FEBS 29966 FEBS Letters 579 (2005) 5443–5453

Elucidation of subfamily segregation and intramolecular coevolution of the olfactomedin-like proteins by comprehensive phylogenetic analysis and gene expression pattern assessment

Ling-Chun Zeng[a,b], Ze-Guang Han[b], Wei-Jun Ma[a,b,*]

[a] *Health Science Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences and Shanghai Second Medical University, 225 South Chongqing Road, Shanghai 200025, China*
[b] *Chinese National Human Genome Center at Shanghai, 351 Guo Shou-Jing Road, Shanghai 201203, China*

Received 18 May 2005; revised 20 August 2005; accepted 23 August 2005

Available online 23 September 2005

Edited by Takashi Gojobori

Abstract The categorization of genes by structural distinctions relevant to biological characteristics is very important for understanding of gene functions and predicting functional implications of uncharacterized genes. It was absolutely necessary to deploy an effective and efficient strategy to deal with the complexity of the large olfactomedin-like (OLF) gene family sharing sequence similarity but playing diversified roles in many important biological processes, as the simple highest-hit homology analysis gave incomprehensive results and led to inappropriate annotation for some uncharacterized OLF members. In light of evolutionary information that may facilitate the classification of the OLF family and proper association of novel OLF genes with characterized homologs, we performed phylogenetic analysis on all 116 OLF proteins currently available, including two novel members cloned by our group. The OLF family segregated into seven subfamilies and members with similar domain compositions or functional properties all fell into relevant subfamilies. Furthermore, our Northern blot analysis and previous studies revealed that the typical human OLF members in each subfamily exhibited tissue-specific expression patterns, which in turn supported the segregation of the OLF subfamilies with functional divergence. Interestingly, the phylogenetic tree topology for the OLF domains alone was almost identical with that of the full-length tree representing the unique phylogenetic feature of full-length OLF proteins and their particular domain compositions. Moreover, each of the major functional domains of OLF proteins kept the same phylogenetic feature in defining similar topology of the tree. It indicates that the OLF domain and the various domains in flanking non-OLF regions have coevolved and are likely to be functionally interdependent. Expanded by a plausible gene duplication and domain couplings scenario, the OLF family comprises seven evolutionarily and functionally distinct subfamilies, in which each member shares similar structural and functional characteristics including the composition of coevolved and interdependent domains. The phylogenetically classified and preliminarily assessed subfamily framework may greatly facilitate the studying on the OLF proteins. Furthermore, it also demonstrated a feasible and reliable strategy to categorize novel genes and predict the functional implications of uncharacterized proteins based on the comprehensive phylogenetic classification of the subfamilies and their relevance to preliminary functional characteristics.

*Corresponding author. Fax: +86 21 63852655.
E-mail address: wjma@sibs.ac.cn (W.-J. Ma).

Keywords: Olfactomedin; Evolution; Phylogenetic analysis; Domain coupling; Intramolecular coevolution

1. Introduction

There is usually significant functional diversity among members in large gene families. Since the cloning of the first identified OLF member, Olfactomedin, which is a major structural component of bullfrog olfactory neuroepithelium and may influence the differentiation of olfactory cilia [1,2], more than 100 known OLF members have been discovered in various species ranging from *Caenorhabditis elegans* to *Homo sapiens*. Some OLF members have been demonstrated to play essential roles in various physiological processes. For example, Myocilin/TIGR was the first member in this family identified to be associated with human disease processes [3], and recurrent mutations in the OLF domain of Myocilin were closely associated with primary open angle glaucoma [4]. Another OLF member, Optimedin/OLFM3, was demonstrated to colocalize with Myocilin in the trabecular meshwork of human eye and may be involved in the disorders of the anterior segment of the eye [5]. Noelin was found to play an important role in neural development [6,7]. A recent report revealed that *C. elegans* UNC-122 was involved in neuromuscular signaling [8]. Although some members in the OLF family were found to have similar activities involved in the development of the nervous system, there still existed significant functional diversity among them. Furthermore, some OLF members were not expressed in the neural tissues at all [9–11], and many new OLF members with no accompanying functional information have poured into this family. Thus, this makes it even more difficult to characterize the ever-expanding OLF family without comprehensive view on structural and functional properties of the OLF proteins through effective and efficient analysis.

In our research on hunting for novel human secreted proteins, we trapped several putative secreted protein sequences with OLF domains, which comprise approximately 260 amino acids and are usually located at the C-terminals. In order to categorize these trapped putative OLF secreted proteins and to find functional clues, we used the common highest-hit homology method to analyze them, but it gave incomprehensive results because of the complexity of the large OLF family,

0014-5793/$30.00
doi:10.1016/j.febslet.2005.08.064

and even some annotation information of them led to incorrect inference. For example, one of our trapped OLF sequences hOLF44, was previously deposited in sequencing database as HNOEL-iso (Accession No. AF201945), and was annotated with the description of OLFML3 (olfactomedin-like 3) as the ortholog of chick OLFM/Noelin (Accession No. AF239804) at the time of analysis. However another human OLF member, named OLFM1 (Accession No. NM_014279), was found to have much higher sequence similarity to chick Noelin than HNOEL-iso. Such orthologous relationships may be resolved by the reciprocal BLAST method only if all the relevant sequences are available at the time of analysis, which is not often the case for genes from most species. Furthermore, it is hard to obtain a comprehensive view on the orthologous and paralogous relationships among the OLF proteins only by the reciprocal BLAST method. Evolutionary information may help to solve this problem when the phylogenetic relationships of the OLF family are analyzed and examined systematically. However, very few studies have been reported on this issue, and those evolutionary or bioinformatic analyses focused on either identifying conserved structural motifs in the Olfactomedin protein [12], searching for the human OLF homologs [13], or hunting for Myocilin-related OLF proteins involved in pathogenesis of ocular disorders [14]. The expanded OLF proteins in sequencing database available now makes it possible to classify the OLF family and obtain the biological implications of the OLF members from a comprehensive view of the orthologous and paralogous relationships by phylogenetic analysis.

To clarify relationships among the OLF proteins we trapped, and to better understand the ever-expanding and complex OLF family, and of particular interest is to find some common biological/functional characteristics of the OLF proteins, we performed phylogenetic classification on all currently available OLF proteins. To further investigate whether our phylogenetic classification of the OLF family has biological relevance to functional implications, we analyzed mRNA expression patterns of the typical human OLF members in different subfamilies, including two novel members cloned by our group. Additionally, the structural properties and evolutionary history of the OLF subfamilies were discussed. Our strategy of grouping family members into structurally and functionally distinct subfamilies is of great value to make sense of the complex OLF family, and it may also be useful to the characterization of other gene families.

2. Materials and methods

2.1. Data collection and taxon sampling

227 OLF protein sequences were obtained using the OLF domain of the first identified OLF member Olfactomedin (Accession No. Q07081) as query to search NCBI non-redundant database. Additionally, because the *Takifugu rubripes* protein sequences were not available in the NCBI database at the time of analysis (February, 2005), we searched for the OLF genes in International Fugu Genome Consortium database (http://www.fugu-sg.org/) and Smart database [15], and this search resulted in 41 *Takifugu rubripes* OLF protein sequences.

The redundant sequences (alternately spliced transcripts and short truncated sequences) were eliminated following the two criteria: (1) if several sequences share the same UniGene number, we select the longest one in length; (2) if not, when the sequence identity is above 90% for 100 or more amino acids, we select the one with UniGene number. After the redundancy was eliminated, 116 OLF members from 20 species were obtained. To briefly present the result, we next extracted the OLF members from the extensively sequenced species. Furthermore, since the protein repertoires between *Homo sapiens* and *Pan troglodytes*, *Mus musculus* and *Rattus norvegicus*, *Tetraodon nigroviridis* and *Takifugu rubripes* are almost identical, only the *H. sapiens*, *M. musculus*, and *T. nigroviridis* OLF members were selected for further analysis. This dataset includes 68 OLF members and the members were named using the abbreviation of genus and species name followed by a gene label. The accession numbers and relevant information of the 68 OLF members are listed in Table 1.

2.2. Multiple alignments and secondary structure analysis

General sequence editing and analysis were performed using the programs of GCG package (Genetics Computer Group, version 10.0). Multiple alignments of protein sequences were created by ClustalW [16] with the default settings. The alignments were then inspected and formatted using the GeneDoc alignment program [17]. The secondary structure analysis was carried out using Jpred [18], a server that integrates information from several programs including PHD, PREDATOR, and ZPRED.

2.3. Phylogenetic tree construction

Phylogenetic trees based on protein sequences were constructed using the MEGA2 program [19]. The Poisson correction model was used for distance matrix calculations and the complete deletion option was set for gaps and missing data. Trees were constructed by both neighbor-joining [20] and minimum evolution [21] methods. Trees were also generated using maximum parsimony [22] methods, as implemented in the MEGA2 program. To assess the confidence of individual nodes, a bootstrap analysis [23] with 1000 replications was performed using the same computer package. All three methods gave virtually identical results for our datasets. Only the neighbor-joining trees were illustrated in this paper, as the neighbor-joining method was known to be quite efficient in obtaining reliable trees from large sets of data [24].

2.4. Expression patterns of human hOLF44 and OLFM1

To characterize the mRNA expression patterns of human hOLF44 and OLFM1, two novel OLF members cloned by our group recently, Northern blot analysis was carried out on Human Multiple Tissue Northern Blot (Clontech). Each lane contained approximately 1.0 μg human poly A^+ RNA. The entire open reading frame (ORF) fragments of hOLF44 and OLFM1 were generated by PCR amplification and used as templates to generate probes which were labeled with [α-^{32}P]dCTP. Experiments were conducted according to the manufacturer's instruction. The blot was hybridized sequentially to different probes after complete striping. The same blot was stripped and rehybridized with a 2 kb human β-actin probe to verify that all lanes contained comparable amounts of mRNA.

3. Results

3.1. Simple sequence alignments revealed the similarity but covered the distinct properties associated with divergent functions of OLF proteins

To identify conserved regions or motifs in the emerging group of OLF proteins, multiple sequence alignments were carried out based on the 68 full-length OLF protein sequences. To date, three-dimensional structural information is not available for any of the OLF proteins, and therefore, the aligned sequences were submitted to Jpred server [18] to predict secondary structure. Part of the most conserved alignments and the corresponding secondary structure motifs were shown in Fig. 1. Although the overall amino acid sequences of the OLF members are relatively divergent, they share similar β-strands motif in their OLF domains and the key residues that stabilize the domain structure are sufficiently conserved during evolution. For example, the boxed sequence

Table 1
The OLF members used in phylogenetic analysis

Accession No.	Description	Species	Name	Chro	AA	Sub-family
NP_055094	Olfactomedin 1 (OLFM1)/hOlfA	*H. sapiens*	Hs-OLFM1	9q34.3	467	**I**
NP_062371	Olfactomedin 1 (OLFM1)	*M. musculus*	Mm-OLFM1	2 A3	485	
AAF43715	NOELIN-2	*Gallus gallus*	Gg-Noelin	17	457	
AAL66226	Noelin-2	*Xenopus laevis*	Xl-Noelin		458	
CAF90469	Unnamed protein	*T. nigroviridis*	Tn-CAF90469		515	
NP_477512	Olfactomedin 2 (OLFM2)/hOlfC	*H. sapiens*	Hs-OLFM2	19p13.2	454	
NP_776138	Hypothetical protein	*M. musculus*	Mm-NP_776138	9 A3	448	
AAH84792	Unknown protein	*X. laevis*	Xl-AAH84792		448	
CAF98141	Unnamed protein	*T. nigroviridis*	Tn-CAF98141	18	482	
AAQ89084	Olfactomedin 3 (OLFM3)	*H. sapiens*	Hs-OLFM3	1p22	478	
P63056	Olfactomedin 3 (OLFM3)	*M. musculus*	Mm-OLFM3	3 G1	478	
XP_422304	Similar to optimedin form A	*G. gallus*	Gg-XP_422304	8	458	
AAH81110	MGC83418 protein	*Xenopus laevis*	Xl-AAH81110		477	
CAG07148	Unnamed protein	*T. nigroviridis*	Tn-CAG07148	15	433	
NP_851376	Latrophilin 3	*Bos taurus*	Bt-Latrophilin-3		1580	**II**
Q9HAR2	Latrophilin 3 precursor	*H. sapiens*	Hs-Latrophilin-3	4q13.1	1447	
XP_205556	Latrophilin 3	*M. musculus*	Mm-Latrophilin-3	5 E1	1268	
XP_420575	Similar to CL3AA	*G. gallus*	Gg-XP_420575	4	2825	
CAG02284	Unnamed protein	*T. nigroviridis*	Tn-CAG02284		1623	
O95490	Latrophilin 2 precursor	*H. sapiens*	Hs-Latrophilin-2	1p31.1	1459	
NP_851356	Lectomedin 2	*B. taurus*	Bt-Latrophilin-2		1478	
XP_131258	Latrophilin 2	*M. musculus*	Mm-Latrophilin-2	3 H3	1218	
XP_422382	Similar to latrophilin 2	*G. gallus*	Gg-XP_422382	8	846	
CAF98480	Unnamed protein	*T. nigroviridis*	Tn-CAF98480	15	1471	
BAA74844	KIAA0821	*H. sapiens*	Hs-Latrophilin-1	19p13.2	1566	
NP_851382	Latrophilin 1	*M. musculus*	Mm-Latrophilin-1	8 C2	1247	
AAD09192	Latrophilin-1	*B. taurus*	Bt-Latrophilin-1		1472	
CAG06092	Unnamed protein	*T. nigroviridis*	Tn-CAG06092	18	1698	
NP_000252	Myocilin	*H. sapiens*	Hs-Myocilin	1q23-q24	504	**III**
NP_034995	Myocilin	*M. musculus*	Mm-Myocilin	1 H2.1	490	
NP_776543	Myocilin	*B. taurus*	Bt-Myocilin		490	
XP_422235	Similar to myocilin	*G. gallus*	Gg-XP_422235	8	668	
CAF89710	Unnamed protein	*T. nigroviridis*	Tn-CAF89710		447	
CAF97613	Unnamed protein	*T. nigroviridis*	Tn-CAF97613	15	454	
BAD38864	Photomedin-1	*H. sapiens*	Hs-Photomedin-1	9q33.3	652	**IV**
NP_766442	Olfactomedin-like 2A	*M. musculus*	Mm-NP_766442	2 B	681	
CAF90128	Unnamed protein	*T. nigroviridis*	Tn-CAF90128		218	
XP_415383	Similar to olfactomedin-like	*G. gallus*	Gg-XP_415383	17	923	
BAD38863	Photomedin-2/hOlfB	*H. sapiens*	Hs-Photomedin-2	1q23.3	750	
NP_796042	Olfactomedin-like 2B	*M. musculus*	Mm-NP_796042	1 H3	746	
XP_422209	Similar to olfactomedin-like 2B	*G. gallus*	Gg-XP_422209	8	867	
CAG01427	Unnamed protein	*T. nigroviridis*	Tn-CAG01427	1	262	
AAQ88930	GW112/hOlfD	*H. sapiens*	Hs-GW112	13q14.3	510	**V**
XP_354831	pDP4	*M. musculus*	Mm-pDP4	14 D3	531	
XP_417022	Similar to pDP4	*G. gallus*	Gg-XP_417022	1	508	
ENSP00000332317	Unnamed protein	*H. sapiens*	Hs-332317	11	468	
NP_766493	E030002O03	*M. musculus*	Mm-NP_766493	7 E2	496	
BAB85495	Tiarin	*X. laevis*	Xl-Tiarin		467	
Q07081	Olfactomedin precursor	*Rana catesbeiana*	Rc-Olfactomedin		464	
CAG10636	Unnamed protein	*T. nigroviridis*	Tn-CAG10636	10	457	
BAD18742	Unnamed protein	*H. sapiens*	Hs-BAD18742	15q21.2	551	**VI**
NP_796324	Collomin	*M. musculus*	Mm-CRG-L2	9 B	549	
XP_425097	Similar to VDLC9339	*G. gallus*	Gg-XP_425097	10	601	
CAF99838	Unnamed protein	*T. nigroviridis*	Tn-CAF99838	5	611	
CAG05536	Unnamed protein	*T. nigroviridis*	Tn-CAG05536	13	599	
NP_573262	CG6867-PA	*D. melanogaster*	Dm-CG6867	X	949	
XP_315876	ENSANGP00000004496	*Anopheles gambiae*	Ag-XP_315876	2L	749	
AAN60526	Colmedin	*C. elegans*	Ce-Colmedin	I	486	
NP_493598	UNC-122	*C. elegans*	Ce-UNC-122	I	598	
XP_426398	Similar to MVAL564	*G. gallus*	Gg-XP_426398	5	380	**VII**
AAQ88954	MVAL564	*H. sapiens*	Hs-AAQ88954	11p15.4	402	
CAG11764	Unnamed protein	*T. nigroviridis*	Tn-CAG11764	11	388	
CAG03805	Unnamed protein	*T. nigroviridis*	Tn-CAG03805	9	260	
XP_418008	Similar to Olfactomedin-like 3	*G. gallus*	Gg-XP_418008	26	392	
AAH84769	Unknown protein	*X. laevis*	Xl-AAH84769		391	
NP_598620	Olfactomedin-like 3	*M. musculus*	Mm-NP_598620	3 F2.2	406	
AAR88262	hOLF44	*H. sapiens*	Hs-hOLF44	1p13.2	406	
NP_999798	Amassin	*Strongylocentrotus purpuratus*	Sp-Amassin		495	

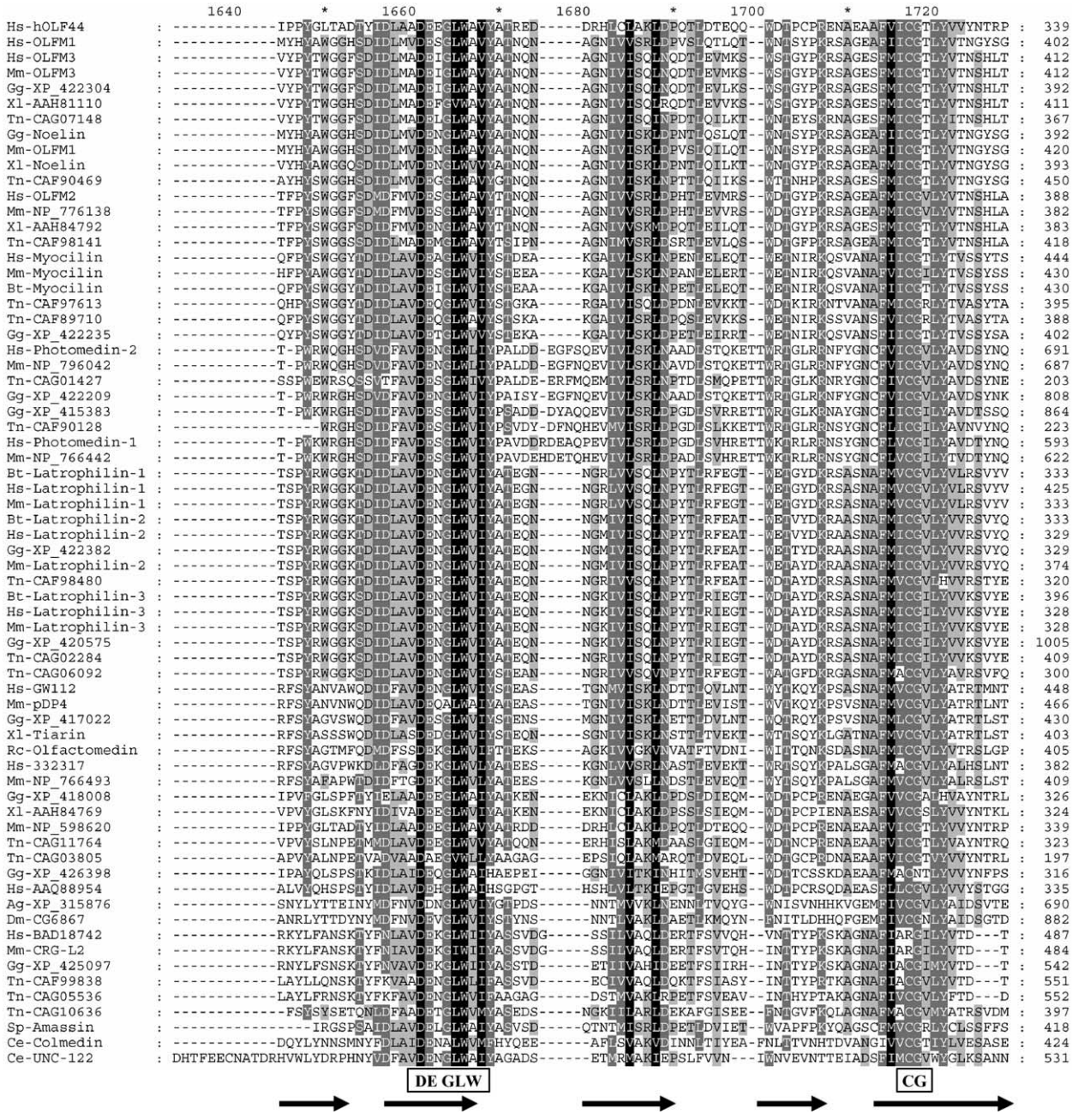

Fig. 1. The most conserved part of multiple amino acid sequence alignments of the 68 OLF members (Table 1). Gaps are indicated with dashes. Identical amino acid residues are dark-shaded and similar residues are shaded at various degrees. The conserved sequence motifs (DExGLW, CG) are boxed, where x can be any amino acid. The β-strands motif is depicted with arrows.

motifs (DExGLW, CG) are highly conserved across all the 68 OLF members (Fig. 1). The secondary structure analyses of the 68 OLF proteins suggested that the N-terminal of OLF members mainly contains α-helix motif, whereas the C-terminal chiefly contains β-strands and is more conserved than the N-terminal in evolution (Data not shown), and these findings were consistent with previous studies done on Olfactomedin [12], Myocilin [25–27], Noelin and Amassin [28]. Thus, the multiple sequence alignments cannot efficiently provide valuable information on apparent structural features that might be associated with divergent biological functions of OLF proteins.

3.2. The OLF family was classified into seven distinct subfamilies by phylogenetic analysis

The resulting unrooted neighbor-joining tree based on the 68 full-length OLF proteins was presented in Fig. 2. This phylogenetic tree strongly suggests that the OLF proteins segregate into seven evolutionarily distinct subgroups. Furthermore, the phylogenetic tree topology for all 116 OLF members is also closely consistent with that outlined in Fig. 2 (Data not shown). Accordingly, the OLF family was classified into seven subfamilies (labeled with Roman numerals) (Fig. 2). The protein sequences used in our analysis are listed in Table 1 and assigned to the corresponding subfamilies. Members in different

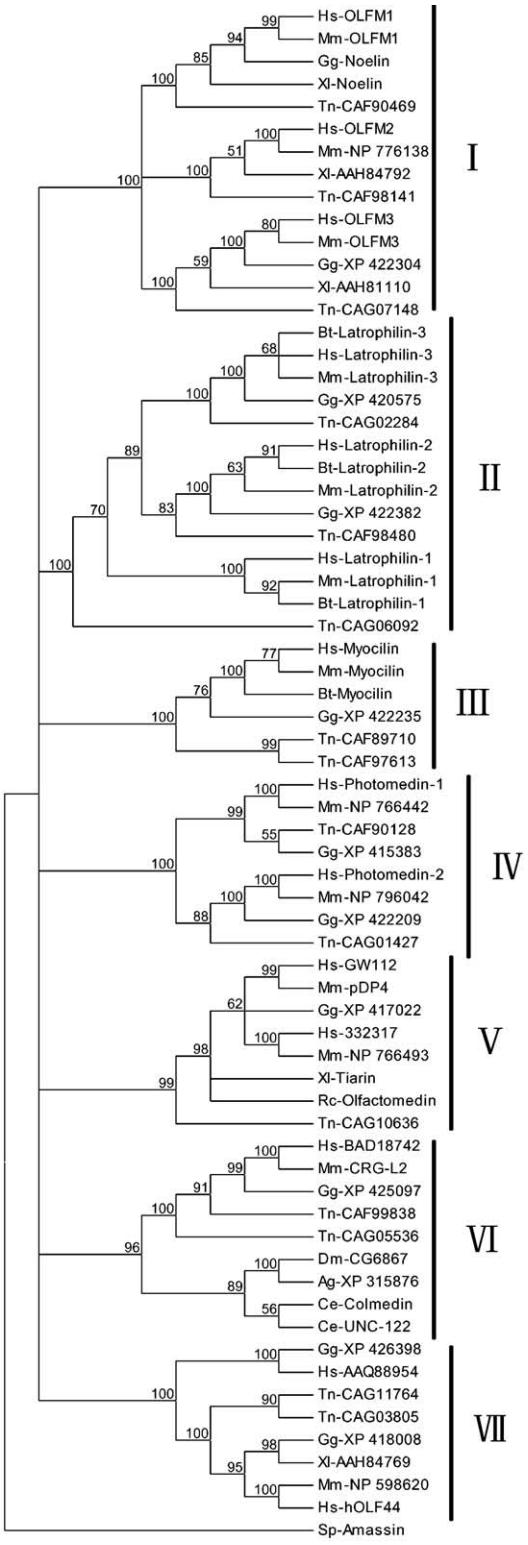

Fig. 2. Phylogenetic classification of the OLF family. The neighbor-joining tree is based on the 68 full-length OLF protein sequences (Table 1). Vertical bars and Roman numerals delineate the seven subfamilies. Bootstrap values based on 1000 replications are shown only when they are greater than 50%.

subfamilies usually contained different domain architectures, and members with similar domain compositions all fell into relevant subfamilies. As shown in Fig. 2, OLFM1 was closely clustered with the characterized Noelin genes (subfamily **I**), and in sharp contrast, hOLF44/HNOEL-iso/OLFML3 dropped into a different subfamily (**VII**) without any Noelin or OLFM3 genes. Therefore, HNOEL-iso or OLFML3 was thus far inappropriately annotated in NCBI GenBank database and the biological features of it may differ from that of Noelin or OLFM3. To clarify this fact, we preferred to use the name hOLF44 (with a predicted protein molecular weight of 44 kDa) instead of HNOEL-iso or OLFML3.

3.3. Human OLF members in the segregated OLF subfamilies showed distinct expression patterns

To investigate whether our phylogenetic classification of the OLF family has biological relevance to functional implications, we compared mRNA expression patterns of the characterized OLF genes in different subfamilies, as the tissue or cellular-specific localization of gene expression is the primary implication for the function of genes. The analysis showed that within a same subfamily, OLF members from different species usually shared relatively conserved expression patterns. For example, in subfamily **I**, the rat, mouse, chick and xenopus Noelin (also named OLFM or Pancortin) genes were almost all exclusively expressed in neural tissues [6,7,29–32]. OLF members in different subfamilies usually exhibited relatively divergent expression patterns, for example, GW112 (also named OhlfD or hGC-1) (subfamily **V**) was predominantly expressed in colon and small intestine [9,10,13], which was greatly different from that of Noelin genes (subfamily **I**).

To get a more comprehensive view of the expression patterns of the typical human OLF members in each subfamily, we specifically supplemented the available human OLF genes expression profile data [5,9,10,13,33,34] with the two uncharacterized genes, OLFM1 and hOLF44, which are the typical human OLF members in subfamilies **I** and **VII**, respectively (Fig. 2). Our Northern blot analyses revealed that approximately 1.8 kb transcript of hOLF44 was detected abundantly in placenta, and moderate expression was observed in liver and heart. Fairly weak signals were also detected in skeletal muscle, small intestine and kidney (Fig. 3, *hOLF44*). In contrast, OLFM1 was almost exclusively expressed in brain, and only very faint signal was detected in skeletal muscle among other nine tissues examined at this level of sensitivity (Fig. 3, *OLFM1*). The expression pattern of OLFM1 is in close agreement with a previous report on hOlfA [13], which is a small partial sequence of OLFM1. Thus, OLFM1 and hOLF44 exhibited distinct expression patterns, which is consistent with their quite divergent positions in the OLF subfamily tree (Fig. 2).

The comprehensive comparison of expression patterns among the typical human OLF members in different subfamilies is listed in Table 2. Human members in each subfamily exhibited distinct expression patterns, and they were all selectively expressed in different set of tissues (Table 2). Therefore, the analysis of our expression profile data and previous reports suggest that the OLF members in different subfamilies may have divergent functions in different tissues where they are expressed, and accordingly, our phylogenetic classification of the OLF family may have biological relevance to functional implications. Thus, such distinct expression patterns in turn strongly supported our classification of the OLF family.

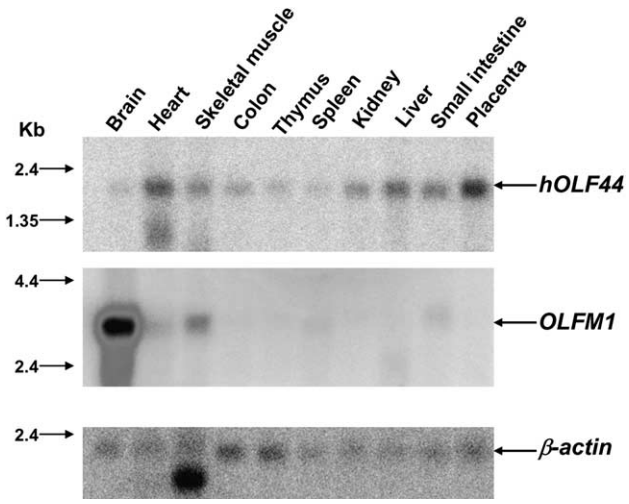

Fig. 3. Distinct mRNA expression patterns of hOLF44 and OLFM1. Northern blot was sequentially hybridized with [α-^{32}P]dCTP labeled hOLF44 and OLFM1 probes, as described in Section 2.4. Size standards and the order of tissues source are marked at the left and above of the blot, respectively. β-Actin was used as an RNA loading control, and the strong band below β-actin in the "skeletal muscle" lane is muscle-preferred α-actin.

3.4. The structural properties of composition and coevolution of domains within OLF proteins provided strong basis for the establishment of the OLF subfamily framework

As mentioned above, the OLF members with similar domain compositions all fell into relevant subfamilies. For example, members in subfamily **II** almost all contain Gal_Lectin, OLF, HormR, GPS, 7tm_2 and Latrophilin domains. In other words, domain composition is the dominant contributor for subfamily classification. To explore the phylogenetic features of the conserved OLF domains alone in defining subfamily classification, we extracted the OLF domains from the 68 protein sequences and reconstructed a second unrooted neighbor-joining tree (Fig. 4B). Interestingly, the tree topology based on the OLF domains alone (Fig. 4B) was virtually identical with that of the full-length tree (Fig. 4A). It indicates that the OLF domain has distinct structural properties in determining the subfamily classification in the absence of coupled non-OLF regions. Thus, we hypothesized that the OLF domain may have the possibility to coevolve with the non-OLF regions containing domains unique to different subfamilies. To verify this hypothesis, we excised the OLF domains from the 68 OLF protein sequences and concatenated the remaining two segments for phylogenetic tree construction, and then another unrooted neighbor-joining tree for the non-OLF regions was constructed. As demonstrated in Fig. 4C, the tree topology for non-OLF regions was very similar to that based on OLF domains alone (Fig. 4B), and there is only one remarkable divergent site between this two trees (labeled with dot). These data suggest that the OLF domain and the non-OLF regions of the OLF proteins have coevolved and are likely to be functionally interdependent.

To extend the analysis on the evolutionary features of each domain regarding duplication and coevolution, we conducted an evolutionary analysis on the six separated functional domains (Gal_Lectin, OLF, HormR, GPS, 7tm_2 and Latrophilin ranging from N-terminal to C-terminal) of Latrophilin proteins within subfamily **II** (Fig. 5A). To compare with the full-length tree, we also reconstructed another two phylogenetic trees, one was based on the full-length Latrophilins, designated as Full-length, and the other was based on the Latrophilins without the OLF domains, designated as OLF-eliminated. As depicted in Fig. 5B, the tree topology is very similar between the full-length tree and all the other seven trees. Except three divergent sites, including one remarkably divergent site in the tree for Gal_Lectin and another two slightly divergent sites in the trees for GPS and 7tm_2 domains (labeled with square, dot and triangle), all the eight phylogenetic trees are almost identical in topology. Thus, each functional domain of Latrophilins generally kept the same phylogenetic feature in defining similar tree topology, and the OLF domain and the other five functional domains of the Latrophilins may have coevolved under some functional constraints.

Taken together, these findings also supported the rationality of our classification on the OLF family.

3.5. Distinct biological features of characterized OLF members are consistent with the classification of the OLF family

To make better sense of the complex OLF family, we further conducted comparative analysis of the functional and

Table 2
Comparisons of the distinct mRNA expression patterns of the typical human members in each OLF subfamily

Sub-family	Human OLF members	mRNA expression patterns of the typical human members											References
		Brain	Colon	Heart	Pancreas	Placenta	Small intestine	Liver	Skeletal Muscle	Kidney	Retina	Spleen	
I	OLFM1[a]	+++	–	–	N/A	–	–	–	+	–	N/A	–	This report, Fig. 3
II	Latrophilin-2	+++	+	++	N/A	+	+	+	N/A	++	N/A	+	[33]
III	Myocilin	–	–	++	–	–	–	+	–	–	++	–	[5,13]
IV	Photomedin-2/hOlfB	–	–	–	++	–	–	–	–	–	N/A	–	[13]
V	GW112 / hOlfD	–	++	–	–	–	+++	–	–	–	N/A	–	[9,10,13]
VI	CRG-L2	+	–	–	–	+	–	++	–	–	N/A	–	[34]
VII	hOLF44[b]	–	+	++	N/A	+++	+	++	+	+	N/A	–	This report, Fig. 3

Signals from hybridizing blots were scored as strong (+++), moderate (++), and very weak (+). N/A represents the data is Not Available at the time of the preparation of this paper. [a] and [b]indicate that this two OLF members are reported in this paper and illustrated in Fig. 3.

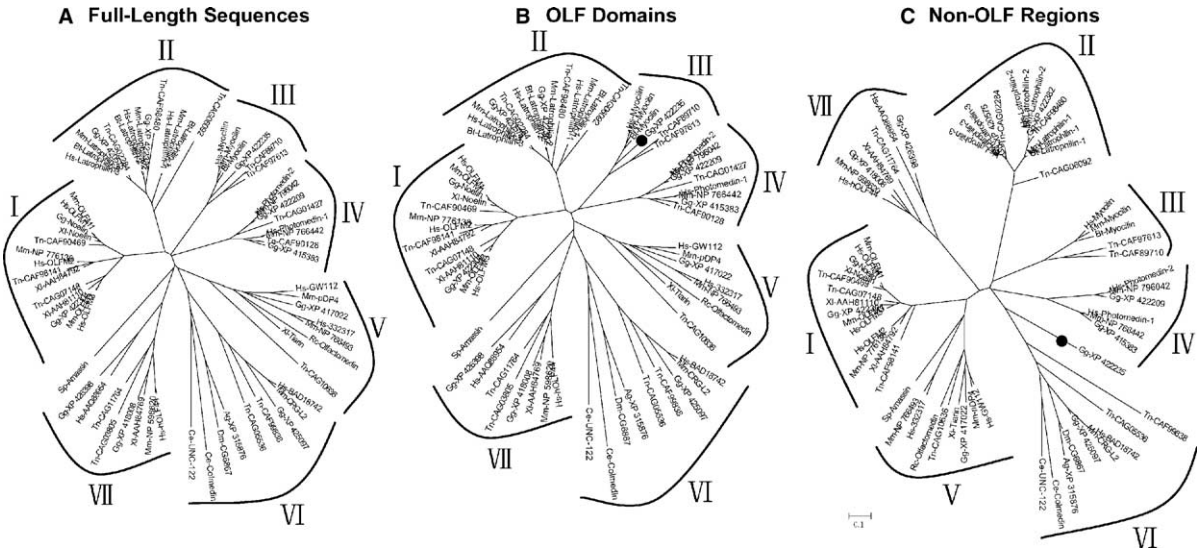

Fig. 4. Comparison of unrooted phylogenetic trees based on the 68 full-length OLF protein sequences (Table 1) (A), the OLF domains alone (B), and the non-OLF regions (C). The major topology difference between the OLF domains tree (B) and the non-OLF regions tree (C) is labeled with symbol (dot) in the corresponding position.

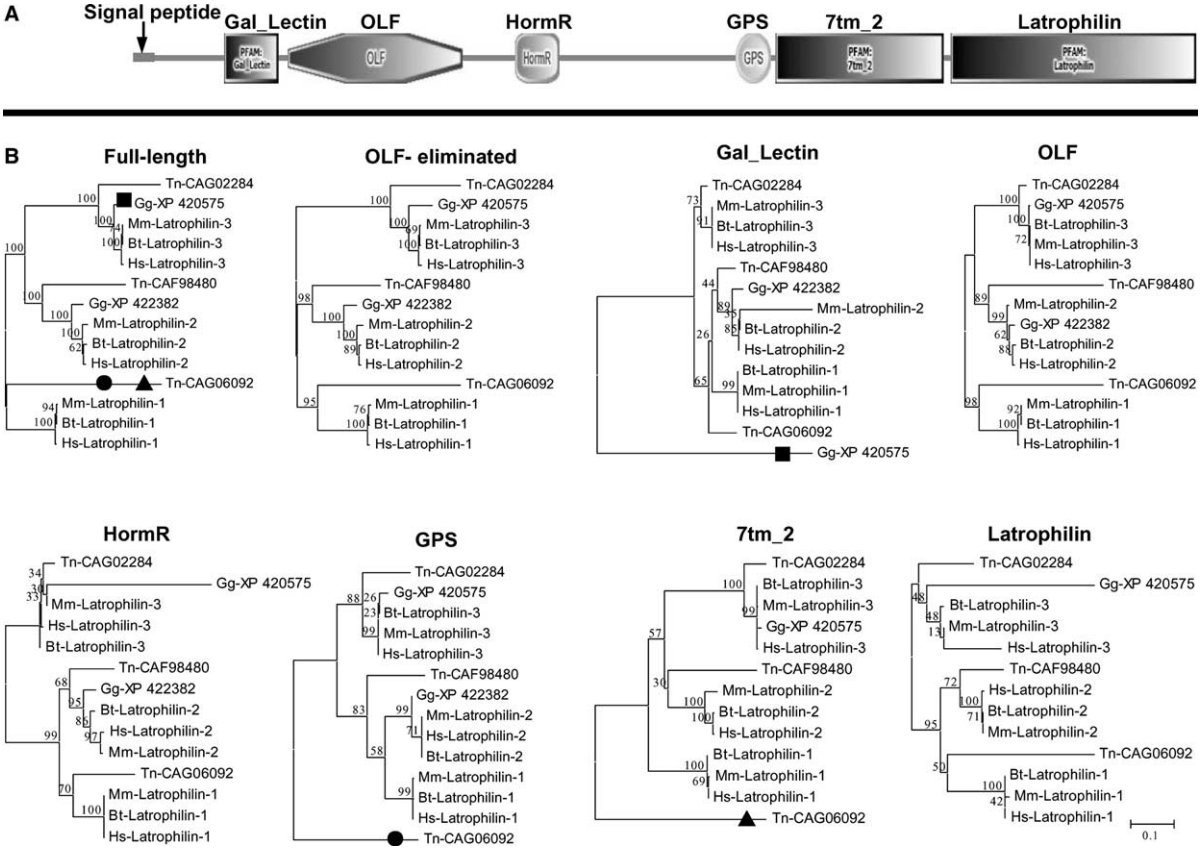

Fig. 5. Coevolution of the OLF domain and the other five functional domains of Latrophilin genes (Table 1). (A) The general domain architecture of Latrophilin protein. (B) Neighbor-joining trees for Latrophilin proteins and different segments of them, respectively. The name of each part is labeled above the corresponding tree. The difference between the full-length tree topology and the other seven trees are labeled with symbols (square, dot and triangle) in the corresponding positions. Branch lengths are shown to scale. Bootstrap values based on 1000 replications are shown above the branches.

evolutionary aspects of the characterized OLF members in each subfamily. The results are briefly summarized in Table 3. As presented in Table 3 and Fig. 2, the OLF members with similar functional/structural characteristics usually fell into relevant subfamilies, and the topology tree of each subfamily was generally consistent with the currently accepted view of

Table 3
The characteristic features of characterized OLF members in each subfamily

Subfamily	General biological features in each subfamily	References
I	Dominantly expressed in brain; involved in brain development and function	[5–7,29–32], and this report, Fig. 3
II	G protein-coupled receptors for α-Latrotoxin, mainly expressed in brain and heart; related to the control of synaptic vesicle exocytosis	[33,35,47–49]
III	Dominantly expressed in retina; associated with primary open angle glaucoma; the most conserved subfamily in the OLF family	[3,5,13]
IV	Expressed in pancreas, prostate, lung, heart, eye; function is unknown, the mouse members are extracellular proteins	[13,39]
V	Containing the first cloned OLF gene, Olfactomedin; mainly expressed in brain, colon, small intestine; extracellular matrix proteins involved in different physiological processes	[1,2,9–11,13,37]
VI	The most primitive clade in the OLF family, containing invertebrate members; containing one or two collagen domains preceding the OLF domain of all the members	[8,34,50]
VII	Human member hOLF44 may have matrix-related function involved in placental and embryonic development, or play a similar role in other physiological processes	[38], and this report, Fig. 3

species tree. Thus these data also strongly supported our phylogenetic classification of the OLF family.

4. Discussion

In this paper, we established a comprehensive subfamily framework for the categorization of all currently available OLF proteins. The assertion that the OLF family is composed of seven evolutionarily and functionally distinct subfamilies is supported by at least five lines of evidence: (1) each was found in trees generated by all the phylogenetic methods used, and each tree has reasonably high bootstrap values; (2) our Northern blot analysis and reported studies showed that in the same subfamily, OLF members from different species usually shared relatively conserved expression patterns; (3) the topology tree of each subfamily is generally coincident with the currently accepted view of species tree; (4) the phylogenetic trees for OLF domains alone and the non-OLF regions of the 68 OLF proteins are either identical or very similar to the full-length tree representing the unique phylogenetic feature of full-length OLF proteins and their particular domain composition; (5) members with similar domain compositions or functional characteristics usually fell into relevant subfamilies.

Although the emerging OLF family is relatively complex, our subfamily classification and structural analysis may provide additional insights into understanding some general characteristics of this family. Firstly, subfamily **I** majorly includes the Noelin proteins, which usually shared conserved brain-preferred expression patterns and therefore functioned in the nervous system [5–7,29–32]. Secondly, subfamily **II** consists of Latrophilin proteins (G protein-coupled-receptor), and all the three Latrophilins paralogs had similar function related to the control of synaptic vesicle exocytosis [35]. Thirdly, subfamily **III** is composed of the most conservative Myocilin genes within this family, and the human Myocilin was associated with the pathogenesis of primary open angle glaucoma [3]. Fourthly, some members in subfamily **V** clearly demonstrated extracellular matrix-related functions in different physiological processes [2,9–11]. Fifthly, subfamily **VI** represents the most primitive clade in this family, as some members within this subfamily belong to invertebrates ranging from *C. elegans* to *Drosophila melanogaster*, whereas members in other subfamilies all belong to vertebrates. Finally, most of the OLF proteins were predicted to possess a signal peptide in their N-terminal, suggesting they may be secreted proteins (except the Latrophilin proteins). Actually, most of the characterized OLF members were demonstrated in vivo or in vitro to secrete out from the cells [2,5–7,11,36–40], and some of them may have extracellular matrix-related functions [2,11,38,39].

To analyze the plausible evolutionary scenarios for the OLF family, we rooted the OLF family phylogenetic tree to *C. elegans*, as to date, the most primitive organism containing OLF genes is *C. elegans*, and this situation should allow an unambiguous rooting for the metazoan OLF proteins. The rooted phylogenetic tree for the OLF family was depicted in Fig. 6, and the domain architecture of the typical human members was illustrated at the right of its corresponding subfamilies.

Up to now, there are only one or two OLF members in all species of invertebrates, such as *C. elegans* and *D. melanogaster*, whereas the number of the OLF members in a relatively early vertebrate *T. nigroviridis* was remarkably increased and then maintained in all higher vertebrates. Furthermore, the domain architecture of the OLF members in invertebrates is relatively simple, whereas it becomes more complex for some OLF members in the vertebrates. For example, vertebrate Latrophilin proteins with OLF domains contain multiple other functional domains, and their domain architectures are much more complex than that of the invertebrate OLF members as well as the invertebrate Latrophilin members without OLF domains. Although it does not exclude the possibility that the invertebrate lineage has lost the OLF domain from an ancient Latrophilin, current data suggests one plausible scenario for the evolution of the OLF family by gene duplication to increase gene content and by domain couplings to produce the multidomain genes. In this scenario, the entire genome duplications or individual gene duplication events are thought to have occurred during vertebrate evolution [41–43]. Actually, we have found that two human OLF genes (Hs-OLFM2 and Hs-Latrophilin-1) were mapped to 19p13.2 (Table 1), and interestingly, their corresponding *T. nigroviridis* counterparts (Tn-CAF98141 and Tn-CAG06092) were also found to be located in a same chromosome (Table 1). So, the gene duplication events may have occurred at least several rounds in early vertebrate leading to more than 10 paralogs in *T. nigroviridis*. These paralogs evolved independently in teleost, amphibian, fowl, and mammal lineages, and that later gave rise to the present seven subfamilies.

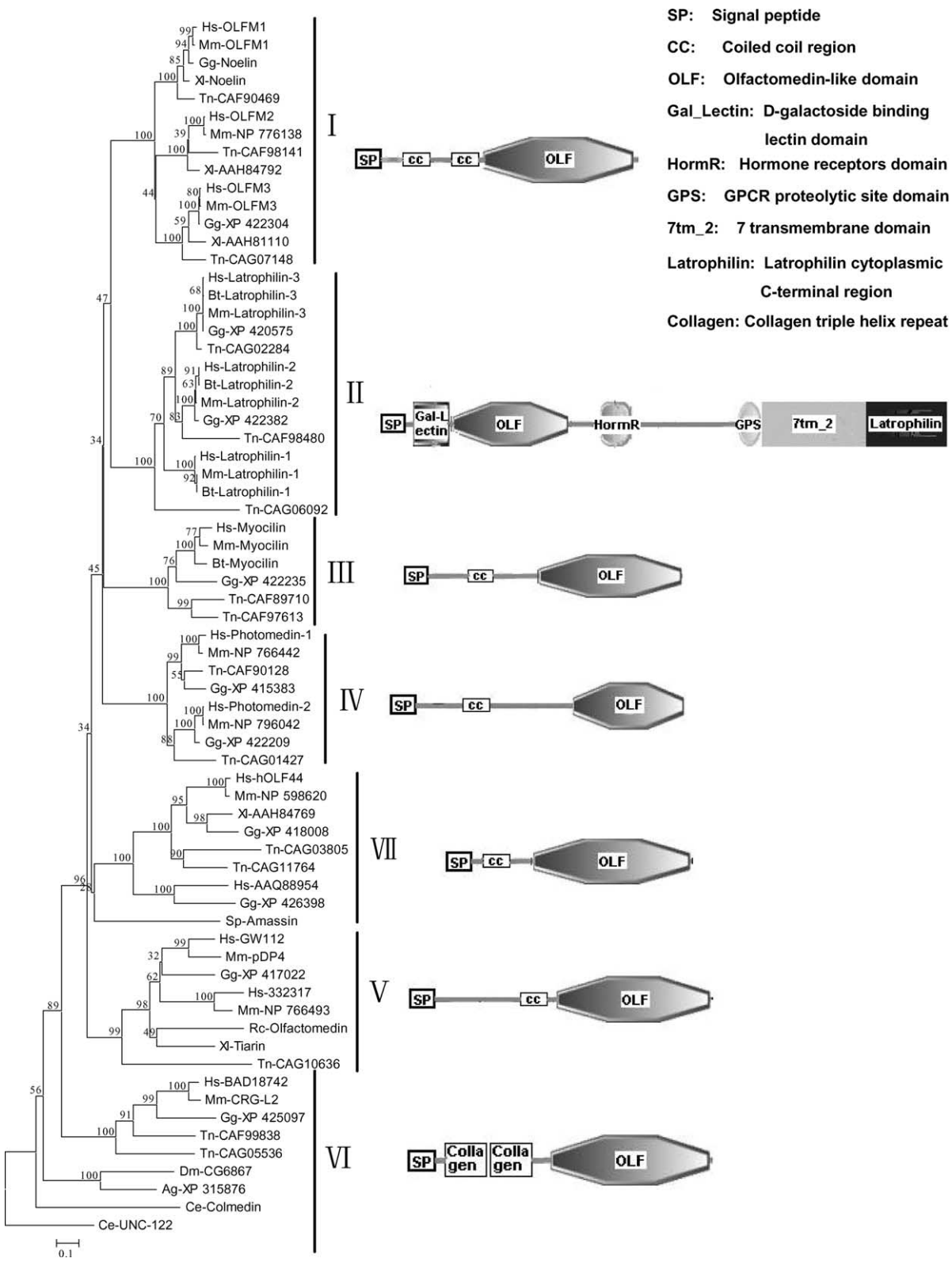

Fig. 6. Rooted neighbor-joining tree for the 68 full-length OLF members (Table 1) and the domain architecture of the typical human members in each subfamily. Vertical bars and Roman numerals delineate the seven subfamilies. Domain architecture of the typical human members is illustrated at the right of its corresponding subfamily. Domain names are noted at the up right corner. Branch lengths are shown to scale. Bootstrap values based on 1000 replications are shown above the branches.

It seemed that the basic function of the OLF domain at molecular level was still retained, such as members in subfamily **VI**, **V**, **III** and **I**, as some characterized members in these subfamilies were usually involved in formation of extracellular matrix. At the same time, some duplicated paralogs also have gained divergent functions and properties during evolution.

The members in subfamily **IV** and **VII** seemed to acquire new features to fulfill their tissue or cellular specific functions in different physiological processes. Thus, during the evolution of vertebrates, especially of mammals, the OLF members have acquired broader functions and gained more specificity in different aspects. It should be possible to determine more accurate evolutionary history of this family once more relevant sequences information is available. However, we believe that the overall topology of the relationships among subfamilies will not change remarkably from that outlined here.

The inferred subfamily framework here can be used to predict functional implications of some uncharacterized OLF members. As we indicated above, biological properties were usually conserved within the OLF subfamilies, and then the functions of some uncharacterized OLF proteins could be predicted by comparison with their identified orthologs or paralogs in the same subfamily. For example, we predicted that human OLFM1 may be expressed abundantly in brain and play an important role in the development of the nervous system, as its orthologs (chick and xenopus Noelins) were all specifically expressed in neural tissues and functioned in nervous system [6,7]. In fact, our preliminary experimental data have provided compelling evidence to support some predictions on OLFM1. For example, our Northern blot analysis revealed that OLFM1 mRNA was abundantly and almost exclusively expressed in brain. We believe that such preliminary functional predictions are also reliable for some other uncharacterized OLF members by our inferred subfamily structure presented here.

Making accurate functional predications for uncharacterized genes is very important in many areas of biological research. Almost all functional prediction methods are based on the sequence similarity, because primary sequence structure is the basis for biological function. However, sequence similarity cannot always ensure the accurate prediction of functions, since significant functional diversity usually exists among the members within large gene family, especially for those containing multiple functional domains [44,45]. Therefore, in this study, we utilized phylogenetic method to better characterize the complex OLF family by grouping the OLF members into specific subfamilies. Although this method is labor intensive, we believe it is worth employing if accuracy is the first to be considered. Furthermore, it is more efficient to make appropriate functional predictions on batch of uncharacterized OLF members based on the inferred subfamily structure, and the phylogenetic framework is also useful for putting functional information into an evolutionary context. Thus, phylogenetic analysis can greatly benefit the highest-hit methods and serve to improve the accuracy of such functional predictions [46].

5. Conclusions

The principal results here revealed that the OLF family comprises seven evolutionarily and functionally distinct subfamilies, and moreover, the OLF domain and the non-OLF regions (or coupled domains) of the OLF proteins may have coevolved and are likely to be functionally interdependent. We also suggested a plausible, even if still incomplete, gene duplication and domain couplings scenario for the evolution of this family. The phylogenetic and subfamily framework presented here will greatly benefit the preliminary functional predictions on some uncharacterized OLF genes, and may be of great value to people studying this biologically important family of proteins. At the same time, our analyses in this paper demonstrated a feasible and reliable strategy to categorize novel genes and predict the functional implications of uncharacterized proteins based on the comprehensive phylogenetic classification of the subfamilies and their relevance to preliminary functional properties.

Acknowledgments: We thank Professor Todd G. Kroll for helpful comments and critical reading of the manuscript. This work was supported by Chinese 863 High Technology R&D Project, grant from Chinese Academy of Sciences, and Grant (04ZR14150) from Shanghai Commission of Science and Technology.

References

[1] Snyder, D.A., Rivers, A.M., Yokoe, H., Menco, B.P. and Anholt, R.R. (1991) Olfactomedin: purification, characterization, and localization of a novel olfactory glycoprotein. Biochemistry 30, 9143–9153.
[2] Yokoe, H. and Anholt, R.R. (1993) Molecular cloning of olfactomedin, an extracellular matrix protein specific to olfactory neuroepithelium. Proc. Natl. Acad. Sci. USA 90, 4655–4659.
[3] Stone, E.M., Fingert, J.H., Alward, W.L., Nguyen, T.D., Polansky, J.R., Sunden, S.L., Nishimura, D., Clark, A.F., Nystuen, A., Nichols, B.E., Mackey, D.A., Ritch, R., Kalenak, J.W., Craven, E.R. and Sheffield, V.C. (1997) Identification of a gene that causes primary open angle glaucoma. Science 275, 668–670.
[4] Adam, M.F., Belmouden, A., Binisti, P., Brezin, A.P., Valtot, F., Bechetoille, A., Dascotte, J.C., Copin, B., Gomez, L., Chaventre, A., Bach, J.F. and Garchon, H.J. (1997) Recurrent mutations in a single exon encoding the evolutionarily conserved olfactomedin-homology domain of TIGR in familial open-angle glaucoma. Hum. Mol. Genet. 6, 2091–2097.
[5] Torrado, M., Trivedi, R., Zinovieva, R., Karavanova, I. and Tomarev, S.I. (2002) Optimedin: a novel olfactomedin-related protein that interacts with myocilin. Hum. Mol. Genet. 11, 1291–1301.
[6] Barembaum, M., Moreno, T.A., LaBonne, C., Sechrist, J. and Bronner-Fraser, M. (2000) Noelin-1 is a secreted glycoprotein involved in generation of the neural crest. Nat. Cell Biol. 2, 219–225.
[7] Moreno, T.A. and Bronner-Fraser, M. (2001) The secreted glycoprotein Noelin-1 promotes neurogenesis in Xenopus. Dev. Biol. 240, 340–360.
[8] Loria, P.M., Hodgkin, J. and Hobert, O. (2004) A conserved postsynaptic transmembrane protein affecting neuromuscular signaling in *Caenorhabditis elegans*. J. Neurosci. 24, 2191–2201.
[9] Shinozaki, S., Nakamura, T., Iimura, M., Kato, Y., Iizuka, B., Kobayashi, M. and Hayashi, N. (2001) Upregulation of Reg 1alpha and GW112 in the epithelium of inflamed colonic mucosa. Gut 48, 623–629.
[10] Zhang, J., Liu, W.L., Tang, D.C., Chen, L., Wang, M., Pack, S.D., Zhuang, Z. and Rodgers, G.P. (2002) Identification and characterization of a novel member of olfactomedin-related protein family, hGC-1, expressed during myeloid lineage development. Gene 283, 83–93.
[11] Rosenbauer, F., Wagne, K., Zhang, P., Knobeloch, K.P., Iwama, A. and Tenen, D.G. (2004) A novel glycoprotein secreted by mature granulocytes, is regulated by transcription factor PU.1. Blood 103, 4294–4301.
[12] Karavanich, C.A. and Anholt, R.R. (1998) Molecular evolution of olfactomedin. Mol. Biol. Evol. 15, 718–726.
[13] Kulkarni, N.H., Karavanich, C.A., Atchley, W.R. and Anholt, R.R. (2000) Characterization and differential expression of a human gene family of olfactomedin-related proteins. Genet. Res. 76, 41–50.
[14] Mukhopadhyay, A., Talukdar, S., Bhattacharjee, A. and Ray, K. (2004) Bioinformatic approaches for identification and character-

ization of olfactomedin related genes with a potential role in pathogenesis of ocular disorders. Mol. Vis. 10, 304–314.
[15] Letunic, I., Goodstadt, L., Dickens, N.J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R.R., Ponting, C.P. and Bork, P. (2002) Recent improvements to the SMART domain-based sequence annotation resource. Nucleic Acids Res. 30, 242–244.
[16] Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22, 4673–4680.
[17] Nicholas, K.B., Nicholas Jr., H.B. and Deerfield II, D.W. (1997) GeneDoc: Analysis and Visualization of Genetic Variation. EMBNEW News 4, 14.
[18] Cuff, J.A., Clamp, M.E., Siddiqui, A.S., Finlay, M. and Barton, G.J. (1998) JPred: a consensus secondary structure prediction server. Bioinformatics 14, 892–893.
[19] Kumar, S., Tamura, K., Jakobsen, I.B. and Nei, M. (2001) MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17, 1244–1245.
[20] Saitou, N. and Nei, M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.
[21] Rzhetsky, A. and Nei, M. (1993) Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol. Biol. Evol. 10, 1073–1095.
[22] Fitch, W.M. (1971) Towards defining the course of evolution: Minimum change for a specific tree topology. Syst. Zool. 20, 406–416.
[23] Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.
[24] Zhang, J. and Nei, M. (1996) Evolution of Antennapedia-class homeobox genes. Genetics 142, 295–303.
[25] Green, M.L. and Klein, T.E. (2002) A multidomain TIGR/olfactomedin protein family with conserved structural similarity in the N-terminal region and conserved motifs in the C-terminal region. Mol. Cell Proteomics 1, 394–403.
[26] Mukhopadhyay, A., Gupta, A., Mukherjee, S., Chaudhuri, K. and Ray, K. (2002) Did myocilin evolve from two different primordial proteins?. Mol. Vis. 8, 271–279.
[27] Nagy, I., Trexler, M. and Patthy, L. (2003) Expression and characterization of the olfactomedin domain of human myocilin. Biochem. Biophys. Res. Commun. 302, 554–561.
[28] Hillier, B.J. and Vacquier, V.D. (2003) Amassin, an olfactomedin protein, mediates the massive intercellular adhesion of sea urchin coelomocytes. J. Cell. Biol. 160, 597–604.
[29] Danielson, P.E., Forss-Petter, S., Battenberg, E.L., deLecea, L., Bloom, F.E. and Sutcliffe, J.G. (1994) Four structurally distinct neuron-specific olfactomedin-related glycoproteins produced by differential promoter utilization and alternative mRNA splicing from a single gene. J. Neurosci. Res. 38, 468–478.
[30] Nagano, T., Nakamura, A., Mori, Y., Maeda, M., Takami, T., Shiosaka, S., Takagi, H. and Sato, M. (1998) Differentially expressed olfactomedin-related glycoproteins (Pancortins) in the brain. Brain Res. Mol. Brain Res. 53, 13–23.
[31] Moreno, T.A. and Bronner-Fraser, M. (2002) Neural expression of mouse Noelin-1/2 and comparison with other vertebrates. Mech. Dev. 119, 121–125.
[32] Ando, K., Nagano, T., Nakamura, A., Konno, D., Yagi, H. and Sato, M. (2005) Expression and characterization of disulfide bond use of oligomerized A2-Pancortins: extracellular matrix constituents in the developing brain. Neuroscience 133, 947–957.
[33] White, G.R., Varley, J.M. and Heighway, J. (2000) Genomic structure and expression profile of LPHH1, a 7TM gene variably expressed in breast cancer cell lines. Biochim. Biophys. Acta 1491, 75–92.
[34] Graveel, C.R., Harkins-Perry, S.R., Acevedo, L.G. and Farnham, P.J. (2003) Identification and characterization of CRG-L2, a new marker for liver tumor development. Oncogene 22, 1730–1736.
[35] Ichtchenko, K., Bittner, M.A., Krasnoperov, V., Little, A.R., Chepurny, O., Holz, RW. and Petrenko, A.G. (1999) A novel ubiquitously expressed alpha-latrotoxin receptor is a member of the CIRL family of G-protein-coupled receptors. J. Biol. Chem. 274, 5491–5498.
[36] Jacobson, N., Andrews, M., Shepard, A.R., Nishimura, D., Searby, C., Fingert, J.H., Hageman, G., Mullins, R., Davidson, B.L., Kwon, Y.H., Alward, W.L., Stone, E.M., Clark, A.F. and Sheffield, V.C. (2001) Non-secretion of mutant proteins of the glaucoma gene myocilin in cultured trabecular meshwork cells and in aqueous humor. Hum. Mol. Genet. 10, 117–125.
[37] Tsuda, H., Sasai, N., Matsuo-Takasaki, M., Sakuragi, M., Murakami, Y. and Sasai, Y. (2002) Dorsalization of the neural tube by Xenopus tiarin, a novel patterning factor secreted by the flanking nonneural head ectoderm. Neuron 33, 515–528.
[38] Zeng, L.C., Liu, F., Zhang, X., Zhu, Z.D., Wang, Z.Q., Han, Z.G. and Ma, W.J. (2004) hOLF44, a secreted glycoprotein with distinct expression pattern, belongs to an uncharacterized olfactomedin-like subfamily newly identified by phylogenetic analysis. FEBS Lett. 571, 74–80.
[39] Furutani, Y., Manabe, R., Tsutsui, K., Yamada, T., Sugimoto, N., Fukuda, S., Kawai, J., Sugiura, N., Kimata, K., Hayashizaki, Y. and Sekiguchi, K. (2005) Identification and characterization of photomedins: novel olfactomedin-domain-containing proteins with chondroitin sulphate-E-binding activity. Biochem J. 389, 675–684.
[40] Cao, Q., Yu, D., Lee, A., Kasai, Y., Tychsen, B., Paus, R., Freedberg, I.M. and Sun, T.T. (2005) Expression of an olfactomedin-related gene in rat hair follicular papilla cells. J. Invest. Dermatol. 125, 24–33.
[41] Holland, P.W., Garcia-Fernandez, J., Williams, N.A. and Sidow, A. (1994) Gene duplications and the origins of vertebrate development. Dev. Suppl., 125–133.
[42] Mazet, F. and Shimeld, S.M. (2002) Gene duplication and divergence in the early evolution of vertebrates. Curr. Opin. Genet. Dev. 12, 393–396.
[43] Jaillon, O., Aury, J.M., Brunet, F., Petit, J.L. and Stange-Thomann, N., et al. (2004) Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 431, 946–957.
[44] van der Bliek, A.M. (1999) Functional diversity in the dynamin family. Trends Cell Biol. 9, 96–102.
[45] Heim, M.A., Jakoby, M., Werber, M., Martin, C., Weisshaar, B. and Bailey, P.C. (2003) The basic helix–loop–helix transcription factor family in plants: a genome-wide study of protein structure and functional diversity. Mol. Biol. Evol. 20, 735–747.
[46] Eisen, J.A. (1998) Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Res. 8, 163–167.
[47] Davletov, B.A., Shamotienko, O.G., Lelianova, V.G., Grishin, E.V. and Ushkaryov, Y.A. (1996) Isolation and biochemical characterization of a Ca2+-independent alpha-latrotoxin-binding protein. J. Biol. Chem. 271, 23239–23245.
[48] Krasnoperov, V.G., Bittner, M.A., Beavis, R., Kuang, Y., Salnikow, K.V., Chepurny, O.G., Little, A.R., Plotnikov, A.N., Wu, D., Holz, R.W. and Petrenko, A.G. (1997) Alpha-Latrotoxin stimulates exocytosis by the interaction with a neuronal G-protein-coupled receptor. Neuron 18, 925–937.
[49] Matsushita, H., Lelianova, V.G. and Ushkaryov, Y.A. (1999) The latrophilin family: multiply spliced G protein-coupled receptors with differential tissue distribution. FEBS Lett. 443, 348–352.
[50] Kramer, J.M. (1994) Structures and functions of collagens in *Caenorhabditis elegans*. FASEB J. 8, 329–336.