Immunogenetics (2010) 62:263–272
DOI 10.1007/s00251-010-0428-1

ORIGINAL PAPER

Domain architecture evolution of pattern-recognition receptors

Qing Zhang · Christian M. Zmasek · Adam Godzik

Received: 18 September 2009 / Accepted: 3 February 2010 / Published online: 2 March 2010

Abstract In animals, the innate immune system is the first line of defense against invading microorganisms, and the pattern-recognition receptors (PRRs) are the key components of this system, detecting microbial invasion and initiating innate immune defenses. Two families of PRRs, the intracellular NOD-like receptors (NLRs) and the transmembrane Toll-like receptors (TLRs), are of particular interest because of their roles in a number of diseases. Understanding the evolutionary history of these families and their pattern of evolutionary changes may lead to new insights into the functioning of this critical system. We found that the evolution of both NLR and TLR families included massive species-specific expansions and domain shuffling in various lineages, which resulted in the same domain architectures evolving independently within different lineages in a process that fits the definition of parallel evolution. This observation illustrates both the dynamics of the innate immune system and the effects of "combinatorially constrained" evolution, where existence of the limited numbers of functionally relevant domains constrains the choices of domain architectures for new members in the family, resulting in the emergence of independently evolved proteins with identical domain architectures, often mistaken for orthologs.

Keywords Parallel evolution · Lineage-specific expansion · Domain shuffling · NOD-like receptor · Toll-like receptor · Innate immunity

Abbreviations

BIR	baculovirus inhibitor of apoptosis repeat
CARD	caspase recruitment domain
CIITA	MHC class II transactivator
DAMPs	damage-associated molecular patterns
Ig	immunoglobulin
IL-1R	interleukin-1 receptor
IPAF	ICE (IL-1β converting enzyme) protease activating factor
LRRs	leucine-rich repeats
MyD88	myeloid differentiation factor 88
NACHT	domain present in NAIP, CIITA, HET-E, and TP1
NAIP	neuronal apoptosis inhibitory protein
NALPs	NACHT, LRR, and PYRIN domain-containing proteins
NLRs	NOD-like receptors
PAMPs	pathogen-associated molecular patterns
PRRs	pattern-recognition receptors
PYRIN (also known as PAAD, PYD, or DAPIN domain)	the N-terminal domain of protein pyrin

Qing Zhang and Christian M. Zmasek contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s00251-010-0428-1) contains supplementary material, which is available to authorized users.

Q. Zhang · C. M. Zmasek · A. Godzik
Burnham Institute for Medical Research,
10901 North Torrey Pines Road,
La Jolla, CA 92037, USA

A. Godzik (✉)
Skaggs School of Pharmacy and Pharmaceutical Sciences,
University of California,
San Diego, 9500 Gilman Drive,
La Jolla, CA 92093, USA
e-mail: adam@burnham.org

SARM	sterile α and HEAT-Armadillo motifs containing protein
TIR	Toll/interleukin-1 receptor
TIRAP	TIR domain containing adaptor protein
TLRs	Toll-like receptors
TRAM	TRIF-related adaptor molecule
TRIF	TIR domain-containing adaptor inducing interferon-β

Introduction

Evolution of eukaryotic genomes is characterized by complex genome rearrangements leading to gene duplication, fusion, fission, recombination, and loss of fragments. These effects play a significant role in the evolution of many gene families and can lead to extensive domain rearrangements and evolution of novel domain architectures (Apic et al. 2001; Bjorklund et al. 2005; Patthy 2003; Weiner et al. 2006). Another phenomenon illustrating the dynamics of genome evolution are lineage-specific protein family expansions (Lespinet et al. 2002), seen first in *Caenorhabditis elegans* (Copley et al. 1999), but coming fully into focus only after sequencing of the sea urchin genome (Sodergren et al. 2006). In sea urchin, several innate immune- and apoptosis-related families underwent an unprecedented expansion as compared to any previously sequenced organism (Hibino et al. 2006; Robertson et al. 2006). There are many questions surrounding these expansions. How did the functions of recently diverged paralogs evolve? Is the number of paralogs after species-specific expansion similar in different species? Is it possible that proteins in such independently evolved groups would converge on similar functions? In this paper, we attempt to answer some of these questions for the several groups of innate immune-related receptors and regulators, which display all the phenomena mentioned in this paragraph.

The regulation of innate immune responses relies on several families of pattern-recognition receptors (PRRs) that recognize pathogen- or damage-associated molecular patterns (PAMPs, DAMPs), which originate from invading pathogens or are released by dying or injured cells. In the absence of adaptive immunity, the number and diversity of PRRs may provide an advantage to an organism living in pathogen-rich environments. Two families of PRRs conserved from early invertebrates to mammals, the intracellular NOD-like receptors (NLRs) (Fritz et al. 2006; Martinon and Tschopp 2005; Ting et al. 2006) and the transmembrane toll-like receptors (TLRs; Beutler et al. 2006; Medzhitov 2001; West et al. 2006), are of particular interest because of their roles in a number of diseases.

The NLR family is a group of cytoplasmic PRRs that are characterized by the presence of a conserved nucleotide binding NACHT domain. The general domain organization of NLRs includes an N-terminal effector domain, such as a caspase recruitment domain (CARD), a PYRIN domain (also known as PAAD, PYD, or DAPIN), or several baculovirus inhibitor of apoptosis repeat (BIR) domains, all of which mediate protein–protein interactions for initiating downstream signaling; a centrally located NACHT domain, which is required for nucleotide binding and self-oligomerization; and an array of C-terminal leucine-rich repeat (LRR) domains, which mediate ligand sensing and autorepression (Kanneganti et al. 2007; Martinon and Tschopp 2005). Human NLRs can be classified into several subgroups according to their N-terminal effector domain: CARD-containing NODs, IPAF, and CIITA; PYRIN-containing NALPs; and BIR-containing NAIP (Kanneganti et al. 2007; Martinon and Tschopp 2005). The second family of innate immune receptors discussed here is the TLR family, which belongs to type I transmembrane receptors and is characterized by its C-terminal signaling domain–Toll/Interleukin-1 receptor (TIR) domain (West et al. 2006)–and N-terminal LRR domains. The TIR domain is also present in the interleukin-1 receptor (IL-1R) family and in the TIR-domain-containing adaptors (Boraschi and Tagliabue 2006; McGettrick and O'Neill 2004). The interleukin-1 receptors use the immunoglobulin (Ig) domain for ligand binding instead of the LRR domain as in TLRs (Boraschi and Tagliabue 2006).

As genomes of more and more organisms become available, the phylogenetic analysis of NLR and TLR families can be done across a large number of species, which is useful for deciphering the evolutionary relationships inside these families and helps us understand the evolutionary dynamics of the innate immune system. We show here that the above receptor families have especially interesting evolutionary histories, undergoing large expansions and extensive domain recombination in various lineages. In particular, several domain architectures, such as Death–NACHT–LRR, CARD–NACHT–LRR, PYRIN–NACHT–LRR, and Ig–TIR, have emerged multiple times in different lineages, suggesting that parallel evolution is a common phenomenon in the evolution of innate immunity.

Materials and methods

Sequence database searches

The v1.0 genome assemblies and related protein sets of amphioxus (*Branchiostoma floridae*) and sea anemone

(*Nematostella vectensis*) were downloaded from the Joint Genome Institute (http://www.jgi.doe.gov). The genome assembly Spur_v2.0 and the GLEAN3 gene models for the sea urchin (*Strongylocentrotus purpuratus*) were obtained from the Baylor College of Medicine Human Genome Sequencing Center (http://www.hgsc.bcm.tmc.edu). The other genome sequences and corresponding protein sets were downloaded from Ensembl (http://www.ensembl.org). Several rounds of PSI-TBLASTN searches (Altschul et al. 1997) were performed against each genome by using known human NACHT or TIR domain amino acid sequences as seeds. The hits were then mapped to the corresponding genome protein set to acquire the full-length protein sequences (for sea urchin and sea anemone, some of the gene models were in addition predicted by GENSCAN (Burge and Karlin 1998)). All identified genes were checked by reciprocal BLAST analysis, Pfam protein searches (Bateman et al. 2004), Conserved Domain Search (CD-Search), and Reverse PSI-BLAST (Marchler-Bauer and Bryant 2004). Domains verified by Pfam and CD-Search are evolutionarily conserved units in proteins (Bateman et al. 2004; Marchler-Bauer et al. 2002). Additionally, two lamprey TIR domain-containing proteins (laTLR14a and laTLR14b) identified in (Ishii et al. 2007) are also included.

Multiple sequence alignments and phylogeny reconstructions

In phylogenetic analysis of multidomain families, for both practical and conceptual reasons, it is critical to analyze each domain separately. In multidomain proteins, variable linker lengths, different mutation rates in different domains, and occasional domain losses, duplications, or substitutions make it oftentimes impossible to build high-quality alignments across more than one domain. At the same time, making alignments and performing phylogenetic analysis on only the subset of protein families with the same domain architectures would likely produce a misleading picture by neglecting the possible gene recombination and domain rearrangement events. In this paper, the phylogenetic analyses of NLR and TLR families are based on the NACHT domain and the TIR domain, accordingly. To ensure alignment of homologous domains, collected protein sequences with NACHT or TIR domains were trimmed according to Pfam 21.0 models (Bateman et al. 2004). Multiple sequence alignments were produced by PROBCONS 1.12 (Do et al. 2005), MAFFT 6.240 (localpair, maxiterate 1000) (Katoh et al. 2005), and hmmalign from HMMER 2.3.2 (Eddy 1998; Nuin et al. 2006). Multiple sequence alignment columns with a gap in more than 50% of sequences were deleted.

Phylogenetic analysis was performed using three different approaches. For the Bayesian inference approach, MrBayes 3.1.2 was used with 4,000,000 generations, 64 chains, a sample frequency of 1,000, a mixture of amino-acid models with fixed-rate matrices and equal rates, and 25% burn in (Ronquist and Huelsenbeck 2003). For the maximum likelihood approach, RAxML 7.0.4 was used with rapid bootstrap analysis (100 steps) and search for the best-scoring ML tree ("-f a" option), the variable time (VT) model and four relative rate substitution categories with empirical base frequencies (Stamatakis 2006). For distance-based approaches, such as FastME 1.1 (Desper and Gascuel 2002), neighbor-joining from PHYLIP 3.66 (Felsenstein 1989; Saitou and Nei 1987), and BIONJ (Gascuel 1997), pair-wise distances were calculated by TREE-PUZZLE 5.2 using the VT model (Schmidt et al. 2002). Phylogenetic trees were drawn using Archaeopteryx 0.901 (http://www.phylosoft.org/archaeopteryx/). All conclusions presented in this work are robust under different multiple sequence alignment and phylogeny reconstruction methods. All sequence, alignment, and phylogeny files are available upon request.

Domain composition analysis

Domains were analyzed with hmmpfam from HMMER 2.3.2 and Pfam 21.0 (Bateman et al. 2004; Eddy 1998).

Structural modeling

The crystal structure of Apaf-1 CARD from Apaf-1/procaspase-9 complex (PDB code 3YGS; Qin et al. 1999) was used as a template for modeling other CARD domains. The SCWRL program (Canutescu et al. 2003) was used for homology modeling, and APBS (Baker et al. 2001) was used for calculating surface potentials. All structure figures were prepared with PyMOL (http://www.pymol.org).

Results

The NACHT protein family

We collected the NACHT domain-containing genes from three recently sequenced marine invertebrate genomes whose sequences became publicly available in the last 2 years, including a cephalochordate (the amphioxus *B. floridae*; Putnam et al. 2008), an echinoderm (the sea urchin *S. purpuratus*; (Sodergren et al. 2006), and a cnidarian (the sea anemone *N. vectensis*; Putnam et al. 2007). We also collected the NLR genes from other animals, including several vertebrates (the human *Homo sapiens*, the mouse *Mus musculus*, the dog *Canis familiaris*,

the chicken *Gallus gallus*, the western clawed frog *Xenopus tropicalis*, the zebrafish *Danio rerio*, the Japanese pufferfish *Fugu rubripes*, and the green pufferfish *Tetraodon nigroviridis*), and a urochordate (the transparent sea squirt *Ciona intestinalis*). No NLR-like genes were found in the arthropod fruit fly *Drosophila melanogaster* and the nematode *C. elegans*—two very popular model organisms.

Previously, it was thought that the NLR genes are vertebrate-/deuterostome-specific, since they were absent in both *Drosophila* and *C. elegans* (Fritz et al. 2006). However, multiple copies of NACHT domain-containing proteins were found in sea anemone, an animal that belongs to the basal phylum Cnidaria, suggesting that this family emerged even before the protostome–deuterostome split (Darling et al. 2005; Putnam et al. 2007) and was lost in the arthropod and nematode lineages. The repertoire of NLR proteins in mammals is fairly stable at around 20, while there is significant (five to ten times) expansion of this receptor gene family in both invertebrate deuterostomes—amphioxus and sea urchin—and to a smaller extent in some fish genomes, such as pufferfish.

A detailed phylogenetic analysis of the NACHT domains of NLR proteins shows that all the invertebrate NLRs belong to the lineage-specific groups (Fig. 1; Supplementary Fig. 1), suggesting that all extant NACHT-containing genes in invertebrates are a result of multiple rounds of duplication of a single ancestral gene. Interestingly, despite that, several recurring domain architectures of NLR proteins can be found throughout the tree. The possible explanation of this is a scenario that includes multiple, independent evolution of similar domain architectures in a process that fits the definition of parallel evolution (West-Eberhard 2003). We do not know what the ancestral domain architecture of NLR protein at the internal node (A) was, as it could have been any of the following: Death–NACHT–LRR, CARD–NACHT–LRR, or NACHT–LRR (or, less likely, NACHT associated with other domains or by itself; Fig. 1). But no matter what the ancestral domain organization looked like, other domain architectures must have evolved independently. For example, if we assume that the Death–NACHT–LRR architecture is ancestral, then the CARD–NACHT–LRR architecture evolved independently in the IPAF branch and in the second amphioxus-specific branch. If the ancestor had the CARD–NACHT–LRR domain architecture, then the Death–NACHT–LRR association appeared later separately in the sea urchin branch and the amphioxus-specific branch. The same situation encountered with the NACHT–LRR or other NACHT domain associations, in such case, both the Death–NACHT–LRR and the CARD–NACHT–LRR architectures should be evolved multiple times in the descendants. At another internal node (B), we have a similar situation in which no matter what the ancestral domain architecture was PYRIN–NACHT–LRR, CARD–NACHT–LRR, NACHT–LRR or other (Fig. 1), one or more architectures had to evolve independently more than once from the same ancestor. This scenario fits the definition of parallel rather than convergent evolution. In parallel evolution, similar traits (here domain architectures) independently evolve from similar ancestral states (here the ancestral NACHT-containing gene); whereas in convergent evolution, similar traits evolve from unrelated (or distantly related) ancestral states. We can only speculate about the driving force behind this independent emergence of identical domain architectures as being related to the pressure elicited by pathogens and the advantage provided by quick response by the innate immunity system.

The TIR protein family

The TIR domain-containing proteins constitute another group of proteins involved in the innate immune response. The TIR domain is present in several groups of proteins with different domain architectures: the transmembrane TLRs and IL-1Rs, as well as in the intracellular TIR domain-containing adaptors (such as myeloid differentiation factor 88 (MyD88), TIR domain-containing adaptor protein (TIRAP), TIR domain-containing adaptor inducing interferon-β (TRIF), TRIF-related adaptor molecule (TRAM), and sterile α and HEAT-Armadillo motifs containing protein (SARM) in human; McGettrick and O'Neill 2004; O'Neill and Bowie 2007).

Similar to the situation with the NACHT domain, the TIR protein family also underwent a large expansion in both amphioxus and sea urchin. There are around 24 TIR domain-containing proteins in mammals, 11 in *Drosophila*, and 2 in *C. elegans*; whereas there are more than 100 copies of such proteins in both amphioxus and sea urchin genomes.

The exact phylogenetic tree for the entire TIR domain family remains elusive because of extreme sequence diversity in this family, but the separation of the main sub-branches is quite robust under different multiple sequence alignment and tree-building methods. The tree can be divided into two parts—the TLR family together with the TIR domain-containing adaptors and the IL-1R family (Fig. 2). Many TIR domain-containing proteins from invertebrates belong to species-specific branches. Similar to the previously discussed case of NLR proteins, while we do not know what the ancestral TIR domain-containing protein looked like, multiple lines of circumstantial evidence suggest the role that parallel evolution played in the evolution of the TIR domain family. In this case, we have an indirect argument that an IL-1R-like domain architecture with an Ig–TIR domain combination probably was not ancestral, as no IL-1- or IL-18-like gene

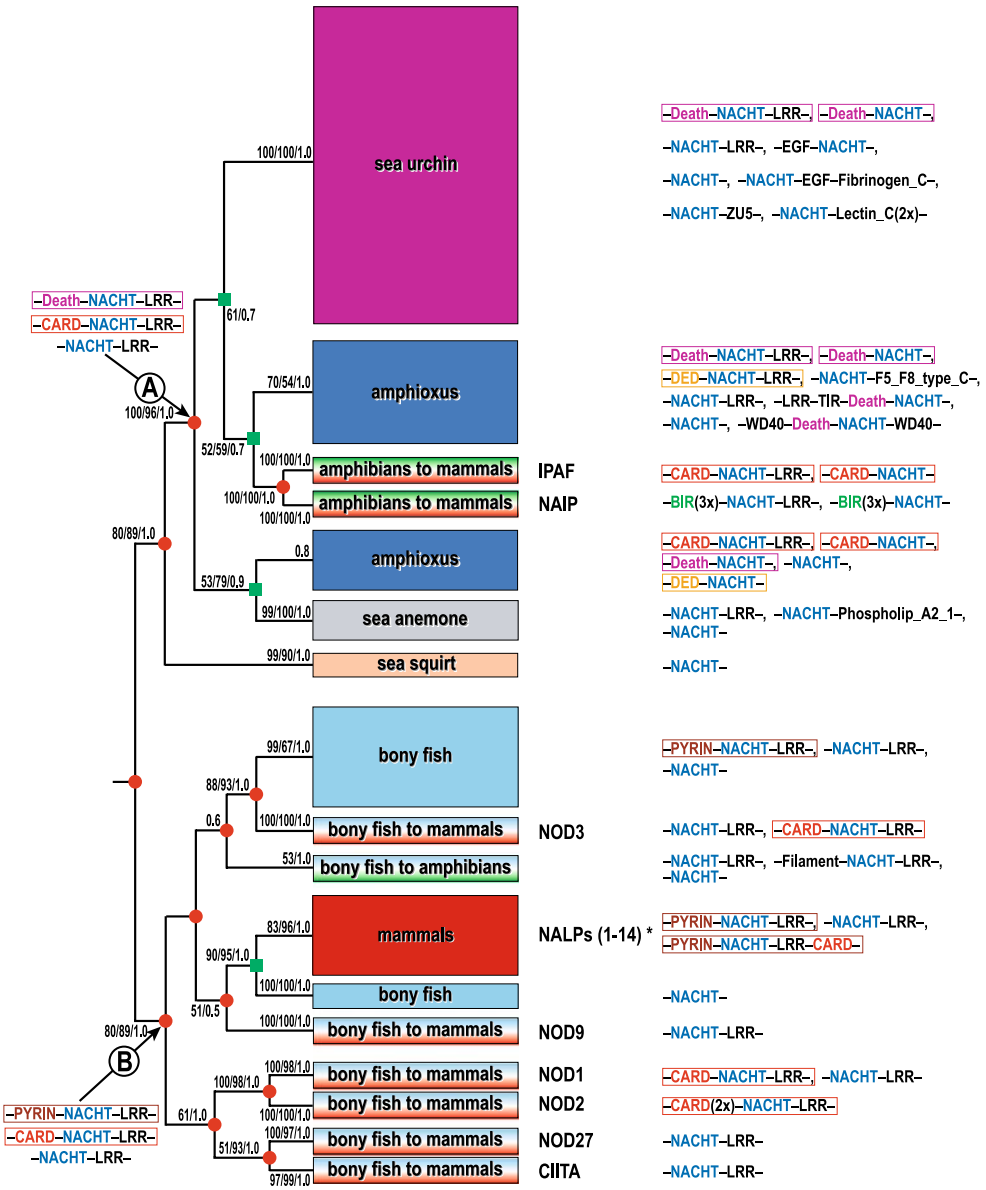


Fig. 1 Overview of the phylogeny and domain organization of the NACHT protein family. On the phylogenetic tree, branches are assigned with *different background colors* according to species. The *numbers associated with each clade* indicate bootstrap values for minimum evolution and maximum likelihood and the posterior probability from Bayesian inference, respectively (values lower than 50% are not shown). The size of each branch is proportional to the number of proteins inside that branch with the domain architectures shown on the *right side*. Duplication event is indicated by a *red circle* and speciation by a *green square*. Domain architectures at the internal nodes **A** and **B** are speculated. The detailed sequence information for each terminal node is shown in Supplementary Fig. 1. Some unusual NACHT gene models in sea urchin and amphioxus require further experimental verification. *The NALPs branch appears to be mammalian-specific, with the exception of one protein from chicken with the canonical PYRIN–NACHT–LRR architecture

(the main ligands for the vertebrate IL-1R family; Boraschi and Tagliabue 2006) was found in amphioxus and sea anemone. Although we cannot rule out the possibility that cytokines in these two animals are too divergent to be recognized, no matter what the ancestral domain architecture was at the internal node (A; Fig. 2), either the Ig–TIR or LRR–TIR domain combination has evolved independently more than once.

Parallel evolution can result in proteins with identical domain architectures, such as amphioxus and sea anemone IL-1R-like proteins, which look like the vertebrate IL-1R family, but most likely have evolved independently (Fig. 2). Another, better-known example can be found in the TLR family, where human and *Drosophila* TLR proteins, despite the similar size of the family and numbering scheme that may suggest one-to-one orthology between individual proteins in fruit fly and human, have also evolved independently (except Toll-9, which is the only *Drosophila* toll family member that groups with the vertebrate TLRs).

Structural features of the associated protein–protein interaction domain

Pattern-recognition receptors use the associated protein–protein interaction domains, such as CARD, Death, and PYRIN, to connect to the downstream part of the signaling cascade. Proteins that evolved by parallel evolution arose independently from each other; even though they have identical domain architectures, their individual domains come from non-orthologous branches and may have different functions.

Phylogenetic analysis cannot tell us more about functional differences or similarities between such proteins. However, for proteins for which the functions are well understood and the three-dimensional structures are available, we can use other tools, such as protein structure modeling and model analysis, to reason about their functional similarities or differences. In the example in Fig. 3, the two amphioxus proteins with similar domain architectures, both containing a CARD–fn3 domain combination, most likely evolved independently. On the other hand, chicken and mouse CARD domains are part of proteins with different domain architectures that both represent a specific vertebrate expansion of the CARD family. For CARD domains, their function (protein–protein binding) is determined by the charge distribution of their surfaces (Qin et al. 1999) and in particular by the details of the dimer interface (see the top panel in Fig. 3). We built three-dimensional models of the four CARD domains mentioned above and calculated the charge distributions of their surfaces using the APBS package (Baker et al. 2001). Surface similarity analysis can be used to compare two protein structures (Binkowski and Joachimiak 2008; Dlugosz and Trylska 2008; Pawlowski and Godzik 2001; Sael et al. 2008; Sasin et al. 2007), similar to using sequence similarity to compare sequences with the difference that the resulting surface similarity score is related directly to function, rather than to evolutionary relation between two proteins. Here, the surface feature differences between the last structure (02_BRAFL_CARD in Fig. 3) and the other three are noticeable even by visual inspection. As seen in Fig. 3, even though the two CARD domains from amphioxus come from proteins with similar domain architectures and are both part of amphioxus-specific expansion of CARD domains, one of them has surface features suggesting functional similarity to CARD domains from mouse or chicken, despite their low sequence similarity (sequence identity around 20%), while the other domain has a very different charge distribution and is likely to interact with a different downstream partner.

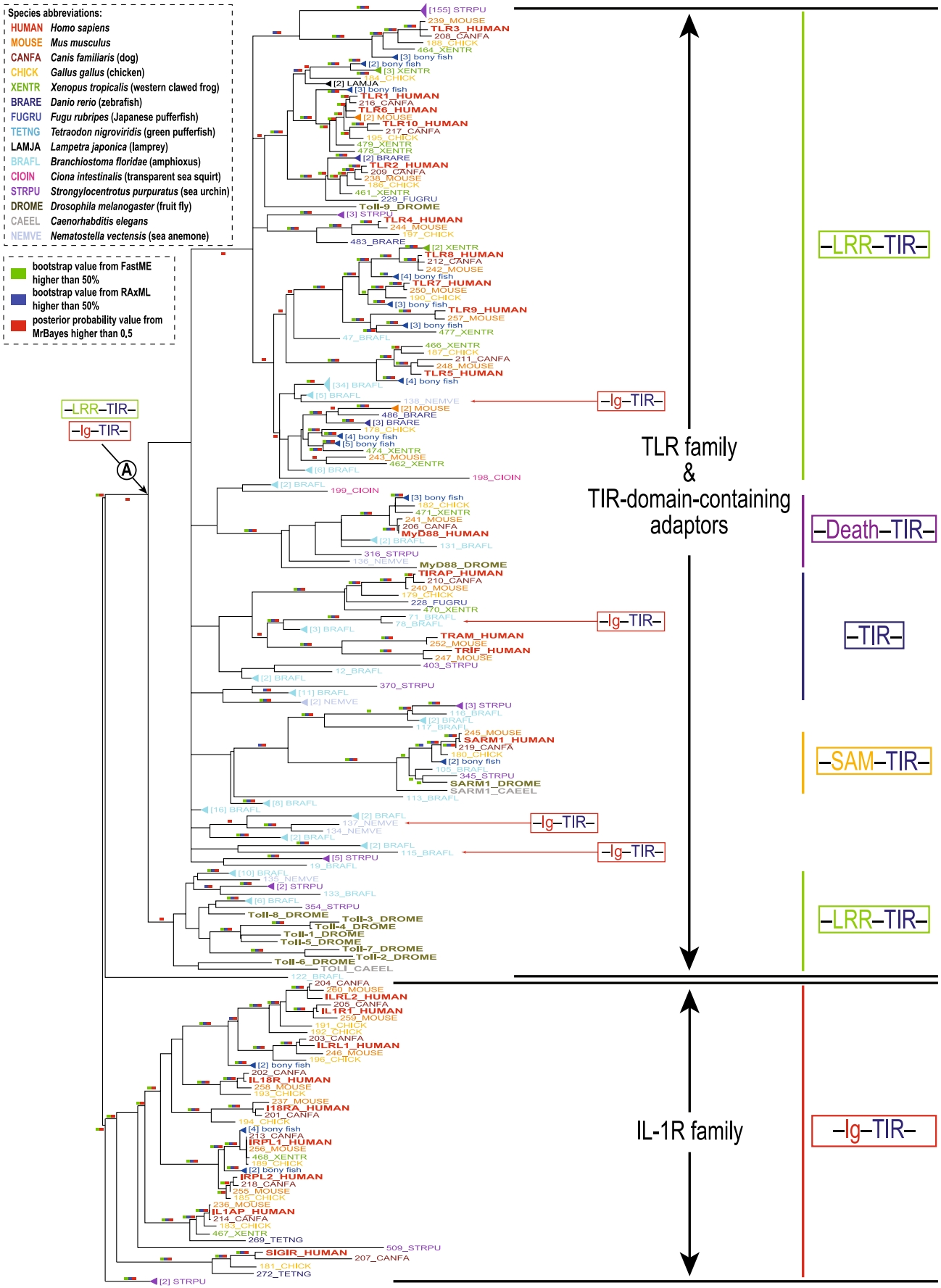
Discussion

The conservation of NLR and TLR receptors from (at least) cnidarians to mammals highlights the importance and the ancient evolutionary history of these important innate immunity families. The presence of multiple proteins with similar domain architectures creates the impression that all these proteins and, by extension, possibly even the specific pathways in which they participate, could have been present in ancestral species. However, we show here that the appearance of conservation hides a very complex evolutionary history of these receptor families, which underwent massive species-specific expansions and independently evolved identical domain architectures. This phenomenon is most obvious in the NACHT protein family, where all invertebrate NLR proteins evolved by species-specific expansions (Fig. 1). However, this phenomenon also plays an important role in the evolution of the TIR protein family (Fig. 2) and possibly other families. The expansions of these protein families in fish and amphioxus were noticed earlier by several independent studies (Laing et al. 2008; Oshiumi et al. 2008; Stein et al. 2007; Zhang et al. 2008). Also, studies for TLRs and other innate immune-related protein families between arthropods and vertebrates reach similar conclusions that members of innate immune systems could have evolved independently (Hughes 1998; Hughes and Piontkivska 2008), which reinforces our parallel evolution hypothesis.

The most interesting observation is that such massive expansions and domain shuffling only resulted in a relatively small number of protein architectures. Clearly, the number of possible solutions must be limited by functional considerations that act as constraints, limiting the potentially huge number of possible domain architectures to the same, independently rediscovered ones. The presence of such constraints limiting the number of functional domain combinations provides a possible alternative explanation for the conservation of domain architectures in eukaryotes, where the majority of the genomic proteins are multidomain proteins (Han et al. 2007), but only a small fraction of all possible domain combinations are present.

Some studies suggested that domain architectures are largely inherited (Gough 2005). However, a more recent study indicates that domain architecture reinvention is a more common phenomenon than previous thought (Forslund et al. 2008). These authors suggested that between 5.6% and 12% of all domain architectures could have been created more than once in different genomes. In this paper we show specific examples of parallel evolution in families of innate immune receptors: the NACHT and TIR protein families. Both NACHT and TIR domain are protein–protein

Fig. 2 Phylogeny of the TIR protein family. Protein names are assigned to *different colors* according to species origin. Lineage-specific subtrees are collapsed for clarity with the number of sequences shown in *brackets*. Support values for each clade are indicated by *differently colored boxes* to increase their visibility. Bootstrap values higher than 50% from minimum evolution and maximum likelihood approaches are colored in *green and blue*, respectively. Posterior probability values from a Bayesian approach higher than 0.5 are colored in *red*. The simplified domain architectures of each branch are presented on the *right side* of the phylogenetic tree. Domain architectures at the internal node **A** are speculated. Ig–TIR domain-containing sequences from amphioxus and sea anemone are separate from the vertebrate IL-1R branch, which has the same domain architecture. The detailed sequence information can be found in Supplementary Table 2


Species abbreviations:
HUMAN Homo sapiens
MOUSE Mus musculus
CANFA Canis familiaris (dog)
CHICK Gallus gallus (chicken)
XENTR Xenopus tropicalis (western clawed frog)
BRARE Danio rerio (zebrafish)
FUGRU Fugu rubripes (Japanese pufferfish)
TETNG Tetraodon nigroviridis (green pufferfish)
LAMJA Lampetra japonica (lamprey)
BRAFL Branchiostoma floridae (amphioxus)
CIOIN Ciona intestinalis (transparent sea squirt)
STRPU Strongylocentrotus purpuratus (sea urchin)
DROME Drosophila melanogaster (fruit fly)
CAEEL Caenorhabditis elegans
NEMVE Nematostella vectensis (sea anemone)
bootstrap value from FastME higher than 50%
bootstrap value from RAxML higher than 50%
posterior probability value from MrBayes higher than 0.5
–LRR–TIR–
–Ig–TIR–
A
TLR family & TIR-domain-containing adaptors
IL-1R family
–LRR–TIR–
–Death–TIR–
–TIR–
–SAM–TIR–
–LRR–TIR–
–Ig–TIR–

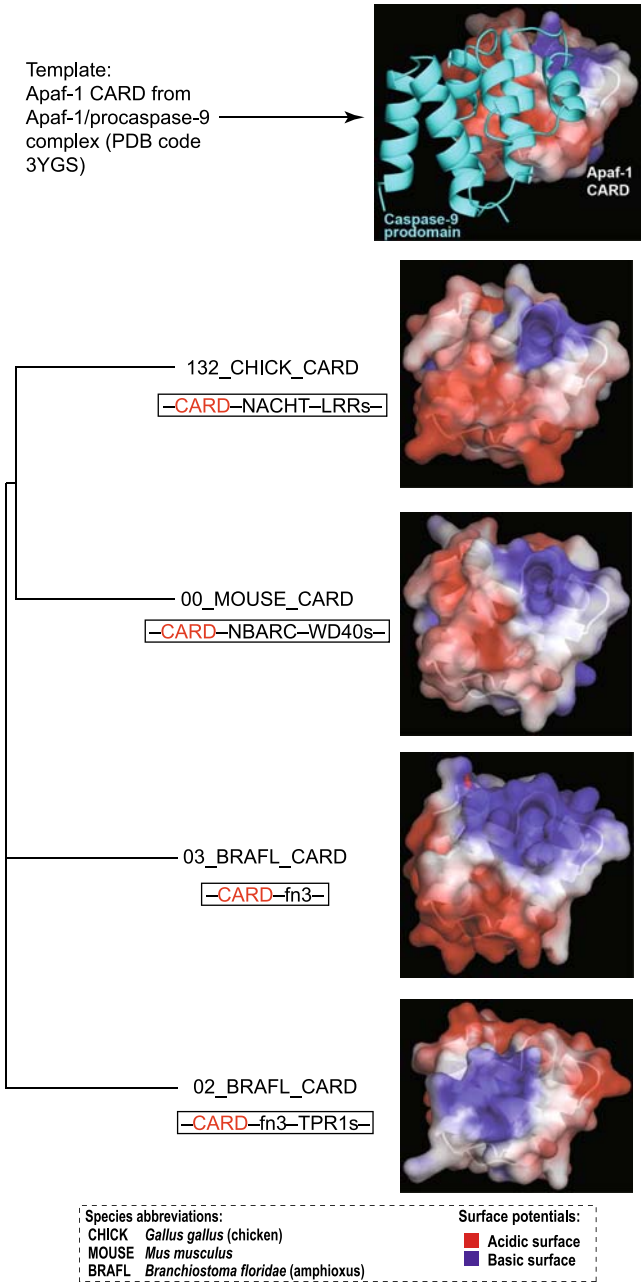


Fig. 3 Structure comparison between different CARD domains. A phylogeny based on the CARD domains is shown on the left, and the related structure models are displayed on the right. The electrostatic potential was mapped on the protein surface of each 3D model. The acidic surface (negatively charged) is colored in *red*, and the basic surface (positively charged) is in *blue*. All protein surfaces are displayed in the same view with the putative binding surface (the surface used by the Apaf-1 CARD that binds the Caspase-9 prodomain (Qin et al. 1999) shown on the top panel) facing the viewer. The potential binding surface of the last structure (02_BRAFL_CARD) is dominated by a positive charge while the other three structures are mainly negatively charged

interaction domains that contribute to signal transduction, and this functional class of proteins was called "promiscuous" because of their tendency to associate with different domains (Basu et al. 2008). When compared with the list of the top 215 highly promiscuous domains in eukaryotes (Basu et al. 2008), it turned out not only the NACHT and TIR domain themselves, but also the domains they associate with, such as Death, CARD and Ig domains, are on that list. However, only a small fraction of the possible domain combinations actually exist in nature, suggesting that domain architectures are under strong evolutionary selection (Han et al. 2007). For the NLR family proteins, where only a limited number of protein–protein interaction domains such as Death, CARD, DED, or PYRIN domains can appear at the amino terminus, provides us with a clue how such selection may be executed. These four domains belong to the death domain superfamily, which has very similar structures and modes of action (Reed et al. 2004). Reshuffling between these domains would not incur much structural conflict with the function of controlled oligomerization facilitated by the NACHT domain. In this context, it may be worth mentioning that the PYRIN domain, which is not found in any currently sequenced invertebrate genomes, has probably evolved from other death domain superfamily members and represents another example of domain reshuffling. It is found in several very different types of protein architectures.

While the emergence of similar domain architectures can be clearly shown by comparing predicted genes identified in genome projects, we still do not know if proteins with the same domain architecture share similar functions in different species. Some examples, including those from the families discussed here, suggest that this is not always true. For instance, while *Drosophila* toll-like receptors mainly carry out roles in embryonic development, their mammalian homologs are key regulators of immune responses (Kambris et al. 2002; Leulier and Lemaitre 2008). For other proteins, we have some indirect arguments about their functional divergence. For example, both amphioxus and sea anemone have Ig–TIR domain-containing sequences, the same architecture as IL-1R family members in vertebrates. These sequences are likely reinvented in various animal lineages by parallel evolution. The function of the IL-1R-like proteins in amphioxus and sea anemone is not clear and could be different from its corresponding sequences in vertebrates, as no IL-1- or IL-18-like genes were found in these two genomes. Further experimental work is needed to unravel the precise roles of these proteins. We also show here that proteins that evolved independently by parallel evolution can have very divergent surface features (Fig. 3). Therefore, extrapolation of protein function based on the domain architecture must be done very carefully.

Acknowledgments This work was supported by grants AI056324 and GM076221 from the National Institutes of Health. *B. floridae* and *N. vectensis* genome data, including gene models and annotations, were produced by the US Department of Energy Joint Genome Institute and downloaded from their web site. *S. purpuratus* genome data were produced by the Sea Urchin Genome Project at the Baylor College of Medicine. The authors acknowledge the JGI, the HGSC, and all other sequencing centers for their efforts in sequencing, assembling, and annotating the genomes that we used for the analysis presented here.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Apic G, Gough J, Teichmann SA (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. J Mol Biol 310:311–325

Baker NA, Sept D, Joseph S, Holst MJ, McCammon JA (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci U S A 98:10037–10041

Basu MK, Carmel L, Rogozin IB, Koonin EV (2008) Evolution of protein domain promiscuity in eukaryotes. Genome Res 18:449–461

Bateman A, Coin L, Durbin R et al (2004) The Pfam protein families database. Nucleic Acids Res 32:D138–D141

Beutler B, Jiang Z, Georgel P, Crozat K, Croker B, Rutschmann S, Du X, Hoebe K (2006) Genetic analysis of host resistance: toll-like receptor signaling and immunity at large. Annu Rev Immunol 24:353–389

Binkowski TA, Joachimiak A (2008) Protein functional surfaces: global shape matching and local spatial alignments of ligand binding sites. BMC Struct Biol 8:45

Bjorklund AK, Ekman D, Light S, Frey-Skott J, Elofsson A (2005) Domain rearrangements in protein evolution. J Mol Biol 353:911–923

Boraschi D, Tagliabue A (2006) The interleukin-1 receptor family. Vitam Horm 74:229–254

Burge CB, Karlin S (1998) Finding the genes in genomic DNA. Curr Opin Struct Biol 8:346–354

Canutescu AA, Shelenkov AA, Dunbrack RL Jr (2003) A graph-theory algorithm for rapid protein side-chain prediction. Protein Sci 12:2001–2014

Copley RR, Schultz J, Ponting CP, Bork P (1999) Protein families in multicellular organisms. Curr Opin Struct Biol 9:408–415

Darling JA, Reitzel AR, Burton PM, Mazza ME, Ryan JF, Sullivan JC, Finnerty JR (2005) Rising starlet: the starlet sea anemone, *Nematostella vectensis*. Bioessays 27:211–221

Desper R, Gascuel O (2002) Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J Comput Biol 9:687–705

Dlugosz M, Trylska J (2008) Electrostatic similarity of proteins: application of three dimensional spherical harmonic decomposition. J Chem Phys 129:015103

Do CB, Mahabhashyam MS, Brudno M, Batzoglou S (2005) ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res 15:330–340

Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14:755–763

Felsenstein J (1989) PHYLIP—Phylogeny Inference Package (Version 3.2). Cladistics 5:164–166

Forslund K, Henricson A, Hollich V, Sonnhammer EL (2008) Domain tree-based analysis of protein architecture evolution. Mol Biol Evol 25:254–264

Fritz JH, Ferrero RL, Philpott DJ, Girardin SE (2006) Nod-like proteins in immunity, inflammation and disease. Nat Immunol 7:1250–1257

Gascuel O (1997) BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol 14:685–695

Gough J (2005) Convergent evolution of domain architectures (is rare). Bioinformatics 21:1464–1471

Han JH, Batey S, Nickson AA, Teichmann SA, Clarke J (2007) The folding and evolution of multidomain proteins. Nat Rev Mol Cell Biol 8:319–330

Hibino T, Loza-Coll M, Messier C et al (2006) The immune gene repertoire encoded in the purple sea urchin genome. Dev Biol 300:349–365

Hughes AL (1998) Protein phylogenies provide evidence of a radical discontinuity between arthropod and vertebrate immune systems. Immunogenetics 47:283–296

Hughes AL, Piontkivska H (2008) Functional diversification of the toll-like receptor gene family. Immunogenetics 60:249–256

Ishii A, Matsuo A, Sawa H, Tsujita T, Shida K, Matsumoto M, Seya T (2007) Lamprey TLRs with properties distinct from those of the variable lymphocyte receptors. J Immunol 178:397–406

Kambris Z, Hoffmann JA, Imler JL, Capovilla M (2002) Tissue and stage-specific expression of the tolls in *Drosophila* embryos. Gene Expr Patterns 2:311–317

Kanneganti TD, Lamkanfi M, Nunez G (2007) Intracellular NOD-like receptors in host defense and disease. Immunity 27:549–559

Katoh K, Kuma K, Toh H, Miyata T (2005) MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33:511–518

Laing KJ, Purcell MK, Winton JR, Hansen JD (2008) A genomic view of the NOD-like receptor family in teleost fish: identification of a novel NLR subfamily in zebrafish. BMC Evol Biol 8:42

Lespinet O, Wolf YI, Koonin EV, Aravind L (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes. Genome Res 12:1048–1059

Leulier F, Lemaitre B (2008) Toll-like receptors—taking an evolutionary approach. Nat Rev Genet 9:165–178

Marchler-Bauer A, Bryant SH (2004) CD-Search: protein domain annotations on the fly. Nucleic Acids Res 32:W327–W331

Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res 30:281–283

Martinon F, Tschopp J (2005) NLRs join TLRs as innate sensors of pathogens. Trends Immunol 26:447–454

McGettrick AF, O'Neill LA (2004) The expanding family of MyD88-like adaptors in toll-like receptor signal transduction. Mol Immunol 41:577–582

Medzhitov R (2001) Toll-like receptors and innate immunity. Nat Rev Immunol 1:135–145

Nuin PA, Wang Z, Tillier ER (2006) The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics 7:471

O'Neill LA, Bowie AG (2007) The family of five: TIR-domain-containing adaptors in toll-like receptor signalling. Nat Rev Immunol 7:353–364

Oshiumi H, Matsuo A, Matsumoto M, Seya T (2008) Panvertebrate toll-like receptors during evolution. Curr Genomics 9:488–493

Patthy L (2003) Modular assembly of genes and the evolution of new functions. Genetica 118:217–231

Pawlowski K, Godzik A (2001) Surface map comparison: studying function diversity of homologous proteins. J Mol Biol 309:793–806

Putnam NH, Srivastava M, Hellsten U et al (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317:86–94

Putnam NH, Butts T, Ferrier DE et al (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 453:1064–1071

Qin H, Srinivasula SM, Wu G, Fernandes-Alnemri T, Alnemri ES, Shi Y (1999) Structural basis of procaspase-9 recruitment by the apoptotic protease-activating factor 1. Nature 399:549–557

Reed, J. C., Doctor, K. S., and Godzik, A.: The domains of apoptosis: a genomics perspective. *Sci STKE 2004:* re9, 2004

Robertson AJ, Croce J, Carbonneau S, Voronina E, Miranda E, McClay DR, Coffman JA (2006) The genomic underpinnings of apoptosis in *Strongylocentrotus purpuratus*. Dev Biol 300:321–334

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574

Sael L, La D, Li B, Rustamov R, Kihara D (2008) Rapid comparison of properties on protein surface. Proteins 73:1–10

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sasin JM, Godzik A, Bujnicki JM (2007) Surf's up!—protein classification by surface comparisons. J Biosci 32:97–100

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) Tree-puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504

Sodergren E, Weinstock GM, Davidson EH et al (2006) The genome of the sea urchin *Strongylocentrotus purpuratus*. Science 314:941–952

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690

Stein C, Caccamo M, Laird G, Leptin M (2007) Conservation and divergence of gene families encoding components of innate immune response systems in zebrafish. Genome Biol 8:R251

Ting JP, Kastner DL, Hoffman HM (2006) Caterpillers, pyrin and hereditary immunological disorders. Nat Rev Immunol 6:183–195

Weiner J 3rd, Beaussart F, Bornberg-Bauer E (2006) Domain deletions and substitutions in the modular protein evolution. Febs J 273:2037–2047

West AP, Koblansky AA, Ghosh S (2006) Recognition and signaling by toll-like receptors. Annu Rev Cell Dev Biol 22:409–437

West-Eberhard MJ (2003) Developmental plasticity and evolution. Oxford University Press, Oxford

Zhang Q, Zmasek CM, Dishaw LJ, Mueller MG, Ye Y, Litman GW, Godzik A (2008) Novel genes dramatically alter regulatory network topology in amphioxus. Genome Biol 9:R123