EVOLUTION OF THE NIMROD GENE SUPERFAMILY

Ph.D. thesis summary

Botond Sipos

Supervisor: Zsolt Pénzes, Ph.D.
Consultant: Kálmán Somogyi, Ph.D.

University of Szeged
Doctoral School of Biology
Faculty of Science and Informatics

HAS Biological Research Center,
Intitute of Genetics

2010

Szeged

Introduction

Many members of the recently described Nimrod gene superfamily have been demonstrated to play a role in phagocytosis and binding of bacteria, which suggests that this superfamily may be an important constituent of innate immunity.

The proteins encoded by the members of the superfamily are characterized by the presence of a sequence motif (CCxGY/W) in their N-terminal region, and of at least one "NIM domain", a special type of EGF domain. They also possess a signal peptide suggesting that they act extracellularly.

The domain structure within the superfamily is highly variable. Based on the domain architecture, three major groups can be defined. The proteins with the Draper-type architecture have an N-terminal EMI domain, and only one copy of the NIM domain, followed by a variable number of "classic" EGF-like domains. The Nimrod C and Nimrod B type ("poly-NIM") proteins posses a variety of N-terminal domains, and an array of NIM domains separated by short linker sequences. The Draper and Nimrod C type proteins, in contrary to the Nimrod B type, generally have a transmembrane domain.

Genes encoding proteins with Draper-type domain architecture have a wide taxonomic distribution, being present, for example, in *Caenorhabditis elegans*, fruit fly, and human genomes. On the other hand, proteins containing an array of repeated NIM domains ("poly-NIM" proteins) have been identified only in insects so far.

Some genes belonging to the Nimrod superfamily, along with the *centaurin gamma* gene, and with genes belonging to the *ance* and CPF gene families, are a part of a gene cluster (the "Nimrod cluster") which is conserved on large evolutionary timescales (300-350 million years).

There is a considerable body of literature suggesting the involvement of various members of the superfamily in innate immunity. Draper-type proteins were described to have a function in phagocytosis in many species, for example, Ced-1 in *C. elegans*, Draper in *Drosophila melanogaster*, as well as MEGF-10 in human.

The role in phagocytosis was also shown for some Nimrod C type genes, like *eater* and *nimrodC1* in *D. melanogaster* or the gene coding for the 120 kDa protein in *Sarcophaga peregrina*. 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. It was also shown, that the *nimrodC4* (*SIMU*) gene is involved in the clearance of apoptotic debris in *D. melanogaster*.

Experimental data also support the role of a Nimrod B type protein as a pattern recognition receptor for bacterial lipopolysaccharide in the beetle *Holotrichia diomphalia*, which was show to facilitate *in vivo* the phagocytosis of *Escherichia coli*.

The "concerted" and "Birth-and-Death" (BD) models are one of the oldest conceptual models for the characterization of gene family evolution, formulated in the early years of the molecular evolutionary studies. These conceptual models basically concern the independence of evolutionary units (genes, domains, etc.), and the effects of the processes (gene conversion, unequal crossing over) violating independence.

These two conceptual models give different predictions on the topology of phylogenetic trees built from sequences coming from closely related species. In the case of units which experienced episodes of concerted evolution, the topology is often incongruent with the evolutionary history of the harboring species. Based on this, episodes of concerted evolution can be detected by using phylogenetic methods.

Goals

- The members of the Nimrod superfamily can be classified into major groups based on the previously published predicted domain structure (Nimrod A/Draper, Nimrod C, Nimrod B), but the domain architectures alone are not informative on the detailed evolutionary history of the superfamily. The first main objective of our study was to reconstruct the detailed evolutionary history of the Nimrod superfamily by using phylogenetic reconstruction and other sequence analysis methods, with the main focus on the insect "poly-NIM" sequences.

- We tried to elucidate the origins of the characteristic domain structure by using methods of sequence analysis.

- The second main objective of our study on the Nimrod was the characterization of the evolution of the domain repeats in the "poly-NIM" type protein in the light of the "concerted" and "Birth-and-Death" conceptual models.

- We intended to study the conserved Nimrod cluster, based on the literature and other publicly available data.

Methods

- The NIM domains were identified and aligned by a profile Hidden Markov Model trained by us, by using HMMER suite version 2.3.2.

- In order to study the evolutionary origins of the NIM domain, we assessed the similarity between the EGF-like and NIM domains by using the pairwise HMM logo method based on the local-local alignment of pHMMs (Logomat-P webserver).

- The amino acid substitution models best fitting the domain alignments were selected by using ProtTest 1.3. The same software was used for the Maximum Likelihood estimation of domain phylogenies. Neigbor-Joining domain phylogenies were built by MEGA 3.1.

- We used five different heuristic multiple alignment software (Clustal W 1.3, Muscle 3.6, T-Coffee 4.45, ProbCons 1.1, Dialign 2.2) to align the Nimrod A, Nimrod B and Nimrod C sequences. The quality of every alignment was evaluated by a couple of criteria (e.g., the placement of the NIM domains and CCXGY/W motifs, consistency scores calculated by T-Coffee) in order to choose the one with the most biological relevance. Sequences with domains supposedly engaged in concerted evolution were excluded from sequence alignments and hence from further phylogenetic analyses.

- Amino acid substitution models best fitting the alignments produced by ProbCons were selected by ProtTest 1.3.

- The phylogenetic signal in the ProbCons alignments was assessed by

performing likelihood mappings by using Tree-Puzzle 5.2.

- By using classical (Maximum Likelihood - PhyML 3.0, Neighbor-Joining - MEGA 3.1) and Bayesian methods (MrBayes 3.1), we reconstructed the phylogenies of the Nimrod A, Nimrod B and Nimrod C sequences. We used a simple method to encode the gaps from the multiple sequence alignments into binary characters, and during the Bayesian analyses we have studied the effect of the inclusion of these characters on the uncertainty of the topology estimation.

- The synonymous and non-synonymous distances between the nucleotide sequences coding the NIM domains were estimated by the Modified Nei-Gojobori method (MEGA 3.1).

- We tried to infer the evolutionary relationship between the Nimrod A/Draper, Nimrod B and Nimrod C sequences by the joint estimation of the alignment and phylogeny of the sequences of the *D. melanogaster* Nimrod paralogs (Bali-Phy 2.0.0).

- We used an approximate analytical method, which also deals with the presence of gene families, to asses the statistical significance of the conservation of the Nimrod cluster.

- Trees were edited by using the iTOL 1.7 web application and the APE R package. The consensus networks used for the evaluation of posterior topology samples were built by using SplitsTree 4.10.

Results and discussion

- Pairwise tests assessing the conservation against the cluster from the *D. melanogaster* genome indicate that this cannot be explained by a neutral null model assuming randomization of the gene order. Based on previously published data, we propose, that the conservation of the Nimrod cluster is maintained by some common regulatory elements requiring the vicinity of the genes, probably acting at the level of chromatin structure. Considering the timescale of the conservation and the available functional and expression data concerning the member gene families, it seems possible that the cluster is an important "functional module" of insect innate immunity.

- During the analysis of phylogenetic trees based on the amino acid sequences of the NIM domains, we found that, in contrary to the majority of the members of the superfamily, the evolution of a group of domains from the proteins encoded by the Drosophila *eater* genes could be described best by the concerted model. This finding was also corroborated by the estimated synonymous distances between the respective domains.

- Based on the domain phylogenies, we also suspect that domains from the *A. mellifera* Nimrod CI, *T. castaneum* Nimrod CI and also *T. castaneum* Nimrod CII may have been involved in episodes of concerted evolution, but in the lack of closely related ortologues, we were not able to confirm that.

- Since the analysis of phylogenetic trees, which is routinely used to study gene family evolution, proved to be less practical in the case of domain repeats, we have developed a new visual method. The method is based on domain

phylogenies, and makes possible the rapid assessment of which conceptual model describes the best the evolution of the domains from the different regions. The basic idea behind the method is to visualize the domains from a pair of sequences and the information relevant to the conceptual models from the phylogeny of the domains. The Perl script implementing the method can be downloaded from http://t2prhd.sf.net.

- By using the newly developed visual method we found that the domains from the middle of the Eater sequences are the ones involved in the concerted evolution. We also found, that more domains avoid homogenization at the N-terminal end of the domain array as compared to the C-terminal region, which we explain as caused by natural selection.

- Based on the pairwise logos and other similarities, we proposed that the EMI and NIM domains originated from two EGF-like domains and a short linker sequence through structural reorganizations in which insertions played an important role. This means that the Draper-type architecture might be the descendant of a "poly-EGF" architecture containing a transmembrane domain and also an array of EGF-like domains. Our hypothesis for the origin of the EMI and NIM domains could be a good example for the recruitment of domain repeats during the emergence of domains with new functions.

- The consensus network built from the posterior sample obtained by the joint estimation of the alignment and phylogeny of the sequences of the *D. melanogaster* Nimrod paralogs suggests that the Nimrod B sequences are part of the Nimrod C2 lineage.

- In conclusion, we have developed a hypothesis discussing the evolutionary history of the Nimrod superfamily from the emergence of the first characteristic domain architecture to the diversification of the individual families, and which we hope will prove to be useful for further studies aiming its member genes.

List of publications

The thesis is based on the following publications

- Somogyi*, K., Sipos*, B., Pénzes, Zs., Kurucz, É., Zsámboki, J., Hultmark, D., Andó, I. (2008): Evolution of genes and repeats in the Nimrod superfamily – *Molecular Biology and Evolution* **25**(11):2337–2347 IF: 7.28.

  * - shared first authorship.

- Sipos, B., Somogyi, K., Andó, I. , Pénzes, Zs. (2008): *t2prhd*: a tool to study the patterns of repeat evolution. – *BMC Bioinformatics* **9**: 27 IF: 3.781.

Other publications

- Szabó, K., Bozsó, M., Sipos, B., Pénzes, Zs.: Genetic diversity of great bustard (*Otis Tarda*) populations in the Carpathian Basin – *Conservation Genetics*, accepted.

- Pénzes, Zs., Melika, G., Bozsóki, Z., Bihari, P., Mikó, I., Tavakoli, M., Pujade-Villar, J., Fehér, B., Fülöp, D., Szabó, K., Bozsó, M., Sipos, B., Somogyi, K., Stone, G. N. (2009): Systematic re-appraisal of the gall-usurping wasp genus Synophrus Hartig, 1843 (Hymenoptera: Cynipidae: Synergini) – *Systematic Entomology* **34**(4):688-711 IF (2008): 1.808.

- Márkus, R., Laurinyecz, B., Kurucz, É, Honti, V, Bajusz, I., Sipos, B., Somogyi, K., Kronham, J., Hultmark, D., Andó, I. (2009): Sessile hemocytes as a hematopoietic compartment in Drosophila melanogaster – *PNAS* **106**(12):4805-4809 IF (2008): 9.38.

- Álmos, P.Z., Horváth, S., Czibula, Á., Raskó, I., Sipos, B., Bihari, P., Béres, J., Juhász, A., Janka, Z., Kálmán, J. (2008): H1 tau haplotype-related genomic variation at 17q21.3 as an Asian heritage of the European Gypsy population. – *Heredity* **101**(5):416–419 IF: 3.823.

- Markó, B., Sipos, B., Csősz, S., Kiss, K., Boros, I., Gallé, L. (2006): A comprehensive list of the ants of Romania (*Hymenoptera*: *Formicidae*). – *Myrmecologische Nachrichten* **9**: 65–76.

- Schlick-Steiner, B. C., Steiner, F. M., Konrad, H., Markó, B., Csősz, S., Heller, G., Ferencz, B., Sipos, B., Christian, E., Stauffer, C. (2006): More than one species of *Messor* harvester ants (*Hymenoptera*: *Formicidae*) in Central Europe. – *European Journal of Entomology* **103**(2):469–476 IF: 0.782.