed, generally allowed, and disallowed regions (based on Ramachandran statistics) and the goodness factor (G factor), which indicates how typical or atypical the residue’s location is on the Ramachandran plot. Additionally, PROSA-web program checks the model quality by the PROSA Z-score of the structure, which is representative of overall model quality and is used to test whether the model structure is within the range of scores usually found in proteins of similar size. The PROCHECK analysis of all dyskerin-like protein models showed that the percentage of residues predicted in the disallowed regions was acceptable (<1.1%) and the total of residues found in the favored regions (core, allowed, and generally allowed) was always higher than 99%. The G factor value should be >-0.5 for a good model and was >-0.19 in the seven models. The PROSA Z-score for all dyskerin models has a range between −8.07 and −7.10, being the value for 3U28A structure of −7.22. PROSA tool also showed that the seven structures were within the acceptable range of X-ray and NMR studies (data not shown). Overall results obtained by PROCHECK and PROSA analyses showed good results for the seven models built by SWISS-MODEL, concluding that they were reliable for further studies. The quality of models was also assessed by comparing predicted structures to the experimentally solved structure (3U28A). By superimposition; a consensus structure was created (Figure 4A) and rmsd was calculated using MATRAS. A model can be considered accurate when its rmsd is ≤2 Å. In all pairs the obtained values were very good (Homo sapiens rmsd = 0.10 Å; Mus musculus rmsd = 0.10 Å; J. Proteome Res. 2015, 14, 874–887).
Gallus gallus \text{rmsd} = 0.16 \text{ Å}; Danio rerio \text{rmsd} = 0.09 \text{ Å};
Drosophila melanogaster \text{rmsd} = 0.10 \text{ Å};
Caenorhabditis elegans \text{rmsd} = 0.09 \text{ Å}; and
Arabidopsis thaliana \text{rmsd} = 0.09 \text{ Å}). In general, structure was more conserved than the protein sequence. Regions with higher structural and sequence conservation are shown in the consensus structure (Figure 4A) and in a multiple alignment form (Figure S1, Supporting Information). The protein region comprised between DKCLD domain and PUA domain has an overall sequence conservation of 81.1% (Figure 4C).

Protein–Protein Interactions of Eukaryotic Dyskerin-Like Proteins

As mentioned in the Introduction, dyskerin is part of the Box H/ACA RNPs but is also involved in other functions such as telomere maintenance,\textsuperscript{6,7} regulation of microRNAs,\textsuperscript{8} and in translational control of mRNA.\textsuperscript{7,8} This supports the concept that dyskerin has dynamic interactions with associated proteins, instead of forming a static complex with a fixed number of members. The involvement of dyskerin in such a variety of processes has led to an enormous research interest in this protein. In recent years, several techniques have been employed to identify complex of proteins that could lead to a better understanding of the protein function. Therefore, an integrative protein–protein interaction (PPI) approach could be useful to reveal the complexity of dyskerin interactions.

It should be stressed that not all dyskerin-like proteins described in literature are well-represented in single protein databases such as IntAct, MINT, GeneMania, and STRING. Here we present a network model of eukaryotic dyskerin interactions based on information extracted from diverse PPI databases (MINT, IntAct, GeneMania, and STRING) using PSICQUICUniversal plug-in in Cytoscape. After integrating all components from database searches, the merged protein interactions network (Figure 5) contains 189 nodes and 5818 edges (representing different kinds of interactions: interacts with in same component; reacts with state change). Some of the main topological concepts of PPI are node degree, that is the number of interacting partners per protein; clustering coefficient, which shows the ratio between the number of neighbors of node that interact with each other and the maximum number of possible interactions between neighboring nodes; and many centrality measures, involving structural attribute nodes that show the importance of them in the network. Stress centrality is an attribute that counts the number of shortest paths passing through a node. Betweenness centrality of the node reflects the amount of control that this node exerts over the interactions of other nodes in the network and favors nodes that join communities. High values of these two parameters (Stress and Betweenness centrality) indicate that the protein has great influence over information flows in the whole network.\textsuperscript{66,67}

Thus, we analyzed the topological features and attributes with the Network Analysis plug-in in Cytoscape and observed that all nodes in this network are interconnected and the average node degree of the extended network is 124.3, and the network clustering coefficient (average of clustering coefficients for all nodes) was 0.84. For dyskerin, clustering coefficient was 0.15 and node degree was 58. Betweenness and stress centrality

**Figure 5.** Eukaryotic dyskerin-like proteins interaction network. (A) Interaction network directly linked to dyskerin. Node size represents the "Betweenness centrality"; Node colors the "Stress centrality" (both of which indicate node importance in the network), Edge size the "Normalized max weight Edge Size", and Edge colors the interaction type: interacts with (light blue edges); in_same_component (gray edges); reacts with (blue edges); state_change (red edges). Networks obtained by STRING for the different eukaryotic organism are also shown: (B) Vertebrata, (C) Invertebrata, (D) Plant and Fungi. The thickness of the lines represents the strength of the association (confidence view). For each node we provided the canonical name and the original STRING name.
values of dyskerin (0.13 and 13 104, respectively) were in the range of SMG5 (0.15 and 18 978), while EST1A (0.37 and 32 114) and TERT (0.26 and 25 724) showed the highest values. All measures indicated significant positive biases for dyskerin, SMG5, EST1A, and TERT (i.e., more betweenness and stress centrality), showing for this proteins higher connectivity and high frequent participation in protein complexes.

ClueGO analysis results showed that dyskerin interactors are clustered into four main groups according to their biological processes and molecular functions: telomere maintenance, double-stranded telomeric DNA binding, Cajal body (CB), and nuclear-transcribed mRNA catabolic process, with the bulk of the direct interactors belonging to the first two groups. CBs are suborganelles found in the nucleus of proliferative cells like embryonic cells and tumor cells. In contrast with cytoplasmic organelles, CBs lack any phospholipid membrane that would separate their content, mainly consisting of proteins and RNA, from the surrounding nucleoplasm. These suborganelles have been implicated in RNA-related metabolic processes such as snRNPs biogenesis, maturation and recycling, histone mRNA processing, and telomere maintenance. Experimental evidence indicates that CBs contribute to the biogenesis of telomerase enzyme. Two of the most important components of telomerase (hTERT and hTR) accumulate within CBs in all cells in which telomerase is active but not in primary or ALT cells (where little or no hTERT is present). CBs are involved in the assembly and function of human telomerase. Additionally, we evaluated coexpression of DKC1 and other proteins at Xq28 using STRING and GeneMania servers (data not shown). This search allowed us to detect coexpression of DKC1 and 48 proteins located at the Xq28. However, we have not found specific publications where interaction between DKC1 and any of them was validated.

Then, we constructed seven additional protein interactions networks using the STRING server, which is a database of known and predicted protein interactions based on the sources derived from the genomic context, high-throughput experiments, and previous knowledge. There are bioinformatics tools (algorithms) that allow predicting PPIs, such as gene neighborhood, gene fusion, phylogenetic profile, and interologs. In the last approach, the associations between two proteins are predicted when they have homologous proteins in other species, and it has been experimentally shown to interact. Thus, we analyzed the undirected protein–protein networks from seven eukaryotic organisms to predict the existence of homologous proteins that interact with dyskerin and evaluated if besides conservation in sequence and structure of proteins there was also a conservation of interactors somehow linked to dyskerin. Figure 5 shows the networks of interactions with a confidence score higher than 0.7, and no more than 10 interactors were shown for chart simplicity. We also performed the same analysis allowing up to 50 interactors and analyzing the interactions with dyskerin that were supported on a confidence score greater than 0.8. The analysis of these results allowed us to infer a group of stably preserved interactors in the eukaryotic organisms studied: GAR1, NAF1, NH2P, NOP56, NOP58, and RUVBL1 (interactions between these proteins and dyskerin were directly found with the STRING server); RUVBL2, FBL, and NOP10 (which have no homologues in G. Gallus); and SHQ1 (for which we did not find a homologue in A. thaliana). Furthermore, KRR1, NOP10, and TERT proteins were present on almost all networks inferred with STRING. Then, we proceeded to find their homologous proteins, being able to detect KRR1 in A. thaliana (AAP64563, expected value: $7 \times 10^{-105}$ and NOP10 in D. rerio (NP_001003868, expected value: $3 \times 10^{-34}$). Finally, TERT homologues in S. cerevisiae (EST2, CAL36058) and D. melanogaster (RT_like, AAR86959) were detected with a very low E value ($4 \times 10^{-17}$ in the first case and 0.012 in the second).

### DISCUSSION

Human dyskerin is a conserved nucleolar protein involved in at least three basic processes: maintenance of telomere integrity, biogenesis and function of the ribosome, and pseudouridylation of various cellular RNAs. Dyskerin is widely studied because its mutations are implicated in DKC, cancer, and other pathological disorders. Therefore, we hypothesized that having a comprehensive bioinformatic characterization of this protein will help us to develop new drugs for treatment of those diseases.

We started with a sequence characterization, recovering the dyskerin-like protein sequences of 20 one different representative species from eukaryotes, archaea, and bacteria. The comprehensive analysis of these sequences allowed us to identify conserved features of shared domains in the three superkingdoms (TruB_N and PUA domains) and shared domains in archaea and eukaryotes (DKCLD). Although present in eukaryotes, archaea, and bacteria, the domain with the lowest sequence conservation is TruB_N. Both DKCLD and PUA domains exhibited the greater sequence conservation, with PUA domain even presenting residues that are conserved in bacteria.

Then, we compared and analyzed the secondary structure with the hydrophobic profile of dyskerin-like proteins. Almost all of them have highly conserved hydrophilic properties between proteins of different superkingdoms. The domain having the greater divergence in the hydrophobic profile was TruB_N, the one with the lower sequence conservation. The greatest difference among all analyzed dyskerin-like proteins was observed in the NTE and CTE regions. These two regions are present only in eukaryotes, and even among them there are differences in length and sequence. Regarding this aspect, all modifications annotated so far by high-throughput mass spectrometry (MS) experiments reside at the NTE and CTE regions. We should mention the elimination of initiator methionine and N-acetylation of residue 2, phosphorylation of serines 21, 451, 453, 455, 485, 494, 513 and of threonine 458, and sumoylation of lysines 16, 39, 46, and 448 evolutionarily preserved at variable degree. Conservation analysis of phosphorylated residues is weakened by the absence of known effector kinases. For example, even if no specific function has been attributed so far to any phosphorylation event, it is interesting to note that three of them (at Ser 21, 485, and 494) are reported to occur specifically at mitosis and thus might be related to cell proliferation. Moreover, in the NTE region of human dyskerin, several pathogenic mutations (Table 1) and two post-translational modifications (N-acetylation and phosphorylation) were described and experimentally determined. One of the mutations (A2V) occurs in the alanine that undergoes post-translational modification (N-acetylation), and the clinical presentation is recessive DKC. It is also very important to take into account the present information about sumoylation, which could be relevant for both localization and assembly of the protein. Modification of dyskerin by SUMO1 and SUMO2/3, first revealed by high-throughput MS experi-
ments at four residues, provided a direct correlation between pathologic mutations and sumoylation defects. Not only was the lack of sumoylation on residues mutated in X-DC patients shown to reduce protein stability, but also it was found that the K39R mutation could be rescued by a chimeric SUMO3-dyskerin product in which this modification was mimicked, demonstrating its fundamental effect on dyskerin functionality.

All of these findings indicate that the NTE region may have functional relevance and could play a role in the regulation of dyskerin, modulating its activity. Predicted secondary structures in the NTE region of eukaryotic dyskerin-like proteins showed a disordered region (only in a few sequences is the existence of an α-helix predicted), while for the CTE region the presence of two conserved α-helix motifs in the same position is revealed. The functional relevance of CTE region is related to the existence of bipartite NLS. Initially, it has been described that dyskerin localized in the nucleoplasm and then translocates to the nucleolus and to the Cbs. However, after a detailed bioinformatics and molecular analysis of the DKC1 transcriptional activity, Angrisani et al. were able to show that DKC1 encodes a new alternatively spliced mRNA (isoform 3) able to direct the synthesis of a variant dyskerin with unexpected cytoplasmic localization that may have biological functions in cell growth and adhesion.

Intriguingly, when overexpressed in HeLa cells, the new isoform promotes cell-to-cell and cell-to-substratum adhesion, increases the cell proliferation rate, and leads to cytokeratin hyperexpression. In addition, the other three alternative dyskerin isoforms, lacking the CTE region and PUA domains, have been described.

Another aspect studied of dyskerin-like proteins was the comparison and analysis of tertiary structures. We built seven models of eukaryotic dyskerin-like proteins by homology modeling using SWISS-MODEL and the tertiary structure stored in the PDB corresponding to yeast dyskerin (3U28A) as a template. Then, the seven models were superimposed to 3U28A using MATRAS program, and this analysis allowed us to confirm that there is a strong structural conservation beyond the sequence differences, mainly in the PUA domain. The function of the PUA domain seems to be that of anchoring in the nucleoplasm and then translocates to the nucleolus and to the Cbs. However, after a detailed bioinformatics and molecular analysis of the DKC1 transcriptional activity, Angrisani et al. were able to show that DKC1 encodes a new alternatively spliced mRNA (isoform 3) able to direct the synthesis of a variant dyskerin with unexpected cytoplasmic localization that may have biological functions in cell growth and adhesion. Intriguingly, when overexpressed in HeLa cells, the new isoform promotes cell-to-cell and cell-to-substratum adhesion, increases the cell proliferation rate, and leads to cytokeratin hyperexpression. In addition, the other three alternative dyskerin isoforms, lacking the CTE region and PUA domains, have been described.

Also, we demonstrated through a phylogenetic study of the complete dyskerin-like proteins its wide distribution and conservation. The existence of dyskerin-like proteins in different species and the presence of highly conserved regions indicate the preservation of the main biological functions of the dyskerin-like proteins. In addition, we explored the evolutionary history of each eukaryotic dyskerin-like protein through the reconstruction of ancestral sequences. Vertebrata, invertebrata, and plant and fungi constitute individual monophyletic clades, and each has a last common ancestor. In the vertebrata clade, C. gallus and R. norvegicus dyskerin-like proteins are the extant sequences that have required the fewest and the most mutational steps to be derived from the last common ancestor, respectively. In the invertebrata clade, C. elegans is the more divergent extant sequence derived from the last common ancestor, requiring between 2.4 and 4 times more mutational steps than the other invertebrates. While in plant and fungi clade the quantity of mutational steps from the last common ancestor are the most homogeneous.

Finally, because dyskerin is involved in a variety of biological processes we performed an integrative protein–protein interaction approach. Most PPI are conserved among homologous eukaryotic and archael proteins, although some adaptive changes have occurred in eukaryotes; for instance, human dyskerin contains additional terminal extensions not found in aCbf5. Only interactions described in eukaryotic organisms were used for the analysis of PPI; the network graphically shows the involvement of dyskerin in different complexes. According to the literature, our PPI results showed that dyskerin interacts with proteins involved in a variety of processes such as telomere maintenance, double-stranded telomeric DNA binding, Cajal body binding, and nuclear-transcribed mRNA catalytic process. Our results also highlight the preservation of proteins interacting with dyskerin because most of them are present in the different analyzed networks. Also, we identified conserved dyskerin interactor proteins among the different eukaryotes organisms analyzed using ortholog-based and comparing PPI network approaches. Likewise, we found that current PPI data of diverse organisms leads to complementary predictions.

**CONCLUSIONS**

As a general conclusion we carried out a novel bioinformatics insight into dyskerin that allowed us to describe the phylogeny of dyskerin-like proteins and PPI of human dyskerin, reflecting their conservation, functional diversity, and therefore the central role that this protein has in the dynamics of cellular processes. Hence, our present study focuses on function, evolution, and structure of dyskerin-like proteins.
The data presented in this work could be of importance because it will allow us to expand the knowledge of dyskerin function, opening the way to the development of new drugs to be employed in the treatment of telomeropathies.

**ASSOCIATED CONTENT**

2 Supporting Information

Alignment of eukaryote structural models of selected dyskerin-like proteins. This material is available free of charge via the Internet at http://pubs.acs.org.

**AUTHOR INFORMATION**

Corresponding Author

*D.E.G. and P.D.G. have equally contributed to the work.

Notes

The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

D.E.G. and P.D.G. are researchers from CONICET. D.E.G. is Director of the National Cancer Institute of Argentina. C.S.C has a postdoctoral fellowship from CONICET. This work was supported by Grants from Quilmes National University, CONICET and MINCyT (Argentina).

**REFERENCES**