MiCroKit 3.0: an integrated database of midbody, centrosome and kinetochore

Jian Ren¹, Zexian Liu¹, Xinjiao Gao¹, Changjiang Jin¹, Mingliang Ye², Hanfa Zou², Longping Wen¹,³, Zhaolei Zhang⁴, Yu Xue¹,³,* and Xuebiao Yao¹,*

¹Hefei National Laboratory for Physical Sciences at Microscale and School of Life Sciences, University of Science & Technology of China, Hefei 230027, ²National Chromatographic R&A Center, Dalian Institute of Chemical Physics, The Chinese Academy of Sciences, Dalian 116023, ³College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China and ⁴Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, Toronto, Ontario, Canada M5S 3E1

Received July 24, 2009; Revised and Accepted September 4, 2009

ABSTRACT

During cell division/mitosis, a specific subset of proteins is spatially and temporally assembled into protein super complexes in three distinct regions, i.e. centrosome/spindle pole, kinetochore/centromere and midbody/cleavage furrow/phragmoplast/bud neck, and modulates cell division process faithfully. Although many experimental efforts have been carried out to investigate the characteristics of these proteins, no integrated database was available. Here, we present the MiCroKit database (http://microkit.biocuckoo.org) of proteins that localize in midbody, centrosome and/or kinetochore. We collected into the MiCroKit database experimentally verified microkit proteins from the scientific literature that have unambiguous supportive evidence for subcellular localization under fluorescent microscope. The current version of MiCroKit 3.0 provides detailed information for 1489 microkit proteins from seven model organisms, including Saccharomyces cerevisiae, Schizasaccharomyces pombe, Caenorhabditis elegans, Drosophila melanogaster, Xenopus laevis, Mus musculus and Homo sapiens. Moreover, the orthologous information was provided for these microkit proteins, and could be a useful resource for further experimental identification. The online service of MiCroKit database was implemented in PHP + MySQL + JavaScript, while the local packages were developed in JAVA 1.5 (J2SE 5.0).

INTRODUCTION

Mitosis, also called as cell division, is the most crucial and fundamental affair of a eukaryotic cell cycle (1), separating and distributing the sister chromatids into two daughter cells equally and faithfully. During cell division, numerous proteins spatially and temporally organize protein super-complexes at the three distinct regions of centrosome/spindle pole body (2–9), kinetochore/centromere (10–17) and cleavage furrow/midbody (18–22), and orchestrate the accomplishment of cell division process. The related or homolog structures of midbody in plants and budding yeast are called as phragmoplast (21) and bud neck (23), respectively.

The centrosome of animal cells, spindle pole body in budding yeast, and related/homolog structures in other organisms share a conserved function to nucleate and organize microtubules, serving as the major MicroTubule-Organizing Centre (MTOC) (2–9). Besides essential functions in mitosis, the centrosome/MTOC also plays important roles in formation of primary cilia (8), fertilization (6) and intracellular trafficking (6). Aberrant organization of centrosome is associated with the dysfunction of cell division and chromosomal aneuploidy, which is implicated in tumorigenesis (2–5). In human, many centrosomal proteins are also involved in genetic diseases (9). Thus, comprehensive identification of centrosomal proteins will be the foundation of understanding the molecular regulatory mechanisms of this organelle and provide potentially important drug targets.

During mitosis and meiosis, a proteinaceous super-complex of kinetochore is assembled on centromeric DNA/centromere in eukaryotes, mediating the attachment and segregation of chromosome through microtubule of
Xenopus laevis we manually collected experimentally identified microkit centrosome and/or kinetochore. From scientific literature, defined a microkit protein that localizes in midbody, to be localized on centrosome, kinetochore or midbody, midbody proteins.

At the last stage of cell division, cytokinesis is crucial for partitioning and distributing intracellular contents into two independent daughter cells (18–22). In animals, an actomyosin-based contractile ring has emerged at the dividing site/cleavage furrow (23), while its similar homolog structure in budding yeast is bud neck (23). Numerous proteins compose a dense complex defined as midbody beneath the cleavage furrow (18–22), while the nearby bi-flanking regions of midbody are called as intracellular bridges (23). Then these cellular structures mediate ingestion and scission of the endo-membrane furrow. Contrast to in animals, in higher plants Golgi-derived vesicles are transported to the equatorial region from the seven organisms. Taken together, the MiCroKit database.

Although many proteins were experimentally identified to be localized on centrosome, kinetochore or midbody, an integrated resource was still not available. First, we defined a microkit protein that localizes in midbody, centrosome and/or kinetochore. From scientific literature, we manually collected experimentally identified microkit proteins from two fungi (Saccharomyces cerevisiae and Schizasaccharomyces pombe) and five animals, including Caenorhabditis elegans, Drosophila melanogaster, Xenopus laevis, Mus musculus and Homo sapiens. To guarantee the quality of data, these proteins were unambiguously observed under fluorescent microscope as directly supportive evidences. Then an integrated and searchable database of MiCroKit—midbody, centrosome and kinetochore was established. The online service and local packages were provided and implemented in PHP + MySQL + JavaScript and JAVA 1.5 (J2SE 5.0), respectively. Currently, the MiCroKit 3.0 contains 1489 unique proteins, and will be regularly updated as new microkit proteins are reported. Furthermore, with previously established approaches (24–26), we computationally detected potentially orthologous hits for these microkit proteins among the seven model organisms. Taken together, the MiCroKit database could be an integrated resource and provide useful information for further experimental identifications.

CONSTRUCTION AND CONTENT

With the aim of a high-quality curated database, we manually collected the proteins localized on midbody, centrosome and/or kinetochore (microkit proteins) from over 8000 scientific articles in PubMed (before 12 June 2009). Due to the information limitation, we only collected microkit proteins from two fungi (S. cerevisiae and S. pombe) and five animals, including C. elegans, D. melanogaster, X. laevis, M. musculus and H. sapiens. In plants, there were only a dozen of proteins identified to be localized on kinetochore (27). In this regard, although information of plant microkit proteins might also be useful, we did not include the very limited data in MiCroKit database.

To search the midbody proteins, we adopted the keywords ‘midbody’, ‘cleavage furrow’, ‘intracellular bridge’ and ‘contractile ring’ to query the PubMed, since all of the four structures are located at the dividing site of the cell. And for S. cerevisiae, we additionally used the term ‘bud neck’. Whereas, to query the centrosomal proteins, we chose the terms ‘centrosome’, ‘centriole’, ‘microtubule-organizing centre’, ‘MTOC’ and ‘centrosomal’. We also used the keyword ‘spindle pole’ to search the related information in S. cerevisiae. In addition, for kinetochore proteins, we employed the terms ‘kinetochore’, ‘centromere’ and ‘centromeric’ for querying. Totally, we collected 1493 microkit proteins from the seven organisms.

After all microkit proteins with unambiguous localization information were collected, we searched the UniProt Knowledgebase (28) to obtain protein sequences and related annotation information. The theoretical $I_p$ (isoelectric point) and $M_w$ (molecular weight) were calculated for each microkit protein (http://www.expasy.org/tools/pi_tool.html) (29,30). Furthermore, the orthologous information was provided. The pairwise orthologous information was determined with the InParanoid program (24,25), while the orthologous group information was further computed based on similar approaches in Clusters of Orthologous Groups of proteins (COGs) (26). The orthologous information was manually checked. Finally, we detected 802 orthologous groups, including 1264 microkit proteins and 2694 unidentified proteins.

The MiCroKit 3.0 database was constructed as an integrated and useful resource, while the online service and local packages were implemented in PHP + MySQL + JavaScript and JAVA 1.5 (J2SE 5.0), separately. The online documentation and a user manual were also provided.

USAGE

The MiCroKit 3.0 database was developed in an easy-to-use mode. The search option (http://microkit.biocuckoo.org/search.php) provides an interface for querying the MiCroKit 3.0 database with one or several keywords or accession numbers (UniProt ID or MiCroKit ID). For example, if the keyword of ‘survivin’ is inputted and submitted (Figure 1A), the results will be shown in a tabular format, with the features of MiCroKit ID, UniProt accession number and protein/gene names/aliases (Figure 1B). By clicking on the MiCroKit ID (MCK-HS-00222), the detailed information for human Survivin protein will be shown (Figure 1C). MiCroKit database supports the sequence information (both protein and nucleotide sequence), Gene Ontology annotation, domain organization, molecular weight, computed/theoretical pI and related references of the protein.
The orthologous information for human Survivin is also provided (Figure 1C).

Furthermore, we provided five additional advance options, including (i) advance search, (ii) browse, (iii) BLAST search, (iv) orthologous group browse and (v) pairwise orthologous browse (Figure 2).

(i) **Advance search.** In this option, users could use relatively complex and combined keywords to locate the precise information, with up to three search terms. The interface of search-engine permits the querying by different database fields and the linking of queries through three operators of ‘and’, ‘or’ and ‘exclude’ (Figure 2A). (ii) **Browse.** Instead of searching for a specific protein, all entries of MiCroKit database could be listed either by species name and/or subcellular localization information (Figure 2B). (iii) **BLAST search.** This option was designed for the propose of finding the related information in MiCroKit database quickly. The blastall program of NCBI BLAST packages (31) was included in MiCroKit 3.0 database (Figure 2B). Users could input a protein sequence in FASTA format for searching identical or homologous proteins. (iv) **Orthologous group browse.** Users could browse the pre-calculated orthologous group information by including or excluding one or several species (Figure 2C). Two examples were provided for this option. (v) **Pairwise orthologous browse.** Users could specifically browse the orthologous information between any two different species (Figure 2C). For example, by clicking on the ‘Submit’ button with default parameters, the orthologous information (identity \(\geq 20\%\)) between *M. musculus* and *H. sapiens* will be shown, with gene names and detailed results of score, E-value, identities and positives from BLAST (Figure 2D).

**RESULTS AND DISCUSSION**

As the first integrated database for proteins localized on midbody, centrosome and/or kinetochore, MiCroKit 3.0 database (Figure 2B). Users could input a protein sequence in FASTA format for searching identical or homologous proteins. (iv) **Orthologous group browse.** Users could browse the pre-calculated orthologous group information by including or excluding one or several species (Figure 2C). Two examples were provided for this option. (v) **Pairwise orthologous browse.** Users could specifically browse the orthologous information between any two different species (Figure 2C). For example, by clicking on the ‘Submit’ button with default parameters, the orthologous information (identity \(\geq 20\%\)) between *M. musculus* and *H. sapiens* will be shown, with gene names and detailed results of score, E-value, identities and positives from BLAST (Figure 2D).
contains 1489 microkit proteins, including 265, 149, 94, 111, 61, 132 and 677 entries in S. cerevisiae, S. pombe, C. elegans, D. melanogaster, X. laevis, M. musculus and H. sapiens, respectively. Previously, there were several proteomic-scale identifications of the potential centrosomal (7) and midbody (22) proteins carried out in human. These results provided a useful reservoir for further experimental verification. Recently, Nogales-Cadenas et al. collected 108 genes identified from the large-scale survey of Anderson et al. (7), and developed a human centrosomal proteins database of CentrosomeDB (32). Also, the human centrosomal proteins in MiCroKit 2.0 were also integrated into CentrosomeDB (32). However, in MiCroKit 3.0, we did not include the potential candidates from the large-scale experiments (7,22), before the proteins were observed under fluorescent microscope with unambiguous localization.

For statistics of the distribution of localizations, we counted the number of proteins classified by subcellular localizations for each organism, respectively (Figure 3). Obviously, our results exhibited that more of the efforts were performed in S. cerevisiae and H. sapiens rather than other organisms. In this regard, we still poorly understand the molecular compositions of midbody, centrosome, or kinetochore in other model organisms. More interestingly, some proteins could have multi-localizations. For example, human Survivin protein (MCK-HS-00222) was identified to be localized on midbody (33), centrosome (34) and kinetochore (35). In S. cerevisiae, there were only 31 (11.7%, 31/265) proteins with more than one localization, while there were 147 (21.7%, 146/677) human proteins with more than one location (P < 0.003, Fisher’s Exact Test, two-tailed). Then an interesting question has emerged that whether proteins could change their profiles of sub-cellular localizations during evolution. For example, the ortholog of human Survivin

Figure 2. Five advance options in MiCroKit 3.0. (A) Advance search allows users to input up to three terms for querying; (B) browse and BLAST search; (C) orthologous group browse and Pairwise orthologous browse; (D) the default example of pairwise orthologous browse. The orthologous information (identity ≥ 20%) between M. musculus and H. sapiens will be shown in details. The proteins collected in the MiCroKit database are marked in bold.

Figure 3. The statistics of localization distributions of microkit proteins from seven organisms, separately.
protein in S. cerevisiae is Bir1 (MCK-SC-00111), which was reported to be localized on kinetochore solely. Can we explain this phenomenon only by the reason of limited information? Or does some protein really get additional functions during evolution that could be localized on more subcellular localizations to play more roles? Further experimental identifications might be necessary to address this question. In addition, since numerous proteins have multi-localizations, these proteins might play important roles to mediate the crosstalk and communication of the three complex structures. Again, this hypothesis still remained to be experimentally dissected.

Taken together, although MiCroKit 3.0 database contains 1493 proteins, we are still far from fully understanding the molecular compositions and regulatory mechanisms of the three complex structures of midbody, centrosome and kinetochore. As an integrated resource, MiCroKit database could be useful for further experimental consideration. Since many novel components still remain to be identified, MiCroKit database will be updated routinely to keep up with the experimental discoveries.

ACKNOWLEDGEMENTS
The authors thank Dr Michael Galperin for his constructive and helpful suggestions.

FUNDING
Funding for open access charge: National Basic Research Program (973 project) (2002CB713700, 2006CB933300, 2007CB914503), National Natural Science Foundation of China (90919001, 30700138, 30900835, 30830036, 30721002, 30871236, 20065022 and 90713017), Chinese Academy of Sciences (KSCX2-YW-R-139, KSCX1-YW-R65, INFO-115-C01-SDB4-36), the China High Technology Research Program (2008ZX1002-020), the Cultivation Fund of the Ministry of Education of China (NO706035) and National Science Foundation for Post-doctoral Scientists (20080430100). Canadian Institutes of Health Research funding to Z.Z.

Conflict of interest statement. None declared.

REFERENCES
