PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2012, 39(4): 359~367
www.pibb.ac.cn

An Integrated Analysis of Lineage-specific Small Proteins Across Eight Eukaryotes Reveals Functional and Evolutionary Significance*

ZHAO Qian[1,2], XIAO Jing-Fa[1,2)**], YU Jun[1,2)**]
([1]) *Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing* 100029, China;
[2]) *Graduate School of Chinese Academy of Sciences, Beijing* 100049, China)

Abbreviations: SPs, small proteins; ORFs, open reading frames; sORFs, short ORFs; ERPs, evolutional related proteins; VNM, vertebrate but not mammal.

Abstract Small proteins (< 100 amino acids) are prevalent in all three domains of life. Earlier studies have been focusing on a limited number of small protein families in specific organisms and developing genome-wide algorithms to identify short open-reading-frames or sORFs. Here the *in silico* analyses on small proteins (SPs) include both known SPs and genes with sORFs. RefSeq proteins that shorter than 100 amino acids in length are defined as SPs and are grouped according to their sequence conservation within lineages of eukaryotes, vertebrates, and mammals. Biological roles of the grouped SPs are found basically performing lineage-specific functions. Tissue-specificity of human SPs are also investigated and showed that a majority of the human-specific SPs are tissue-specific and that most of the human SPs originated after the split of vertebrates and invertebrates are mostly universally expressed. In addition, the results indicated that some of the eukaryotic SPs perform lineage-specific functions and they evolve and express in certain unique ways.

Key words eukaryotic small proteins, selective pressure, lineage-specific, tissue-specific expression
DOI: 10.3724/SP.J.1206.2011.00290

Small proteins (< 100 amino acids in length), representing an untapped source of important biological information, exist ubiquitously in the three domains of life. Some of the known SPs include a number of important functional classes, such as mating pheromones and proteins involved in energy metabolism, proteolipids, chaperonins, stress proteins, transporters, transcriptional regulators, nucleases, ribosomal proteins, thioredoxins, and metal ion chelators [1]. Among multicellular organisms, a rich diversity of short polypeptides has been investigated, including peptide hormones, antibacterial defensins, cecropins, and magainins [1]. Moreover, SPs provide simple model systems to study deterministic elements of protein folding and stability [2] and can serve as candidates for novel drug design and screening [3]. In addition, microbial SPs play important roles in the response to specific biotic and abiotic stresses [4–7], indicating that they are subject to strong natural selection.

Although SPs are biologically important, computational and experimental challenges still exist in studying their structure, evolution, and function. First, computational identification of sORFs (short open-reading-frames) is difficult because they are

*This work was supported by grants from The National Natural Science Foundation of China (31071163) and National Basic Research Program of China (2010CB126604).
**Corresponding author.
Xiao Jing-Fa. Tel: 86-10-82995384, E-mail: xiaojingfa@big.ac.cn
Yu Jun. Tel: 86-10-82995357, E-mail: junyu@big.ac.cn
Received: June 27, 2011 Accepted: July 29, 2011

mingled among abundant sORFs that are supported by evidence for transcription but missed by *ab initio* prediction [6, 8-9]. Second, sORFs are not favourable targets for random mutagenesis [6], and its functional study is more difficult than larger proteins with regards to routine biochemical assays and molecular methods. However, high-throughput technologies, such as expression-based analysis [10-14], gene-trapping [15], and homology searching [10-11, 13, 15-18] have helped in solving these problems. Tuskan and his colleagues used genomics, proteomics, and computational approaches to discover and annotate SPs of *Populus deltoids* [14]; Shiu *et al*. identified many novel small coding ORFs in *Arabidopsis thaliana* genome, and suggested that these novel sORFs are transcribed and/or under purifying selection [19]. In addition, several algorithms are developed to predict sORFs in more reliable ways [14, 20].

Although large-scale studies of SPs and/or sORFs have drawn some attentions recently, studies focusing on functional significance in an evolutionary context of lineage-specific SPs have yet to be accounted for. In this study, we selected eight well-studied eukaryotes, including a fungus (*Saccharomyces cerevisiae*), a worm (*Caenorhabditis elegans*), an insect (*Drosophila melanogaster*), a bony fish (*Danio rerio*), a bird (*Gallus gallus*), and three mammals (*Bos taurus*, *Mus musculus*, and *Homo sapiens*), and characterized their SPs in terms of sequence conservation, functional classification in lineage-specific ways, and tissue-distributions. We developed very stringent sequence alignment criteria and assembled different datasets for evolutionary and functional studies. Our results demonstrated that SPs play important roles in the evolution of eukaryotes.

1 Materials and methods

1.1 Datasets

We downloaded the RefSeq proteins of eight eukaryotes (*Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Danio rerio*, *Gallus gallus*, *Bos taurus*, *Mus musculus* and *Homo sapiens*) from NCBI (release 34, ftp://ftp.ncbi.nih.gov/refseq/) and retrieved human EST(expressed sequence tag) data from dbEST regarding hg19 of UCSC's databases [21], which contain about 8 million human ESTs from over 200 human tissue types.

1.2 Conservation of SPs

For each species, we selected RefSeq proteins that are less than 100 amino acids (aa) in size as SPs and used BLAST [22] to search against the RefSeq protein collection. Then we applied high-stringency criteria (Table 1) to ensure that SPs and their matched proteins share high levels of similarity (identity reflects the ratio of exact matches to total alignment length; the ratio of alignment length to query length ensures matched proteins preserve the majority part of query SPs; the ratio of alignment length to subject length helps in identifying matched proteins that share a similar length with query SPs). To make sure that the three essential conserved groups are biologically meaningful, we set a moderate-stringency and unconditional criteria to facilitate the screening process (Table 1). The SPs whose alignments are not shared exclusively by 8 eukaryotes, 5 vertebrates, or 3 mammals were excluded. The shared data are the most reliable SPs and evolutionarily conserved among eukaryotic lineages. Moreover, using unconditional criteria, we defined SPs unique to a single species as species-specific.

Table 1 Three criteria to screen BLAST result

	Ratio of alignment length to query length	Ratio of alignment length to subject length	Identity
High-stringency criteria (0.8_0.6_40)	> 0.8	> 0.6	> 40
Moderate-stringency criteria (0.5_0.3_40)	> 0.5	> 0.3	> 40
Unconditional criteria (0_0_0)	> 0	> 0	> 0

1.3 Annotation of SPs

According to our definition for lineage-specificity, we classified SPs and their homologs (according to the high-stringency criteria) into groups. Furthermore, if two SPs have the same homolog in common, we merged the two homolog groups into one. All of the RefSeq-listed proteins in each homolog group were considered to have similar functions and the functions of each homology group were derived from DAVID (Database for Annotation, Visualization and Integrated Discovery) [23] (DAVID Bioinformatics Resources 6.7). We also attempted to simplify InterPro annotation [24] of these homology groups by merging what share a common function term into one function cluster.

1.4 Tissue-associated expression of human SPs

Although the total number of human EST seems enormous, this is still considered a poor sampling of human transcriptomes. In addition, most of the available tissue samples are anatomically heterogeneous, and precise tissue definition requires the advancement of micro-dissection tools and single-cell techniques, which were not available to apply to the present RNA sample preparation. Therefore, we consolidated the same organ or tissue samples into a single sample to enlarge the sampling depth and avoid redundancy with the assistance of MeSH (September 1, 2009 update). This procedure yielded 29 tissue categories that are frequently studied and contain more data. We selected 24 better-studied tissues, and each seemed to express more than 150 SPs.

After extracting human RefSeq transcripts and protein-transcript relationship files from NCBI RefSeq (release 34, ftp://ftp.ncbi.nih.gov/refseq/), we assigned EST to support tissue expression based on the following four steps: (i) excluding RefSeq transcripts aligned to hg19 genome sequences with an identity of < 95%; (ii) excluding ESTs aligned to genome sequences with an identity of < 95%; (iii) excluding ESTs that are not uniquely mapped to re-annotated RefSeq transcripts; and (iv) assigning tissue types for human SPs based on EST data.

1.5 The evolutionary origin of SPs

For each representative species, we performed BLAST alignments for its SPs in a one-against-all fashion against all RefSeq proteins from organisms in our collection in an "evolutionary order" (e.g., *S. cerevisiae*, *C. elegans*, and *D. melanogaster* exist before *D. rerio* diverged from their common ancester), and a significant match must have an *E*-value < 10^{-5}. We found an average of 24.9% of total RefSeq proteins that have significant alignments with SPs are at least 25 aa longer than their aligned SPs.

We also performed BLAST alignment between 74 SPs that conserved in all 8 eukaryotes and 180 879 prokaryotic RefSeq proteins (length < 100 amino acid), which were compiled from our previous work (*E*-value < 10^{-5}, identity > 60%)[25]. Mega4[26] was used for multiple sequence alignments and phylogeny.

2 Results

2.1 An overall profile of eukaryotic SPs in eight eukaryotes

We can summarize the overall profile of eukaryotic SPs as follows. First, the percentage of SPs in total RefSeq proteins was higher in invertebrates (about 5%) than that in vertebrates (about 2%, Figure 1a). The difference is quite small as compared to 10.99% reported for bacterial and archaeal SPs [25]. Second, the numbers of SPs in invertebrates, especially in multicellular invertebrates, were greater than those among vertebrates in general (316 in *S. Cerevisiae*, 1 395 in *C. elegans*, 795 in *D. melanogaster*, 255 in *D. rerio*, 103 in *G. gallus*, 307 in *B. Taurus*, 583 in *M. musculus*, and 794 in *H. Sapiens*). These two observations suggest that SPs may be unique to different eukaryotic lineages. Third, when the distribution of amino acids was compared between SPs and total RefSeq-defined proteins in the analyzed eukaryotes (Figure 1b), two biased groups were noticed: (i) M (methionine) and C (cysteine) exhibit a stronger usage bias preferences between SPs and the control (Wilcoxon $P < 0.001$) and (ii) S (serine), H (histidine), D (Aspartic acid), and K (lysine) have a moderate usage biases (Wilcoxon *P*-value < 0.01). Because M (methionine) is usually the first residue on translational grounds, it is expected to have a higher occurrence in shorter proteins, and the usage bias of C (cysteine) has just supported our understanding that generally the structure of SPs are more stable than large proteins, because disulfide bonds provide extra stability (such as conotoxins) [19]. Moreover, usage biases of other amino acids (S, H, D, and K) also provide a clue that SPs are a specialized group of proteins in terms of their molecular origins and structures.

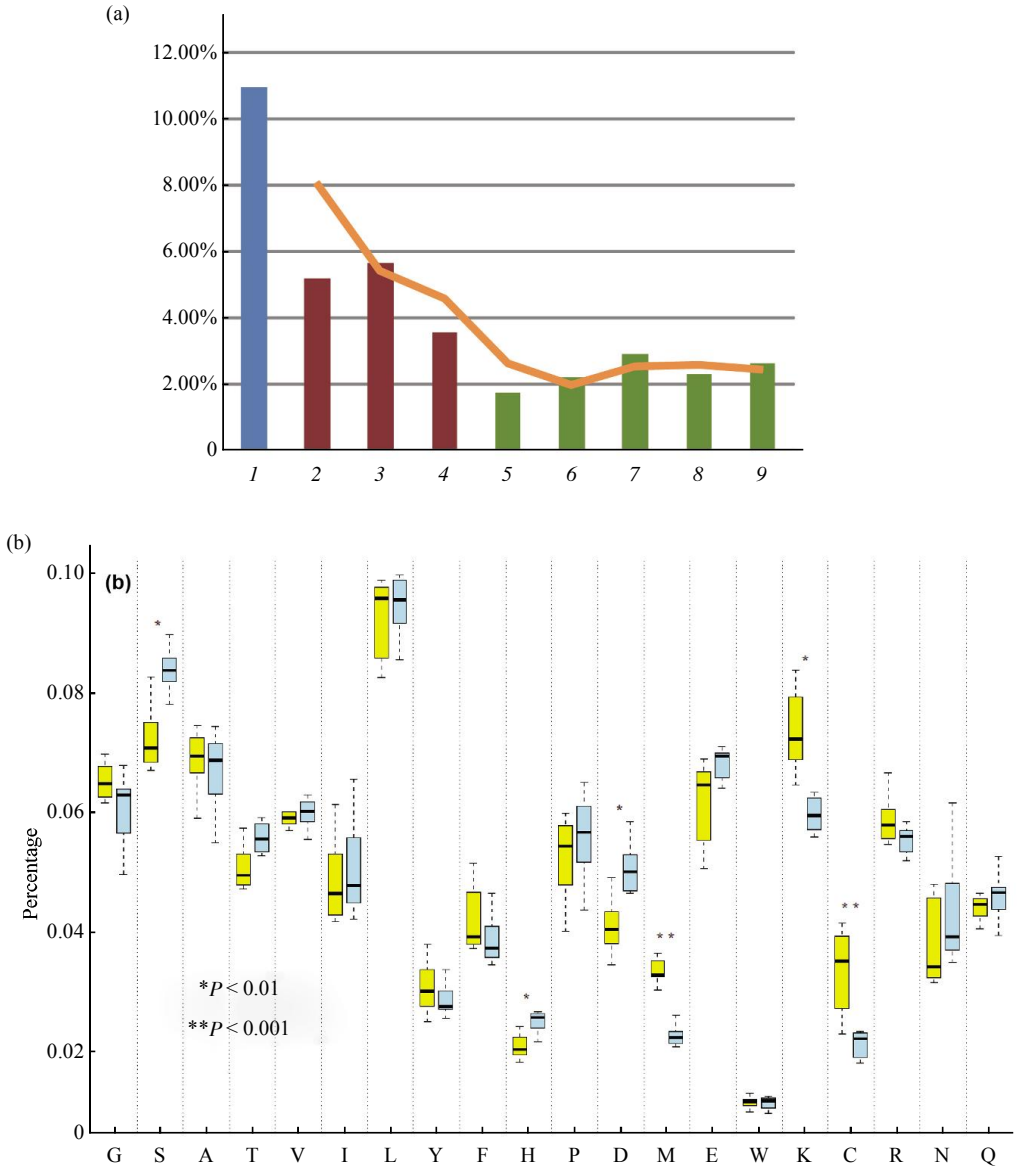

Fig. 1　The profile of eukaryotic SPs

(a) The percentages of SPs in the total RefSeq proteins of 8 eukaryotic species. The species are color-coded: blue for Bacteria and Archaea, red for invertebrates (*S. cerevisiae*, *C. elegans*, and *D. melanogaster*), and green for vertebrates (*D. rerio*, *B. taurus*, *G. gallus*, *M. musculus*, and *H. sapiens*). A curve was simulated to exhibit the obvious trend. *1*: *Bacteria and archaea*; *2*: *Saccharomyces cerevisiae*; *3*: *Caenorhabditis elegans*; *4*: *Drosophila melanogaster*; *5*: *Danio rerio*; *6*: *Gallus gallus*; *7*: *Bos taurus*; *8*: *Mus musculus*; *9*: *Homo sapiens*. (b) Comparison of the amino acid distributions of SPs and total RefSeq proteins. Yellow indicates SPs, whereas blue indicates all RefSeq proteins.

2.2　The functional and evolutionary significance of SPs

We defined the conservation of SPs across three categories: conserved in (i) all 8 representative eukaryotic species (74); (ii) only 5 vertebrates (102), and (iii) only 3 mammals (123). Moreover, we used unconditional criteria to select SPs that failed to align with sequences in other species, and defined these proteins as species-specific (Table 2). We found that these species-specific SPs are much more abundant than the conserved SPs, possibly because organisms tend to enrich abundant SPs to perform specialized functions.

Table 2 Statistics of small proteins identified by conservation study

	Conserved in 8 species	Conserved in 5 vertebrates	Conserved in 3 mammals	Species specific
Saccharomyces cerevisiae	7	–	–	258
Caenorhabditis elegans	8	–	–	1238
Drosophila melanogaster	14	–	–	568
Danio rerio	10	17	–	40
Gallus gallus	7	12	–	25
Bos taurus	8	17	33	27
Mus musculus	8	27	43	104
Homo sapiens	12	29	47	159

Four categories of small proteins were identified, and the count of each classified small proteins was shown.

We conducted a functional analysis of the three conserved groups to uncover their major roles in different eukaryotic lineages. We generated homolog groups in the 3 conserved groups, yielding 133 in total (9 in all 8 species, 26 in 5 vertebrates, and 98 in 3 mammals). We also compiled 6, 13, and 30 functional clusters shared by 8 species, 5 vertebrates, and 3 mammals, respectively. We summarized the top 5 largest functional clusters for each conserved group in Figure 2. SPs conserved among all 8 eukaryotic species include histone H4 [27], like-Sm ribonucleoprotein [28–29], ribosomal proteins, ubiquitin[30], and Acyl-CoA-binding protein [31]. These proteins are known to play essential roles in nucleo-compartmentalization, protein synthesis, and post-translational modification (Figure 2a). SPs conserved in vertebrates are S100 [32], CaBP-9k [33],

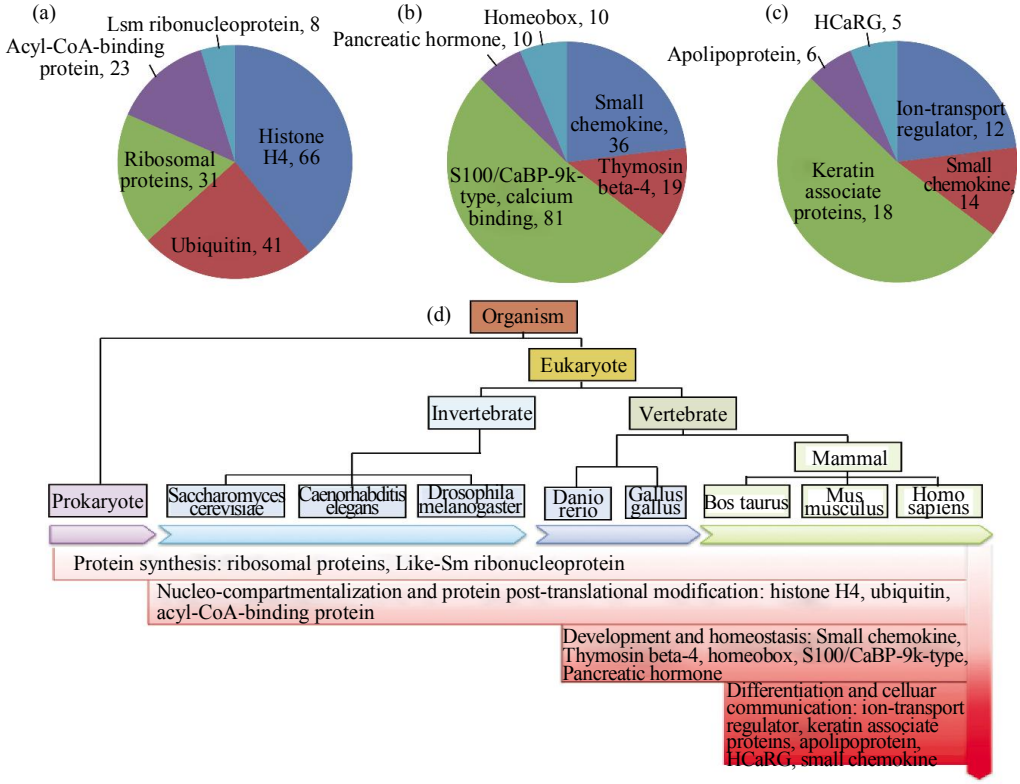

Fig. 2 Functional study of SPs

Numbers are SPs counts. The total counts of the functional clusters of related proteins were summed up to 100% in the pie chart. (a) Top 5 functional SP clusters conserved among all 8 eukaryotic species. (b) Top 5 functional SP clusters conserved only among 5 vertebrate species. (c) Top 5 functional SP clusters of s that were conserved among all 8 eukaryotic species. (d) The schema displays life evolving from simple to complex (from left to right) and time (arrows). The red downward arrow indicates increasing functional complexity (right). Each step corresponds to an important evolutionary leap forward. At each step, novel SPs correlate with lineage-specific functions (as listed in the frames).

homeobox[34], thymosin beta-4[35], small chemokine[36–38], and pancreatic hormone [39], and they are related to development and homeostasis (Figure 2b). The mammal-specific SPs are FXYD protein [40], keratin-associated protein[41–42], apolipoprotein[43–44], HCaRG[45], and small chemokines [36–38], and they are related to differentiation and cellular communication (Figure 2c). Interestingly, we found all 3 groups of conserved SPs perform lineage-specific functions and this result indicates SPs are functionally important and thus selected during the eukaryotic evolution (Figure 2d).

2.3 The expression of human SPs

Evolutionarily ancient genes are known to be mostly housekeeping and universally expressed [46–47]. To understand the expression of SPs in relation to their evolutionary ages, we compiled human dbEST data for evidence of transcription and classified their cDNA libraries into 24 integrated tissue types, yielding 524 (66.0%) human SPs in 4 categories according to their phyletic distribution: (i) 77 (9.7%) of SPs do not have any ERP (evolutionary related proteins, defined as BLAST-matched SPs from other species with *E*-values $<10^{-5}$) as unique to human; (ii) 128 (16.1%) have ERPs unique to mammals; (iii) 139 (17.5%) have at least one mammalian ERP and one VNM (indicates SPs that are vertebrate-specific but not mammal-specific) ERP but did not have any invertebrate ERP (indicates SPs that originated after the split of non-VNM and VNM); and (iv) 180 (22.7%) had at least one mammalian ERP, one VNM ERP, and one invertebrate ERP (indicates their origin after the split of invertebrates and vertebrates).

We found that ancient human SPs tended to be widely expressed in the 24 tissues (Figure 3a). Of the human SPs, 55.8% human-specific and 6.1% of invertebrate-originated were found expressed in less than 5 tissues, but 9.4% human-specific and 42.2% of invertebrate- originated were expressed in more than 20 tissues. Furthermore, human SPs originating after the split of non-VNM and VNM and after the split of non-mammals and mammals were expressed with modest tissue specificity (Figure 3b).

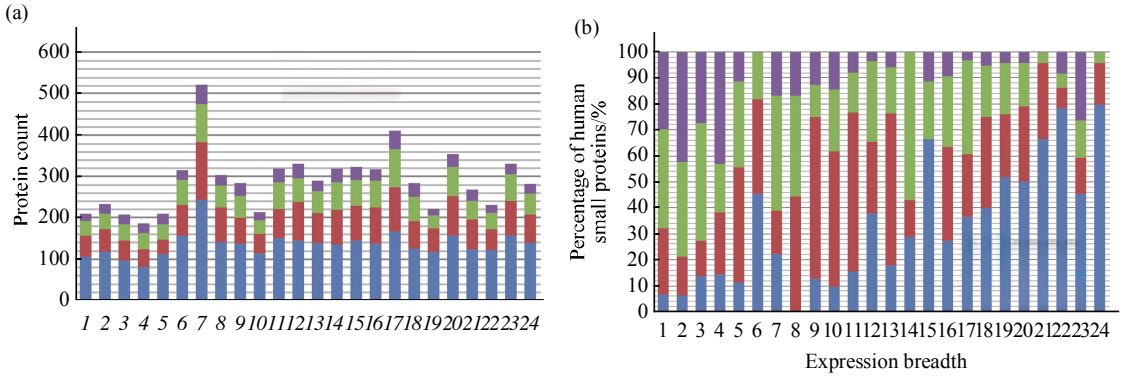

Fig. 3 The expression of human SPs in 24 tissues

(a) Tissues and protein counts in 24 well-studied human tissues. *1*: Heart; *2*: Thyroid; *3*: Thymus; *4*: Spleen; *5*: Ovary; *6*: Muscle; *7*: Intestine; *8*: Lymphnode; *9*: Bone_marrow; *10*: Blood; *11*: Prostate; *12*: Liver; *13*: Kidney; *14*: Uterus; *15*: Placenta; *16*: Lung; *17*: Brain; *18*: Testis; *19*: Stomach; *20*: Breast; *21*: Bone; *22*: Adrenal_gland; *23*: Stem_cell; *24*: Nervous_system. (b) 524 human SPs were classified into 4 phyletic categories to approximately represent different evolutionary origins. More ancient SPs tended to be expressed in a broader range of tissues. ■: Human specific; ■: Originated from mammal; ■: Originated from VNM; ■: Originated from invertebrate.

3 Discussion

3.1 The origin of SPs

New genes emerge as consequences of gene genesis, mutation, and horizontal transfer, among others[14]. To investigate the genesis of eukaryotic SPs in general, we used BLAST alignment to assess evolutionary processes that generate novel SPs among 8 species representing a large evolutionary timescale. For each species, we found approximately 25% of total RefSeq proteins are at least 25aa (25aa was considered to be the minimum domain length) longer than SPs. Because domains are functional and evolutionary units of proteins and most SPs contain only one domain, we speculate that the majority of eukaryotic SPs emerged and evolved independently as individual domains

rather than through complex processes such as sequence mutations. This speculation agrees with our previous findings based on a survey of prokaryotic mini-proteins[25].

Furthermore, though SPs that conserved in the lineages of vertebrates and mammals are not only lineage-specific but also functionally relevant, the origin of 74 SPs that conserved in all 8 eukaryotes deserves further discussion. We conducted a thorough search for these proteins against all bacterial and archaeal SPs. The BLAST result showed that 6 out of 8 homolog groups were conserved in either the eukaryote-bacterial or the eukaryote-archaeal groups. Because horizontal gene transfer is prevalent among unicellular organisms, we next considered the possibility of gene transfer between prokaryotes and unicellular eukaryotes. According to the phylogenetic analyses (data not shown) we speculate: (i) the Acyl-CoA-binding protein has probably undergone horizontal gene transfer and being acquired by bacteria, which is also proved by Knudsen and his colleagues [48]; (ii) ribosomal proteins and Like-Sm ribonucleoprotein probably originated from Archaea; and (iii) Archaea and eukaryotes shared more ribosomal proteins than the same groups shared with bacteria. We further speculate that bacteria separated from a common ancestor much earlier than Archaea, even before the translation machinery had reached maturity. Our speculations based on phylogenetic analyses support Carl Woese's three-domain system[49]. We also performed a thorough comparison between bacterial and archaeal SPs (data not shown), and our results suggest that these proteins experienced frequent horizontal gene transfer events, except in the case of the Gas vesicle protein (GvpA, IPR000638 and IPR018493).

3.2 The lineage specificity of SPs

In our study, newly emerged SPs by definition are lineage-specific, and each of the new classes correlates with new features of the specific lineages according to our study (Figure 2d). Furthermore, we noticed that species-specific SPs are vital for microbial survival under environmental pressure [4–7]. Thus, we speculate that these SPs may arise for a variety of reasons. First, SPs are readily generated because organisms tend to minimize the cost of protein biosynthesis [50]. Second, because the majority of SPs contain only a single protein domain, they probably perform straightforward functions (not simple in terms of interacting with other more complex proteins) and through direct protein-protein interactions or binding to DNA/RNA sequences, which is particularly important for stimuli responding.

Another point is that ancient SPs are mostly structural proteins, such as histone H4, ribosomal proteins, and like-Sm ribonucleoprotein, whereas relatively young SPs are either homeostasis- or communication-related proteins; these functions are characteristics of both multicellularity and complex regulatory networks. In addition, the standard deviations of protein length in each lineage-specific function cluster are excellent indicators of evolutionary pressure (Figure 4).

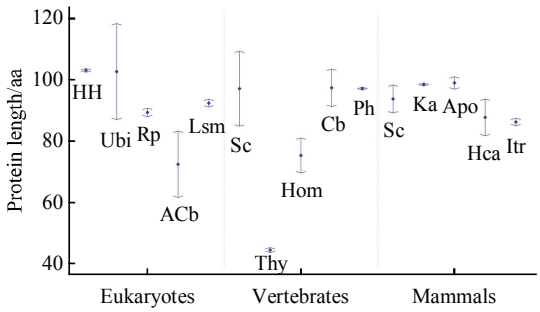

Fig. 4 The mean and standard variation of protein length in different function clusters

Columns separated by broken lines represent (from left to right) present our 3 conserved categories (conserved in eukaryotes, vertebrates, and mammals, respectively). Each data point shows the average length of related proteins in a function cluster. Standard deviations are indicated by arrows. The abbreviations used are: HH, Histone H4; Ubi, Ubiquitin; Rb, Ribosomal protein; ACb, Acyl-CoA-binding protein; Lsm, Like-Sm ribonucleoprotein; Sc, Small chemokine; Thy, Thymosin beta-4; Hom, Homeobox; Cb, S100/CaBP-9k-type calcium binding; Ph, Pancreatic hormone; Sc, Small chemokine; Ka, Keratin-associated proteins; Apo, Apolipoprotein; Hca, HcaRG; and Itr, Ion-transport regulator.

In summary, SPs are ancient; some may be relics of the RNA world and some are created new along with different lineages. In the prokaryotic domain of life, proteins are not as compact as what of eukaryotes, and therefore SPs are prevalent and readily reorganized, manifested as the phenomenon we know as genetic complementation. As SPs evolve, two destinies are obvious: folded into compact multifunctional proteins and select to remain as small as originally created. The former often disappear eventually and the latter should expand over time with new functionality. Since SPs perform simple functions,

they may be an easy class of proteins to be created *de novo* as compared to large proteins. Furthermore, such newly created proteins may be cell-specific or tissue-specific initially and then evolve to become universal as the functionality becomes essential to all cell types. Eukaryotic SPs should therefore be systematically identified, classified, and functionally characterized in a thorough manner.

References

[1] Basrai M A, Hieter P, Boeke J D. Small open reading frames: beautiful needles in the haystack. Genome Res, 1997, **7**(8): 768-771

[2] Imperiali B, Ottesen J J. Uniquely folded mini-protein motifs. J Pept Res, 1999, **54**(3): 177-184

[3] Martin L,Vita C. Engineering novel bioactive mini-proteins from small size natural and de novo designed scaffolds. Curr Protein Pept Sci, 2000, **1**(4): 403-430

[4] Wu W, Jin S. PtrB of Pseudomonas aeruginosa suppresses the type Ⅲ secretion system under the stress of DNA damage. J Bacteriol, 2005, **187**(17): 6058-6068

[5] Ha U H, Kim J, Badrane H, *et al*. An *in vivo* inducible gene of *Pseudomonas aeruginosa* encodes an anti-ExsA to suppress the type Ⅲ secretion system. Mol Microbiol, 2004, **54**(2): 307-320

[6] Kastenmayer J P, Ni L, Chu A, *et al*. Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. Genome Res, 2006, **16**(3): 365-373

[7] Setlow P. I will survive: DNA protection in bacterial spores. Trends Microbiol, 2007, **15**(4): 172-180

[8] Ghaemmaghami S, Huh W K, Bower K, *et al*. Global analysis of protein expression in yeast. Nature, 2003, **425**(6959): 737-741

[9] Huh W K, Falvo J V, Gerke L C, *et al*. Global analysis of protein localization in budding yeast. Nature, 2003, **425**(6959): 686-691

[10] Oshiro G, Wodicka L M, Washburn M P, *et al*. Parallel identification of new genes in *Saccharomyces cerevisiae*. Genome Res, 2002, **12**(8): 1210-1220

[11] Velculescu V E, Zhang L, Zhou W, *et al*. Characterization of the yeast transcriptome. Cell, 1997, **88**(2): 243-251

[12] Basrai M A, Hieter P. Transcriptome analysis of *Saccharomyces cerevisiae* using serial analysis of gene expression. Methods Enzymol, 2002, **350**: 414-444

[13] Kessler M M, Zeng Q, Hogan S, *et al*. Systematic discovery of new genes in the *Saccharomyces cerevisiae* genome. Genome Res, 2003, **13**(2): 264-271

[14] Yang X, Tschaplinski T J, Hurst G B, *et al*. Discovery and annotation of small proteins using genomics, proteomics, and computational approaches. Genome Res, 2011, **21**(4): 634-641

[15] Kumar A, Harrison P M, Cheung K H, *et al*. An integrated approach for finding overlooked genes in yeast. Nat Biotechnol, 2002, **20**(1): 58-63

[16] Souciet J, Aigle M, Artiguenave F, *et al*. Genomic exploration of the hemiascomycetous yeasts: 1. A set of yeast species for molecular evolution studies. FEBS Lett, 2000, **487**(1): 3-12

[17] Brachat S, Dietrich F S, Voegeli S, *et al*. Reinvestigation of the *Saccharomyces cerevisiae* genome annotation by comparison to the genome of a related fungus: *Ashbya gossypii*. Genome Biol, 2003, **4**(7): R45

[18] Cliften P, Sudarsanam P, Desikan A, *et al*. Finding functional features in *Saccharomyces genomes* by phylogenetic footprinting. Science, 2003, **301**(5629): 71-76

[19] Hanada K, Zhang X, Borevitz J O, *et al*. A large number of novel coding small open reading frames in the intergenic regions of the *Arabidopsis thaliana* genome are transcribed and/or under purifying selection. Genome Res, 2007, **17**(5): 632-640

[20] Hanada K, Akiyama K, Sakurai T, *et al*. sORF finder: a program package to identify small open reading frames with high coding potential. Bioinformatics, 2010, **26**(3): 399-400

[21] Kuhn R M, Karolchik D, Zweig A S, *et al*. The UCSC Genome Browser Database: update 2009. Nucleic Acids Res, 2009, **37**(Database issue): D755-761

[22] Altschul S F, Madden T L, Schaffer A A, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res, 1997, **25**(17): 3389-3402

[23] Dennis G, Jr. Sherman B T, Hosack D A, *et al*. DAVID: Database for annotation, visualization, and integrated discovery. Genome Biol, 2003, **4**(5): P3

[24] Hunter S, Apweiler R, Attwood T K, *et al*. InterPro: the integrative protein signature database. Nucleic Acids Res, 2009, **37**(Database issue): D211-215

[25] Wang F, Xiao J, Pan L, *et al*. A systematic survey of mini-proteins in bacteria and archaea. PLoS ONE, 2008, **3**(12): e4027

[26] Tamura K, Dudley J, Nei M, *et al*. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol, 2007, **24**(8): 1596-1599

[27] Thatcher T H, Gorovsky M A. Phylogenetic analysis of the core histones H2A, H2B, H3, and H4. Nucleic Acids Res, 1994, **22**(2): 174-179

[28] Kufel J, Allmang C, Petfalski E, *et al*. Lsm Proteins are required for normal processing and stability of ribosomal RNAs. J Biol Chem, 2003, **278**(4): 2147-2156

[29] He W, Parker R. Functions of Lsm proteins in mRNA degradation and splicing. Curr Opin Cell Biol, 2000, **12**(3): 346-350

[30] Ross C A, Pickart C M. The ubiquitin-proteasome pathway in Parkinson's disease and other neurodegenerative diseases. Trends Cell Biol, 2004, **14**(12): 703-711

[31] Rose T M, Schultz E R, Todaro G J. Molecular cloning of the gene for the yeast homolog (ACB) of diazepam binding inhibitor/endozepine/acyl-CoA-binding protein. Proc Natl Acad Sci USA, 1992, **89**(23): 11287-11291

[32] Rothermundt M, Ponath G, Arolt V. S100B in schizophrenic psychosis. Int Rev Neurobiol, 2004, **59**: 445-470

[33] Bronner F. Mechanisms of intestinal calcium absorption. J Cell Biochem, 2003, **88**(2): 387-393

[34] Alonso C R. Hox proteins: sculpting body parts by activating localized cell death. Curr Biol, 2002, **12**(22): R776-778

[35] Weeds A, Way M. Is thymosin-beta4 the missing link?. Curr Biol, 1991, **1**(5): 307−308

[36] Oppenheim J J, Zachariae C O, Mukaida N, *et al*. Properties of the novel proinflammatory supergene "intercrine" cytokine family. Annu Rev Immunol, 1991, **9**: 617−648

[37] Stoeckle M Y, Barker K A. Two burgeoning families of platelet factor 4-related proteins: mediators of the inflammatory response. New Biol, 1990, **2**(4): 313−323

[38] Wolpe S D, Cerami A. Macrophage inflammatory proteins 1 and 2: members of a novel superfamily of cytokines. Faseb J, 1989, **3**(14): 2565−2573

[39] Blundell T L, Humbel R E. Hormone families: pancreatic hormones and homologous growth factors. Nature, 1980, **287**(5785): 781−787

[40] Crambert G, Geering K. FXYD proteins: new tissue-specific regulators of the ubiquitous Na,K-ATPase. Sci STKE, 2003, **2003**(166): RE1

[41] Elleman T C. The amino acid sequence of protein SCMK-B2C from the high-sulphur fraction of wool keratin. Biochem J, 1972, **128**(5): 1229−1239

[42] Mitsui S, Ohuchi A, Adachi-Yamada T, *et al*. Structure and hair follicle-specific expression of genes encoding the rat high sulfur protein B2 family. Gene, 1998, **208**(2): 123−129

[43] Storjohann R, Rozek A, Sparrow J T, *et al*. Structure of a biologically active fragment of human serum apolipoprotein C-Ⅱ in the presence of sodium dodecyl sulfate and dodecylphosphocholine. Biochim Biophys Acta, 2000, **1486**(2−3): 253−264

[44] Lins L, Flore C, Chapelle L, *et al*. Lipid-interacting properties of the N-terminal domain of human apolipoprotein C-Ⅲ. Protein Eng, 2002, **15**(6): 513−520

[45] Solban N, Jia H P, Richard S, *et al*. HCaRG, a novel calcium-regulated gene coding for a nuclear protein, is potentially involved in the regulation of cell proliferation. J Biol Chem, 2000, **275**(41): 32234−32243

[46] Freilich S, Massingham T, Bhattacharyya S, *et al*. Relationship between the tissue-specificity of mouse gene expression and the evolutionary origin and function of the proteins. Genome Biol, 2005, **6**(7): R56

[47] Zhu J, He F, Hu S, *et al*. On the nature of human housekeeping genes. Trends Genet, 2008, **24**(10): 481−484

[48] Burton M, Rose T M, Faergeman N J, *et al*. Evolution of the acyl-CoA binding protein (ACBP). Biochem J, 2005, **392** (Pt 2): 299−307

[49] Woese C R, Kandler O, Wheelis M L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci USA, 1990, **87**(12): 4576−4579

[50] Seligmann H. Cost-minimization of amino acid usage. J Mol Evol, 2003, **56**(2): 151−161

基于 8 种真核生物的整合分析揭示种属特异性小蛋白的功能和进化特征 *

赵 倩[1, 2] 肖景发[1)**] 于 军[1)**]

([1)] 中国科学院北京基因组研究所，北京 100029；[2)] 中国科学院研究生院，北京 100049)

摘要 小蛋白 (< 100 个氨基酸) 广泛存在于三界生命中，具有重要生物功能. 早期涉及小蛋白的研究主要集中于少量特殊物种中的蛋白质家族，以及在全基因组尺度预测短小开放读码框(sORFs)的算法开发，但并无跨真核物种的大规模组学分析来揭示小蛋白的功能和进化特征. 通过对已知小蛋白和拥有短小开放读码框的基因进行全基因组尺度的计算分析，长度小于 100 个氨基酸的 RefSeq proteins 按照其序列保守性被划分为存在于所有 8 种真核生物、只存在于脊椎动物和只存在于哺乳动物三个进化分类中，此三个进化分类所对应的生物学功能揭示了小蛋白行使种属特异性功能的特征. 进一步研究发现，大多数人类特有的小蛋白也是组织表达特异性的，并且绝大多数古老的小蛋白在人体内普遍表达. 因此认为，一些真核小蛋白出现并在自然选择压力下富集，行使种属特异性功能，并且以特殊的方式进化和表达.

关键词 真核小蛋白，选择压力，种属特异性，组织特异性表达

学科分类号 Q3 **DOI**: 10.3724/SP.J.1206.2011.00290

* 国家自然科学基金(31071163) 和国家重点基础研究发展计划(973)(2010CB126604)资助项目.

** 通讯联系人.

肖景发. Tel: 010-82995384, E-mail: xiaojingfa@big.ac.cn

于 军. Tel: 010-82995357, E-mail: junyu@big.ac.cn

收稿日期：2011-06-27，接受日期：2011-07-29