TOLL EVOLUTION: A PERSPECTIVE FROM REGULATORY REGIONS

Rajakumar Sankula

Submitted to the faculty of Indiana University in partial fulfillment of the requirements for the degree Master of Science in the department of Bioinformatics in School of Informatics of Indiana University

29 January, 2004

Accepted by the Graduate Faculty, Indiana University in partial fulfillment of the requirements for the degree of Master of Science

_____

(Dr. Narayanan Perumal)

Master's Thesis Committee

_____

(Dr. Lang Li)

_____

(Dr. Chen Su)

Dedicated to all those who has done good for the faculty of science

Table of contents

ACKNOWLEDGEMENTS

I give my sincere gratitude to the people who do science with pure, unselfish and honest passion as they are the people who made me grow and appreciate the world in the way I see. This thesis is a result of one and half years of work where I was accompanied and supported by not many but few people. It is a pleasant opportunity to express my sincere appreciation to all of them.

The first person I would like to express my heartfelt gratitude is Dr. Perumal whose support and supervision has been unlimited throughout the thesis. I am indebted to the opportunity he has provided me with. He molded me in the right direction and taught me to be positive always.

Dr. Lang Li has been a great source of inspiration with his diplomatic and tactful approach towards any situation.

I would like to thank Dr. Chen Su for her kindness and being a part of the thesis committee.

The chain of gratitude will be incomplete without Kalyan for his help in various aspects.

I deeply appreciate and acknowledge all the love and support I received from my family and their belief in me.

ABSTRACT

Background: Toll and Toll-related proteins play an important role in antibacterial innate immunity and are widespread in insects, plants, and mammals. The completion of new genomes such as Anopheles gambiae has provided an avenue for a deeper understanding of Toll evolution. While most evolutionary analyses are performed on protein sequences, here, we present a unique phylogenetic analysis of Toll genes from the perspective of upstream regulatory regions so as to study the importance of evolutionary information inherited in such sequences.

Results: In a comparative study, phylogeny on the protein products of Toll like genes showed consistency with earlier literature except for the single point of divergence between insects and mammals. On the other hand, the phylogeny based on upstream regulatory sequences (-3000 to +10) showed a broader distinction between the plants and the rest, though the tree was not well resolved probably due to poor alignment of these sequences. The phylogeny based on TFBs necessitated the development of a supervised statistical approach to determine their "evolutionary informativeness". Employing the frequency of evolutionarily informative TFBs, a phylogeny was derived using pair-wise distances. It suggested a closer relationship between Anopheles and plants than to Drosophila and a significant homology among mammalian TLRs.

Conclusions: A unique approach of using TFBs in studying evolution of Toll genes has been developed. Broadly, this approach showed results similar to the protein phylogeny. The inclusion of the evolutionary information from TFBs may be relevant to such analyses due to the selective pressure of conservation in upstream sequences.

I. INTRODUCTION

A. Introduction to the subject

We live in a potentially hostile world filled with an array of infectious agents of diverse shape, size, composition and subversive character which would use us as rich sanctuaries for propagating their 'selfish genes' had we not also developed a series of defense mechanisms at least their equal in effectiveness and ingenuity. The immune system comprises of these defense mechanisms, which can establish a state of immunity against infection (Latin *immunitas*, freedom from). The immune system has traditionally been divided into innate and adaptive components, each with a different function and role. Overwhelmingly, however, studies on immunity during the last few decades have concentrated on the adaptive response and its hallmarks, that is, the generation of a large repertoire of antigen-recognition receptors and immunological memory. Only quite recently has innate immunity gained renewed interest, particularly as it became apparent that it is an evolutionary, ancient defense mechanism[1, 2].

The innate immune system is an ancient mechanism of host defense found in essentially every multicellular organism from plants to humans. In invertebrates, it is the only mechanism of defense. Vertebrates also developed adaptive immune response; however, the innate immune system is essential for instructing the cells of the adaptive system (T and B cells) by presenting antigen in the context of an appropriate costimulatory molecule. The innate immune system developed to not only discriminate self from non-self but more importantly, it can discriminate *infectious* non-self from *innocuous* non-self.

Innate Immunity – an evolutionary perspective: Innate immunity is the first-line host defense of multicellular organisms that operates to limit infection upon exposure to microorganisms. Research from the last decade has shown that there is a strong evolutionary relationship in the regulation of innate immunity among plants, insects, and mammals. The mechanistic elucidation and insight into innate immunity required very different methods and different perspectives, drawn from two separate domains of biology that involve insects and mammals. Innate immunity preceded adaptive immunity in evolution as it has been supported by the presence of conserved signaling pathway components in organisms lacking the typical adaptive immunity of vertebrates[3]. The last common ancestor of insects and mammals is thought to have lived more than half a billion years ago[4]. During that time, the physiology of insects and mammals has diverged considerably: innate immune defense based on antimicrobial peptides predominates in insects, whereas cytokine-mediated inflammation seems to hold sway in mammals.

Toll genes and Toll-related proteins in Innate Immunity: Toll genes and Toll-related proteins play an important role in antibacterial innate immunity and are widespread in insects, plants, and mammals[5]. Toll is a singlepass transmembrane receptor with an ectodomain marked by leucine-rich repeat motifs. Toll was originally identified as a Drosophila gene required for ontogenesis and anti-microbial resistance[6]. Genetic analysis revealed that this gene controls dorsoventral polarization in the fruit fly as well as immunity against fungal infection. The recognition of sequence similarity between the cytoplasmic portion of Toll and that of signalling interleukin-1 (IL-1) receptor (IL-1R) components (the Toll/IL-1R module, or TIR module) of mammals represented the merging point of Drosophila work with more conventional innate-immunity research[7].

The family of Toll-like receptors recognizes pathogen-associated molecular patterns (PAMPs) as nonself. PAMP recognition by TLRs then leads to cytokine production and expression of costimulatory molecules such as CD40, which can shape up protective innate and adaptive immunity[8].

Some of the first phylogenetic analyses suggested that the insect Toll family of proteins and their mammalian counterparts evolved independently and that they shared one common ancestor[9]. This result is consistent with the conclusion of discontinuous evolution of innate immunity between invertebrates and vertebrates, obtained through phylogenetic analyses of various proteins[10]. The protein motif shared by Toll and IL-1R, or TIR, is evident in plant proteins as well as in animal proteins[11], so this motif might be traceable to the origins of eukaryotic life, that is, between one and two billion years ago. However, according to a recent phylogenetic analysis of Toll interleukin-1-receptor domain as reported[4], Toll-9 receptor of Drosophila resembles the mammalian TLRs more closely than do any of other Drosophila Tolls. This phylogenetic analysis has been done by including plant disease resistant genes, mammalian TLR genes, IL-1/IL-18/ST2/SIGIRR/MYD88 genes, and Drosophila Toll genes. This approach makes way for a comprehensive phylogenetic analysis by including the Toll genes of Anopheles in addition to the genes above mentioned and that is likely to answer the question of Toll evolution in more comprehensible manner.

Drosophila and Mammals: Drosophila genome sequencing project reveals important similarities between the functioning of mammalian and invertebrate immune systems[12]. These similarities made Drosophila well suited to the study of innate immunity. The important functional similarity is between the transmembrane receptors.

These receptors, Toll in Drosophila and the IL-1 receptor in mammals, share an intracytoplasmic homology domain (referred to as Toll/IL-1 receptor or 'TIR' domain) that associates with adaptor molecules, leading to the activation of the homologous protein kinases Pelle and IRAK (IL-1 receptor-associated kinase), respectively[5, 13]. Drosophila relies for its host defense on both cellular and humoral reactions. The hallmark of the humoral response is the challenge-induced synthesis and secretion by the fat body (a functional equivalent of the vertebrate liver) of a battery of small cationic polypeptides. These are induced in response to immune challenge and have potent antimicrobial activities directed against either fungal pathogens (drosomycin, metchnikowin) or bacteria (diptericin, drosocin, cecropin, attacin, defensin)[14, 15]. The genes encoding these peptides have in their upstream regions nucleotide sequence motifs similar to mammalian NF-κB-binding sites. Establishment of transgenic fly lines with various reporter constructs demonstrated that these sequence motifs confer the immune-inducibility to the corresponding genes[16, 17]. In the early 1990s, it was recognized that there were striking similarities between the control of dorsoventral patterning by the Rel transcription factor Dorsal in Drosophila embryos and the activation of the Rel protein NF-κB by the cytokine interleukin-1 (IL-1) in mammalian cells and this was interpreted as a point of common ancestry for Drosophila and mammals[5].

Anopheles and Drosophila: Anopheles mosquito is capable of mounting a robust innate immune response against Plasmodium infection. It was first shown that a set of diverse immune genes is transcriptionally activated both systemically and locally in epithelial tissues in the course of Plasmodium infection[18]. Recent completion of the *Anopheles gambiae* genome sequencing project resulted in an interesting revelation that

242 genes from 18 gene families implicated in innate immunity with marked diversification relative to *Drosophila melanogaster*[19]. Out of the 18 families, the signaling receptor family showed the modest diversification. Anopheles has 11 Toll genes, of which four (Toll 6, 7, 8, and 9) are unambiguous orthologs of *Drosophila melanogaster*[20]. This advocates the necessity of looking for the position of Anopheles in the evolution of innate immunity.

Plants: Studies of receptors and signal-transduction components that play a role in plant disease resistance have revealed remarkable similarities with innate immunity pathways in insects and mammals. Many responses involved in plant disease resistance are dependent upon interaction of pathogenic effector molecules with specific plant resistance (R) proteins[21]. Although the signalling pathways initiated by these interactions are just beginning to be unraveled, the past few years have seen dramatic advances in the understanding of the molecular principles of plant disease resistance. It has become clear that some of the molecular mechanisms involved in innate immunity in mammalian and insect systems are remarkably similar to the molecular mechanisms underlying plant disease resistance responses[3].

Phylogenetic Analysis of Protein Sequences: Ever since the pioneering work of Zuckerkandl and Pauling[22], protein sequences and structures have been used extensively to infer organismal phylogeny and to predict biochemical functions. As these models cannot be easily reconstructed using conventional phylogenetic methods alone, diverse methodologies were applied to glean the requisite information. As a result, the most conserved sets of proteins and constituent domains, traceable to the Last Universal

Common Ancestor (LUCA) of all life forms was identified. Through a comparison of the homologous domains, the pre-LUCA stages of protein evolution were reconstructed. One of the conclusions that became apparent was that even before the extant translation apparatus was in place, complex protein domains, resembling extant forms, were already being synthesized.

Phylogenetic Analysis with Regulatory Regions: Complex eukaryotic genomes are composed of large amounts of DNA that do not code for protein. The presence of non-protein-coding DNA in introns and intergenic regions represents one of the major differences in genome structure between prokaryotes and eukaryotes, and is likely responsible in part for the major transition in complexity between these groups of organisms. Little is known about the structure or function of non-protein-coding DNA, although it is thought that some fraction contributes to the regulation of gene expression. Gene regulation has been speculated to play a major role in the evolution of animal and plant morphology, and thus finding the keys to understand non-protein-coding DNA evolution may be an important step in understanding the morphological diversity of life on earth.

Phylogenetic analyses, in general, employ as traditional input sequences either entire gene sequences (including both coding and non-coding regions) or protein sequences. The importance of utilizing regulatory regions in phylogenetic analysis is not well known and hence considerably less well studied. However, the alignments of regulatory regions of human and rodent genes often reveal blocks of highly conserved sequences[23]. In a

recently reported study by Thomas *et al.*[24], 302 intergenic 'multi-species conserved sequences' (MCS) have been found in the sequences upstream of the gene encoding cystic fibrosis transmembrane conductance regulator and nine other genes. These MCS overlap with 63% of functionally validated regulatory elements and some of these regulatory elements may be specific to the primate lineage. More recently, at a wider evolutionary scale, Dermitzakis *et al.*[25] have observed the conservation of non-genic sequences among 14 mammalian species from primates to monotremes to marsupials. Observations of such strong sequence conservation suggest conserved function, thereby generating testable hypotheses that can be verified.

B. Importance of the Subject

Innate defense is so fundamental that vertebrates, invertebrates and plants have many similarities. With the given understanding of Toll genes and Toll-related proteins in innate immunity, we further proceed to elucidate the importance of studying these receptors across mammals, Drosophila, Anopheles, and plants from evolutionary point of view, which may unearth important patterns that could explain the basis and divergence of these organisms with respect to innate immunity. For a while, researchers have been exploring these genomes from the phylogenetic perspective and the efforts have been increased manifold with the completion of new genome projects. However, the explored similarities between Drosophila and Anopheles Toll genes result in an opportunity for the above-mentioned study. Undoubtedly, the intellectual input from phylogenetic studies on innate immunity will be invaluable in advancing the field, to the point that intervention through the vector immune system can be considered as part of an integrated approach to

the control of diseases like malaria (casual organism Anopheles) and other parasitic diseases.

The usage of conserved non-coding sequences including regulatory elements for comparative genomic studies is gaining importance now, particularly for exploring the mechanism of transcriptional gene regulation across multiple organisms. But the major bottle neck for these studies is the small size of these conserved regulatory elements (~14 bp). So far, efforts have been made to look at non-coding sequences including regulatory elements, mostly through phylogenetic foot printing. However, these efforts aimed at studying closely related species such as either insects or mammals alone, as there is considerable alignment given the evolutionary distance among those species is rather less. But for looking at distantly related species such as plants, insects, mammals together, irrespective of presence of conserved regulatory elements, these non-coding regulatory regions result in poor alignments and making it difficult to arrive at accurate phylogenetic analyses of these non-coding regulatory regions across multiple organisms. Hence it calls for a computational approach to extract the evolutionary information that is inherent in conserved non-coding regulatory regions. It holds a great potential to look at conserved regulatory elements with an evolutionary perspective in case of none but Toll genes and Toll-related proteins as they are wide spread and conserved across mammals, insects and functionally much similar to plant disease resistance genes.

Protein based phylogeny of Toll genes and Toll-related proteins across mammals, insects and plants would give greater insight into functional conservation as did earlier phylogenetic analyses, which did not have an opportunity to include Anopheles Toll

genes. Such a comprehensive phylogenetic analysis would invariably test the hypothesis of single point of divergence between mammals and insects with respect to evolution of Toll genes and Toll-related proteins.

In the event of similarities between protein based phylogeny and phylogeny based on upstream regulatory regions using a novel computational approach, one can find the validity of conservation of non-coding regulatory regions in studying the evolution and that may well explain discrepancies between phylogenetic analyses with both nucleotide sequences and protein sequences.

C. Knowledge Gap

Large amount of research has been performed regarding the evolution of Toll genes and Toll-related proteins with respect to innate immunity. It was mostly involving mammals, insects (Drosophila) and plants. However, the recent completion of Anopheles genome and its exploration by researchers revealed that there are four unambiguous orthologs of Drosophila Toll genes in Anopheles Toll genes. Given the importance of their presence, it calls for a comprehensive phylogenetic study of Toll genes and Toll-related proteins by including Anopheles Toll genes, which forms one of the major goals of this research project. It would answer some important questions regarding the evolutionary position of Anopheles Toll genes that could be helpful in developing an integrated approach to the control of diseases like malaria (casual organism Anopheles).

Though there is a general agreement on the importance of conserved non-coding regions with respect to evolution, not much has been done to exploit that inherent

evolutionarily important information. Conventional phylogenetic approaches were not successful given the smaller size of (<14 bp) of the conserved regulatory elements among these non-coding regions especially upstream regulatory regions. Towards this end the main goal of this research project is to develop a computational approach to utilize evolutionarily important information in upstream regulatory regions of highly conserved Toll genes and Toll-related proteins. This approach would have higher probability to study the evolution of transcriptional regulatory mechanism of Toll genes and Toll-related proteins with a new insight. In addition, it would lead to the development of advanced computational approaches for phylogenetics that assumes a lot of potential when more and more researchers are convinced with the fact that in addition to protein sequences there is important evolutionary information in the non-coding regions, which is worth exploring.

II. BACK GROUND

A. Related research in evolutionary significance of non-coding DNA

In a recently published study, Dermitzakis *et al.*[25] have quantified levels and patterns of conservation of 191 conserved non-geneic sequences (CNGs) of human chromosome 21 in 14 mammalian species. These CNGs were significantly more conserved than protein coding genes and non-coding RNAS (ncRNAs) within the mammalian class from primates to monotremes to marsupials. The pattern of substitutions in CNGs differed from that seen in protein-coding and ncRNA genes and resembled that of protein-binding regions. About 0.3% to 1% of the human genome corresponds to a previously unknown class of extremely constrained CNGs shared among mammals.

Thomas *et al.*[23] reported substantial numbers of conserved non-coding segments beyond those previously identified experimentally, most of which were not detectable by pair-wise sequence comparisons alone. This study involved generation and analysis of over 12 megabases (Mb) of sequence from 12 species, all derived from the genomic region orthologous to a segment of about 1.8Mb on human chromosome 7 containing ten genes, including the gene mutated in cystic fibrosis. It resulted in an interesting revelation that 302 intergenic 'multi-species conserved sequences' (MCS) have been found in the sequences upstream of the gene encoding cystic fibrosis transmembrane conductance regulator and nine other genes. These MCS overlap with 63% of functionally validated

regulatory elements and some of these regulatory elements may be specific to the primate lineage.

Related research in evolution of Toll genes and Toll-related proteins

Looking at the some of the earlier research in this direction which offered valuable insight, Luo and Zheng[9] studied the phylogeny of Toll family of proteins from mammals, insects and plants with amino acid sequences as input.

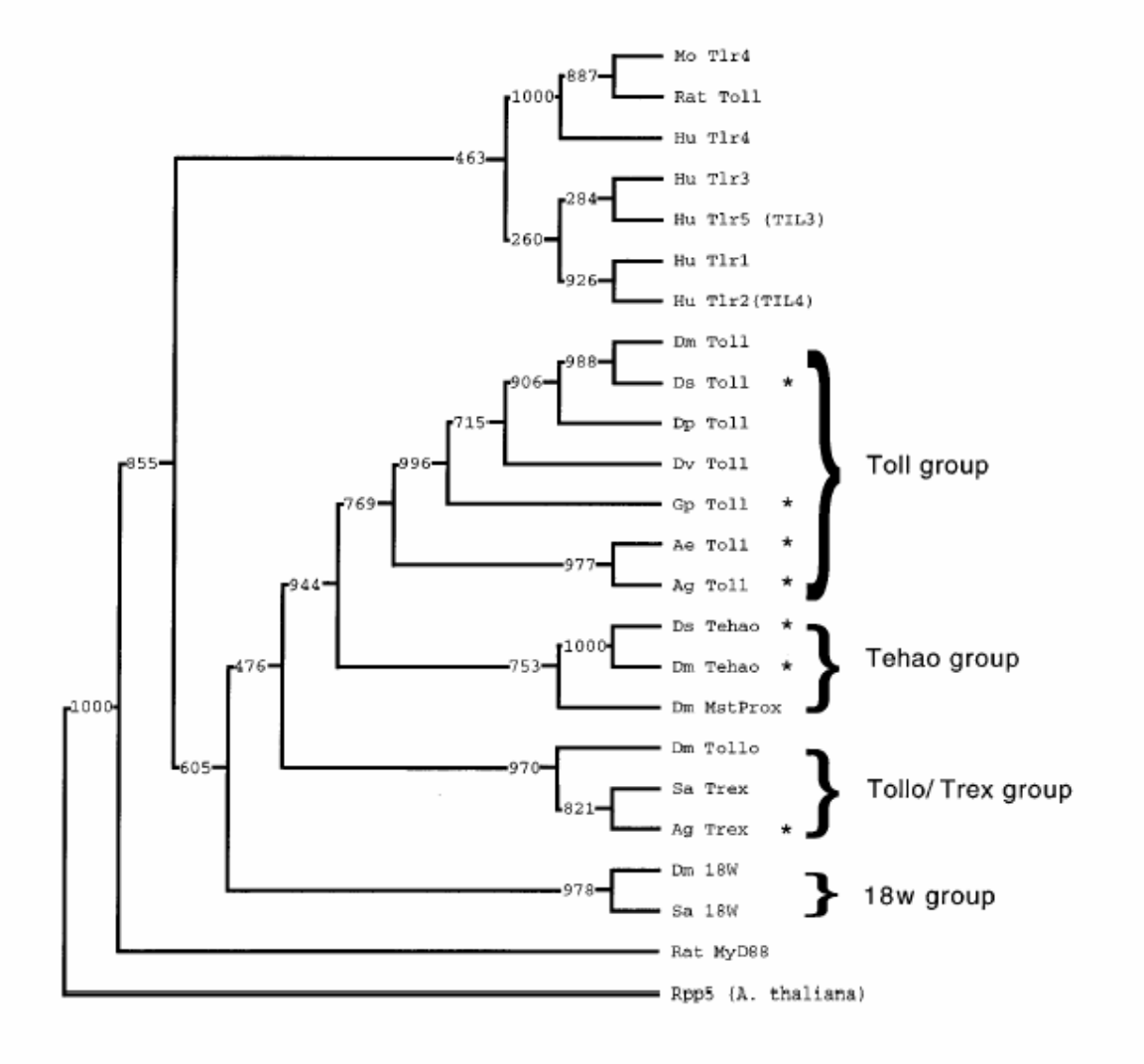

FIGURE 1: Phylogenetic tree of Toll family of proteins. (Luo and Zheng, 2000)

This study revealed that valuable patterns explained in Fig. 1; the insects and mammalian Toll-related proteins are clustered into separate groups and rest of the Toll-related proteins of plants clustered distantly from mammals compared to insects, which suggest that there may be a single point of divergence between mammals and insects with respect to Toll-related proteins. However the placement of MyD88 of rat as a separate group is not clearly explained. The insect Toll- related proteins are further subdivided into four subgroups: Toll, Tehao, Tollo/Trex, and 18W. Dm Toll and Dm 18W have been implicated in the innate immunity of Drosophila; the functions of the other two groups are not yet known.

Another genomic sequence analysis did not identify a Toll/IL-1R homologue in *Caenorhabditis elegans*[26]. Proteins with leucine-rich repeats (leucine-rich repeats are characteristic of Toll-like receptors) are present in *C. elegans*, however. This raises the question of the origin and evolution of innate immunity in higher organisms. Was a novel signal domain utilized after the appearance of segmented organisms? Or, were there several independent evolution events, with the conservation of Toll/IL-1R-like proteins resulting from convergent evolution? This question can be addressed looking for the Toll family of proteins outside dipteran insects and Insecta and further down the evolutionary scale[10].

A similar but recent study reported by Kimbrell and Beutler[4] looked at the evolution of the Toll interleukin-1-receptor domain revealed interesting patterns as explained in Fig.2.

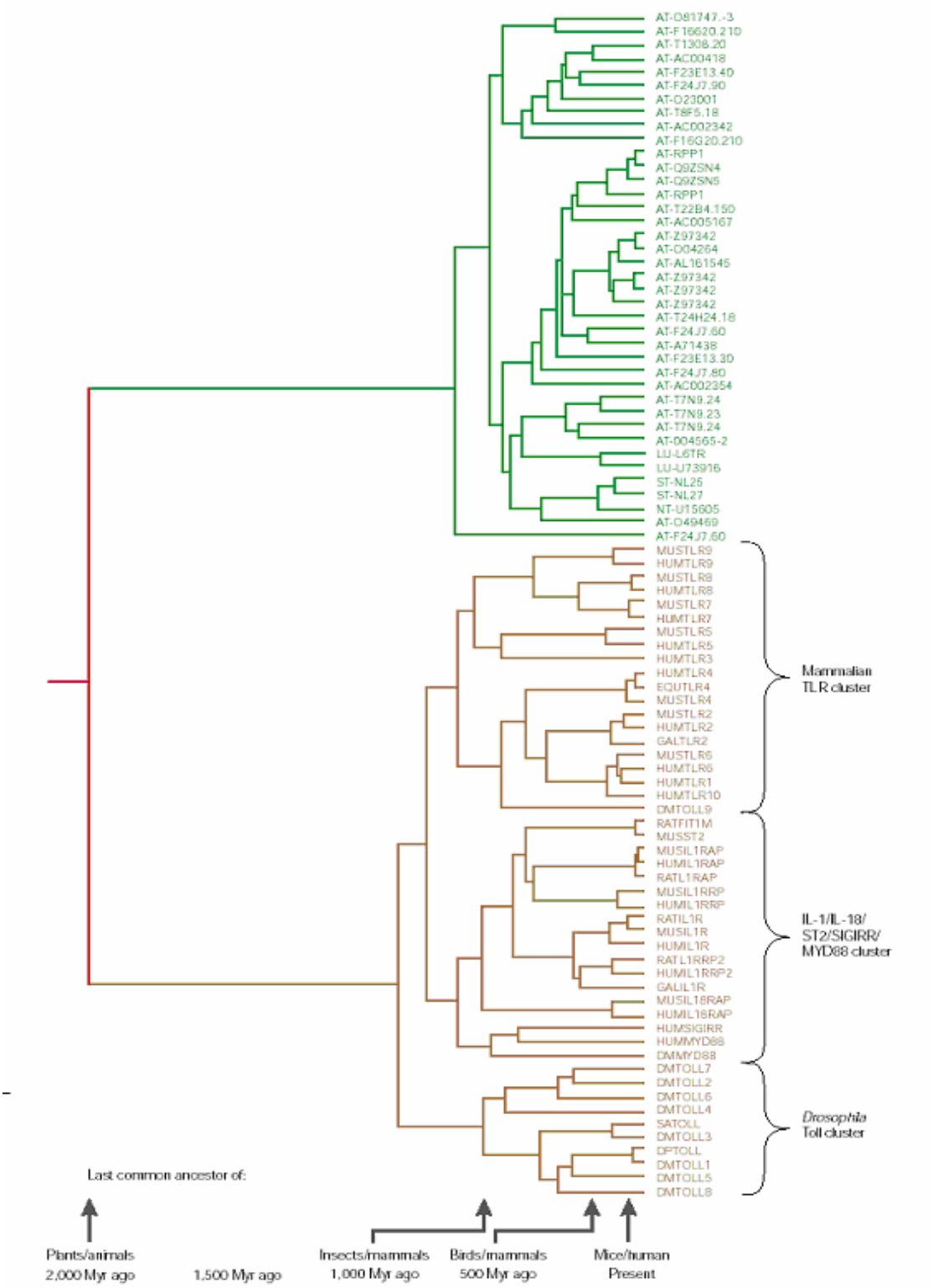

FIGURE 2. Evolution of the Toll interleukin-1-receptor domain (Kimbrell and Beutler, 2001)

Two great divisions in the evolution of the Toll interleukin-1-receptor (TIR) domain are immediately evident in the phylogenetic tree, as Metazoa (bottom) and Viridiplantae (top) show only weak TIR homology at the protein level. TIR domains are found in most plant and animal taxa. The basis of this study was that the protein motif shared by Toll and IL-1R, or TIR, is evident in plant proteins as well as in animal proteins[11], so this motif might be traceable to the origins of eukaryotic life, that is, between one and two billion years ago.

As stated earlier, the *Anopheles gambiae* genome sequencing project resulted in a significant identification of 242 genes from 18 gene families implicated in innate immunity and their marked diversification relative to *Drosophila melanogaster*[19]. Out of the reported 18 families, the signaling receptor family showed the modest diversification. Anopheles has 11 Toll genes, of which four (Toll 6, 7, 8, and 9) are unambiguous orthologs of *Drosophila melanogaster*[20].

B. Current understanding of the subject

The phylogenetic analysis reported by Luo and Zheng[9] suggested that the insect Toll family of proteins and their mammalian counterparts evolved independently and that they shared one common ancestor. This result is consistent with the conclusion of discontinuous evolution of innate immunity between invertebrates and vertebrates, obtained through phylogenetic analyses of various proteins[10].

This study by Kimbrell and Beutler[4] offers an explanation of placement of rat MyD88 as separate group as reported by Luo and Zheng[9]; MYD88 shares a proximal

common ancestor with the IL-1R/IL-18R/ST2 family of TIR-bearing receptors, which were not involved in the earlier study. Another interesting observation is that Toll-9 receptor of Drosophila resembles the mammalian TLRs more closely than do any of other Drosophila Tolls.

The presence of four unambiguous orthologs of Drosophila Toll genes in Anopheles Toll genes further supports the evolutionary importance of Anopheles with respect to evolution of innate immunity. This resulted in an opportunity for a comprehensive phylogenetic study with similar approaches of earlier researchers with addition of Anopheles Toll genes to those of mammals, Drosophila and plants.

In addition there is significant accumulation of evidence for the importance of conserved non-coding regulatory sequences in studying evolution [24, 25]. Being widely dispersed across plants, insects and mammals, Toll genes offer an excellent opportunity to explore the evolution of transcriptional regulatory mechanisms by using non-coding regulatory sequences.

C. Research Question:

This research project is mainly intended to answer the following questions.

1. Does the addition of *Anopheles gambiae* genes to the phylogenetic analysis prove/disprove the earlier hypothesis[4, 9] of divergence between insect and mammalian Toll and Toll-related proteins?

2. Although Toll and Toll-related proteins are well conserved across diverse species, is there any conservation of transcriptional regulatory sequences upstream of the coding sequences? In highly conserved genes, regions of homology have been shown to extend

into upstream transcriptional regulatory sequences indicating selective pressure against random changes in these regions of functional conservation[23-25].

D. Intended Research Project:

To evaluate the significance of evolution of Toll genes in general as well as from the perspective of upstream regulatory regions (-3000 to +10), phylogenetic analyses would be conducted in the following two approaches:

(1) A comprehensive phylogenetic analysis with protein sequences and upstream regulatory regions (-3000 to +10) of insect (Anopheles and Drosophila) Toll genes, mammalian Toll – like receptors (TLRs), mammalian Toll interleukin-1-receptor (TIR) domain, and plant disease resistance genes.

(2) An approach based on the presence of "**evolutionarily informative**" transcription factor binding sites (TFBs) in upstream regions (-3000 to +10) of the genes involved in the first approach. The "evolutionary informativeness" of TFBs would be determined by a supervised statistical approach.

The resultant phylogenies would be used for comparative analysis to analyze the similarities and dissimilarities between the two approaches and to look for evolutionarily interesting patterns as well to answer the proposed research questions.

III. METHODS

A. Materials and Instruments:

The sequence data for upstream regions (-3000 to +10) and proteins of mammalian TLR, mammalian TIR domain, Drosophila Toll genes and MyD gene, Anopheles Toll genes, and plant disease resistance genes were obtained from GENBANK[27]. Upstream sequences (-3000 to +10) were retrieved based on assigning the transcription start site (TSS) as the +1 site. The list of genes is provided in Table 1.

TABLE 1: List of genes involved in the study.

Group	Gene Name	TFBs presence in upstream region (-3000 to +10) as predicted by EZ-retrieve (%)
Anopheles Toll genes*	Anatoll10	25.78
	Anatoll11	20.86
	Anatoll1B	23.82
	Anatoll5A	18.83
	Anatoll5B	19.40
	Anatoll6	21.69
	Anatoll7	24.18
	Anatoll8	17.14
	Anatoll9	20.59
Drosophila Toll genes**	DMTOLL1	15.34
	DMTOLL2	16.74
	DMTOLL3	18.57
	DMTOLL4	16.91
	DMTOLL5	17.40
	DMTOLL6	24.15
	DMTOLL7	20.29
	DMTOLL8	26.04
	DMTOLL9	21.99
Plant disease Resistance genes**	At1g27170 (T7N9.23; similar to N protein from Nicotiana glutinosa)	22.29
	At1g27180 (T7N9.24; similar to flax rust resistance protein)	25.28
	At1g65390 (T8F5.18; Disease resistance)	20.83
	At2g17050 (AC002354 putative disease resistance protein)	20.83
	At3g44410 (T22K7.90 RPP1)	18.33
	At4g04110 (T24H24.18; domain signature TIR exists, suggestive of a disease resistance protein)	20.13
	At4g11170 (T22B4.150; domain signature TIR-NBS-LRR, suggestive of a disease resistance protein)	20.29
	AT4g16860 (AL161545) (strong similarity to Downy Mildew Resistance Protein RPP5)	25.31
	At4g19500 (F24J7.60; downy mildew resistance protein RPP5)	19.56
	At4g19520 (F24J7.80; downy mildew resistance protein RPP5)	20.20
	At4g19530 (F24J7.90; TMV resistance protein N, Nicotiana)	20.16

Group	Gene Name	TFBs presence in upstream region (-3000 to +10) as predicted by EZ-retrieve (%)
	At4g23510 (F16G20.210)	22.92
	At4g36140 (F23E13.40; domain signature TIR-NBS-LRR disease resistance protein)	21.10
	ATAC002342 (T19K24.2 Disease resistance)	17.90
	ATAC005167 (F12A24.5 disease resistance)	22.12
	Z97342 (dl4460c disease resistance RPP5 like protein)	24.91
	Z97342 (dl4470c similarity to Downy mildew resistance protein RPP5)	21.66
	Z97342 (dl4475c strong similarity to Downy mildew resistance protein RPP5)	25.78
	Z97342 (dl4480c disease resistance RPP5 like protein)	23.05
Mammalian TLR Domain**	HUMTLR1	22.49
	HUMTLR2	23.72
	HUMTLR3	17.54
	HUMTLR4	20.00
	HUMTLR5	25.35
	HUMTLR6	24.58
	HUMTLR7	24.48
	HUMTLR8	18.81
	HUMTLR9	23.02
	HUMTLR10	17.24
	MUSTLR2	19.63
	MUSTLR4	23.78
	MUSTLR5	25.78
	MUSTLR6	20.43
	MUSTLR7	22.03
	MUSTLR8	21.79
	MUSTLR9	21.89
Mammalian TIR domain**	HUMIL1R1	23.89
	HUMIL1R2	23.02
	HUMIL1RAP	21.03
	HUMIL18RAP	25.41
	HUMMYD88	22.16
	MUSIL1RAP	17.80
	MUSIL18RAP	20.60
	RATIL1R1	21.16
	RATIL1R2	27.77
	RATIL1RAP	22.86
Drosophila MYD gene**	DMMYD88	19.07

* Christophides *et al.*[19], **Kimbrell and Beutler[4]
The percentage of TFB presence in the upstream regulatory regions (-3000 to +10) in included in the table.

The following software and web tools were used at various stages of this study for different purposes.

ClustalX[28] – This software is used for multiple sequence alignment of both upstream regulatory regions and protein sequences.

EZ-Retrieve[29] – This web based tool is used to predict the presence of TFBs in upstream regulatory regions of genes involved in the study. This web tool is available at the following web address: http://www.cag.icph.org/bioinformatics.html

PAUP 4.0[30] – This software is used for phylogenetic analysis of protein sequences involved in the study using Neighbor Joining method.

RASA[31] – This web based tool is used for noise reduction in order to improve the phylogenetic signal in the upstream regulatory regions of all the genes involved in the study. This web tool is available at the following web address: http://bioinformatics.upmc.edu/RASA.html

MEGA 2.1[32] – This software is used for phylogenetic analysis of upstream regulatory regions of all genes involved in the study under Neighbor Joining method as well as for phylogenetic analysis using evolutionarily informative TFBs based on pair-wise distance matrix using Neighbor Joining method.

MATLAB 6.5.1 Statistics Tool Box – This software is used to calculate the pair-wise distance matrix from the frequencies of highly evolutionary TFBs using Euclidean distance.

Neighbor Joining Method: It is the most common phylogenetic method developed by Saitou and Nei[33] to construct the phylogenetic trees from evolutionary distance data. In this study this method is applied for all the phylogenetic analyses. The principle of this method is to find pairs of operational taxonomic units (OTUs [= neighbors]) that minimize the total branch length at each stage of clustering of OTUs starting with a star-like tree. The branch lengths as well as the topology of a parsimonious tree can quickly be obtained by using this method.

B. Samples and Subjects:

The data involved in this study are upstream regulatory sequences (-3000 to +10) and protein sequences of 65 genes (mammalian TLR (16 genes), mammalian TIR domain (11 genes), Drosophila Toll genes (9 genes) and MyD gene, Anopheles Toll genes (9 genes), and plant disease resistance genes (19 genes)) that were obtained from GENBANK[27]. The frequencies of 126 TFBs across the genes were obtained by using EZ-Retrieve[29].

C. Procedures:

The complete methodology followed in this study was presented in the following Figure 3.

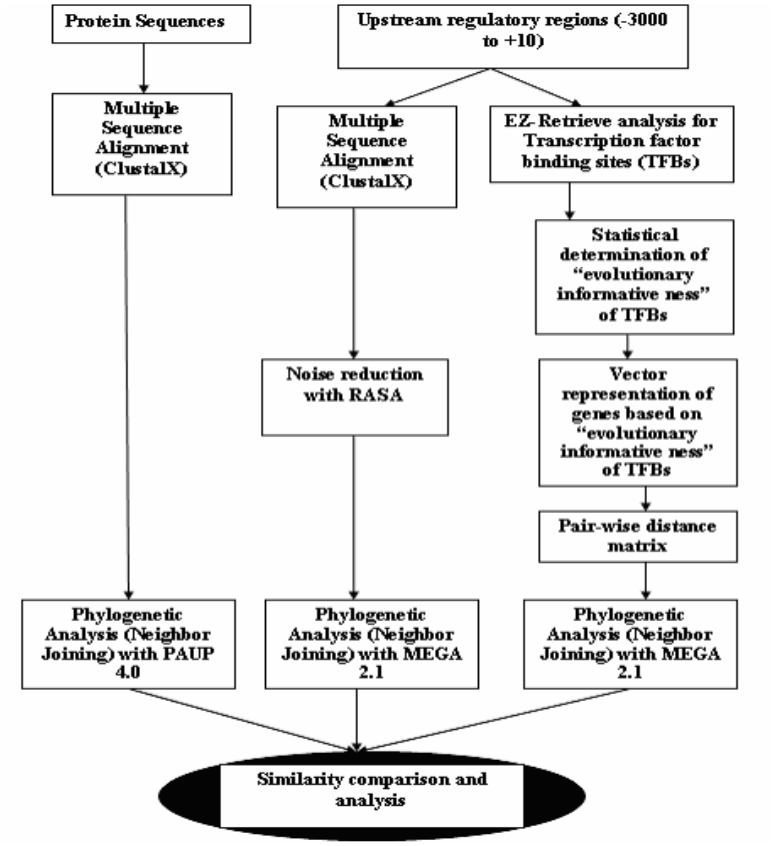

FIGURE 3: Methodology for comparative phylogenetic analysis.

Protein Phylogeny: In this conventional approach, protein sequences of all genes were obtained from GENBANK[27] and multiple sequence alignment was performed with ClustalX[27]. Then, the phylogenetic analysis was done with PAUP 4.0[30] using the Neighbor Joining method with p-distance; 1000 bootstrap replicates were done and a bootstrap consensus tree was obtained.

Phylogeny based on upstream regulatory regions (-3000 to +10): Upstream regions (-3000 to +10) of all genes were obtained from GENBANK[27] and multiple sequence alignment was performed with ClustalX[28]. However, there was not much of a significant alignment observed. The aligned data was subjected to noise reduction employing the RASA[31] (Relative Apparent Synapomorphy Analysis) algorithm to reduce the noisy data and improve the phylogenetic signal. The resultant data was subjected to phylogenetic analysis with MEGA 2.1[32] using Neighbor Joining Method with p-distance; 1000 bootstrap replicates were done and a bootstrap consensus tree was obtained.

Identification of TFBs among genes: Upstream regulatory regions (-3000 to +10) of all genes were arranged in a FASTA format and searched for TFBs with EZ-Retrieve[29] using a threshold of 90%. The resultant frequencies of each TFB were tabulated.

Selection of evolutionary informative TFBs: We followed a supervised statistical approach to calculate the "evolutionary informativeness" of TFBs. All genes involved in the study were divided into four groups as follows: (1) insect Toll genes (Anopheles and Drosophila), (2) mammalian TLR genes, (3) mammalian TIR Domain, and (4) plant disease resistance genes. From EZ-Retrieve[29], frequencies of TFBs were

obtained and tabulated per each gene in these groups. These frequencies of TFBs were normalized by using the maximum frequency value for each TFB across all genes.

Selection of evolutionarily informative TFBs was performed by a supervised statistical approach[34]. For these normalized frequencies of TFBs, Between Sum of Squares (BSS) and Within Sum of Squares (WSS) were calculated across the four groups based on formulas given below.

$$BSS = \sum_{i=1,...,4} n_i (\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2, \qquad WSS = \sum_{i=1,...,4} \sum_{j=1,...,n_i} (\bar{x}_{i\bullet} - x_{ij})^2,$$

where, x_{ij} is the *jth* normalized frequency of TFB in group $i = 1,...,4$, $\bar{x}_{i\bullet} = \sum_{j=1,...n_i} x_{ij} / n_i$ is the average normalized frequency of TFB in group i, and $\bar{x}_{\bullet\bullet} = \sum_{i=1,...,4} \sum_{j=1,...n_i} x_{ij} / n$ is the overall average normalized frequency. In particular, (n_1, n_2, n_3, n_4) represent the number of TFBs in the insect toll genes (Anopheles and Drosophila), mammalian TLR genes, mammalian TIR genes, and plant disease resistance genes respectively. The selection of evolutionary informative of TFBs was based on the BSS/WSS ratio of TFBs among the four groups. For this study, we used the TFBs with a BSS/WSS ratio above 0.2 as highly evolutionarily informative TFBs with respect to Toll-like genes.

TFB based Phylogeny: Twelve highly evolutionarily informative TFBs were chosen and all genes were represented in vector form based on the normalized frequencies of these TFBs. For these vectors, pair-wise distances were calculated employing Euclidean distance (as distance between vectors) using MATLAB 6.5.1 Statistics Tool Box.

Euclidean distance: $d(x, y) = |x - y| = [(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots\ldots\ldots + (x_n - y_n)^2]^{1/2}$

Pair-wise distances were calculated by using "pdist function (Y = pdist (X, 'Euclid')", where pair-wise distance matrix Y was computed for vectors in data matrix X by employing Euclidean distance. Using the resultant pair-wise distance matrix, phylogenetic analysis was performed with MEGA 2.1[32] using the Neighbor Joining Method.

D. Statistical Analysis:

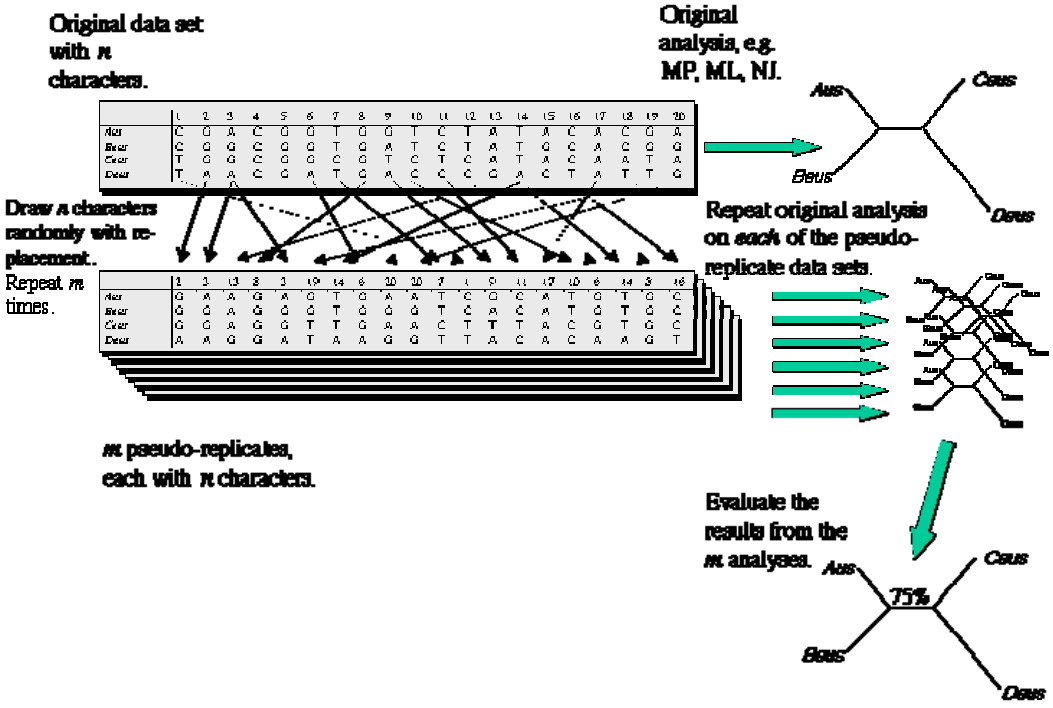

FIGURE 4: Non-Parametric Bootstrap Analysis

(Source:http://artedi.ebc.uu.se/course/bioinfo/Phylogeny/Phylogeny-Credibility/Phylogeny-Credibility.html)

Bootstrap analysis was done by PAUP 4.0[30] as well as MEGA 2.1[32] as a statistical procedure during the construction of phylogeny in this study. This bootstrap analysis was done when the phylogenetic tree was constructed from sequences of both upstream regulatory regions and proteins. In case of the TFB based phylogeny, bootstrap analysis could not be done as the input data for constructing phylogenetic tree was distance matrix rather than sequences. The bootstrap procedure (Fig. 4) starts by creating a new data set by randomly drawing (with replacement) the same number of characters as in the original data set. Thus, some of the original characters (sites) may be represented more than once while others may be missing completely from then new data set. This new data set is than analyzed in the same way as the original data; we get a tree. The procedure is repeated a number of times (100-1000) and each pseudo-replicate produce one (or several) trees as result. At the end, one has to evaluate all these trees in some way, usually by a majority-rule consensus tree.

E. Expected Results:

This study is partially discovery driven and partially hypothesis driven; hence the results of this study could either confirm results from earlier relevant research or may deviate. However, there would be some findings that are of our prime interest. If the phylogenetic analysis with protein sequences result in clustering of Anopheles and Drosophila Toll genes and if this cluster joined to the cluster of mammalian counter parts, while being distant from all the plant disease resistance genes; that would lend the support for single point of divergence between insects and mammals with respect to Toll genes and Toll like proteins. In that case addition of Anopheles Toll genes to the

phylogenetic analysis would not have resulted in significant difference and it could be suggesting that Anopheles Toll genes also have similar function of antibacterial immunity and ontogenesis. In case of a contrary result, it would be worth while to analyze more into the function of Anopheles Toll genes to see if they have only antibacterial immunity as their function.

Another important result would be the similarity between phylogeny based on proteins and phylogeny based on upstream regulatory regions. However, we could not expect well-resolved phylogenetic tree from upstream regulatory regions due to the poor alignment owing to the smaller size (<14 bp) of conserved regulatory elements. The evolutionarily distant nature of involved species (plants, insects and mammals) could be an additive factor for such poor resolution. Under such circumstances, emphasis would be laid on the perceived similarities between phylogeny based on proteins and TFB based phylogeny. Given the greater difference in the methodology as well as the input data, the results may not be highly similar. However, in the event of any similarity, that would not only validate evolutionary importance of upstream regulatory regions by suggesting the conserved transcriptional elements could be an extension of conserved function, but also would lend credibility for the novel supervised statistical approach to determine the evolutionary informativeness of TFBs in this study.

F. Alternate Plans

The Neighbor Joining method that was used in this study for all the phylogenetic analyses is mainly for comparative purposes as the TFB based phylogeny can be done only using distance method. In the event of poorly resolved phylogeny from protein

phylogeny from this method, we planned to do the phylogenetic analysis in Maximum Parsimony and Maximum Likelihood method, while latter being computationally intensive.

For the TFB based phylogeny, our approach is mainly based on the frequency of TFBs across the involved genes, while the information of their location and the distances between different TFBs in a sequence could not be exploited. In the event of complete dissimilarity between phylogeny based on proteins and TFB based phylogeny, we would be further planning to devise a more evolutionarily pertinent distance rather than Euclidean distance, which is commonly being used in many clustering algorithms. Such more evolutionarily pertinent distance between vectors should incorporate frequency of TFBs along with location and the distances between different TFBs in a sequence. However, it is expected that calculating that type of distance would be computationally more intensive as incorporating the information of location of TFBs demands very high number of combinations of the order of presence of the TFBs in the upstream regulatory sequences.

IV. RESULTS

A. Introduction

In comparison to earlier research in Toll evolution, this project resulted in some similar as well as dissimilar, but interesting outcomes that are capable of offering new insights; hence these outcomes are worth deep exploration. In this section, the results of this exploration are presented.

Protein Phylogeny:

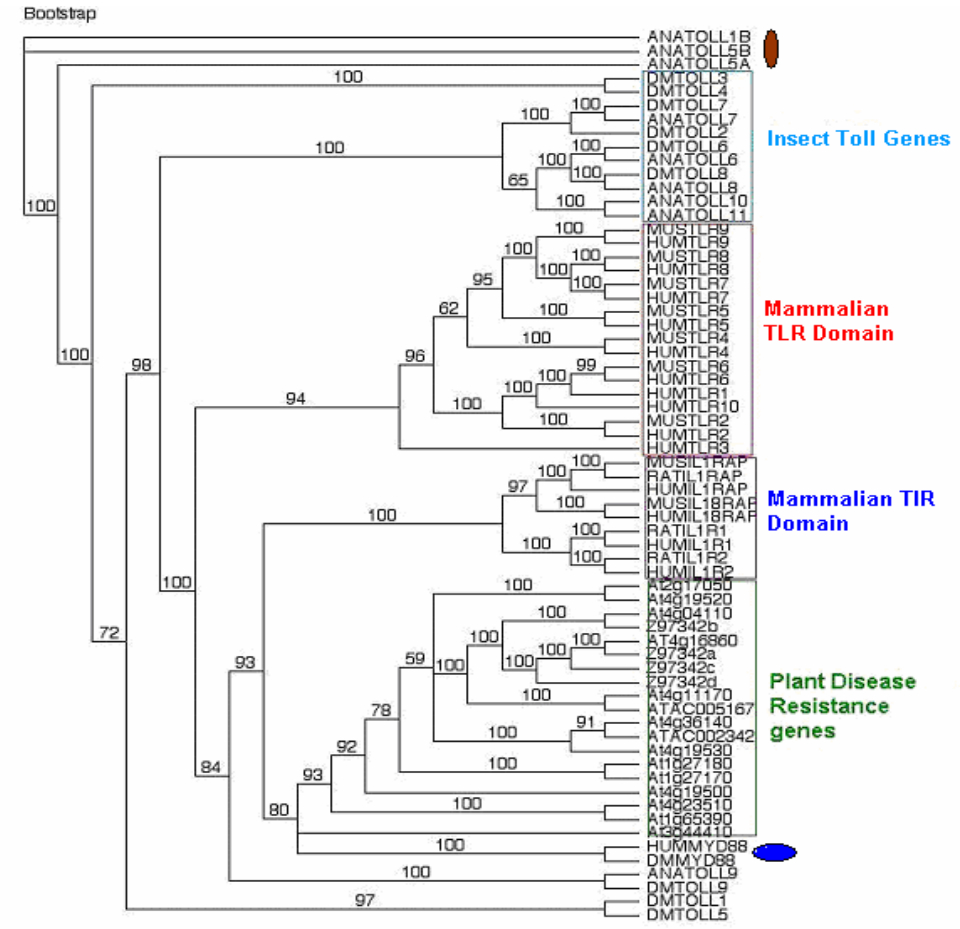

FIGURE 5: Phylogeny based on protein sequences of Toll-like genes. The phylogenetic analysis has been performed by PAUP 4.0 using Neighbor Joining method with p-distance. The bootstrap consensus tree was obtained from 1000 bootstrap replicates. The numbers on the branch nodes are the bootstrap values.

The bootstrap consensus tree (Fig 5) constructed from the protein sequences of Toll like genes suggested some important inferences. As expected, the tree resolved into four major clusters namely, (1) insect Toll genes (Anopheles and Drosophila), (2) mammalian TLR genes, (3) mammalian TIR Domain, and (4) plant disease resistance genes. However, all insect Toll genes were not completely clustered together and both mammalian and Drosophila MyD88 genes clustered along with plant disease resistance genes. Unlike the earlier results of Luo and Zheng[9], this phylogeny contradicts the hypothesis of single point of divergence between mammals and insects with respect to Toll-related proteins. As the tree shows, all insect Toll genes were clustered distantly to that of both their mammal and plant counter parts. This difference might have been induced due to the addition of Anopheles Toll genes to the analysis in this study. The fact that the ANATOLL1B, 5B, and 5A genes clustered distantly to all other genes is an interesting deviation. The distant appearance of ANATOLL1B, 5B, and 5A to the rest of insect Toll genes might be due to the reduplication of single type 1 and 5 genes in Anopheles. The mammalian TIR domain is closer to the plant disease resistance genes than the mammalian TLR domain suggesting significant homology between the plant genes and the mammalian TIR domain at protein level. This result is consistent with the fact that the protein motif shared by Toll and IL-1R, or TIR, was evident in plant disease resistance proteins as well as in animal proteins as reported by Meyers *et al*[11]. The clustering of human MyD88 and Drosophila MyD88 with plant disease resistance genes is another contradictory result observed as both these genes clustered with mammalian TIR domain in an earlier phylogenetic analysis using protein sequences[4].

Phylogeny based on upstream regulatory regions (-3000 to +10):

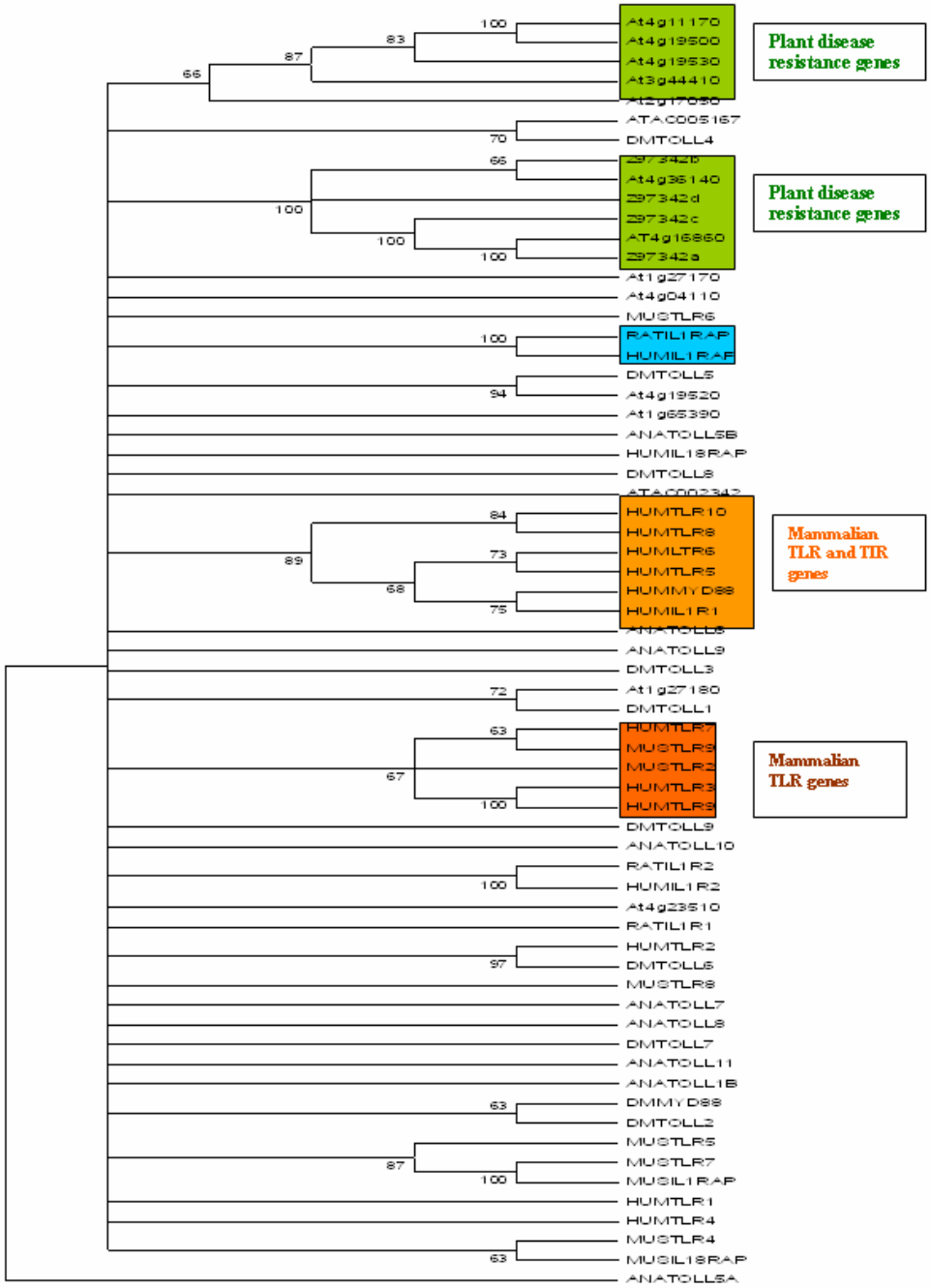

FIGURE 6: Phylogeny based on upstream regulatory sequences (-3000 to +10) of Toll-like genes. The phylogenetic analysis has been performed by MEGA 2.1 using Neighbor Joining method with p-distance. The bootstrap consensus tree was obtained from 1000 bootstrap replicates. The numbers on the branch nodes are the bootstrap values.

The phylogeny (Fig 6) resulting from the upstream region sequences (-3000 to +10) was not well resolved even after noise reduction This is most likely due to poor alignment, which might be the result of the predominantly short (<14 base pairs) transcription factor binding site (TFBs) sequences being swamped by the unaligned background of ~3 kb upstream sequences. The percentages of sequences present as TFBs (as predicted by EZ-Retrieve[29]) in the upstream regions (-3000 to +10) of all genes involved in the study are shown in Table 1 and range from 15 – 28% indicating long unaligned stretches. In spite of these unaligned sequences, the resultant tree (Fig 6) shows some significant clustering with higher bootstrap values for plant disease resistance genes, the TLR domain and rat and human IL1RAP genes, which suggests strong conservation of upstream regulatory regions for these genes. A broader assessment of the phylogenetic tree (Fig 6) reveals that plant disease resistance genes dominate the upper one-third of the tree, while lower two-thirds comprises of mammalian Toll-like and insect Toll genes. This suggests that mammalian and insect Toll genes have some homology at the level of transcriptional regulatory sequences, while plant disease resistance genes appear distant to both these groups. This suggests some extension of the homology of the antibacterial immunity into upstream sequences as the Toll genes have significant homology at protein level between mammals and insects.

Selection of evolutionarily informative TFBS: As discussed in Section III (Methods), we followed a supervised statistical approach[34] to select the evolutionarily informative TFBs across the four groups of Toll-like genes in the study. The selection of evolutionarily informativeness of TFBs was based on the BSS/WSS ratio of TFBs among

the four groups, where the BSS and WSS were determined by a supervised statistical approach[34]. For this study we used the TFBs with a BSS/WSS ratio above 0.2 as highly evolutionarily informative with respect to Toll like genes. The list of highly evolutionarily informative TFBs and their corresponding BSS/WSS values were shown in Table 2 as well as in Figure 6 below.

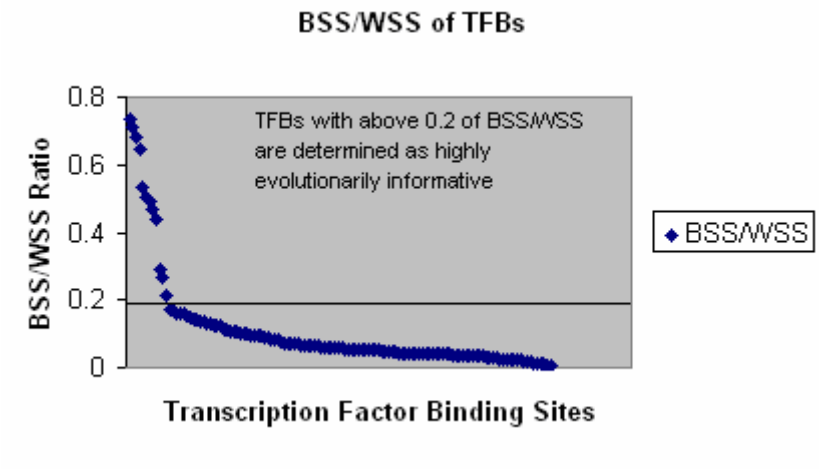

FIGURE 7: BSS/WSS ratio of TFBs in the upstream regulatory sequences (-3000 to +10) of Toll-like genes.

TABLE 2: List of highly evolutionarily informative TFBs

Transcription factor binding site		BSS/WSS
ADR1	(alcohol dehydrogenase gene regulator 1)	0.733019922
HSF	(heat shock transcription factor (Drosophila))	0.708738594
STRE	(stress-response element)	0.683149481
StuAp	(Aspergillus Stunted protein)	0.64874741
v-Myb	(AMV, avian myeloblastosis virus)	0.535117365
SRY	(sex-determining region Y gene product)	0.50357773
MZF1	(the myeloid zinc finger protein MZF1)	0.491608776
cap	(cap signal for transcription initiation)	0.468023143
E2F	(E2F transcription factor)	0.440753429
GATA-1	(GATA-binding factor 1)	0.290438638
MyoD	(myoblast determining factor)	0.264242705
Sox-5	(SRY-related HMG-box gene 5)	0.212790012

* The cut-off for determining highly evolutionarily informative TFBs is 0.2 (BSS/WSS)

TFB based Phylogeny:

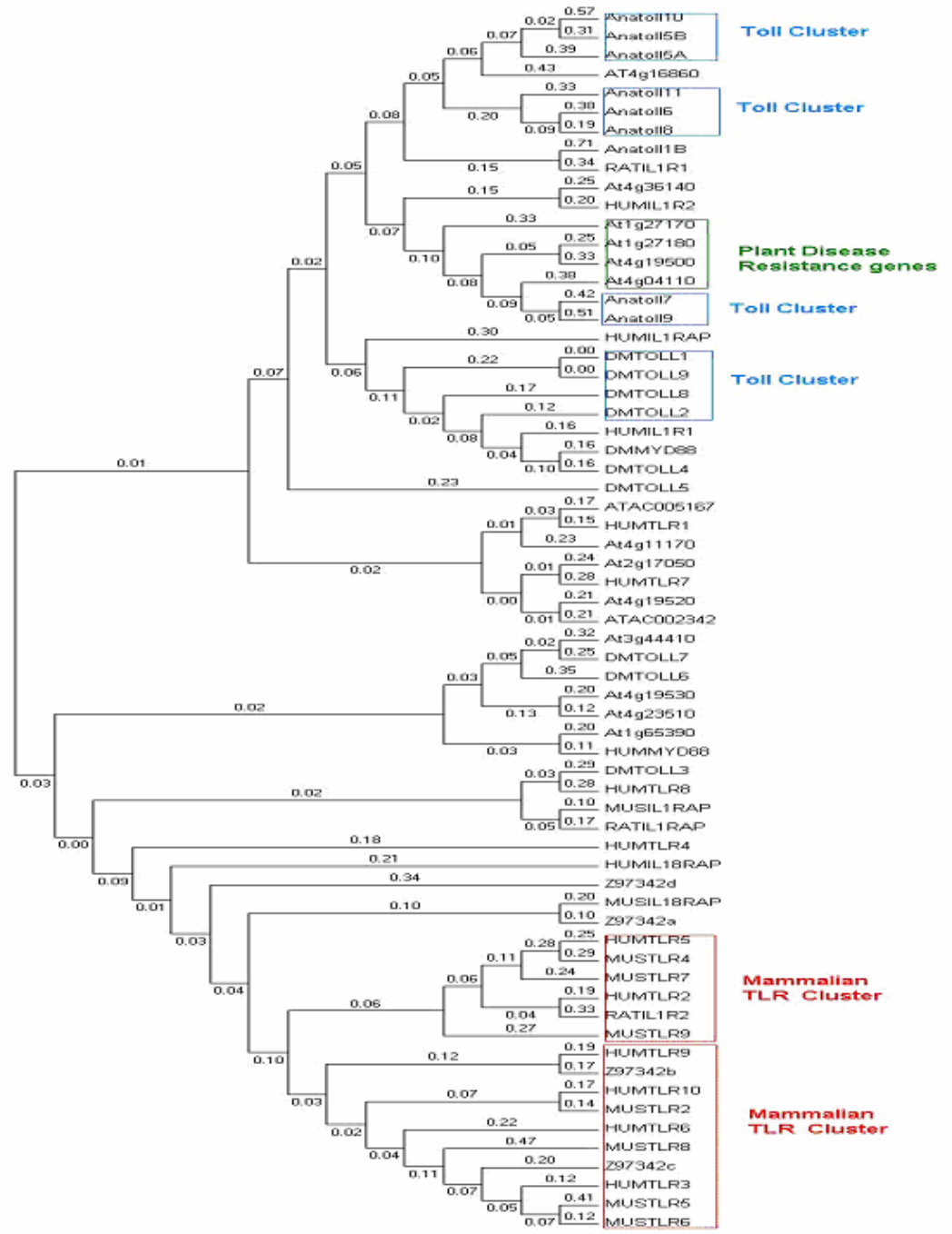

FIGURE 8: Phylogeny based on "evolutionary informativeness" of TFBs in the upstream regulatory sequences (-3000 to +10) of Toll-like genes. The phylogenetic analysis has been performed by MEGA 2.1 using Neighbor Joining method with pair-wise distance (Euclidean) matrix, which is derived from the frequencies of highly evolutionarily informative TFBs in the upstream regulatory sequences (-3000 to +10) of Toll like genes. The numbers on the branch nodes are corresponding to branch lengths.

The phylogenetic tree based on the TFBs (Fig 8) resolved into two major clusters, one of which is dominated by insect Toll genes and the other by mammalian TLR domain. It shows only a weak homology between mammalian TLRs and insect Toll genes with respect to the presence of TFBs, and this result is similar to the distance between these two groups as revealed at the protein level. From a broader look at the tree (Fig 8), it can be inferred that the insect Toll genes are much closer to the plant genes than to mammalian TLRs at the level of transcriptional regulatory sequences. Unlike in protein phylogeny, the TFBs of Drosophila Toll genes are comparatively distant from those of Anopheles. The upstream regions of Drosophila Toll genes may have two different sets of TFBs, one controlling the expression for antibacterial immunity and the second for ontogenesis. So far, no such ontogenic role for Toll genes has been demonstrated in Anopheles. It is possible that the TFBs of Anopheles Toll genes are primarily responsible for controlling antibacterial immunity, which is a role closer in function to controlling plant disease resistance and this may explain the closer appearance of Anopheles to plants in this analysis.

Plant disease resistance genes and mammalian TIR genes appeared in both clusters. The clustering of At4g36140 (a plant disease resistance gene) with HUMIL1R2 was concurrent with the homology between amino acid motifs of plant disease resistance genes and the cytoplasmic region of mammalian interleukin-1 receptor[35]. In this case, such a clustering may suggest the extension of the amino acid conservation in the TIR domain into the upstream regulatory sequences of these two genes. Compared to the phylogeny from upstream regulatory regions (Fig 6), where plant disease resistance genes

showed significant clustering, phylogeny based purely on TFBs showed no such significant conservation except for four plant disease resistance genes. This suggests that there maybe significant homology among plant disease resistance genes with respect to upstream regulatory regions rather than the presence of common TFBs.

Overall, the mammalian TLRs are clustered distantly from all plant and insect genes with respect to the TFBs of evolutionary importance. This may be due to sets of TFB sequence segments that have diverged from those of plants and insects to perform functionally different transcriptional controlling events in mammals, though the downstream innate immunity function is similar.

B. Important Highlights

Important highlights of this project are the major similarities and dissimilarities in comparison with earlier research in the evolution of Toll like genes. Phylogeny based on amino acid sequences of Toll-like genes (Fig 5) results in four distinctly resolved clusters namely, (1) insect Toll genes (Anopheles and Drosophila), (2) mammalian TLR genes, (3) mammalian TIR Domain, and (4) plant disease resistance genes. While this being an expected result, the distant clustering of insect Toll genes from their mammalian and plant counterparts suggested a major contradiction to earlier hypothesis of single point of divergence[9] between insects and mammals. This observation is again corroborated by the closer homology between plant disease resistance genes and mammalian TIR domain. The closer homology between Anopheles and Drosophila Toll genes[19] could be the major reason for this distant clustering of insect Toll genes, while the functional conservation[11]

between plant disease resistance genes and mammalian TIR domain cannot be discounted.

The phylogeny (Fig 6) resulting from the upstream regulatory sequences (-3000 to +10) of Toll like genes is highlighted by insignificant resolution. Poorly resolved phylogeny can be attributed to the significantly lower presence (15 – 28%) of small (<14bp) but conserved regulatory elements, which leads to poor alignment of the ~3kb upstream sequences. However, the broader assessment of this phylogeny suggests some homology between mammalian and insect Toll - like genes as well as among plant disease resistant genes at the level of transcriptional regulatory regions.

The major highlight of TFB based phylogenetic approach is that it showed only a weaker homology between mammalian TLRs and insect Toll genes, while showing similarity of distance between these two groups at protein level. However, in totality, the mammalian TLRs are clustered distantly from all plant and insect genes with respect to the TFBs of evolutionary importance. Comparatively farther distance between Drosophila and Anopheles based merely on the presence of TFBs is an interesting revelation as it suggests possible conservation of only the transcriptional apparatus that regulates antibacterial immunity in the Toll genes of these two organisms.

C. Specific Findings

In case of protein phylogeny, the distant appearance of ANATOLL1B, 5B, and 5A to the rest of insect Toll genes is an interesting deviation, which might be due to the

reduplication of single type 1 and 5 genes in Anopheles. The clustering of human MyD88 and Drosophila MyD88 with plant disease resistance genes is another contradictory result observed as both these genes clustered with mammalian TIR domain in an earlier phylogenetic analysis using protein sequences[4]. This lends support to the functional conservation[11] between plant disease resistance genes and mammalian TIR domain, but this specific finding had been revealed by the addition of Anopheles Toll genes in our study to the earlier phylogenetic analyses of Toll like genes[4].

In case of evolutionarily informative TFB based phylogeny, the clustering of At4g36140 (a plant disease resistance gene) with HUMIL1R2 was concurrent with the homology between amino acid motifs of plant disease resistance genes and the cytoplasmic region of mammalian interleukin-1 receptor[35]. In this case, such a clustering may suggest the extension of the amino acid conservation in the TIR domain into the upstream regulatory sequences of these two genes.

D. Summary

In an attempt to summarize the results from three different phylogenetic approaches to determine evolution of Toll like genes, it can be reported that the protein phylogeny is consistent with most of the earlier studies of Toll evolution. However, the resulted deviations are most likely due to the addition of Anopheles Toll genes.

Though there is accumulated evidence for conservation of non-coding regulatory regions, the phylogeny based on the upstream regulatory sequences (-3000 to +10)

resulted in a poorly resolved tree. In spite of this poor resolution, it suggests some homology between mammalian and insect Toll-like genes as well as among plant disease resistant genes at the level of transcriptional regulatory regions. The considerable homology between mammalian and insect Toll-like genes suggests some extension of the function of antibacterial immunity as the Toll genes have significant homology at protein level between mammals and insects.

The phylogeny based on evolutionarily informative TFBs resulted in some significant observations regarding conservation of transcriptional regulation between Drosophila and Anopheles Toll genes as well as across all four groups involved in the study. Though this approach may not give a complete picture of Toll evolution, the broader similarity of this tree with the phylogeny from protein sequences advocates the importance of the contribution of regulatory regions in studying evolution.

V. CONCLUSION

A. Overview of significant findings

This study was undertaken to analyze the phylogeny of Toll and Toll-like genes based on the evolutionary informativeness of TFBs in upstream regulatory sequences in comparison to protein sequences. Toll and Toll-like genes are well conserved across multiple organisms at the protein level. In addition, a significant amount of evidence has been accumulating in regards to the conservation of upstream and intergenic sequences in evolutionarily related organisms and orthologous genes. Our rationale for this analysis was to analyze the evolutionary importance of upstream regulatory sequences of Toll genes, particularly TFBs.

Except for the single point of divergence between insects and mammals, protein phylogeny of Toll genes (Figure 5) was consistent with results published earlier and this deviation is most likely due to the addition of Anopheles genes. On the other hand, the phylogeny based on upstream regulatory sequences (-3000 to +10) showed a broader distinction between the plants and the rest (Figure 6) , though the tree was not well resolved probably due to poor alignment of these sequences. This poor alignment could be attributed to the predominantly short (<14 base pairs) TFB sequences being swamped by the unaligned background of ~3 kb upstream sequences. Finally, a phylogeny was derived using pair-wise distances employing the frequencies of evolutionarily informative TFBs. Broadly, this TFB based phylogeny (Figure 8) showed results similar to the protein phylogeny. It suggested a closer relationship between plants and Anopheles

than that between plants and Drosophila and this may be due to the fact that the upstream regions of Drosophila Toll genes contain two different sets of TFBs, one controlling the expression for antibacterial immunity and the second for ontogenesis. The mammalian TLRs are clustered distantly from all plant and insect genes probably due to the evolutionary divergence of their TFB sequence segments with respect to the transcriptional controlling function. Though this approach may not give a complete picture of Toll evolution, the broader similarity of this tree with the phylogeny from protein sequences advocates the importance of the contribution of regulatory regions in studying evolution, particularly with instances of highly conserved gene families.

B. Consideration of findings in context of current knowledge

The closer association of mammalian TLRs with insect Toll genes, while mammalian TIR domain is closer to plant disease resistance genes is a significant observation given that much of the related research has supported the single point of divergence between mammals and insects. As this could be due to the result of the addition of Anopheles Toll genes to the phylogenetic analysis, it could be helpful in developing a therapeutic solution for malaria that is caused by Anopheles. However, it requires a deeper understanding of the functional similarities among these organisms with respect to antibacterial immunity.

The suggested extension of amino acid conservation into upstream regulatory regions adds to the evolutionary importance of non coding regulatory regions. The approach of using evolutionarily informative TFBs for phylogenetic analysis offers a new

computational direction to extract vital information in regulatory regions. This approach provides an excellent computational framework since more evidence is being accumulating for the conservation of these non coding regulatory regions and not many such computational solutions have been established given the smaller size of these conserved regulatory elements.

C. Theoretical implications of the findings

The significant homology between mammalian TLRs and insect Toll genes at protein level coupled with similar observation from the approach of employing evolutionarily informative TFBs suggests functional conservation. However, it remains unclear whether antibacterial immunity is the only function that is conserved across these organisms as there was no evidence regards to conservation of ontogenic function of Drosophila Toll genes in mammals. In this context, the weaker but significant homology between mammalian and insect Toll genes in the upstream regulatory regions suggests that only the transcriptional apparatus that regulates antibacterial immunity might have been conserved.

The homology between mammalian TIR domain and plant disease resistance genes assumes greater importance as there are some recent unpublished reports suggesting that rice genes have the TIR signature in their amino acid sequences. A deeper understanding of such homology might be helpful in mitigating bacterial diseases of rice as well as other monocot staple food crops.

VI. DISCUSSION

A. Limitations of the study

While this approach is based on the frequency of the presence of evolutionarily informative TFBs, it does not take into account the location of those TFBs in the upstream region. We view this as a limitation and by resolving this limitation, it may be possible to arrive at a result that offers better insight into the similarities and differences between phylogenies based on regulatory regions and protein sequences. In addition, this approach may require a more evolutionarily pertinent distance measure(s) that can be applied between the gene vectors. Such a distance measure should incorporate frequency of TFBs along with location and the distances between different TFBs in a sequence, which would be computationally more intensive.

This study employed only the neighbor joining method of phylogeny for comparative analysis as evolutionarily informative TFB based phylogeny could be performed by employing a distance method only. The other computationally intensive methods such as maximum likelihood and maximum parsimony should be explored, but with more developed computational approaches to capture the essential evolutionary information from non coding regulatory regions.

B. Recommendations for further research

The logical extension of this project would be to develop a evolutionarily more pertinent distance that can be applied between gene vectors, which should include the location of TFBs as well as the distance that separates them in the upstream regulatory

sequence. With the advent of such a distance it could be expected that the similarity of conservation between protein sequences and upstream regulatory sequences can be precisely captured. In addition, more advanced TFB predicting algorithms would add to the accuracy of analysis as not all TFBs can be predicted by EZ-Retrieve.

It is worthwhile to include additional insect species to explore functional separation between antibacterial immunity and ontogenesis as in Drosophila since so far only the antibacterial immunity of Drosophila Toll genes seems to be conserved. Exploration of silencing elements of the Toll genes across various organisms would be another area of potential research.

Development and use of a better noise (evolutionarily irrelevant, background information) reduction algorithm other than RASA would be beneficial since computationally intensive phylogenetic methods can applied for analysis and this may result in an accurate and complete picture of Toll evolution from the perspective of regulatory regions.

REFERENCES

1. Janeway CAJ (1989): Evolution and revolution in immunology. ***Cold Spring Harbor Symp. Quant.Biol.*** 54:1–13
2. Fearon DT, Locksley RM (1996): The instructive role of innate immunity in the acquired immune response. ***Science*** 272:50-54
3. Hoffman JA, Kafatos FC, Janeway CA, Ezekowitz RAB (1999): Phylogenetic perspectives in innate immunity. ***Science*** 284: 1313-1318
4. Kimbrell DA, Beutler B (2001): The evolution and genetics of innate immunity. ***Nature Rev. Genetics*** 2: 256-267
5. Belvin MP, Anderson KV (1996): A conserved signaling pathway: The *Drosophila* toll-dorsal pathway. ***Annu Rev Cell Dev Biol*** 12:393–416
6. Anderson KV, Jurgens G, Nusslein-Volhard C (1985): Establishment of dorsal-ventral polarity in the *Drosophila* embryo: genetic studies on the role of the *Toll* gene product. ***Cell*** 42: 779-789
7. Gay NJ, Keith FJ (1991): *Drosophila* Toll and IL-1 receptor. ***Nature*** 351: 355–356
8. Kaisho T, Akira S (2001): Toll-like receptors and their signaling mechanism in innate immunity. ***Acta Odontol Scand*** 59:124 –130
9. Luo C, Zheng L (2000): Independent evolution of *Toll* and related genes in insects and mammals. ***Immunogenetics*** 51:92–98
10. Hughes AL (1998): Protein phylogenies provide evidence of a radical discontinuity between arthropod and vertebrate immune systems. ***Immunogenetics*** 47:283–296

11. Meyers BC, Dickerman AW, Michelmore RW, Pecherer RM, Sivaramakrishnan S, Sobral BWS, Young ND (1999): Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. ***Plant J.*** 20:317-332
12. Medzitov R, Janeway CAJ (2000): Fly immunity:great expectations. ***Genome Biology*** 1:reviews 106.1-106.4
13. Kopp EB, Medzhitov R (1999): The Tollreceptor family and control of innate immunity. ***Curr. Opin. Immunol***. 11: 13–18
14. Hoffmann JA, Reichart JM, Hetru C (1996): Innate immunity in higher insects. ***Curr. Opin. Immunol***. 8: 8-13
15. Hoffmann J, Reichhart J (1997): *Drosophila* immunity. ***Trends Cell Biol.*** 7: 309–316
16. Khush RS, Lemaitre B (2000): Genes that fight infection: what the *Drosophila* genome says about animal immunity. ***Trends Genet.*** 16: 442–449
17. Engstrom Y (1999): Induction and regulation of antimicrobial peptides in *Drosophila*. ***Dev. Comp.Immunol.*** 23: 345–358
18. Richman A, Dimopoulos G, Seeley D, Kafatos FC (1997): Plasmodium̧ activates the innate immune response of Anopheles gambiae mosquitoes. ***EMBO J*** 16:6114-6119
19. Christophides GK *et al.* (2002) Immunity-related genes and gene families in *Anopheles gambiae.* ***Science*** 298: 159-165

20. Luna C, Wang X, Huang Y, Zhang J, Zheng L (2002): Characterization of four Toll related genes during development and immune responses in Anopheles gambiae. ***Insect Biochem Mol Biol.*** 32:1171-1179

21. Hammonnd-Kosack K, Jones JDG (1997): Plant disease resistance genes. ***Annu Rev Plant Physiol Plant Mol Biol*** 48:575-607

22. Zuckerkandl E, Pauling L (1962): Molecular disease, evolution, and genic heterogeneity. In Kasha/Pullman, ***Horizons in Biochemistry,*** 189-225

23. Hardison RC, Oeltjen, Miller W (1997): Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. ***Genome Res*** 7:959- 966

24. Thomas JW, Touchman JW, Blakesley RW, Bouffard GG, Beckstrom-Sternberg SM, Margulies EH, Blanchette M, Siepel AC, Thomas PJ, Mcdowell JC *et al.* (2003)*:* Comparative analyses of multi-species sequences from targeted genomic regions. ***Nature*** 424**:**788-793

25. Dermitzakis ET, Reymond A, Scamuffa N, Ucla C, Kirkness E, Rossier C, Antonarakis SE (2003): Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs). ***Science Express Online*** [http://www.sciencexpress.org/] 10.1126/science.1087047.

26. Ruvkun G, Hobert O (1998): The taxonomy of developmental control in *Caenorhabditis elegans*. ***Science*** 282:2033–2041

27. GENBANK: **NIH genetic sequence database** [http://www.ncbi.nlm.nih.gov]

28. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) : The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. ***Nucl Acids Res.*** 25**:**4876–4882

29. Zhang H, Ramanathan Y, Soteropoulos P, Recce M, Tolias PP (2002): EZ-Retrieve: A web server for batch retrieval of coordinate-specified human DNA sequences and underscoring putative transcription factor-binding sites. ***Nucl Acids Res.*** 30**:** e121

30. Swofford DL: PAUP*: phylogenetic analysis using parsimony (and other methods), version 4. Sinauer, Sunderland, Mass. 1998.

31. Lyons-Weiler J, Hoelzer GA, Tausch RJ (1996): Relative Apparent Synapomorphy Analysis (RASA) I: the statistical measurement of phylogenetic signal. ***Mol Biol Evol.*** 13**:** 749-757

32. Kumar S, Tamura K, Jakobsen IB, Nei M: MEGA2: Molecular Evolutionary Genetics Analysis software**.** Arizona State University, Tempe, Arizona, USA. 2001.

33. Saitou N, Nei M (1987): The neighbor-joining method: a new method for reconstructing phylogenetic trees. ***Mol Biol Evol.*** 4:406-25

34. Dudoit S, Fridlyand J, Speed T (2002): Comparison of discrimination methods for the classification of tumors using gene expression data. ***J Amer Stat Assoc*** 97:77-87

35. Shamrai SN (2003): Plant resistance genes: molecular and genetic organization, function and evolution. ***Zh Obshch Biol.*** 64(3):195-214.