Evolution of Genes and Repeats in the Nimrod Superfamily

Kálmán Somogyi,[1] Botond Sipos,*[1] Zsolt Pénzes,*† Éva Kurucz,* János Zsámboki,* Dan Hultmark,‡ and István Andó**

*Institute of Genetics, Biological Research Center of the Hungarian Academy of Sciences, Szeged, Hungary; †Department of Ecology, University of Szeged, Szeged, Hungary; and ‡Umea Centre for Molecular Pathogenesis, Umea University, Umea, Sweden

The recently identified Nimrod superfamily is characterized by the presence of a special type of EGF repeat, the NIM repeat, located right after a typical CCXGY/W amino acid motif. On the basis of structural features, *nimrod* genes can be divided into three types. The proteins encoded by Draper-type genes have an EMI domain at the N-terminal part and only one copy of the NIM motif, followed by a variable number of EGF-like repeats. The products of Nimrod B-type and Nimrod C-type genes (including the *eater* gene) have different kinds of N-terminal domains, and lack EGF-like repeats but contain a variable number of NIM repeats. Draper and Nimrod C-type (but not Nimrod B-type) proteins carry a transmembrane domain.

Several members of the superfamily were claimed to function as receptors in phagocytosis and/or binding of bacteria, which indicates an important role in the cellular immunity and the elimination of apoptotic cells. In this paper, the evolution of the Nimrod superfamily is studied with various methods on the level of genes and repeats. A hypothesis is presented in which the NIM repeat, along with the EMI domain, emerged by structural reorganizations at the end of an EGF-like repeat chain, suggesting a mechanism for the formation of novel types of repeats. The analyses revealed diverse evolutionary patterns in the sequences containing multiple NIM repeats. Although in the Nimrod B and Nimrod C proteins show characteristics of independent evolution, many internal NIM repeats in Eater sequences seem to have undergone concerted evolution. An analysis of the *nimrod* genes has been performed using phylogenetic and other methods and an evolutionary scenario of the origin and diversification of the Nimrod superfamily is proposed.

Our study presents an intriguing example how the evolution of multigene families may contribute to the complexity of the innate immune response.

Introduction

Genes of the recently described Nimrod superfamily (Kurucz et al. 2007) encode proteins containing various number of NIM repeats. The NIM repeat is a special type of the EGF domain (Pfam clan: CL0001), which is a—frequently repetitive—structural unit of a wide range of extracellular proteins in eukaryotic (mostly animal) organisms (Bork 1991; Bork et al. 1996). The consensus sequence of the NIM repeat (CXPXCXXXCXNGXCXXPXXCXCXXGY) is shifted by one cysteine unit relative to the typical EGF repeat consensus ($xxxxCx_{2–7}Cx_{1–4}(G/A)xCx_{1–13}ttaxCx\text{-}CxxGax_{1–6}GxxCx$) (Kurucz et al. 2007). Nimrod proteins have a characteristic structure (Kurucz et al. 2007). They all contain a signal peptide followed by N-terminal motifs of different type. The first NIM repeat is always located right after a typical CCxGY amino acid sequence motif. Based on other features, *nimrod* genes can be divided into three types.

Proteins encoded by Draper-type genes (e.g., *nimrod A*, *draper* in *Drosophila melanogaster*) have an EMI domain (Callebaut et al. 2003) at the N-terminal part and only one copy of the NIM motif, followed by a variable number of EGF domains. This type has wide taxonomic distribution, being present for example in *Caenorhabditis elegans* (Mangahas and Zhou 2005), fruit fly (Manaka et al. 2004), and human (Hamon et al. 2006) genomes.

On the other hand, proteins containing many NIM repeats ("poly-NIM" proteins) have been identified only in insects so far. The poly-NIM genes can be divided into two subgroups: Nimrod C- and Nimrod B-types. The products of Nimrod C-type genes (e.g., *nimrod C1-4*, *eater* in *D. melanogaster*) are transmembrane proteins lacking EGF repeats but containing a variable number of NIM repeats. Nimrod B-type genes (e.g., *nimrod B1-5* in *D. melanogaster*) differ from Nimrod C genes in that they lack the transmembrane domain in the encoded protein (Kurucz et al. 2007).

Draper-type proteins were described to have a function in phagocytosis in many species, for example, Ced-1 in *C. elegans* (Mangahas and Zhou 2005), Draper in *D. melanogaster* (Manaka et al. 2004), as well as MEGF-10 in human (Hamon et al. 2006). The role in phagocytosis was also shown for some Nimrod C-type genes, like *eater* (Kocks et al. 2005) and *nimrod C1* (Kurucz et al. 2007) in *D. melanogaster* or 120 kDa protein in *Sarcophaga peregrina* (Nishikawa and Natori 2001). Each of these Nimrod C-type genes are expressed in hemocytes, Nimrod C1, and Eater proteins were demonstrated to be involved in phagocytosis, Eater being a bacterium binding protein. Experimental data support the role of a Nimrod B-type protein as a pattern recognition receptor for bacterial lipopolysaccharide (Ju et al. 2006). These data suggest that the whole superfamily might be a remarkable component of the innate immune response.

The duplication and subsequent diversification of genes is one of the major factors leading to formation of gene families of variable size (Ohta 1994; Zhang 2003). Different models have been proposed to improve our understanding as to how gene families evolve (reviewed in Nei and Ronney 2005). Many examples are known where high sequence similarities among the members of a gene family were maintained during evolution, thereby making the member genes within a species (paralogs) more similar to each other than to the orthologous ones even in closely related species (Liao 1999; Nei and Rooney 2005; Eickbush JD and Eickbush DG 2007). A model, "concerted evolution," was proposed in order to explain these observations: the members of a gene family evolve as a unit and changes occurring

[1] K.S. and B.S. contributed equally to this work.

Key words: gene family, concerted evolution, birth-and-death evolution, repeat evolution, innate immunity, insect, Drosophila.

E-mail: somogyi@brc.hu.

Mol. Biol. Evol. 25(11):2337–2347. 2008
doi:10.1093/molbev/msn180
Advance Access publication August 14, 2008

in a single gene can spread throughout the whole family by mechanisms like gene conversion and/or unequal crossing over (Liao 1999; Nei and Rooney 2005; Eickbush JD and Eickbush DG 2007). These processes maintain a high degree of sequence similarity within a family. Often, other patterns may emerge: some genes become deleted or lose their functions, others duplicate further, and some may acquire novel functions. These processes lead to the formation of gene families whose members show more similarity to their orthologs than to their paralogs. The so-called "birth-and-death evolution" model describes that process (Ota and Nei 1994): genes are "born" by duplications, can exist for a long time, their sequences and functions might change, and finally, they "die" by inactivation or deletion. However, high similarities of paralogous sequences can be maintained also under birth-and-death evolution if strong purifying selection acts (Nei and Rooney 2005; Nei et al. 2000). The third model of gene family evolution is the "divergent" model. Both the divergent and the birth-and-death models imply the independent evolution of the units.

A significant fraction of proteins contain a variable number of the domain of the same type, generally thought to be the result of internal duplications (Björklund et al. 2006). The evolutionary processes described above can also be observed in the case of these repeats. The sequence similarity between the duplicated repeats may decrease over time as they independently accumulate mutations, as for example in HEAT repeat containing proteins (Andrade et al. 2001). In other situations, like in the case of the sequence repeats in VERL protein of abalone (Haliotis) species (Swanson and Vacquier 1998), in tenascins of mammals (Hughes 1999), in SOWpg protein of the human pathogenic fungi Coccidioides species (Johannesson et al. 2005), or in Dumpy protein of the fruit fly *D. melanogaster* (Carmon et al. 2007) repeats undergo concerted evolution. The members of the Nimrod gene superfamily contain variable numbers of repeats providing an excellent opportunity to study evolutionary processes on the levels of genes and repeats. In this paper, we have analyzed evolutionary processes of *nimrod* genes using the sequences collected from genomes of the following insect species: *D. melanogaster*, *D. pseudoobscura*, *D. sechellia*, *D. yakuba*, *D. virilis*, *Anopheles gambiae*, *Tribolium castaneum*, and *Apis mellifera*.

Materials and Methods

Sequence Data

Sequences of *nimrod* genes were collected from genomes of following species: *D. melanogaster*, *D. pseudoobscura*, *D. sechellia*, *D. yakuba*, *D. virilis*, *A. gambiae*, *T. castaneum*, and *A. mellifera*. The *nimrod*-related genes in the Drosophila, Apis, and Tribolium species have been described (Kurucz et al. 2007; Evans et al. 2006; Zou et al. 2007; Sackton et al. 2007). Homologous genes were identified in a similar way in the *Anopheles* genome (Holt et al. 2002). Briefly, we used TblastN to search the sequenced genomes for conserved CCXGY motifs, followed by at least one NIM repeat. Computer-generated gene models in the identified regions (sequence/gene identifiers in supplementary table 40, Supplementary Material online) were manually curated to include additional conserved sequences and to split artificially fused genes. The gene models were further refined by comparison to available EST and cDNA sequences and by cross-species comparisons. In a few cases, we had to correct frameshift errors by rechecking raw sequence data from the trace archives. The amino acid sequences used in our analysis are listed in supplementary text 1 (Supplementary Material online).

Detection of NIM Repeats

NIM repeats were identified by profile Hidden Markov model (HMM) search using HMMER suite, version 2.3.2 (http://hmmer.janelia.org). NIM repeats from *nimrodline* and *eaterline* genes of *D. melanogaster* were extracted using regular expressions by capturing string units containing six cysteine residues from sequences following the CCXG motif. These repeats were compared with the repeats resulting from the manually annotated genes. After removing the sequences considered "atypical" (very long or short repeats relative to the length of the NIM consensus), the whole data set was aligned with ProbCons (Do et al. 2005). Positions after GY motif (from the end of NIM repeat) were deleted. From this alignment, a profile HMM was built with *hmmbuild* (default parameters, ls mode) and calibrated with *hmmcalibrate*. This first "preliminary" HMM was used with *hmmsearch* (default parameters) to identify repeats in the whole data set. Raw search results were converted into Fasta format by a Perl script, using BioPerl modules (Stajich et al. 2002). Repeats were aligned with *hmmalign* and were used to build a new "refined" profile HMM. HMM logos built from the two models showed no major differences, but the refined HMM could identify more NIM repeats in some genes of our data set suggesting that it might be more sensitive. The refined model was used in further analyses for identification and alignment of repeats (using *hmmalign*), except when building Neighbor-Joining (NJ) trees to detect repeat-level concerted evolution before aligning the sequences (the preliminary HMM was used in these cases).

Multiple-Sequence Alignments

Alignments of sequence regions containing mostly repeats evolving in a concerted manner cannot be expected to have any biological significance. Because of this, genes which contain repeats evolving in a concerted fashion were detected. NJ trees (complete and pairwise deletion) were built based on an alignment containing all repeats identified by the preliminary profile HMM. The following genes were excluded from the multiple-sequence alignments as they are considered to have repeats which evolved in concerted fashion: *nimrod CI* and *nimrod CII* of *T. castaneum*, *nimrod CI* of *A. mellifera*, and all *eater* genes of the Drosophila species.

The relatively high divergence, the significant variation in the length, domain structure, and the overall size of *nimrod* genes make it a difficult task to find the best method and parameters to build a biologically meaningful alignment. The repetitive regions also represent a special problem because it is difficult to find an ideal alignment for them (Higgins 2003). Results obtained with different methods can differ significantly, so five software packages implementing different heuristic multiple-sequence alignment algorithms were used: ClustalW 1.83 (Thompson et al. 1994), Dialign 2.2 (Morgenstern 1999), Muscle 3.6 (Edgar 2004), T-Coffee 4.45 (Notredame et al. 2000), and Probcons 1.1 (Do et al. 2005). The sequences were also aligned by using Dialign 2.2 with a bonus given for aligning together CXPXCXXXCXXGXCXXPXXCXCXXGX (a relaxed NIM consensus) motifs. The quality of every alignment was evaluated under the following criteria: the placement of gaps and CCXGY motifs, the handling of terminal indels, and the consistency scores calculated by T-Coffee. NJ trees were also constructed (data not shown) by PHYLIP 3.66 (using default parameters) (Felsenstein 1989) and the topologies were compared in order to assess the effect of the alignment method used on the phylogenetic reconstructions. No major differences were found between the topologies of the trees calculated from the alignments with the highest consistency scores (T-coffee and ProbCons). After evaluating the alignments in the case of each family, the alignments produced by ProbCons were chosen for further analyses.

The likelihood mappings of the ProbCons alignments were performed by Tree-Puzzle 5.2 (Strimmer and von Haeseler 1997) with the amino acid substitution models selected by ProtTest (standalone version 1.3 or web server at http://darwin.uvigo.es/software/prottest_server.html; Abascal et al. 2005). The among-sites rate variation was modeled by a discrete gamma distribution with four categories; amino acid frequencies were estimated from the data and exact parameter estimates were used. In all, 100,000 random quartets were sampled, except for the Nimrod A alignment where all of the possible quartets (70) were considered. The likelihood mappings indicated a sufficient tree-like phylogenetic signal in the alignments (supplementary figs. 32, Supplementary Material online).

Phylogenetic Methods

At first amino acid matrix best fitting the respective alignment was selected by using ProtTest in BIC framework using the alignment length as sample size parameter. NJ trees were built using MEGA 3.1 (Kumar et al. 2004), both with complete and pairwise deletion of sites containing gaps. Tree construction under maximum likelihood (ML) criterion was performed with PhyML version 2.4.4 (Guindon and Gascuel 2003) using a discrete gamma distribution with four categories to model rate variation across sites if applicable. When building NJ and ML trees, the best fitting amino acid matrix (according to the BIC score) implemented in the respective software and model-averaged estimates of alpha and invariant site parameters were used. Branch support was assessed by nonparametric bootstrap (1,000 replications). Some ML trees were calculated using ProtTest with best model and parameters estimated.

Bayesian reconstructions were performed using MrBayes version 3.1.2 (Huelsenbeck and Ronquist 2001). Phylogenetic trees were reconstructed from sequence alignments only and also from mixed data sets containing a separate data partition of gap information. Gap information was coded as variable coding restriction site characters with a simple gap-coding method (Simmons and Ochoterena 2000; MrBayes wiki: http://mrbayes.csit.fsu.edu/wiki), implemented in a Perl script. To assess the phylogenetic information content of the gap data, trees were also reconstructed using the gap information only.

Analyses were run with default priors and parameters, except that a uniform prior was applied over the fixed rate amino acid models (*prset aamodelpr=mixed*), so the matrices were included as parameters in the analyses.

Likelihood plots, standard deviation of split frequencies, and PSRF values were used to diagnose convergence. Some of the runs were performed using a parallel version of MrBayes (Altekar et al. 2004) on a Linux cluster.

For the MrBayes blocks with the exact parameters used, see the supplementary text 2 part 3 (Supplementary Material online).

Because of the lack of suitable outgroup sequences, rooting of selected gene trees were achieved in with software Notung 2.1 (Chen et al. 2000; default parameters) using the rooting analysis feature after reconciliation with the species tree. Topology of the species tree used for rooting (supplementary fig. 33, Supplementary Material online) was built based on the accepted phylogeny of Drosophila species (Russo et al. 1995) and a species tree published recently (Zdobnov and Bork 2007).

The simultaneous Bayesian estimation of alignment and phylogeny was performed by using Bali-Phy version 2.0.0 (Suchard and Redelings 2006) with default settings using the WAG + G amino acid matrix (with four gamma categories) and the default "fragment-based indels + T" indel model. Three independent runs were performed, each with 30,000 iterations starting from unaligned sequences. Convergence was assessed by examining the sample likelihoods and the cumulative split frequency plot calculated by the online version of AWTY (Nylander et al. 2007). Based on these analyses, the first 10,000 samples were discarded as burn in. The similarity of the topologies of the MAP trees and the 80% consensus trees indicated convergence of the three independent runs.

Trees were edited in MEGA 3.1 and iTOL (Letunic and Bork 2007). The trees on the figures 4 and 5 were manually redrawn for better visibility.

Construction and Comparison of Profile HMMs

HMM logos were generated by Logomat-M server (http://www.sanger.ac.uk/cgi-bin/software/analysis/logomat-m.cgi) and pairwise HMM logos were created using Logomat-P server (http://www.sanger.ac.uk/Software/analysis/logomat-p).

Calculation of Pairwise Repeat Distances, Pairwise Repeat Homology Diagrams, and Saturation Plots

The repeat amino acid sequence alignments were converted into corresponding "repeat DNA" alignments (supplementary text 2 part 1, Supplementary Material online),

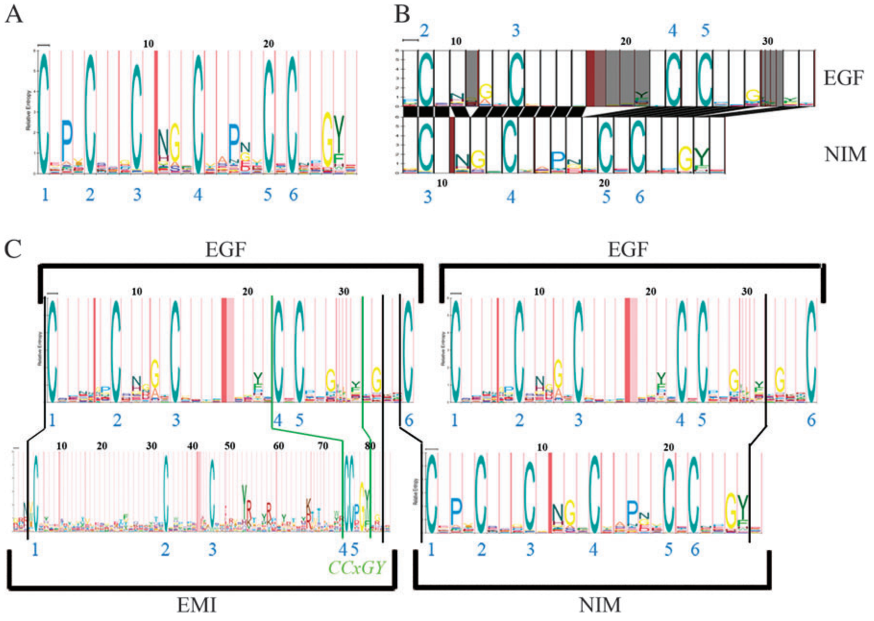

FIG. 1.—Characteristics and proposed origin of the NIM repeat. (*A*) Logo of the NIM repeat profile HMM. (*B*) Comparison of the EGF-like (top, Pfam accession: PF00008.17) and the NIM repeat (bottom) on a pairwise HMM logo showing the aligned HMM states (EGF: 7.-34.; NIM: 8.-27. states). (*C*) Hypothetical model of the origin of the NIM repeat and the EMI domain (Pfam accession: PF07546): Rearrangements of the first two EGF-like domains of a poly-EGF protein (top) results in a protein containing an EMI domain and a NIM repeat (bottom). For more details, see text. On each panel, the characteristic cysteines (blue) and every tenth state of the HMM logos (black) are numbered.

based on the amino acid sequence and coding sequence of the respective gene using a custom Perl script. Conversion was verified by aligning the DNA sequences made by concatenation of ordered repeat DNA sequences from the alignment with the full coding sequences (supplementary text 2 part 2, Supplementary Material online). This alignment also helped to identify linker regions at DNA level. Synonymous and nonsynonymous distances (corrected by Jukes–Cantor formula) were calculated with MEGA 3.1 using the Modified Nei–Gojobori method (Nei and Gojobori 1986; Zhang et al. 1998). Transition/transversion ratio parameters needed for distance calculations were calculated using PAUP* 4.0b10 (Swofford 2003) and Modeltest 3.7 (Posada and Crandall 1998). Model-averaged parameter estimates, averaged on the basis of Akaike weights were used. Pairwise repeat homology diagrams were calculated with t2prhd version 1.7 (Sipos et al. 2008) using ClustalW as backend (with default parameters) and the -*w* parameter set to 2. Saturation plots were calculated with DAMBE 4.2.13 (Xia and Xie 2001) using the F84 model.

Results and Discussion

Origin of NIM Repeats

To analyze the characteristics of NIM repeats, a profile HMM logo (Schuster-Böckler et al. 2004) was built (fig. 1*A*). That was in agreement with the consensus sequence of the NIM repeat (Kurucz et al. 2007). In Draper-type proteins, the NIM motif is followed by a variable number of EGF domains (Kurucz et al. 2007). In the present study, we have used the tools and nomenclature provided by the Pfam (Finn et al. 2006; version 22.0) database to discuss the EGF domains. Domains of the EGF-like clan (Pfam clan: CL0001) might be hard to model due to many similar but different subtypes. The EGF-like domain (Pfam accession: PF00008) containing six conserved cysteine residues is very similar to the Laminin EGF-like domain (Pfam accession: PF00053) containing eight cysteines (URL: http://pfam.sanger.ac.uk/family?acc=PF00008). We matched the sequences of Draper-type genes against the profile HMMs (ls models) of the EGF-like Pfam clan (Pfam clan: CL0001) using *hmmpfam* from the HMMER package (http://hmmer.janelia.org/). The domains containing six conserved cysteines had more hits with high scores in each sequence (EGF2—Pfam accession: PF07974—and EGF-like, with EGF2 having the best scores) as compared with the Laminin EGF-like domain.

The EGF-like domain is slightly more similar to the typical EGF consensus, so its profile HMM was chosen to be compared with the NIM profile HMM by generating a pairwise HMM logo (Schuster-Böckler and Bateman 2005) in order to gain insights into the relationship between the EGF domains and the NIM repeat. This comparison (fig. 1*B*) highlights the similarity between large portions of the EGF-like domain (ca., from the second conserved cysteine residue to the conserved tyrosine residue) and the NIM repeat (ca., from the third cysteine to the conserved GY motif).

The N-terminal part of Draper-type proteins typically contains an EMI domain (Doliana et al. 2000; HMM logo: fig. 1*C*; Pfam accession: PF07546) closely linked to a single NIM repeat (fig. 1*C*). This structure, if compared with two subsequent HMM logos of the EGF-like domain, shows intriguing similarities. Also, when comparing the EMI domain with the EGF-like domain on a pairwise

HMM logo, seven emission states after the states corresponding to the conserved "CC" residues in the EMI domain are aligned with the EGF-like HMM (data not shown).

These observations suggest a possible scenario in which the Draper-type genes originated from a gene encoding a protein with a poly-EGF part. During structural reorganization and further sequence changes (including duplications, insertions, and deletions), a part of the first EGF domain (containing the first five cysteine residues) might have become an EMI domain with the CCXGY motif. The larger size of the EMI domain suggests that insertions were the major events leading to this structural novelty. This idea is in agreement with the findings of Jiang and Blouin (2007), who claim that structural innovation is possible via nested insertions and rapid evolution within variable regions. The last cysteine residue of the first EGF repeat and the following linker region together with a part of the second EGF repeat (which contains the first five cysteines) might have formed the NIM repeat. This process could have given birth simultaneously both to the NIM repeat and to the EMI domain and it can also explain the one cysteine unit shift of NIM consensus relative to the EGF consensus. Supposing that the Draper-type genes are the most ancient forms as suggested by their wide taxonomic distribution, subsequent duplications could turn the single NIM repeat into a repetitive unit of Nimrod C- and B-type (poly-NIM) proteins. During the evolution of the gene superfamily, subsequent sequence changes might have significantly modified the EMI domain giving birth to the N-terminal parts of Nimrod C and B proteins. The CCXGY/W motif appears to be a key component, remaining a conserved characteristic for the whole superfamily.

The mechanism of the formation of new protein repeat types is not fully understood. Andrade et al. (2001), for example, emphasized a contradiction: All members of a repeat family evolved from a common ancestor, which necessarily have contained only a single repeat, but it is unexpected that a single repeat could exist in isolation, as a single folded functional unit. To resolve this problem, a hypothesis was suggested: New repeat types can arise as modified monomers in a multichain oligomeric system. To date, however, there are few, if any, known examples where homologous multirepeat assemblies are formed both from oligomers of single repeats and from a single chain of multiple repeats (Andrade et al. 2001). The scenario described above for the origin of NIM repeat might suggest a simpler mechanism: New repeat types can arise by modification of terminal repeats of a homogeneous repeat chain. The cooperative nature of the folding process may have a lower impact on these repeats because they are neighboring only one other repeat. Presumably, sequence changes happen more easily in such circumstances.

Patterns in NIM Repeat Evolution

Phylogenetic reconstruction represents an established approach to study the mode of evolution of multigene families (Nei et al. 1997; Nei and Rooney 2005; Quesada et al. 2005) and also repeats (Johannesson et al. 2005; Carmon et al. 2007). Following this strategy, phylogenetic trees of repeats from the Nimrod B and C and Eater amino acid sequences of the analyzed species were built (supplementary figs. 7–9, Supplementary Material online). The large size of these trees makes the interpretation difficult but some relevant observations can be made. In most cases, the positionally homologous repeats of Nimrod B orthologous proteins form clades (supplementary fig. 7, Supplementary Material online). Many of these clades reflect the currently accepted Drosophila phylogeny (Russo et al. 1995). Branching patterns of Nimrod B-type repeats support their independent evolution. Many repeats of Nimrod C-type proteins (mainly of Drosophila Eaters), however, show a different pattern (supplementary fig. 9, Supplementary Material online): The repeats of the same protein form clades with short branches irrespective to the species phylogeny, indicating their high similarity to each other, a phenomenon consistent with the concerted model of evolution.

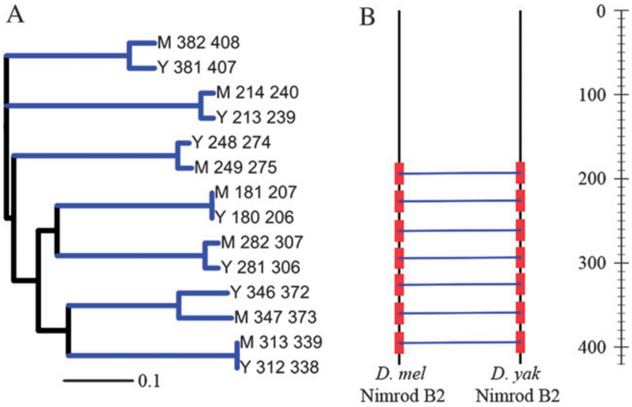

FIG. 2.—Pattern of repeat evolution in Nimrod B2 sequences. (*A*) Neighbor-Joining tree of the repeats from Nimrod B2 proteins of *Drosophila melanogaster* (M) and *Drosophila yakuba* (Y). The branch lengths are drawn to scale; the repeats are identified by starting and ending positions. Clades representing orthology relations are in blue. (*B*) The pairwise repeat homology diagram (PRHD) built from the phylogenetic tree. The NIM repeats are indicated with red rectangles on the protein sequence schemes. The identified homology relations are represented by connecting the respective repeats (with blue lines between the repeats from different protein sequences—"orthology"). The color scale bar it is not shown. The sequence positions are shown on the right of the scheme. Dmel, *D. melanogaster*; Dyak, *D. yakuba*.

Pairwise repeat homology diagrams (Sipos et al. 2008) were applied to analyze patterns of repeat evolution in detail. On these diagrams, homology relations identified by phylogenetic methods are represented by connecting the respective repeats from different protein sequences ("orthology") or within the same species ("paralogy"). Analysis of orthologous Nimrod B sequences of Drosophila species (fig. 2; supplementary figs. 10–13, Supplementary Material online) reveals a clear pattern expected from independent evolution. For example, NIM repeats of *D. melanogaster* and *D. yakuba* Nimrod B2 form a clade on the phylogenetic tree according to their positions inside the amino acid sequence producing a perfect ladder-like figure on pairwise repeat homology diagrams (fig. 2*B*). This suggests the independent evolution of single repeat units so as in the case of C3 and C4 (supplementary figs. 14–15, Supplementary Material online). On the contrary, on Nimrod C2 diagrams the ladder-like connection pattern is less pronounced (supplementary fig. 16, Supplementary Material online).

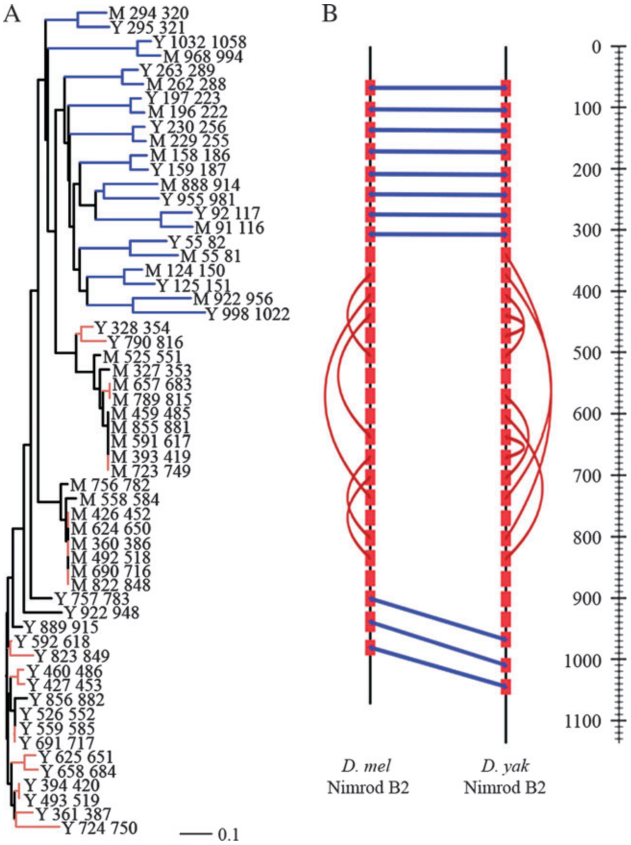

FIG. 3.—Pattern of repeat evolution in Eater sequences. (*A*) Neighbor-Joining tree of the repeats from Eater proteins of *Drosophila melanogaster* (M) and *Drosophila yakuba* (Y). The branch lengths are drawn to scale; the repeats are identified by starting and ending positions. Clades representing orthology relations are in blue, those representing paralogy are in brown. (*B*) The pairwise repeat homology diagram built from the phylogenetic tree. The NIM repeats are indicated with red rectangles on the protein sequence schemes. The identified homology relations are represented by connecting the respective repeats (with blue lines between repeats from different protein sequences—"orthology" and with brown arcs in case of internal relations—"paralogy"). For further explanations, see figure 2.

Nimrod C1 sequences show a more disrupted pattern. Here, even the number of repeats is variable and there are many missing orthology relations (supplementary fig. 17, Supplementary Material online), particularly in analyses of orthologous sequence pairs from more distantly related species (as, e.g., *D. virilis* vs. the other species). Because of this and based on the comparisons of the synonymous and nonsynonymous repeat distances (see later), we interpret this pattern as a sign of fast evolutionary changes in *nimrod C1* gene obscuring repeat relationships rather than a repeat homogenization process. This is in agreement with the findings of Sackton et al. (2007) who found evidence for positive selection acting on *nimrod C1* besides *nimrod B1* and *B4* genes of Drosophila species.

Analysis of Drosophila Eater sequences revealed a complex pattern indicative of both independent and concerted evolution of the NIM repeats. For example, on the pairwise repeat, homology diagram of Eater sequences of *D. melanogaster* (containing 28 repeats) and *D. yakuba* (30 repeats) different regions can be recognized (fig. 3). Counting from the N-terminus, the first eight and last three NIM repeats are connected in a ladder-like manner suggesting independent evolution, whereas inner repeats have only internal if any connections, as expected under concerted evolution. Similar results were obtained in analyses of other Drosophila Eater sequences (supplementary fig. 18, Supplementary Material online). In principle, it is possible that the observed similarity patterns and repeat numbers are the results of independent duplications of the internal repeats in each analyzed Drosophila species after speciation. However, we do not consider that a parsimonious and hence acceptable explanation because it would imply the parallel evolution from the point of view of the repeat numbers and duplication/deletion rates in all terminal lineages.

High similarity of repeated amino acid sequences can arise also under birth-and-death evolution when strong purifying selection acts. To test this possibility, the repeat similarity was investigated on the level of DNA sequences. Alternatively to the concerted mode of evolution, similar patterns of pairwise synonymous and nonsynonymous distances between repeats in Eater sequences (supplementary tables 1–10, Supplementary Material online) suggest that the homogeneity of the internal NIM repeats is not maintained by a strong purifying selection acting on the amino acid sequences, which would imply larger synonymous distances relative to nonsynonymous distances (Nei et al. 2000). The same analysis also supports the conclusions about the repeat evolution patterns based on pairwise repeat homology diagrams, including the lack of repeat homogenization in the case of Nimrod C1 (supplementary tables 11–30, Supplementary Material online).

In either mechanisms believed to be involved in concerted evolution (unequal crossing over by unequal sister chromatid exchange and/or gene conversion), flanking repeats expectedly do not participate intensively in the homogenization process because of the influence of unrelated flanking sequences, as in the case of repeats of *SOWpg* (Johannesson et al. 2005) and *dumpy* (Carmon et al. 2007) genes and other tandemly repeated elements (McAllister and Werren 1999). This "margin effect" alone, however, cannot account for the asymmetric distribution of conserved repeats (namely, the dissimilar number on the N-terminus and the C-terminus) in Eater proteins. When the protein repeats are encoded by different exons, their homogenization could be inhibited by the presence of introns. It was also reported that regions near the intron–exon boundaries containing splicing enhancer sites have lower evolutionary rates (Parmley et al. 2007). None of these factors can possibly explain the asymmetric evolutionary pattern of Eater repeats because all but one repeats are encoded by a single large exon. It seems more plausible that this asymmetric pattern is maintained by a functional constraint. The generally larger pairwise synonymous distances (compared with nonsynonymous distances) found in the case of the first and last repeats suggest that they evolve under purifying selection. This idea is supported by the results of Kocks et al. (2005) who found that the first few repeats of the *D. melanogaster* Eater protein are sufficient for bacterial binding, and some of these repeats, similar to the ones near the transmembrane region, have predicted N-glycosylation sites which argues for an importance in ligand binding. Therefore, these repeats must keep their sequence "identity," required for their function (binding of bacteria). The homogeneous part of the repeat chain might serve only as a structural element, a "stalk" (Kocks et al. 2005) which must only maintain the basic sequence properties in order to remain functional.

Analysis of pairwise repeat homology diagrams, distance matrices (supplementary tables 1–10, Supplementary Material online), and alignments of repeat DNA sequences (SI text part 1) revealed that in many cases the units of concerted evolution are not individual NIM repeats. For example, in the homogeneous region of *D. melanogaster* Eater, the highest similarity is perceptible between every second repeat (supplementary tables 1–10 and text part 1, Supplementary Material online) arguing for two-repeat units of evolution which is verified by the homogeneity of the corresponding linker regions (supplementary text part 2, Supplementary Material online).

Elsewhere, like in the cases of the repeats from *T. castaneum* Nimrod C-type proteins and *A. mellifera* Nimrod CI, sequence similarities show complex patterns (supplementary tables 11–38, Supplementary Material online). With the lack of close orthologs, however, we cannot safely rule out the possibility that the observed high similarities are the results of independent internal duplications.

Among the poly-NIM genes of Drosophila species, patterns of concerted evolution can be found only in *eater* genes and never in *nimrod C* genes of similar size. In *T. castaneum* repeat homogenization is observed in a relatively short *nimrod CI* gene but not in the much larger *nimrod CII*. It seems, therefore, that gene size (i.e., repeat number) is not a major factor influencing the repeat evolution in the Nimrod superfamily. The chromosomal environment of a certain gene might also have an influence on the homogenization process because the local chromatin structure was shown to regulate gene conversion (Cummings et al. 2007), though the impact of this factor on *nimrod* genes has not been studied yet.

The transition/transversion ratios estimated from alignments of the repeat DNA sequences from single C-type genes of Drosophila species revealed that with the only exception of *nimrod C1* of *D. yakuba*, these ratios appeared to be substantially higher in *eater* than in *nimrod C1* or *C2* (supplementary table 39, Supplementary Material online). This suggests a higher proportion of transitions among repeat sequences evolving in concert. Because transitions are generally believed to occur in greater frequency, among constantly homogenized sequences, they are expected to be observed more prevalently because the more rarely occurring transversions are obscured by the rapid homogenization process. In the case of sequences that have evolved independently for a longer period, both transitions and transversions might have already reached saturation producing a more balanced ratio. This explanation is supported by saturation plots generated for the repeat DNA alignments (supplementary fig. 19, Supplementary Material online), where the ones that do not contain repeats evolving in concert (Nimrod C1 and C2 repeat alignments of Drosophila species) are closer to saturation.

The Phylogeny of the Nimrod Genes

The repetitive structure and the remarkable diversity in size and domain composition of the proteins make a reconstruction of the phylogeny of Nimrod gene families a complicated issue. Because the alignment quality can seriously affect the result of phylogenetic reconstructions (Kumar and Filipski 2007), several methods were used to align sequences and to evaluate the alignment quality.

Repeats evolving in concert can produce misleading alignment results; therefore, the genes were excluded from the full sequence analyses when such a phenomenon was indicated. The large variation in length of the sequences expectedly led to alignments with many gaps, and these may represent significant phylogenetic information that can be used to get a more resolved phylogeny. Pairwise repeat homology diagrams developed for the study of repeat evolution are applicable here, too. If the possibility of intensive concerted evolution can be ruled out, it is expected that as long as two sequences diverged more recently (or have slower evolutionary rates), repeat orthology relationships are more readily detected via sequence similarity producing a ladder-like pattern on the diagram. In sequences which are less closely related, the loss of the phylogenetic signal due to the high number of fixed mutations may cause the repeats not to find their counterparts and the pattern becomes disrupted to some extent (as in the case of Nimrod C1 orthologs). This approach does not depend on multiple-sequence alignment of the full sequences.

Topologies of Nimrod B trees produced by the NJ with pairwise deletion, ML, and Bayesian methods agree in the relative placement of Drosophila Nimrod B1-B5 clades (fig. 4*A*; supplementary fig. 20, Supplementary Material online). The association of orthologous genes on the trees indicates independent evolution in the family. The inclusion of gap information in the Bayesian phylogeny did not alter the relationship between the Drosophila B1–B5 clades; it affected only the placement of *T. castaneum* Nimrod B and in general had a positive effect on the posterior probabilities of the clades, resulting in a more resolved phylogeny. The tree built using only gap information alone was able to resolve relationships between the Drosophila B1–B5 groups (supplementary fig. 20F, Supplementary Material online). In all trees, the Drosophila Nimrod B2 clade is frequently associated with the Nimrod B sequences of *T. castaneum* and *A. gambiae* suggesting that this is the most basal Drosophila Nimrod B form. The trees indicate that Nimrod B3 split off before Nimrod B5 and the Nimrod B1 and B4 split most recently.

Intraspecies comparisons of Drosophila paralogous Nimrod B sequences with pairwise repeat homology diagrams gave results that are more or less concordant with the phylogenetic trees (fig. 4*B*; supplementary fig. 21–24, Supplementary Material online). Between Nimrod B1 and B4 sequences in the same species, most of the repeats are connected with repeats from the other sequence providing a ladder-like pattern, as expected from closely related paralogs. Less solid is the pattern between these sequences

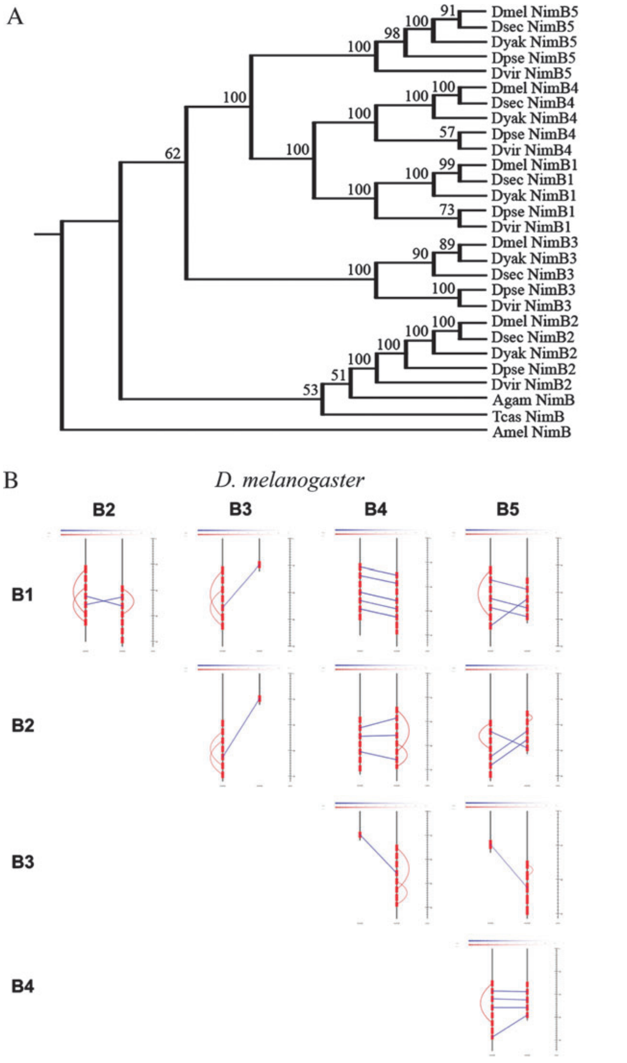

FIG. 4.—(*A*) Bayesian tree of the Nimrod B protein sequences. The tree was obtained by the rooting of the tree on the supplementary figure 21B (Supplementary Material online) using the rooting analysis feature in Notung 2.1 after reconciliation with the species tree (supplementary figure 33, Supplementary Material online). The numbers above the branches indicate the Bayesian posterior probabilities (in percent), the branch lengths are not drawn to scale. The unrooted Bayesian tree was built from the mixed data set containing besides the sequence alignment also the gap information in a separate partition. (*B*) The PRHDs of the *Drosophila melanogaster* Nimrod B sequences. The names of the sequences on the right of the PRHD analyses are shown above.

and Nimrod B5; furthermore, the repeats of the more distantly related Nimrod B2 sequences produce the smallest number of orthology connections with them.

Contrary to Nimrod B, on the pairwise repeat homology diagrams for intraspecies analyses of Nimrod C paralogs (supplementary fig. 25–29, Supplementary Material online), the number of identified repeat orthology relations was small and no relation was consistently present in each of the five intraspecies analyses. Under such circumstances, no further information about the similarity between the specific paralogs can be obtained with this method. These analyses suggest that the Nimrod C genes probably radiated earlier or have higher evolutionary rates than the Nimrod B genes. This may also explain the difficulties in recovering the correct Nimrod C phylogeny as the multiple-sequence alignments of highly diverged sequences are problematic.

The Nimrod C ML tree (fig. 5; supplementary fig. 30A, Supplementary Material online), the Bayesian tree built using only sequence information (supplementary fig. 30B, Supplementary Material online) and the NJ trees (supplementary fig. 30C-D, Supplementary Material online), agrees in placing Nimrod C3 and C4 clades close to each other and in a common clade with the C1 sequences, whereas C2 sequences always form a different clade. The unrooted Bayesian Nimrod C tree constructed only from the sequence alignment (supplementary fig. 30B, Supplementary Material online) contained a polytomy, but by resolving that using the rearrangement feature in Notung and after rooting, the resulting topology was in complete agreement with the topology of the rooted ML tree (fig. 5). The topology of the Nimrod C Bayesian tree built from the mixed data set containing also the gap information (supplementary fig. 30E, Supplementary Material online), however, does not agree with the other trees in the placement of the Drosophila C1-C4 clades, on this tree C1 and C2 form a clade. The tree inferred from the gap information only (supplementary fig. 30F, Supplementary Material online) supports this topology with high posterior probabilities, so the disagreement is probably caused by the inclusion of the gap information.

Because of the known biases affecting the placement of the gaps during multiple-sequence alignment (Golubchik et al. 2007) and the shortcomings of the simple method used to code the gap information, in this case, we prefer the trees built from sequence information only.

In all but one trees, the Nimrod C sequences of *A. gambiae* form a clade, suggesting their common origin by duplications after divergence from the lineage leading to Drosophilidae.

The *nimrod A* genes do not show any signs of duplication events in any of the taxa examined. All trees (supplementary fig. 31, Supplementary Material online) except NJ complete deletion are in a complete topological agreement and they also agree with the species tree used for the rooting analysis.

The Evolutionary History of the Nimrod Superfamily

The characteristics of Nimrod sequences discussed above made it impossible to obtain acceptable quality alignments for the phylogenetic analysis of the whole superfamily using heuristic methods. Because of this, the Bali-Phy (Redelings and Suchard 2005; Suchard and Redelings 2006) software was used for the simultaneous Bayesian estimation of alignment and phylogeny. Due to resource constraints, only the *D. melanogaster* Draper, Nimrod A, B, and C sequences were analyzed. The three independent runs gave the same MAP tree topology and the topologies of the 80% consensus trees agreed with them with one exception containing a polytomy. The superfamily-level 80% Bayesian consensus tree (fig. 6) is consistent with the family-level trees (figs. 4 and 5) except in placing the Nimrod B3 in the

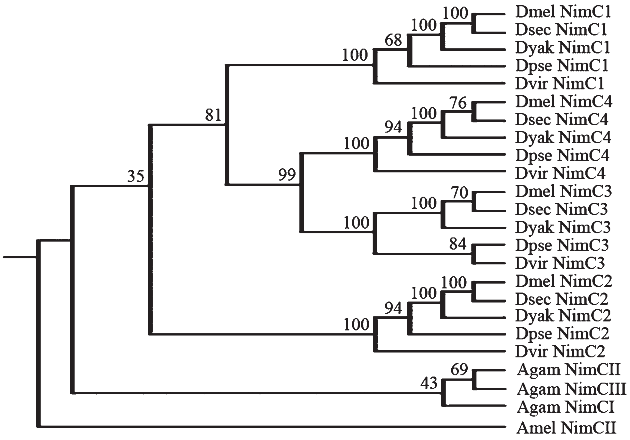

FIG. 5.—ML tree of Nimrod C protein sequences obtained by the rooting of the tree on the supplementary figure 31A (Supplementary Material online) via reconciliation with the species tree (supplementary fig. 33, Supplementary Material online). The numbers above the branches indicate bootstrap values, the branch lengths are not drawn to scale. For abbreviations see figure 4.

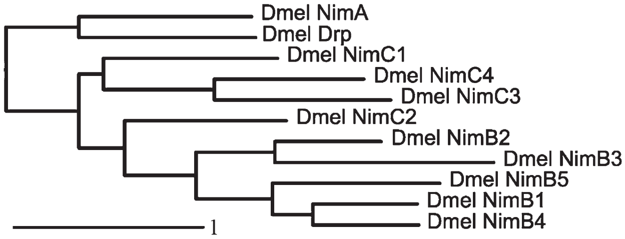

FIG. 6.—Bayesian consensus tree (at 80% level) of the *Drosophila melanogaster* Draper, Nimrod A, B, and C sequences obtained by simultaneous estimation of alignment and phylogeny. The branch lengths are drawn to scale.

same clade as the Nimrod B2. The topology of the tree supports that the B family is the descendant of the Nimrod C family.

Based on our present knowledge, including the domain structure of the Nimrod proteins, a scenario can be drawn. First, the first Draper-type molecule evolved from a protein with a poly-EGF run. This gave rise to the two basic sequence elements, the CCXGY/W motif (in the EMI domain) and NIM motif, characteristic of the whole Nimrod superfamily. Draper-type proteins have a wide taxonomic distribution and are thought to be involved in phagocytosis (e.g., Mangahas and Zhou 2005; Manaka et al. 2004; Hamon et al. 2006). In insects, with the loss of the EGF domains, the modification of the EMI domain and duplications of NIM repeat, the Nimrod C-type emerged as the first poly-NIM gene.

Exon shuffling is a possible mechanism for domain repeat duplication (Björklund et al. 2006), but this does not seem to be the case for the NIM repeats. Rather, most of the NIM repeats are typically encoded by one or less frequently two large exons, indicating a general trend of intra-exon duplications. The Nimrod C-type genes seem to retain the basic functional properties because phagocytosis was proposed as a major function for members of this family (Kocks et al. 2005; Kurucz et al. 2007; Nishikawa and Natori 2001). Deletion of sequences encoding the transmembrane part of a Nimrod C-type protein may have led to the formation of the *nimrod B* genes, whose role in recognition of bacteria is supported by experimental data (Ju et al. 2006).

Conclusions

The experimental (e.g., Mangahas and Zhou 2005; Manaka et al. 2004; Hamon et al. 2006; Kocks et al. 2005; Kurucz et al. 2007; Nishikawa and Natori 2001; Ju et al. 2006) and our in silico results outline a complex history of a gene superfamily, from the birth of the first member by generation of a characteristic domain structure through formation of families by changes of domain composition to the expansion of gene families leading to many recent members. Phylogenetic trees indicate that the *nimrod* genes have undergone birth-and-death evolution. As regards the evolution of the NIM repeats, both concerted and independent evolution were observed, producing various patterns inside the protein sequences of the Nimrod superfamily. For a significant fraction of the proteins encoded by the genes of this superfamily, a function in immune response was suspected or even experimentally shown. The duplications of the poly-NIM genes in insects could create many transmembrane (Nimrod C-type) and extracellular (Nimrod B-type) receptors which, by subsequent sequence changes, acquired slightly modified binding properties broadening the recognition spectrum or increasing the efficiency by subfunctionalization (Zhang 2003), a process that can contribute to improvement of insect innate immunity.

Proteins containing repetitive domain units represent a significant proportion of the proteomes of living organisms (Björklund et al. 2006). Sequence information carried by the repeats is informative to understand the origin and evolution of the harboring genes. This study is a detailed analysis on gene and repeat level of the evolution of a gene superfamily which presumably has an important role in innate immune responses and might present an example as to how evolution of multigene families contributes to the creation of new genetic systems.

Supplementary Material

Supplementary text 1 and 2, tables 1–6, and figures 7–33 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the following sources: FEBS Long-Term Fellowship to K.S.; joint doctoral fellowship of The National Centre for Scholarship Abroad of the Romanian Ministry of Education and Research and the Hungarian Scholarship Board to B.S.; Hungarian Scientific Research Fund grant (NI60442) to I.A.; Bolyai Research Scholarship of the Hungarian Academy of Sciences to Z.P. Swedish Foundation for Strategic Research, the Wallenberg Consortium North, and the Swedish Research Council to D.H. Further support was provided by the OMAI DT-D/4/005/2006 and BIO-00126 funds of the Hungarian

Ministry of Education. Balázs Leitgeb, Ádám Kerényi, and Gábor Rákhely provided assistance in using the Linux cluster environment. Péter B. Kós gave useful suggestions on the manuscript. We thank two anonymous referees for their comments.

Literature Cited

Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. Bioinformatics. 21:2104–2105.

Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F. 2004. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics. 20:407–415.

Andrade MA, Perez-Iratxeta C, Ponting CP. 2001. Protein repeats: structures, functions, and evolution. J Struct Biol. 134:117–131.

Björklund AK, Ekman D, Elofsson A. 2006. Expansion of Protein Domain Repeats. PLoS Comput Biol. 2:e114.

Bork P. 1991. Shuffled domains in extracellular proteins. FEBS Lett. 286:47–54.

Bork P, Downing AK, Kieffer B, Campbell ID. 1996. Structure and distribution of modules in extracellular proteins. Q Rev Biophys. 29:119–167.

Callebaut I, Mignotte V, Souchet M, Mornon JP. 2003. EMI domains are widespread and reveal the probable orthologs of the Caenorhabditis elegans CED-1 protein. Biochem Biophys Res Commun. 300:619–623.

Carmon A, Wilkin M, Hassan J, Baron M, MacIntyre R. 2007. Concerted evolution within the Drosophila dumpy gene. Genetics. 176:309–325.

Chen K, Durand D, Farach-Colton M. 2000. Notung: a program for dating gene duplications and optimizing gene family trees. J Comput Biol. 7:429–447.

Cummings WJ, Yabuki M, Ordinario EC, Bednarski DW, Quay S, Maizels N. 2007. Chromatin structure regulates gene conversion. PLoS Biol. 18:5.

Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S. 2005. PROBCONS: probabilistic consistency-based multiple sequence alignment. Genome Res. 15:330–340.

Doliana R, Bot S, Bonaldo P, Colombatti A. 2000. EMI, a novel cysteine-rich domain of EMILINs and other extracellular proteins, interacts with the gC1q domains and participates in multimerization. FEBS Lett. 484:164–168.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Eickbush TH, Eickbush DG. 2007. Finely orchestrated movements: evolution of the ribosomal RNA genes. Genetics. 175:477–485.

Evans JD, Aronstein K, Chen YP, Hetru C, Imler JL, Jiang H, Kanost M, Thompson GJ, Zou Z, Hultmark D. 2006. Immune pathways and defense mechanisms in honey bees, Apis mellifera. Insect Mol Biol. 15:645–656.

Felsenstein J. 1989. Phylogeny inference package. Version 3.2. Cladistics. 5:164–166.

Finn RD, Mistry J, Schuster-Böckler B, et al. (13 co-authors). 2006. Pfam: clans, web tools and services. Nucleic Acids Res. 34:D247–D251.

Golubchik T, Wise MJ, Easteal S, Jermiin LS. 2007. Mind the gaps: evidence of bias in estimates of multiple sequence alignment. Mol Biol Evol. 11:2433–2442.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Hamon Y, Trompier D, Ma Z, Venegas V, Pophillat M, Mignotte V, Zhou Z, Chimini G. 2006. Cooperation between engulfment receptors: the case of ABCA1 and MEGF10. PLoS One. 1:e120.

Higgins D. 2003. Multiple alignment. In: Salemi M, Vandamme AM, editors. The phylogenetic handbook: a practical approach to DNA and protein phylogeny. Cambridge: Cambridge University Press. p. 45–71.

Holt RA, Subramanian GM, Halpern A, et al. (123 co-authors). 2002. The genome sequence of the malaria mosquito Anopheles gambiae. Science. 298:129–149.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755.

Hughes AL. 1999. Concerted evolution of exons and introns in the MHC-linked tenascin-X gene of mammals. Mol Biol Evol. 16:1558–1567.

Jiang H, Blouin C. 2007. Insertions and the emergence of novel protein structure: a structure-based phylogenetic study of insertions. BMC Bioinformatics. 15:444.

Johannesson H, Townsend JP, Hung CY, Cole GT, John W. 2005. Concerted evolution in the repeats of an immunomodulating cell surface protein, SOWgp, of the human pathogenic fungi Coccidioides immitis and C. posadasii. Genetics. 171:109–117.

Ju JS, Cho MH, Brade L, Kim JH, Park JW, Ha NC, Söderhäll I, Söderhäll K, Brade H, Lee BL. 2006. A novel 40-kDa protein containing six repeats of an epidermal growth factor-like domain functions as a pattern recognition protein for lipopolysaccharide. J Immunol. 177: 1838–1845.

Kocks C, Cho JH, Nehme N, et al. (13 co-authors). 2005. Eater, a transmembrane protein mediating phagocytosis of bacterial pathogens in Drosophila. Cell. 123:335–346.

Kumar S, Filipski A. 2007. Multiple sequence alignment: in pursuit of homologous DNA positions. Genome Res. 17:127–135.

Kumar S, Tamura K, Nei M. 2004. MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. Brief Bioinform. 5:150–163.

Kurucz E, Markus R, Zsamboki J, et al. (13 co-authors). 2007. Nimrod, a putative phagocytosis receptor with EGF repeats in Drosophila plasmatocytes. Curr Biol. 17:649–654.

Letunic I, Bork P. 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics. 23:127–128.

Liao D. 1999. Concerted evolution: molecular mechanism and biological implications. Am J Hum Genet. 64:24–30.

Manaka J, Kuraishi T, Shiratsuchi A, Nakai Y, Higashida H, Henson P, Nakanishi Y. 2004. Draper-mediated and phosphatidylserine-independent phagocytosis of apoptotic cells by Drosophila hemocytes/macrophages. J Biol Chem. 279: 48466–48476.

Mangahas PM, Zhou Z. 2005. Clearance of apoptotic cells in Caenorhabditis elegans. Semin Cell Dev Biol. 16:295–306.

McAllister BF, Werren JH. 1999. Evolution of tandemly repeated sequences: what happens at the end of an array? J Mol Evol. 48:469–481.

Morgenstern B. 1999. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics. 15:211–218.

Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3:418–426.

Nei M, Gu X, Sitnikova T. 1997. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci USA. 94:7799–7806.

Nei M, Rogozin IB, Piontkivska H. 2000. Purifying selection and birth-and-death evolution in the ubiquitin gene family. Proc Natl Acad Sci USA. 97:10866–10871.

Nei M, Rooney AP. 2005. Concerted and birth-and-death evolution of multigene families. Annu Rev Genet. 39:121–152.

Nishikawa T, Natori S. 2001. Targeted disruption of a pupal hemocyte protein of Sarcophaga by RNA interference. Eur J Biochem. 268:5295–5299.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol. 302:205–217.

Nylander JA, Wilgenbusch JC, Warren DL, Swofford DL. 2007. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics. 24:581–583.

Ohta T. 1994. Evolution of gene families: a clue to some problems of Neo-Darwinism. In: Levin SA, editor. Frontiers in mathematical biology. Berlin: Springer. p. 174–185.

Ota T, Nei M. 1994. Divergent evolution and evolution by the birth-and-death process in the immunoglobulin VH gene family. Mol Biol Evol. 11:469–482.

Parmley JL, Urrutia AO, Potrzebowski L, Kaessmann H, Hurst LD. 2007. Splicing and the evolution of proteins in mammals. PLoS Biol. 5:e14.

Posada D, Crandall KA. 1998. Modeltest: testing the model of DNA substitution. Bioinformatics. 14:817–818.

Quesada H, Ramos-Onsins SE, Aguade M. 2005. Birth-and-death evolution of the Cecropin multigene family in Drosophila. J Mol Evol. 60:1–11.

Redelings BD, Suchard MA. 2005. Joint Bayesian estimation of alignment and phylogeny. Syst Biol. 54:401–418.

Russo CA, Takezaki N, Nei M. 1995. Molecular phylogeny and divergence times of drosophilid species. Mol Biol Evol. 12:391–404.

Sackton TB, Lazzaro BP, Schlenke TA, Evans JD, Hultmark D, Clark AG. 2007. Dynamic evolution of the innate immune system in Drosophila. Nat Genet. 39:1461–1468.

Schuster-Böckler B, Bateman A. 2005. Visualizing profile-profile alignment: pairwise HMM Logos. Bioinformatics. 21:2912–2913.

Schuster-Böckler B, Schultz J, Rahmann S. 2004. HMM Logos for visualization of protein families. BMC Bioinformatics. 5:7.

Simmons MP, Ochoterena H. 2000. Gaps as characters in sequence-based phylogenetic analyses. Syst Biol. 49:369–381.

Sipos B, Somogyi K, Andó I, Pénzes Z. 2008. t2prhd: a tool to study the patterns of repeat evolution. BMC Bioinfo. 9:27.

Stajich JE, Block D, Boulez K, et al. (21 co-authors). 2002. The Bioperl toolkit: perl modules for the life sciences. Genome Res. 12:1611–1618.

Strimmer K, von Haeseler A. 1997. Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. Proc Natl Acad Sci USA. 94:6815–6819.

Suchard MA, Redelings BD. 2006. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioinformatics. 22:2047–2048.

Swanson WJ, Vacquier VD. 1998. Concerted evolution in an egg receptor for a rapidly evolving abalone sperm protein. Science. 281:710–712.

Swofford DL. 2003. PAUP*: phylogenetic analysis using parsimony (*and other methods). Version 4. Sunderland (MA): Sinauer Associates.

Thompson JD, Higgins DG, Gibson TJ. 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.

Xia X, Xie Z. 2001. DAMBE: data analysis in molecular biology and evolution. J Heredity. 92:371–373.

Zdobnov EM, Bork P. 2007. Quantification of insect genome divergence. Trends Genet. 23:16–20.

Zhang J, Rosenberg HF, Nei M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci USA. 95:3708–3713.

Zhang J. 2003. Evolution by gene duplication: an update. Trends Ecol Evol. 18:292–298.

Zou Z, Evans JD, Lu Z, Zhao P, Williams M, Sumathipala N, Hetru C, Hultmark D, Jiang H. 2007. Comparative genomic analysis of the Tribolium immune system. Genome Biol. 8:R177.

Jianzhi Zang, Associate Editor

Accepted August 5, 2008