Sequence and Comparative Genomic Analysis of Actin-related Proteins

Jean Muller,*† Yukako Oma,‡§ Laurent Vallar,† Evelyne Friederich,† Olivier Poch,* and Barbara Winsor‡

*Laboratoire de Biologie et Génomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire, CNRS/INSERM/ULP, BP 163, 67404 Illkirch Cedex, France; †Laboratoire de Biologie Moléculaire, d’Analyse Génique et de Modélisation, Centre de Recherche Public-Santé, L-1911, Luxembourg, Luxembourg; and ‡UMR 7156 “Génétique Moléculaire, Génomique et Microbiologie,” Institut de Physiologie et de Chimie Biologique, 67084 Strasbourg Cedex, France

Submitted June 8, 2005; Revised September 16, 2005; Accepted September 21, 2005

Monitoring Editor: Thomas Pollard

Actin-related proteins (ARPs) are key players in cytoskeleton activities and nuclear functions. Two complexes, ARP2/3 and ARPI/11, also known as dyactin, are implicated in actin dynamics and in microtubule-based trafficking, respectively. ARP4 to ARP9 are components of many chromatin-modulating complexes. Conventional actins and ARPs codefine a large family of homologous proteins, the actin superfamily, with a tertiary structure known as the actin fold. Because ARPs and actin share high sequence conservation, clear family definition requires distinct features to easily and systematically identify each subfamily. In this study we performed an in depth sequence and comparative genomic analysis of ARP subfamilies. A high-quality multiple alignment of ~700 complete protein sequences homologous to actin, including 148 ARP sequences, allowed us to extend the ARP classification to new organisms. Sequence alignments revealed conserved residues, motifs, and inserted sequence signatures to define each ARP subfamily. These discriminative characteristics allowed us to develop ARPAnno (http://bips.u-strasbg.fr/ARPAnno), a new web server dedicated to the annotation of ARP sequences. Analyses of sequence conservation among actins and ARPs highlight part of the actin fold and suggest interactions between ARPs and actin-binding proteins. Finally, analysis of ARP distribution across eukaryotic phyla emphasizes the central importance of nuclear ARPs, particularly the multifunctional ARP4.

INTRODUCTION

Since the discovery in the early 1990s of the first genes coding for actin-related proteins (ARPs) called ACT2, now known as ARP2 (Schwob and Martin, 1992) in Saccharomyces cerevisiae and ARP3 (also called ACT2; Lees-Miller et al., 1992b) and ARP1 (called actin-RPV; Lees-Miller et al., 1992a) in Schizosaccharomyces pombe, many new ARPs have been described from unicellular organisms to plants and humans. Sustained investigation of ARP function(s) in yeast, plant, and animal cells has demonstrated that different ARPs, in combination with actin, are required for cytoplasmic or nuclear cellular functions.

The unified classification of ARPs, initially proposed in 1994 (Schroer et al., 1994) was extended in 1997 (Poch and Winsor, 1997). The second study led to the definition of 10 distinct ARP subfamilies according to their relative identity and similarity to conventional actin sequences, where ARP1 is the most similar and ARP10 the least similar. In contrast to the ARPI to ARP3 subfamily classifications, which were based on multiple sequences from diverse organisms, the ARP4–ARP10 subfamilies were proposed on the basis of only 1 or 2 sequences, in particular from the complete genome of S. cerevisiae (Goffeau et al., 1996). Since then, only one new subfamily, ARP11, has been described (Eckley et al., 1999). This suggested nomenclature has been assessed for major model organisms (Eckley et al., 1999; Harata et al., 2001; Goodson and Hawse, 2002), and a certain number of organisms possess additional “orphan” ARPs that do not group into any of the known subfamilies (Goodson and Hawse, 2002). In fact, had the classification been based on a different organism, ARP7 and ARP9, the yeast specific subfamilies (Blessing et al., 2004), would have been considered as orphans. In this classification, ARP1–ARP3 (and more recently ARP10 and ARP11) are localized in the cytoplasm and perform key functions in the spatiotemporal control of actin assembly and movement of vesicles along microtubules in the cytoplasm (Schafer and Schroer, 1999; Machesky and May, 2001; McKinney et al., 2002). In addition to these well-documented functions, a growing body of evidence supports nuclear functions for ARP4–ARP9 participating in processes like chromatin modulation, regulation of transcription, and DNA repair (Weber et al., 1995; Grava et al., 2000; Harata et al., 2000; Olave et al., 2002; Blessing et al., 2004). This has expanded the palette of actin function and kept ARPs in the limelight of investigative biology.
Together with conventional actins, ARPs define a large family of homologous proteins, the actin superfamily, which share the same structural architecture, known as the “actin fold” (Bork et al., 1992; Holmes et al., 1993; Kabsch and Holmes, 1995). This architecture is also found in heat-shock protein Hsc70, sugar kinases, and bacterial proteins (Bork et al., 1992; Holmes et al., 1993). Although some of these bacterial proteins have recently been shown to retain some actin-like functions (Amos et al., 2004; Rodal et al., 2004). In contrast, ARP7 and ARP9 are not essential for RSC protein Hsc70, sugar kinases, and bacterial proteins (Bork et al., 1992; Holmes et al., 1993). In most cases, the ARP subfamily has been characterized as part of one or more multisubunit complexes, many of which also contain at least one actin molecule. ARP1, the only ARP known to form a filament (Bingham and Schroer, 1999), is an essential part of the 11-subunit dynactin complex that functions in transport of cargoes and organelles on microtubules. In human cells, this complex also contains the distantly related ARP11 as well as globular actin and the ABP, CapZ (Eckley et al., 1999; Eckley and Schroer, 2003). The ARP2 and ARP3 dimer is part of a seven-subunit complex that nucleates polymerization of de novo actin filaments and branched networks beneath the plasma membrane (reviewed in Pollard et al., 2000). The 3D structure of the ARP2/3 complex has recently been solved in different states (Robinson et al., 2001; Volkmann et al., 2001; Nolen et al., 2004; Rodal et al., 2005).

Nuclear ARPs and actin have been isolated from many nuclear complexes (see Supplementary Data 1). In chromatin-remodeling complexes, ARP4 is generally present with conventional actin in SWI2/SNF2 complexes, with ARP5 and ARP8 in INO80 complexes, and with ARP6 in SWR1 complexes (reviewed in Mohrmann et al., 2004). ARP4 and actin are also components of histone acetyltransferase (HAT) complexes (Doyon et al., 2004). In most cases, the ARP subunits are important for the enzymatic activity of these complexes (Galarneau et al., 2000; Gorzer et al., 2003; Shen et al., 2003). In contrast, ARP7 and ARP9 are not essential for RSC chromatin remodeling complex activity (Szerlong et al., 2003).

With the availability of numerous new ARP sequences in protein databases and recently sequenced genomes, we propose an in-depth comparative genomic analysis of ARP members. We built a new high-quality Multiple Alignment of Complete Sequences (MACS; Lecompte et al., 2001) of ARPs available at http://bips.u-strasbg.fr/ARPAnno/ARPMACS.html. This alignment provides the basis for the extension of the characterization of ARP subfamilies and the classification of ARPs from new organisms. Our sequence analysis differentiates ARPs by determining conserved Family features such as discriminating residues and insertion or deletion (INDEL) sequence signatures. On the basis of our multiple alignment and new discriminating characteristics, we implemented ARPAnno, a freely available web server to annotate ARP sequences http://bips.u-strasbg.fr/ARPAnno. Analysis of the conservation of residues involved in actin ATP-binding capacity predicts weak (or no) interaction for nuclear ARP4–9, ARP10, and ARP11. Furthermore, sequence conservation in the actin fold highlights the importance of the hydrophobic cleft in ARPs, opening new perspectives for interactions between ARPs and ABPs. Finally, the availability of complete genome sequences allowed us to define the distribution of ARPs among eukaryotic phyla and reveals the central importance of the nuclear ARPs, especially ARP4.

**MATERIALS AND METHODS**

**Sequence Searches and Alignment**

To cover the maximum diversity of ARPs, sequence searches were performed using as sequence queries: three actin sequences, representative sequences from each ARP subfamily selected from three distantly related organisms (S. cerevisiae, Drosophila melanogaster, and Homo sapiens when available), and five orphan sequences. The sequences were retrieved from Uniprot using SRS (Sequence Retrieval System; Ettzold et al., 1996) with their identification (ID) or accession (AC) numbers. This resulted in an initial set of 37 reference proteins shown in Supplementary Data 2.

For each reference protein, a BlastP (Altschul et al., 1990, 1997) search of the Uniprot database (Bairoch et al., 2005; July 2004) was performed and the sequences detected with E ≤ 10−2 were multiply aligned using the PipeAlign program without the clustering step (Plewniak et al., 2003). The PipeAlign web server is available at http://bips.u-strasbg.fr/PipeAlign. These 37 MACS were then merged into a single multiple alignment. Unrelated sequences were removed from the final alignment with LEON (multiple alignment evaluation of neighbors; Thompson et al., 2003). This composite alignment was then refined using RASCAL (rapid scanning and correction of multiple sequence alignment; Thompson et al., 2003) to automatically correct local alignment errors. Finally, manual verification and correction, paying attention to secondary structures, was performed using the seqlab alignment viewer and editor (KCC, 2001). The quality of the final alignment was objectively evaluated using NorMD (normalized mean distance; Thompson et al., 2001). Subfamilies were defined automatically using DPC (density of point clustering; Wicker et al., 2002) and validated by human expertise. Furthermore, a phylogenetic tree based on the final alignment was built with the neighbor joining method (see Supplementary Figure 1). The analysis of this tree confirmed the defined subfamilies.

Overall, the BlastP similarity searches yielded 73,340 proteins, representing 4200 nonredundant and “nonfragment” proteins. Sequences with <15% amino acid identity, notably some bacterial actin-like proteins, were not included in the final alignment. To obtain an objective evaluation of the true number of ARP sequences in Uniprot, we removed database redundancy by counting only nonidentical sequences for each different organism. The final version of the complete multiple alignment of ARPs (ARP-MACS) contains more than 700 proteins (sequence list included in Supplementary Data 3) clustered into one subfamily, 11 ARP subfamilies, and orphans. ARP-MACS is available at http://bips.u-strasbg.fr/ARPAnno/ARPMACS.html.

**Sequence Analysis**

Two statistics were used to characterize each ARP subfamily, RefID and FamID. First, the RefID is defined to compare the current subfamily classification with the one used in 1997 (Poch and Winsor, 1997; called IniID here). This composite alignment was then refined using RASCAL (rapid scanning and correction of multiple sequence alignment; Thompson et al., 2003) to automatically correct local alignment errors. Finally, manual verification and correction, paying attention to secondary structures, was performed using the seqlab alignment viewer and editor (KCC, 2001). The quality of the final alignment was objectively evaluated using NorMD (normalized mean distance; Thompson et al., 2001). Subfamilies were defined automatically using DPC (density of point clustering; Wicker et al., 2002) and validated by human expertise. Furthermore, a phylogenetic tree based on the final alignment was built with the neighbor joining method (see Supplementary Figure 1). The analysis of this tree confirmed the defined subfamilies.

Overall, the BlastP similarity searches yielded 73,340 proteins, representing 4200 nonredundant and “nonfragment” proteins. Sequences with <15% amino acid identity, notably some bacterial actin-like proteins, were not included in the final alignment. To obtain an objective evaluation of the true number of ARP sequences in Uniprot, we removed database redundancy by counting only nonidentical sequences for each different organism. The final version of the complete multiple alignment of ARPs (ARP-MACS) contains more than 700 proteins (sequence list included in Supplementary Data 3) clustered into one subfamily, 11 ARP subfamilies, and orphans. ARP-MACS is available at http://bips.u-strasbg.fr/ARPAnno/ARPMACS.html.

**Sequence Analysis**

Two statistics were used to characterize each ARP subfamily, RefID and FamID. First, the RefID is defined to compare the current subfamily classification with the one used in 1997 (Poch and Winsor, 1997; called IniID here). This composite alignment was then refined using RASCAL (rapid scanning and correction of multiple sequence alignment; Thompson et al., 2003) to automatically correct local alignment errors. Finally, manual verification and correction, paying attention to secondary structures, was performed using the seqlab alignment viewer and editor (KCC, 2001). The quality of the final alignment was objectively evaluated using NorMD (normalized mean distance; Thompson et al., 2001). Subfamilies were defined automatically using DPC (density of point clustering; Wicker et al., 2002) and validated by human expertise. Furthermore, a phylogenetic tree based on the final alignment was built with the neighbor joining method (see Supplementary Figure 1). The analysis of this tree confirmed the defined subfamilies.

Overall, the BlastP similarity searches yielded 73,340 proteins, representing 4200 nonredundant and “nonfragment” proteins. Sequences with <15% amino acid identity, notably some bacterial actin-like proteins, were not included in the final alignment. To obtain an objective evaluation of the true number of ARP sequences in Uniprot, we removed database redundancy by counting only nonidentical sequences for each different organism. The final version of the complete multiple alignment of ARPs (ARP-MACS) contains more than 700 proteins (sequence list included in Supplementary Data 3) clustered into one subfamily, 11 ARP subfamilies, and orphans. ARP-MACS is available at http://bips.u-strasbg.fr/ARPAnno/ARPMACS.html.

**Sequence Analysis**

Two statistics were used to characterize each ARP subfamily, RefID and FamID. First, the RefID is defined to compare the current subfamily classification with the one used in 1997 (Poch and Winsor, 1997; called IniID here). This composite alignment was then refined using RASCAL (rapid scanning and correction of multiple sequence alignment; Thompson et al., 2003) to automatically correct local alignment errors. Finally, manual verification and correction, paying attention to secondary structures, was performed using the seqlab alignment viewer and editor (KCC, 2001). The quality of the final alignment was objectively evaluated using NorMD (normalized mean distance; Thompson et al., 2001). Subfamilies were defined automatically using DPC (density of point clustering; Wicker et al., 2002) and validated by human expertise. Furthermore, a phylogenetic tree based on the final alignment was built with the neighbor joining method (see Supplementary Figure 1). The analysis of this tree confirmed the defined subfamilies.

Overall, the BlastP similarity searches yielded 73,340 proteins, representing 4200 nonredundant and “nonfragment” proteins. Sequences with <15% amino acid identity, notably some bacterial actin-like proteins, were not included in the final alignment. To obtain an objective evaluation of the true number of ARP sequences in Uniprot, we removed database redundancy by counting only nonidentical sequences for each different organism. The final version of the complete multiple alignment of ARPs (ARP-MACS) contains more than 700 proteins (sequence list included in Supplementary Data 3) clustered into one subfamily, 11 ARP subfamilies, and orphans. ARP-MACS is available at http://bips.u-strasbg.fr/ARPAnno/ARPMACS.html.
above, positions in the alignment corresponding to gaps within the subset were excluded from the calculation.

\[
\sum_{i \in \text{query}} I_{D_{\text{SIM}}^{i}} \cdot \text{FamID} = 2 \cdot \frac{\text{identical residues} + \text{similar residues}}{n - 1}
\]

where: \( n \) is the total number of sequence tested, \( S \) and \( S' \) are the ith and jth sequence, and \( I_{D_{\text{SIM}}^{i}} \) are pairwise percent identity between the ith and jth sequence, excluding gapped regions.

Limitations of the RefID and FamID calculations are the absence of certain subfamilies in different organisms and the incomplete representation of ARPs in protein databases. Taking this into account, the two statistics were calculated as a minimum of 17 key reference sequence alignments from the reference actin sequence (Supplementary Data 5). To characterize the INDELS, the entry point was defined as a single position where at least one ARP has an INDEL and the “hot spot” as a short sequence stretch in which many different ARPs have INDELS. The discriminating residues, motifs, and INDELS constitute a knowledge filter used to characterize the ARP subfamilies.

**ARPAnno Web Server**

To make our results easily available to the scientific community, a web server ARPAnno (actin-related proteins annotation server) has been developed to allow reliable classification and annotation of newly sequenced potential actinlike proteins. ARPAnno is written in Tcl/Tk script or in ANSI C for some functions. ARPAnno also requires the Blast and ClustalW programs. The strategy of ARPAnno is based on a three-step process:

1. First, ARPAnno compares query sequences en passant with Blast against dedicated databases of each subfamily contained in ARP-MACS (actin, 11 ARP subfamilies and orphans). Eligible subfamilies which are the most suitable for further investigation are then determined by the calculation of two cutoffs. First, a global percent identity (GID) is defined as the ratio of the number of identical residues to the total number of residues in all HSPs (high scoring pairs) of the query. Second, a percent coverage (pCover) is defined as the ratio of the number of identical residues to the number of residues that could be aligned between the two sequences.

2. The query is then aligned against the eligible subfamilies in the ARP-MACS using the ClustalW global multiple alignment program (Thompson et al., 1994) in profile mode and filtered according to the knowledge-based criteria (residues and INDELs signatures) defined above. For each eligible subfamily, two scores are calculated: the number of discriminating INDELs and motifs, and at least one sequence from mammals, insects, worms, fungi, and plants for each subfamily, except ARP7, ARP9, and ARP10, restricted to yeasts (Accession numbers of sequences used are available upon request). The relative weights of each score were determined experimentally to best consolidate the ARP4 – ARP11 subfamilies classification. In agreement with previous studies, the major ARP subfamilies are ARP1–6, ARP8, and ARP11, whereas fewer sequences are available for subfamilies ARP7, ARP9, and ARP10. As illustrated below, these subfamilies are restricted to certain phyla. Twenty-seven orphan protein sequences were found defined by subdomains 1 and 3 excluding helices H15, H19, and H20, as well as helix H11 and the bottom part of helices H8 and H9 in subdomain 4, with no contribution from subdomain 2 (Kabsch and Holmes, 1995). One major actin-binding interface of actin, known as the “hydrophobic clef” is defined essentially by residues in three helices (H18, H19, and H20) in subdomain 1 ( Dominguez, 2004). The mean percent identity to the reference actin in ARP-MACS was calculated using a sliding-window corresponding to each secondary structure. This statistic was used to replace the temperature factor field in the PDB file. Figure 4A represents the mean percent identity of all ARP subfamilies and Figure 4B that of each ARP subfamily individually.

The sequence conservation is mapped onto the structure with colors ranging from dark blue to red, corresponding to 0–65% identity (id.; loops excluded) in Figure 4A and to 0–100% id. in Figure 4B.

**Phylogenetic Distribution of ARPs in Complete Genomes**

The ARP distribution was examined in 20 eukaryotic organisms for which the complete genome sequences are available. The presence/absence of each ARP was cross-validated at both the proteomic and genomic levels. Inspection of recently reported genomic sequences identified potential new ARP genes predicted during the gene prediction process. A table summarizing proteomic and genomic searches is included in Supplementary Data 6. Where available, the nucleotide sequence was retrieved from the NCBI nucleotide sequence database known as GenBank (Benson et al., 2005) and RefSeq (Pruitt et al., 2003) and queried with the 37 reference sequences using the TBLASTN program.

The 20 complete eukaryotic genomes used are: *Oryza sativa* (Goff et al., 2002), *Arabidopsis thaliana* (Arabidopsis Genome Initiative, 2000), *Plasmodium falciparum* (Gardner et al., 2002), *Escherichia coli* (Katinka et al., 2001), *Neurospora crassa* (Galagan et al., 2003), *S. cerevisiae* (Goffeau et al., 1996), * Candida galliata*, *Yarrowia lipolytica* (Dyson et al., 2004), *S. pombe* (Wood et al., 2002), *Aspergillus fumigatus* (Booth et al., 2002), *D. melanogaster* (Adams et al., 2000), *Caenorhabditis elegans* (Bertuccio et al., 1998; C. elegans Sequencing Consortium, 1998), *Ciona intestinalis* (Dehal et al., 2002), *Tetrahymena pyriformis* (Jaillon et al., 2004), *M. musculus* (Waterston et al., 2002), *H. sapiens* (Lander et al., 2001; Venter et al., 2001). We also used dedicated websites in order to retrieve the latest sequence version for *Thalassiosira pseudonana* (http://genepool.jgi-psf.org/thalps1/Armbrust et al., 2004), *Dictyosteliun discoideum* (http://dictybase.org; Koppel et al., 2004; Eichinger et al., 2005), and other sites for additional Blast searches for *Cryptosporidium parvum* (http://cryptodb.org/CryptDB.shtml; Abrahamsen et al., 2004; Puitt et al., 2004), and *Caenobacter c/eon melor* (http://merolae.biologie.uni-tokyo.ac.jp/; Matsuzaki et al., 2004).


**RESULTS**

**Sequence Analysis and Subfamily Definition**

We built ARP-MACS, a new high quality multiple alignment of complete sequences of all ARPs and actins available in Uniprot (July 2004) as the basis for an extended characterization of ARP subfamilies. In our earlier study, the previously defined ARP1–ARP3 subfamilies (Schroer et al., 1994) were confirmed on the basis of 5–8 sequences, and the remaining ARP subfamilies were proposed essentially on the basis of *S. cerevisiae* sequences (Poch and Winsor, 1997). Later these subfamilies were established by phylogenetic analyses (Eckley et al., 1999; Harata et al., 2001; Goodson and Hawse, 2002). Since 1997, including this analysis, the only new major ARP that was identified is ARP11 (Eckley et al., 1999). The growing number of ARP proteins available in protein databases and classified in ARP-MACS (Table 1) consolidates the ARP4–ARP11 subfamilies classification. In agreement with previous studies, the major ARP subfamilies are ARP1–6, ARP8, and ARP11, whereas fewer sequences are available for subfamilies ARP7, ARP9, and ARP10. As illustrated below, these subfamilies are restricted to certain phyla.

**Structural Studies**

The actin molecule is represented by the 3D structure of yeast actin (Uniprot ID ACT_YEAST, PDB 1YAG; Vorobiev et al., 2003) and secondary structures are named according to the PDB data (see Figure 4). The actin fold is mainly
in Metazoa (H. sapiens, M. fascicularis, M. musculus, and C. elegans), plants (O. sativa and A. thaliana) and in the parasites E. cuniculi, P. falciparum, and P. yoelii. These sequences range in size from 328 to 1207 amino acids, and share from 21 to 49% Id. with the reference actin. They have been included in the overall alignment but are not considered as a subfamily because they lack common defining characteristics.

To validate the reliability of the ARP classification based on the mean percent identity between ARP sequences and the reference actin (RefID), we compared the ranking obtained here with the ARP ranking based on initial percent identities (IniID) deduced from the data available in 1997 (Poch and Winsor, 1997). IniID and RefID are highly correlated (Figure 1); ARP1 is the closest to actin and ARP10 and ARP11 are the most distant in both cases, reinforcing the universal classification. Nevertheless, with our recent refinements and definitions, the relative order of some subfamilies could have been exchanged, e.g., ARP5 with ARP6. In spite of these small variations, to avoid confusion in naming genes and proteins, we do not recommend changing the existing nomenclature. The growing number of sequences per ARP subfamily allows an evaluation of the intrasubfamily conservation (FamID). Three groups of proteins were distinguished (Figure 1). As expected, the conventional actins (546 sequences) are the most conserved subfamily (FamID > 80%). The second group is composed of cytoplasmic ARPs (ARP1–ARP3), and shares significantly more intrasubfamily conservation (50% FamID > 80%) than the last group including all nuclear ARPs and ARP10 and ARP11 (FamID < 40%).

**ARP Subfamily Characterization**

Because of high sequence identity and similarity between ARPs and actin sequences, it is frequently difficult to unambiguously detect and classify an ARP sequence from BlastP database searches. Indeed the Blast score and ranking of ARP homologous sequences is perturbed by the presence of insertions and deletions and the existence of a very limited number of discriminating residues (see Materials and Methods). As an example, the search of homologues (BlastP) for the human ARP1 in Uniprot leads to 1653 protein “hits” exhibiting a significant E-value (E \( \approx 10^{-22} \)). Among these, ARP1 sequences are dispersed among conventional actin and other ARPs. The last ARP1 detected was the yeast ARP1 at rank 769, lower than many non-ARP1 sequences. This prompted us to define discriminating criteria, i.e., sequence features conserved in a given subfamily and strictly absent in any other, for each ARP subfamily using specific residues, motifs or INDELs as shown in Figure 2 and Supplementary Data 5.

![Figure 1](image1.png)

**Figure 1.** Actin and ARP conservation. The initial percent identity (IniID) used in 1997 (Poch and Winsor, 1997) to classify the ARPs is represented as open bars. A new percent identity (RefID) is shown as hatched bars. The closed bars are the mean percent identity inside a given subfamily (FamID). Error bars, SDs.

<table>
<thead>
<tr>
<th>Subfamilies</th>
<th>Poch and Winsor (1997)</th>
<th>No redundancy</th>
<th>Total</th>
<th>20 complete genomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actin</td>
<td>29a</td>
<td>—</td>
<td>546</td>
<td>20</td>
</tr>
<tr>
<td>ARP1</td>
<td>8</td>
<td>20</td>
<td>26</td>
<td>16</td>
</tr>
<tr>
<td>ARP2</td>
<td>5</td>
<td>19</td>
<td>30</td>
<td>15</td>
</tr>
<tr>
<td>ARP3</td>
<td>7</td>
<td>23</td>
<td>34</td>
<td>15</td>
</tr>
<tr>
<td>ARP4</td>
<td>2</td>
<td>21</td>
<td>39</td>
<td>21b</td>
</tr>
<tr>
<td>ARP5</td>
<td>1</td>
<td>12</td>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td>ARP6</td>
<td>2</td>
<td>16</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>ARP7</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>ARP8</td>
<td>1</td>
<td>11</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>ARP9</td>
<td>1</td>
<td>7</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>ARP10</td>
<td>1</td>
<td>6</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>ARP11</td>
<td>0</td>
<td>8</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>29</td>
<td>—</td>
<td>756</td>
<td>152</td>
</tr>
<tr>
<td>Total ARP</td>
<td>29</td>
<td>148</td>
<td>210</td>
<td>132</td>
</tr>
</tbody>
</table>

Sequences were collected and analyzed as described in Materials and Methods.

a Twenty-nine actin sequences out of 194 available (Poch and Winsor, 1997).

b Second ARP4 in Y. lipolytica and S. pombe.

**Table 1. Evolution of the number of actin and ARP sequences since 1997**

<table>
<thead>
<tr>
<th>Subfamilies</th>
<th>Poch and Winsor (1997)</th>
<th>No redundancy</th>
<th>Total</th>
<th>20 complete genomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actin</td>
<td>29a</td>
<td>—</td>
<td>546</td>
<td>20</td>
</tr>
<tr>
<td>ARP1</td>
<td>8</td>
<td>20</td>
<td>26</td>
<td>16</td>
</tr>
<tr>
<td>ARP2</td>
<td>5</td>
<td>19</td>
<td>30</td>
<td>15</td>
</tr>
<tr>
<td>ARP3</td>
<td>7</td>
<td>23</td>
<td>34</td>
<td>15</td>
</tr>
<tr>
<td>ARP4</td>
<td>2</td>
<td>21</td>
<td>39</td>
<td>21b</td>
</tr>
<tr>
<td>ARP5</td>
<td>1</td>
<td>12</td>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td>ARP6</td>
<td>2</td>
<td>16</td>
<td>21</td>
<td>19</td>
</tr>
<tr>
<td>ARP7</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>ARP8</td>
<td>1</td>
<td>11</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>ARP9</td>
<td>1</td>
<td>7</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>ARP10</td>
<td>1</td>
<td>6</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>ARP11</td>
<td>0</td>
<td>8</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td>29</td>
<td>—</td>
<td>756</td>
<td>152</td>
</tr>
<tr>
<td>Total ARP</td>
<td>29</td>
<td>148</td>
<td>210</td>
<td>132</td>
</tr>
</tbody>
</table>
four deletions. Excluding the N- and C-terminal extensions, a total of 16 entry points was observed. Four hot spots (named A, B, C, and D in Figure 2) are found at positions 42–63, 239–247, 271–288, and 323–334. INDELS have a maximum size of 330 amino acids and occur mainly in loops. No large insertions were identified in the core structure of the actin fold. In terms of domain distribution, the discontinuous subdomain 1 is the least susceptible to INDELs with only ARP4, ARP6, and ARP10 having intrasubdomain insertions. In contrast, the smallest subdomain 2 is the most sensitive to adaptation, with the main hot spot (A) comprising 11 INDELs distributed in ARP3–ARP6 and ARP8–ARP11. Hot spot (A) includes one deletion from each of ARP6, ARP10, and ARP11 but none of these are characteristic of all subfamily members. The largest deletion is observed in subdomain 4 of *S. cerevisiae* ARP10 and results in the complete loss of almost all the subdomain. Another remarkable feature, if we consider the ARPs cellular localization, is the paucity of insertions larger than 10 residues in cytoplasmic ARP1, ARP2, and ARP3, in contrast to the nuclear ARPs, which contain from 1 to 5 insertions.

Although many entry points are common to different ARP subfamilies, it is noteworthy that no sequence similarity was found between the insertions from different ARP subfamilies. Thus, ARP characterization can be described by an insertion common to all members of a given ARP subfamily (Family Insertion, highlighted in yellow in Figure 2). However, ARP1 has no insertion ($\geq$ 10 aa), and ARP10 and ARP11 have many different INDels but none are conserved in all members of the subfamily. We also found that the N-terminal motif MS[G/A][G/A][Y/L]YGG in ARP4 (Choi et al., 2001), previously described as characteristic, is absent from 6 ARP4 sequences from different organisms (plasmodia and yeast). Two other Family Insertions are of particular interest. The largest insertion in ARP5 (position 246) is rich in charged residues, and the ARP9 Family Insertion at position 333 contains a pattern rich in rare aromatic amino acids [P/S][D/E]YF[P/S][E/S]WK. Taken together, the specific residues, motifs, and Family Insertions constitute a knowledge filter that defines at least one discriminative feature for each ARP subfamily except for ARP10 and ARP11, which are defined only by sequence similarity.

**ARPAnno Web Server**

Our approach, based on ARP-MACS, combines three complementary strategies with local and global sequence information and a knowledge filter (see Materials and Methods). Based on this, we implemented a web server to annotate ARP sequences. The web server, called ARPAnno, is available at http://bips.u-strasbg.fr/ARPAnno and allows the user to submit a sequence in FASTA format. The analysis of actin and ARP conservation (Figure 1) shows that a query is identified as an actin if it has a GID $\geq$ 80% and a pCover $\geq$ 80% compared with any conventional actin sequence (see Materials and Methods). To estimate the accuracy and reliability of the ARPAnno annotations, we submitted each of the 700 previously identified actin and ARP proteins in ARP-MACS for automatic classification. In this large-scale test, all proteins were assigned to the correct subfamilies. To evalu-
ARP11) has 90% similar residues, whereas the second group (ARP4 – cytoplasmic ARPs (ARP1–ARP3), has 60% identical and 46% similar residues. The first group, composed of conventional actin and/or through other residues. The mean conservation of 17 key reference residues in- teracting with the exception of H20, are part of the previously defined actin fold (Kabsch and Holmes, 1995). The observed conservation points are localized in the bottom half of the actin fold and more precisely, in the hydrophobic cleft (Dominguez, 2004), a key region for actin dimerization and for interaction with ABPs. The analyses of individual ARP subfamily conservation highlight specific patterns. As expected in view of the FamID, the main cytoplasmic ARP1–ARP3 share more conserved elements than nuclear ARPs. Surprisingly, ARP2 is less conserved in the helix H18 and H19 involved in the hydrophobic cleft than in either ARP1 or ARP3. Additional features can be observed in subdomains 2 and 4 for ARP1 and ARP2. We noticed that ARP1 and ARP2 reveal better sequence conservation in helix H9 and in strand S4 and S10 than ARP3. Within the nuclear ARPs, ARP4 unexpectedly maintains high conservation in the lower part of subdomain 1 (H18, H19, and H20). This observation underlines functional perspectives for ARP4 through its hydrophobic cleft. Finally, with regard to other secondary structures that are part of subdomain 1, S1 is highly conserved in ARP5, ARP6, and ARP11; S6 in ARP5; and S2 in ARP7 and ARP11.

Phylogenetic Distribution
The growing number of completely sequenced genomes available allows us to define the edges of the distribution of eukaryotic ARPs by in depth analysis of the proteomes and genomes of 20 organisms ranging across eukaryotic phyla. As observed in many organisms (T. pseudonana, D. discoideum . . . ; Supplementary Data 6), the genomic validation is essential to assess the presence of a given ARP, considering that a certain number of genes present have not been annotated as proteins. The phylogenetic distribution of ARP subfamilies and conventional actin is represented in Figure 5. According to defined ARP signatures, we detected 132 ARP proteins in 11 subfamilies from algae to mammals, and at least 1 actin and 1 ARP in each organism analyzed. It is noteworthy that the organisms with limited numbers of ARP (E. cuniculi, C. merolae) have no detectable cytoplasmic ARPs but include at least one nuclear ARP. In all other organisms, both nuclear and cytoplasmic ARPs are present. Remarkably, the examination of the presence/absence profiles led to the definition of pairs of copresent/coabsent ARPs such as ARP2 with ARP3, ARP4 with ARP6, ARP5 with ARP8, ARP7, with ARP9, and to lesser extent ARP1 with ARP10 or ARP11. Surprisingly, the most widely distributed ARPs in evolution, copresent in all organisms studied with the exception of the small obligate parasite E. cuniculi, are the nuclear ARPs, ARP4, and ARP6. This result was unexpected and leads to the conclusion that ARP4 and ARP6 represent the most universal ARPs conserved throughout the eukaryotic phyla. The second most widely distributed pair of proteins is ARP2 and ARP3, well studied components of the actin nucleation complex. They are copresent in plants, fungi, and Metazoa but are coabsent in algae and in Apicomplexa.

**Figure 3.** Conservation pattern of the 17 residues (D13, S16, G17, L18, K20, Q139, D156, D159, G160, V161, K215, E216, G304, T305, M307, Y308, and K338) known to participate in nucleotide binding to actin. For the 11 ARP subfamilies and actin, percent identity is represented as hatched bars and percent similarity as closed bars. Error bars, SDs.
Figure 4.
ARP1, the closest ARP to conventional actin, is individually more widely distributed than ARP2 and ARP3. However, when one considers the functional complex dynactin where the ARP1 filament is capped by ARP11 (Eckley et al., 1999), the pattern of presence/absence appears more complex than other pairs. In fact, although ARP11 is not present without ARP1, it is not found in every organism bearing ARP1. It is interesting to notice that ARP10, restricted to fungi, only partially complements the ARP11 pattern. Furthermore, our extended exploration of fungi (see Materials and Methods) confirms the presence of ARP1 in 30 out of 31 organisms (except E. cuniculi) and restricts ARP10 to only 5 Ascomycota Saccharomycotina (D. hansenii, E. gossypii, K. lactis, S. cerevisiae, and Y. lipolytica) and 1 Ascomycota Schizosaccharomyces (S. pombe). One ARP11 was found in Ascomycota Pezizomycotina (N. crassa).

Figure 4 (facing page). Actin amino acid conservation in secondary structure throughout ARP subfamilies. Actin 3D structure is drawn from the yeast PDB data 1YAG in standard orientation with secondary structures labeled H for helices and S for strands, numbered in order of appearance from N- to C-terminus. The secondary structures were colored according to percent identity (Id.) with the reference human α-actin (Uniprot ID ACTS_HUMAN and AC P02568) by replacing the temperature factor field in the PDB file. The figures were made with PyMol (Delano, 2002). (A) Mean ARP subfamilies’ global conservation. The conservation scale 0–65% Id. is colored from dark blue to red. An ATP molecule is represented in dark gray. The four circled letters indicate with an arrowhead the four hot spot positions of insertions. The four subdomains of actin are indicated as Ia (1), Ib (2), IIA (3), and IIB (4). (B) Individual ARP subfamily conservation structures. Arrows mark specific details for each subfamily as described in results. The conservation scale 0–100% Id. is colored from dark blue to red.

Figure 5. Schematic representation of ARP distribution among the eukaryotic phyla. The columns represent different organisms with a completely sequenced genome: T. pseudonana (TP), C. merolae (CM), O. sativa (OS), A. thaliana (AT), P. falciparum (PF), C. parvum (CP), D. discoideum (DD), E. cuniculi (EC), N. crassa (NC), S. cerevisiae (SC), C. glabrata (CG), Y. lipolytica (YL), S. pombe (SP), A. gambiae (AG), D. melanogaster (DM), C. elegans (CE), C. intestinalis (CI), T. nigroviridis (TN), M. musculus (MM), and H. sapiens (HS). The existence of a colored rectangle indicates the presence of the protein in the organism considered. Conventional actin is represented as green rectangles in between the two groups of ARPs, the cytoplasmic ARPs (ARP1–ARP3, ARP10, and ARP11) as indicated by orange rectangles and the nuclear ARPs (ARP4–ARP9) indicated by blue rectangles. The four rectangles outlined in red highlight the presence of a second distinct ARP4, named ARP4* in Y. lipolytica (YL) and in S. pombe (SP).
The coabsence profile of ARP5 and ARP8 is puzzling since they are missing in a number of different phyla such as the algae, the Apicomplexa, and two Metazoan phyla, C. elegans and C. intestinalis. Our results also confirm that the functionally obligate heterodimeric partners, ARP7 and ARP9 (Szerlong et al., 2003), were restricted to fungi as previously suggested (Goodson and Hawse, 2002; Blessing et al., 2004). The presence of ARP7 and ARP9 has been assessed in the 31 fungi genomes available at NCBI and we could clearly restrict ARP7 and ARP9 to Ascomycota Saccharomycotina and Ascomycota Schizosaccharomyces. Neither ARP7 nor ARP9 were found in the Ascomycota Pezizomycotina, Basidiomycota, or Microsporidia. Surprisingly, the copresence of ARP7 and ARP9 is not observed in two completely sequenced organisms of Ascomycota, Y. lipolytica, and S. pombe, where ARP9 is present but ARP7 is absent. In this context it is noteworthy that these two organisms are the only fungi that encode an additional and distinct ARP4 (red box in Figure 5, Uniprot accession numbers Q6C0A9 and Q90849; annotated here as ARP4*). This strongly suggests that ARP4* may complement the lack of ARP7 in these yeasts.

**DISCUSSION**

The alignment of all available actin and ARP sequences in Uniprot combined with a detailed comparative analysis of 20 completely sequenced eukaryotic genomes reinforces the existing ARP subfamilies and finds them present in more organisms. Our calculation of conservation of ARPs to actin led to a classification in strong agreement with previous studies (Poch and Winsor, 1997; Eckley et al., 1999; Harata et al., 2001; Goodson and Hawse, 2002). It has been proposed that yeast ARP10 and metazoan ARP11 subfamilies might form only one highly divergent ARP family, based on the presence of an ARP11 in an Ascomycota (N. crassa) present of ARP11 in an Ascomycota, Y. lipolytica, and S. pombe, where ARP9 is present but ARP7 is absent. In this context it is noteworthy that these two organisms are the only fungi that encode an additional and distinct ARP4 (red box in Figure 5, Uniprot accession numbers Q6C0A9 and Q90849; annotated here as ARP4*). This strongly suggests that ARP4* may complement the lack of ARP7 in these yeasts.

The specificity of the ARP2 and ARP3 function relies on the fact that neither is able to homopolymerize. Indeed, these ARPs heterodimerize to bind the first actin monomer of a new filament. ARP2 and ARP3 compared to actin high-light secondary structures that are differentially conserved; ARP2 is more conserved at the pointed end in subdomains 2 and 4 (H9, S4, and S10), whereas ARP3 is better conserved at the barbed end in subdomain 1 (H18 and H19; Figure 4B). In keeping with this, we did not detect any sequences similar to any other subunits of the human dynactin complex.

The specificity of the ARP2 and ARP3 function relies on the fact that neither is able to homopolymerize. Indeed, these ARPs heterodimerize to bind the first actin monomer of a new filament. ARP2 and ARP3 compared to actin highlight secondary structures that are differentially conserved; ARP2 is more conserved at the pointed end in subdomains 2 and 4 (H9, S4, and S10), whereas ARP3 is better conserved at the barbed end in subdomain 1 (H18 and H19; Figure 4B). This differential conservation is in agreement with their inability to self-polymerize and with their role in the ARP2/3 complex. According to the recent analysis of the ARP2/3 complex (Beltzner and Pollard, 2004), ARP2 interacts through its pointed end with ARP3, and ARP3 is in contact through its barbed end with the first actin monomer sequences. Ambiguous situations can often be clarified by direct comparison to the MACS of the closest subfamilies. Our strategy of combining local and global approaches, together with a knowledge-based filter is essential to the study of this family of proteins. The ARP-MACS will be updated and the discriminating criteria revised regularly. It would be neof interest to build an automatic procedure to mine the ARPs in new genomes and combine prediction, extraction, and annotation.

The ARP-MACS has allowed us to better characterize ARP subfamilies with specific residues, motifs, and/or INDELs found in all subfamilies except ARP10 and ARP11. Considering the high number of insertions, the strength and plasticity of the actin fold is remarkable, further illustrated by the peripheral positions of the hot spots of insertion (see Figure 4A). A restricted part of the fold is highlighted by the average structural conservation of all ARP subfamilies (Figure 4A). Furthermore, high conservation in the hydrophobic cleft (helices H18, H19, and H20), which forms an actin-binding interface (Dominguez, 2004), opens new perspectives for possible interactions between ARPs and ABPs. In line with this, ARP4 stands out as having high sequence conservation in H18 and H19. In fact, recent exploration across ARP subfamilies of 27 actin residues involved in gelsolin binding showed the best conservation in ARP4 and ARP1 (Archer et al., 2004). Analysis of a larger pool of ARPs from the ARP-MACS (148 vs. 63) confirmed this result (our unpublished results). Indeed, it has been reported that the gelsolinlike domain in the ABP, Fli-I, binds to ARP4 in the SWI/SNF complex, and contributes to transcriptional activation (Lee et al., 2004).

The most prominent property of actin is to self-assemble into a filamentous structure that implicates actin-actin interfaces distributed over the surfaces of the four subdomains. Presumed actin-actin interfaces in ARP1-ARP3 harbor only a few short insertions proposed to be compatible with their role in actin assembly (Robinson et al., 2001). However, ARP1 is the sole ARP able to form a homopolymer filament in vivo. In vitro, mammalian ARP1 has been shown to polymerize with actin (Eckley and Schroer, 2003) and D. dicostostelium ARP1 was found with ARP2 and an orphan ARP (Uniprot ID Q54HE9_DICDI) in a filament (Gomez-Garcia and Kornberg, 2004). In the dynactin complex, ARP11 is found at the pointed end of the ARP1 filament, but in certain organisms (P. falciparum, C. parvum, and most of the fungi), genomic inspection revealed no ARP11 (or ARP10), suggesting that ARP11 is not an obligate partner for ARP1. This is consistent with weak conservation in the ARP11 subfamily. Like the situation in green plants (Kandasamy et al., 2004), we found that the C. merolae genome lacks ARP1 and ARP11 (Figure 5). In keeping with this, we did not detect any sequences similar to any other subunits of the human dynactin complex.

Here, we propose that ARPAnno can be used to assign new actinlike genes to an ARP subfamily before assigning a name. A protein sequence can be submitted either as an existing entry in a protein database or as a translated open reading frame (ORF) prediction. In this latter case caution should be used because some mispredicted ORFs (wrong insertions or deletions) will give incorrect scores. According to our analysis an S_{ARP, > 55} is highly reliable for complete
of the nascent filament. Our comparative analyses show that ARP2 and ARP3 are absent from the genomes of Apicomplexa, Algae and from Macrostrosporidia. Thus, it is tempting to speculate that other nucleators might replace these ARPs. For example, formins nucleate ARP2/3-independent actin polymerization (Pruyne et al., 2002; Sagot et al., 2002; Moseley et al., 2004). Inspection of the five genomes lacking ARP2 and ARP3 revealed one or more genes coding for formin-like proteins (Higgs and Peterson, 2005 and our unpublished results).

Nuclear ARP4-ARP9 have been isolated in many complexes involved in chromatin modulating functions and localized predominantly in the nucleus (reviewed in Olave et al., 2002; Blessing et al., 2004). The nuclear ARPs show less intrasubfamily conservation (FamID) than actin and major cytoplasmic ARPs. Additionally, ARP-MACs revealed that nuclear ARPs have many insertions, which when conserved characterize the subfamily (Family Insertion). For example, the sole ARP5 large Family Insertion contains several bipartite NLS sequences and shows an overall negative charge. Thus, it might interact with positively charged molecules such as histones. For ARP4, inspection of the ARP-MACs shows that the Family Insertion at position 203 (Insertion I in yeast; Harata et al., 1999) also contains putative NLS sequences (Weber et al., 1995). As an example of a function in a limited number of organisms, S. cerevisiae ARP4 contains a nonconserved insertion (known as Insertion II, position 271) shown to bind core histones (Harata et al., 1999). We found this insertion in only four organisms, S. cerevisiae, S. pombe, P. falciparum, and P. yoelii. The absence of this functionally characterized insertion in the majority of organisms opens the question of whether other ARP4 proteins bind histones. Recent studies in A. thaliana found no evidence of AtARP4 binding putative histone H2B (Kandasamy et al., 2003). Thus, characterization of insertions might serve as a guide for future in vivo studies to determine whether conserved insertions bear subfamily functions.

The decreasing sequence conservation in ARPs relative to actin raises the important question of nucleotide binding capacity. Indeed, for actin and ARP1–ARP3 the binding and hydrolysis of ATP (Bingham and Schroer, 1999; Dayel et al., 2001; Le Clainche et al., 2001) is important for their functions. Although the measured binding affinities differ between different reports, the fact that ARP2 has greater ATP binding affinity than ARP3 is consistent with their percent identity difference of nucleotide binding residues (Figure 3). According to the average conservation of nucleotide contacts, our analysis predicts that nuclear ARPs, if able to bind ATP, would do so with much less affinity than cytoplasmic ARPs. Indeed, a recent report suggests that yeast ARP4 binds weakly to ATP because 5 mM ATP precluded the binding of a competitive inhibitor. Mutation in the ATP-binding pocket appeared to increase ARP4 occupancy in complexes, whereas excess ATP released it from the wild-type NuA4 complex, suggesting a role in modulating complex dynamics (Sunada et al., 2005). To date, no direct ATP binding data have been reported for nuclear ARPs other than ARP4.

In apparent contrast, some nuclear ARPs may not bind ATP for their most important functions because they remain functional when mutated in nucleotide contact residues. Mutations in D13 and/or S16 in ARP5 and ARP8 did not change either inositol growth rate or transcriptional activations (Shen et al., 2003). Similarly, in yeast ARP7 and ARP9, mutations (D13, S16, Q139, D156, D159, G304) did not alter the activity of the RSC complex (Cairns et al., 1998; Szerlong et al., 2003). These results correspond to our predictions. Given that yeast ARP5, ARP7, and ARP9 are essential in certain strains, the lack of ATPase phenotypes implies roles other than direct implication in remodeling. Notably, our results predict that the best conservation of nucleotide contact residues should be in ARP6. Other open questions are whether a restricted number of conserved residues might be sufficient for weak binding and whether other residues present in the nuclear ARPs could contribute to an ATP-binding site. These predictions await validation with biological tests on individual proteins.

Presence/absence patterns of proteins over a wide range of phyla are a potent tool to predict the partners involved in complexes. ARP subfamilies represent a textbook case where identical patterns correspond to their copresence in functional complexes: ARP2 and ARP3 in ARP2/3 complex, ARP4 and ARP6 in SWR1 complex, and ARP5 and ARP8 in INO80 complex (Supplementary Data 1). Thus, ARP distribution should predict the presence of the corresponding complexes in different organisms. Effectively, the lack of ARP5 and ARP8 is correlated with the lack of the catalytic subunit INO80 in the corresponding genomes (Bakshi et al., 2004 and our unpublished results). Although their absence from the lower phyla is understandable during evolution, the mystery of their disappearance from certain Metazoa remains. As observed recently (Goodson and Hawse, 2002), ARP7 and ARP9 are only present in fungi. We further restricted them to Saccharomyces and Schizosaccharomyces subphyla. Surprisingly, the absence of ARP7 and the uneartiness of a second ARP4 (ARP4*) in Y. lipolytica and S. pombe make ARP4* a good candidate for functional replacement of ARP7. In light of this hypothesis, it is noteworthy that the SWI/SNF complexes of vertebrates and Drosophila contain actin and ARP4 (Tapoulas et al., 1998; Zhao et al., 1998), whereas the SWI/SNF and RSC complexes of S. cerevisiae contain ARP7 and an ARP9 (Cairns et al., 1998; Peterson et al., 1998). Do the complexes from Y. lipolytica and S. pombe contain the newly described ARP4*?

Perhaps the most unexpected result of our analyses is the revelation of the omnipresence of the nuclear ARPs, in particular ARP4 and ARP6. ARP4 stands out as the family cornerstone, because it is present in all phyla, except the parasite E. cuniculi, and in almost all complexes that contain nuclear ARPs, both ATP-dependent chromatin remodeling complexes, and in HAT complexes (Supplementary Data 1). As such, ARP4 is implicated in multiple functions such as chromatin remodeling (reviewed in Olave et al., 2002), transcriptional activation (Harata et al., 2002; Percipalle et al., 2003), DNA double-stranded break repair (van Attikum et al., 2004), apoptosis (Ikura et al., 2000), tumor suppression (Medjkane et al., 2004), histone acetylation (Galarneau et al., 2000), histone chaperone activity (Shen et al., 2003), kinesochrome-spindle attachment, and gene silencing at centromeres (Minoda et al., 2005). Because ARP4 is a primary subunit with actin, with ARP6 or with ARP5 and ARP8 in different complexes, ARP4 may be an ancestor of nuclear ARPs. In support of this hypothesis, our results suggest that ARP4* may replace ARP7 in Y. lipolytica and S. pombe. When other ARP-containing complexes are isolated, it would not be surprising to find an ARP4.

In conclusion, comparative genomics revealing the copresence or coabsence of ARP subfamilies among eukaryotic phyla largely confirms the biological data that ARPs are associated in multiprotein complexes. A major and unexpected finding of our study is that the major ARPs and the minimum package for eukaryotic organisms are the nuclear ARPs, ARP4 and ARP6.
ACKNOWLEDGMENTS

We thank Raymond Ripp and Luc Moulinier for stimulating discussions, Julie Thompson and Frederic Chalmel for their useful programs, Arnaud Muller for the mirror website, and Agatha Schulte for the representation of Figure 5. We are also grateful to two very careful reviewers and the monitoring editor whose comments significantly helped remodelling of our manuscript. This work was supported by the National Research Fund, the “Fondation Luxembourgoise Contre le Cancer,” the Fondation “Aide aux Enfants Atteints d’ un Cancer,” Luxembourg, the Centre National de Recherche Scientifique and the Direction des Relations Internationales of the CNRS. This work was achieved using the RING (Réseau National de Génopoles) Strasbourg Bioinformatics Platform infrastructures and competencies, J. Muller is supported by a fellowship from the MCSR, Luxembourg. Y.O. thanks the Japan Society for the Promotion of Science (JSPS) for a Young Scientist Fellowship.

REFERENCES


