Evolutionary selection on toll-like receptor genes

THESIS SUBMITTED FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY (SCIENCE)

DEPARTMENT OF LIFE SCIENCE AND BIOTECHNOLOGY
JADAVPUR UNIVERSITY
2024

MANISHA GHOSH
Index No. 26/22/Life Sc./27
Registration no. SLSBT1102622

Division of Bioinformatics
ICMR-National Institute for Research in Bacterial Infections
(formerly, ICMR-National Institute of Cholera and Enteric Diseases)
Kolkata, India

आई. सी. एम. आर. - राष्ट्रीय जीवाणु संक्रमण अनुसंधान संस्थान
ICMR - NATIONAL INSTITUTE FOR RESEARCH IN BACTERIAL INFECTIONS
Formerly, ICMR-National Institute of Cholera and Enteric Diseases (ICMR-NICED)
स्वास्थ्य अनुसंधान विभाग, स्वास्थ्य एवं परिवार कल्याण मंत्रालय, भारत सरकार
Department of Health Research, Ministry of Health & Family Welfare, Govt. of India
WHO COLLABORATING CENTRE FOR RESEARCH AND TRAINING ON DIARRHOEAL DISEASES

CERTIFICATE FROM THE SUPERVISOR

This is to certify that the thesis entitled **"Evolutionary selection on toll-like receptor genes"** submitted by Smt. **Manisha Ghosh** who got her name registered on **11th February, 2022** for the award of Ph.D. (Science) Degree of Jadavpur University, is absolutely based upon her own work under the supervision of **Dr. Surajit Basak**, Scientist-D, Division of Bioinformatics, **ICMR - National Institute for Research in Bacterial Infections** (formerly, ICMR-National Institute of Cholera and Enteric Diseases)**, Kolkata, India** and that neither this thesis nor any part of it has been submitted for either any degree/diploma or any other academic award anywhere before.

Surajit Basak 11/07/2024.

(Signature of Supervisor(s) with date and official seal)

डॉ. सुरजीत बसाक, पीएच.डी/Dr. Surajit Basak, Ph.D
वैज्ञानिक डी/ Scientist D
आईसीएमआर-राष्ट्रीय जीवाणु संक्रमण अनुसंधान संस्थान
ICMR-National Institute For Research in Bacterial Infections (NIRBI)
पी-33, सीआईटी रोड, स्कीम-एक्सएम,
बेलियाघाटा,कोलकाता-७०००१०/
P-33, CIT Road, Scheme-XM, Beliaghata / Kolkata-700070

Preface

This Ph.D. thesis was prepared at the **Division of Bioinformatics, ICMR - National Institute for Research in Bacterial Infections** (formerly, ICMR-National Institute of Cholera and Enteric Diseases)**, Kolkata, India** to fulfil the requirements for obtaining a Ph.D. degree. The thesis is titled **"Evolutionary selection on toll-like receptor genes"** and the research was conducted at **ICMR - National Institute for Research in Bacterial Infections** (formerly, ICMR-National Institute of Cholera and Enteric Diseases)**, Kolkata, India.** The thesis was completed under the guidance of **Dr. Surajit Basak**, Scientist-D, Division of Bioinformatics, **ICMR - National Institute for Research in Bacterial Infections** (formerly, ICMR-National Institute of Cholera and Enteric Diseases)**, Kolkata, India**.

The thesis has been solely composed by the candidate and has not been submitted for any other degree, except where specifically acknowledged.

The work described in this thesis was conducted from July 2019 to June 24.

Date: 11-07-2024

Manisha Ghosh

Place: Kolkata

MANISHA GHOSH

Acknowledgements

Completing this PhD thesis has been a significant milestone in my academic journey, and it would not have been possible without the support and guidance of many individuals and institutions. I would like to take this opportunity to express my heartfelt gratitude to those who have contributed to this journey.

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Surajit Basak, Scientist-D, ICMR - National Institute for Research in Bacterial Infections (formerly, ICMR-National Institute of Cholera and Enteric Diseases), Kolkata for his unwavering support, insightful guidance, and invaluable encouragement have been instrumental in successful completion of this research. His expertise and commitment to excellence have profoundly shaped my academic development.

I am also deeply grateful to the members of my research advisory committee Professor Tapash Chandra Ghosh, Emeritus Professor, Department of Zoology, Ramakrishna Mission Vivekananda Centenary College, Rahara and Professor Parimal Karmakar, Head, Department of Life Science and Biotechnology, Jadavpur University for their constructive feedback, encouragement, and time. Their diverse perspectives and expertise have enriched this work.

A special thanks to my colleagues and lab members at the institute for creating a stimulating and supportive research environment. Their companionship and intellectual exchange have been instrumental in overcoming the challenges faced during this journey. My heartfelt thanks go to my family for their endless love, patience, and support. Their support has been crucial in enabling me to focus on my studies and research.

List of Figures and Tables

Chapter - VI: Structural and functional objectivity of TLR evolution

List of Abbreviation

AAU, Amino acid usage;
AF, AlphaFold;
AT content, adenosine thiamine content;
BLAST, Basic Local Alignment Search Tool;
CATH, Class, Architecture, Topography and Homology;
CD14, Cluster of differentiation 14;
COA, Correspondence analysis;
CpG DNA, unmethylated cytosine-guanine dinucleotide sequences;
CpG, cytosine phosphate guanosine;
DAMPs, damage-associated molecular patterns;
DL, deep learning;
dN, number of non-synonymous substitutions per non-synonymous site;
DNA, Deoxyribonucleic acid;
dS, number of synonymous substitutions per synonymous site;
dsRNA, double-stranded RNA;
ECD, N-terminal horseshoe-like extracellular domain, or extracellular domain;
EGF-like domains, Epidermal growth factor (EGF)-like domains;
ENC, effective number of codons;
FEL, fixed-effect likelihood;
GC, Guanine-cytosine content;
GC3, Guanine and cytosine content at the third codon position;
GO terms, Gene Ontology terms
HEK293 cells, Human Embryonic Kidney cells;
HSP60 and HSP70, heat shock proteins;
hTLR3, human Toll-like receptor3
IFN1, type-I interferons;
IKK, inhibitor of nuclear factor-κB (IκB) kinase;
IL, interleukin;
IRAK, interleukin-1-receptor-associated kinase;
IRF, interferon-regulated factor;
Ka, nonsynonymous substitution;
Ks, synonymous substitution;
LP, lipoproteins;
LPS, lipopolysaccharide;
LRRCT, leucine rich repeat C-terminal;
LRRNT, leucine rich repeats N-terminal;
LRRs, leucine rich repeats;
MAL, MyD88 adaptor-like protein;
MAMPs, microbe associated molecular patterns;
MD-2, lymphocyte antigen 96;
MHC, major histocompatibility complex;
ML, maximum likelihood;

MP, maximum parsimony;
MSAs, multiple sequence alignments;
MyD88, Myeloid differentiation factor 88;
NCBI, National Center for Biotechnology Information;
NF-*κ*B, Nuclear factor *κ*B;
NLRs, NOD-like receptors;
NOD, nucleotide-binding oligomerization domain;
NPm, TLRs from mammal other than primates;
PAMPs, pathogen-associated molecular patterns;
PDB, Protein Data Bank;
PG, peptidoglycan;
Pm, TLRs from primates;
PRR, Pattern recognition receptor;
PRRs, pattern recognition receptors;
PSI-BLAST, Position-Specific Iterative Basic Local Alignment Search Tool;
RAAU, relative amino acid usage;
RIG-I, retionic acid-inducible gene I;
RIPK1, receptor-interacting serine/threonine kinase 1;
RLRs, RIG-I like receptors;
RNA, Ribonucleic acid
rRNA, Ribosomal ribonucleic acid;
SCOP, Structural Classification of Proteins
SLAC, single likelihood ancestor counting;
SNPs, single nucleotide polymorphisms;
ssRNA, single stranded viral RNA;
TAK1/TGF-β-activated kinase (TAB) complex;
TBK1, TANK-binding kinase 1;
TICAM1, TIR domain containing adaptor molecule 1;
TICAM2, TIR domain containing adaptor molecule 2;
TIR, Toll/interleukin-1 receptor;
TIRAP, TIR-associated protein;
TLR, Toll-like receptor;
TLRs, Toll-like receptors;
TNF, tumour necrosis factor;
TRAF, TNF-receptor-associated factor;
TRAF6, tumour necrosis factor (TNF) receptor-associated factor 6;
TRAM, TRIF-related adaptor protein;
TRIF, TIR domain-containing adaptor-inducing IFNβ;
TRIF, TIR domain-containing adaptor-inducing IFNβ;
tRNA, Transfer ribonucleic acid;
UPGMA, Unweighted Pair Group Method Using Arithmetic Average;
3D structure, three-dimensional structure;

Table of Contents

Chapter VI

Structural and functional objectivity of TLR evolution

Chapter VII

Conclusion

References

Publication and Conferences

Publication related to thesis work

Abstract

Infectious diseases have posed the greatest threat to survival and wellbeing during human evolution. Natural selection is thus expected to exert a major influence on host defence genes, specifically on the genes involve in innate immunity, whose products intervene direct interactions between the host and the pathogen. Toll-like receptors (TLRs) are well-known for their roles in innate immunity, where they recognise pathogens and initiate a signalling response. These receptors can recognize a different types of pathogen-associated molecular patterns (PAMPs) as their ligands and are implicated in immunological response, signalling process development, and cell adhesion. Mammalian TLRs recognise molecular signatures linked with infections and trigger an innate immune response. This study emphasised the significance of evolutionary selection on the diverse mutation of TLR genes from mammals.

In my study I have noted difference in amino acid usage between primate and non-primate mammalian TLR genes. The GC content of TLR genes and the hydrophobicity of encoded proteins are the important factors in determining the distinct pattern of amino acid usage. The GC-content was found to be consistent evolutionary force throughout the course of evolution of TLR genes between primate and non-primate mammalian species. I have observed TLR genes are generally under purifying selection, however several positively selected sites have been found in the ligand binding domain. My study also presented that the amino acid usage pattern of TLRs are influenced by their subcellular location. Different branching patterns of primate and non-primate mammalian TLRs have also been demonstrated through phylogenetic tree. These findings clearly indicate that natural selection influenced the evolution of primate and non-primate mammalian TLR genes.

Following these findings, an amino acid usage analysis of all mammalian TLRs was done to investigate the evolutionary diversity of mammalian TLRs and differences in immunological response. A detailed examination of mammalian TLRs found that TLR9 evolved in a completely different way compared to other mammalian TLRs. Different sequence-based features, including amino acid usage, hydrophobicity, GC content, and evolutionary parameters, have been identified to impact the divergence of TLR9 from other TLRs. Reconstructing ancestral sequences is an important component of molecular evolution of TLR because it allows to follow changes across genes. Ancestral sequence reconstruction study also demonstrated that TLR genes evolved gradually across numerous ancestral lineages, resulting in the distinct TLR9 pattern. It exhibits evolutionary divergence, with the gradual accumulation of mutations resulting in the specific pattern of TLR9.

The evolutionary genetics approach to determine the magnitude of natural selection operating on TLR genes and the progressive changes that lead to divergence will help us better understand the mechanism of host defence mediated by TLRs.

Chapter - I

Introduction

The immune system comprises of two components such as innate immunity and acquired immunity. Both of these components are responsible for host defence against invading microbial pathogens by triggering immune responses to remove the invading pathogen that is identified as non-self. So far both components have been characterised individually, and the majority of research in the immunology area has focused on acquired immunity. In acquired immunity, B and T lymphocytes recognise non-self by using antigen receptors such as immunoglobulin and T cell receptors. The processes by which these antigen receptors recognise foreign antigens have been extensively studied. The major mechanisms are diversity, clonality, and memory being well understood. Though, these receptors are predominantly found in vertebrates, and in less evolved organisms the recognition process of non-self is not well identified.

Since their emergence, multicellular hosts have developed defence mechanisms to survive in optimal symbiosis with parasitizing microbes. On the other hand, microbes have evolved constantly to escape the protective host barriers. The host has evolved a highly developed immune system, known as the innate immune system, that is encoded with germlines as a result of this continuous evolutionary arms race. The innate immune system uses a vast array of pattern recognition receptors (PRR) to identify and respond to threats in the environment as well as to distinguish between beneficial and harmful bacteria. Toll-like receptors (TLRs), RIG-I-like receptors, NOD-like receptors, and C-type lectin receptors are some of the germline-encoded receptors that regulate pathogen detection and host-microbiome balance.

These PRRs have developed the ability to identify highly conserved microbe associated molecular patterns (MAMPs) as a result of host-microbe coevolution. Nucleic acid or cell-wall structures are necessary for microbial survival and their alteration by microbes are challenging. It is possible to identify a variety of microorganisms with a minimal number of receptors by detecting the MAMPs. The innate immune system depends on host cell receptors to detect both advantageous and pathogenic microbes by recognising definite MAMPs and pathogen-

associated molecular patterns (PAMPs) including nucleic acids, proteins, lipids, and lipoproteins. Among these, TLRs are intensively investigated as main mediators of innate immunity in species ranging from insects to humans (*Brennan & Gilmore, 2018*).

One of the well explored family of PRRs, TLRs are the type I membrane-spanning glycoproteins usually contain three domains: extracellular domain (ECD), transmembrane domain and intracellular signalling domain. Though TLR genes are conserved across the animal kingdom (*Leulier & Lemaitre, 2008*), their structural and functional evolution have occured in response to varying environmental conditions and habitats. Presence of Toll protein in fruit fly *Drosophila melanogaster* led to the discovery of TLRs. Toll was found to be a regulator in the developing embryo (*Anderson et al. 1985*). Spätzle, the natural ligand of the Toll protein, was later discovered to be responsible for activating the protein after a fungal infection. In *D. melanogaster (Lemaitre et al.* 1996) such activation triggered the synthesis of antimicrobial peptides, deliberating immunity to fungi. Exploration of proteins similar to Toll in other species lead to in the detection of murine Toll-like receptors (TLR4). It has been established that TLR4 is essential for the natural identification of bacterial lipopolysaccharide (LPS). Numerous TLRs and their corresponding microbial ligands have been recognized and characterised in a widespread range of species since TLR4 was known as the LPS receptor (*Poltorak et al. 1998*). Studies on TLR evolution across many phyla are now possible because of the remarkable developments in whole genome sequencing. Bioinformatics analysis of whole genome data showed fungi and prokaryotes lack TLR orthologs. Receptors with low sequence resemblance to TLRs are found in the plant kingdom; these receptors are known as Receptor-like kinases or Nucleotide-binding site LRRs. They contain LRR motifs attached to different signalling domains. Comparing these plant receptors to animal TLRs, functional studies reveal that they respond to distinct microbial patterns and use fundamentally different signalling networks. This suggests that the plant receptors that contain LRRs are not ancient orthologs of TLRs, but rather belong to distinct classes of plant-specific receptors that have undergone convergent evolution and developed a function similar to that of TLRs (*Ausubel,*

2005; Boller & Felix 2009). TLRs are consequently originated from the animal kingdom (Metazoa).

There have been 16 TLRs found in the jawless vertebrate (lamprey), 13 TLRs in mammals, 10 TLRs in birds, 21 TLRs in amphibians and 20 TLRs in teleost fish. It is predicted that reptiles have a minimum of 9 TLR genes (*Rauta et al. 2014, Alcaide & Edwards, 2011, Babik et al. 2015, Kasamatsu et al. 2010*). Vertebrate TLRs have been categorised into six major families based on their sequence homology (*Roach et al. 2005*). In general, these TLRs have managed to retain their capability to identify unique ligands. The large family of TLR1 contain TLR1, TLR2, TLR6, TLR10, TLR14, TLR15, TLR16, TLR18 and TLR25 responsible for the recognition of lipoproteins (such as di- and triacylated lipopeptides). TLR15 is the members of this family is activated by proteolytic cleavage of pathogen and TLR10 negatively regulate of TLR2 (*Zoete et al. 2011, Oosting et al. 2014*). Double-stranded RNA, LPS and bacterial agents are recognised by the TLR3, TLR15 activated by microbial proteolytic cleavage TLR4 and TLR5 families. TLR7 family includes TLR7, TLR8, TLR9 are able to identify nucleic acid motifs (*Quiniou et al. 2013, Kucera et al. 2010*). The members of TLR11, TLR12, TLR13, TLR19, TLR20, TLR21, TLR22, TLR23, and TLR26 comprise the sixth major family. The functionally characterised receptors in this family detect either nucleic acid patterns or protein. Certain TLR genes, particularly those belonging to the extensive TLR1 and TLR11 families, seem to have disappeared in different lineages, possibly as a result of functional redundancy. Nevertheless, practically every species of vertebrate possesses minimum one gene each from the main families of TLR, highlighting significance of innate immune recognition of a wide variety of microbial ligands (*Raetz M et al. 2013, Keestra et al. 2010*).

Lack of TLR4 in certain teleost fish like *Takifugu rubripes* cause a prominent deviation from the conservation of TLRs in vertebrate TLR groups. The ability of TLR4 to recognise LPS, together with its coreceptors MD-2 and CD14, is crucial for the response of the mammalian immune system to bacterial infections. Certain fish, such as common carp (*Cyprinus carpio*) and zebrafish (*Danio rerio*), do have several copies of TLR4, but they do not have TLR4 coreceptor genes (*Kanwal et al. 2014*). For this reason, TLR4 does not facilitate LPS detection

in fish. Rather, it seems that fish TLR4 negatively regulate of the transcription factor NF-kB, which promotes inflammation (*Sepulcre et al. 2009*). The reasons behind this divergent evolution is not clear and might be anticipated by the analysis of TLR4 in intermediary amphibian and reptile species.

TLR15 in birds and reptiles provides another illustration of dynamic TLR evolution in vertebrates. The TIR domain of TLR15 is related to members of the TLR1 family and it is exclusively found in the genomes of birds and reptiles (*Boyd et al. 2012*). However, substantial sequence variation of LRR motifs of TLR15 has resulted in the unusual capability of this receptor to be triggered by bacterial proteases unlike TLR1 family members that recognise lipopeptides. The reason for the development of this characteristic among diapsid animals and whether it offers any major immunological benefit to these animals is not clear (*Zoete et al. 2011*).

During evolution microbes and their hosts compete together in order to prevent their extinction. Microbes develop ways to overcome host defences to survive and proliferate, while hosts must retaliate these strategies in order to avoid being overexploited. For all TLRs this fact is correct. One of the key functions of TLRs is the detection of microorganisms and limiting their numbers by stimulating the immune system. In contrast, microbes also have advanced different strategies to get around the TLR system.

The diversity of microorganisms and the evasion strategies of TLR exert selection pressure on the evolution of the TLR system. Using phylogeny-based analysis of site-specific codon substitutions, one can ascertain the "direction" of this selective pressure. Comparison of TLR sequence among species identified sites subjected to positive selection when the ratio of nonsynonymous over synonymous codon substitutions is more than 1. This suggests that a site has maintained its polymorphism and could offer a fitness benefit as a result of adaptive evolution. A codon is considered to have experienced purifying selection if the ratio of nonsynonymous to synonymous codon substitutions is less than 1. This suggests that polymorphisms would typically be harmful in such a site, and therefore the site evolves under

functional constraints (*Yang et al. 2002*). Because of their nonredundant roles in signal transduction, TLR adaptor proteins, particularly MyD88 and TRIF evolve under functional constraints (*Nakajima et al. 2008, Fornarino et al. 2011*).

Since the TLR adaptors interact with a variety of proteins, polymorphisms would most likely affect their interaction with some of these proteins. Because the TIR domain exhibits a high degree of similarity across a wide range of species and can become inactive by substituting even a single critical site, maintaining function also controls the evolution of the TIR domain (*Nakajima et al. 2008, Mikami et al. 2012*). Moreover, polymorphisms are present in the ligand-binding region of the ECD of nucleic acid detecting TLRs (such as TLR3, TLR7, TLR8, and TLR9) although they are almost ever detected there, suggesting influence of functional restrictions on the ligand binding by these TLRs (*Keestra et al. 2008*). This restriction is most likely caused by the extremely similar structures of host and microbial nucleic acids, which poses the risk of triggering autoimmune reactions. Detrimental mutations that would have enhanced binding to self-nucleic acids probably been eliminated from the population by purifying selection, reducing the likelihood of identifying self-nucleic acids while preserving sufficient detection of microbial nucleic acids (*Wlasiuk & Nachman, 2010; Vinkler et al. 2014; Fornůsková et al. 2013; Webb et al. 2015*).

The ECD of surface expressed TLRs (such as TLR2, TLR4 and TLR5) shows a robust diversified evolution propelled by positive selection of beneficial mutations, in contrast to nucleic acid-sensing intracellular TLRs. Positively selected sites in TLR genes from a variety of species including fish, cattle, pigs, birds, rodents and primates have been identified from the genomic data (*Werling et al. 2009*). The majority of these sites are situated inside the ligand-binding domain or quite close to it. The need to distinguish between host-specific commensals and pathogens, as well as antagonistic coevolution with host-specific pathogens, may have contributed to the highly variable nature of TLR ligand-binding domains among hosts. The polymorphic nature of ligand-binding domains of TLR among hosts might have been driven by antagonistic coevolution through host-specific pathogens and/or the need to distinguish among host-specific commensals and pathogens.

TLRs are among the widely investigated innate immune receptors. More studies are yet to be done about the evolutionary aspect of this receptor family. Substantially broader functional studies incorporating ligands from a wide range of microorganisms would greatly benefit in our understanding of the evolution of TLRs. Residues with possible importance for TLR function can be predicted using phylogeny-based assessments of the evolution of molecular TLRs. Functional analyses could reveal the selective factors underlying the purifying or diversifying selection of TLRs and offer experimental support for their findings. As a whole, these analyses may be crucial in understanding the molecular foundation of antagonistic host-specific coevolution with microorganisms and the ensuing natural resistance to disease.

Chapter - II

Innate immune cells in mammal including dendritic cells and macrophages are activated by the microbial components recognised as nonself such as lipopolysaccharide (LPS) from Gram negative bacteria. Toll was discovered during the end of the 20th century as a crucial receptor for host defence against fungal infection in Drosophila species having only innate immunity (*Lemaitre et al. 1996*). One year later, a homolog of Toll receptor (now known as TLR4) in mammal of the was found to trigger the gene expression implicated in inflammatory responses (*Medzhitov et al. 1997*). Furthermore, a point mutation in the TLR4 gene has been discovered in a mouse strain that is unresponsive to LPS (*Poltorak et al. 1998*). These results have made innate immunity an interesting research topic, and during recent time, significant progress has been made to understand that the innate immune system has a complex strategy that detects microbial pathogen invasion via Toll-like receptors (TLRs). Furthermore, innate immunity activation is essential for the establishment of acquired immunity for specific antigen.

Identification of the Toll like Receptor (TLR) family

Following its identification of TLR4, the first mammalian TLR, numerous proteins with structural similarity to TLR4 were discovered and termed Toll-like receptors (*Rock et al. 1998*). Mammalian TLRs form a broad family with 11 members. In humans and mice TLR1-TLR9 are conserved. Though it has been thought that in humans TLR10 is functional, substitution with a dissimilar and non-productive region at the C terminal of the mouse TLR10 gene indicated non-functionality of mouse TLR10. Likewise, TLR11 in mouse is functional, but in human TLR11 is absent due to presence of a stop codon in gene (*Zhang et al. 2004*). The cytoplasmic part of TLRs is highly similar to the IL-1 receptor family, and is known as a Toll/IL-1 receptor (TIR) domain. In spite of their similarities, structural differences have been found in extracellular part of both receptors. An immunoglobulin-like domain is found in IL-1 receptors, while the TLRs have leucine-rich repeats (LRRs) in the extracellular domain. Functional role of TLR4 in recognising the microbial component LPS was first characterised (*Poltorak et al. 1998*). Individual TLRs are now known to play key roles in recognising distinct

microbial components generated from pathogens such as bacteria, fungus, protozoa, and viruses.

Toll-like Receptors (TLRs) in Invertebrates

TLR types and numbers across invertebrates differ by species, ranging from one to hundreds. Two types of TLRs have been categorized depending on the CF motif numbers (LRRCT, containing C terminal end of LRRs with cysteine clusters), protostome type (P-type or mccTLR) and vertebrate type (V-type or sccTLR) (*Maaser et al. 2004*). P-type TLRs include just one cluster of cysteine at LRRCT, whereas V-type TLRs contain numerous clusters of cysteine at LRRCT and, in certain cases LRRNT at the N terminal end. P-type TLRs have only been found in invertebrates, suggesting that they are an old variety of TLR. In contrast, most vertebrate and some invertebrate TLRs are V-type (*Hawn et al. 2003*). It has been proposed that, unlike vertebrate V-type TLRs, P-type TLRs do not directly bind PAMPs, as evidenced by Drosophila Toll-1, the most well-studied P-type TLR (*Anderson et al. 1985*). Most TLRs have been detected in the invertebrate phyla such as Porifera, Coelenterata, Platyhelminthes, Nematoda, Annelida, Echinodermata, Mollusca, and Arthropoda.

Porifera

TLRs from porifera have primarily been recorded on *Amphimedon queenslandica* and *Suberites domuncula*. *A. queenslandica* has been found to harbour two TIR domain containing proteins with N terminal IL-1R-like Ig domains and an LRR domain-containing protein with Ig and epidermal growth factor (EGF) like domains (*Gauthier et al. 2010; Hentschel et al. 2012; Srivastava et al. 2010*). Similarly, the sponge *S. domuncula* has been found to harbour a TIR only protein (Sd-TLR) with a transmembrane domain; however, no proteins containing an LRR domain have been found in this species (*Wiens et al. 2007*). The presence of NF-κB homologs and Myeloid differentiation primary response protein 88 in the TLR to NF-κB route in *A. queenslandica* and *S. domuncula* suggests that the MyD88-mediated TLR signalling pathway already has been observed in poriferans (*Gauthier et al. 2010; Gilmore & Wolenski, 2012; Song et al. 2012*). Furthermore, during the early stages of *A. queenslandica* development,

expressions of additional adaptor proteins implicated in the TLR-to-NF-κB pathway are seen, indicating that this route is related to development (*Gauthier et al. 2010*). Sd-TLR has been shown to engage in ongoing interactions with microorganisms and may have a role in *S. domuncula* immune modulation (*Wiens et al. 2005; Wiens et al. 2007*).

Cnidaria

Around 10,000 aquatic organisms that comprise the phylum Cnidaria that are morphologically primeval outgroup to bilaterians include corals, Hydra, sea anemones, and jellyfish (*Putnam et al. 2007*). No classical TLRs, but a large number of proteins have been found in Hydra species connected to the TLR to NF-κB pathway. Furthermore, Hydra has been found to have two transmembrane TIR domain-containing proteins and two LRR domain-containing proteins (*Bosch et al. 2009; Augustin et al. 2010*). In HEK293 cells, a chimeric protein called HyLRR-2, can activate the NF-κB reporter in response to flagellin by combining the human TIR domain with LRR protein of Hydra, but not LPS (*Putnam et al. 2007*). Thus, flagellin could be the HyLRR-2 ligand that starts innate immune signalling. The genome of the sea anemone *Exaiptasia pallida* contains two TIR domain-only genes that may encode the same protein (*Poole & Weis, 2014; Baumgarten et al. 2015*). Recent transcriptome study has shown that a number of additional cnidarians, such as the corals *Acropora digitifera*, *Acropora millepora*, and *Orbicella faveolata* express classical TLR members and elements linked to NF-κB activation (*Miller et al. 2007; Rauta et al. 2014*).

Platyhelminthes

The functions of TLRs in platyhelminth has been explored on planarians, turbellarians, and rotifers. Since they are non-parasitic flatworms, Planarians are evolutionary important to the study of immune responses triggered by injury and the process of metazoan regeneration (*Riutort et al. 2012; Sánchez, 2003*). Many proteins implicated in the TLR to NF-κB pathway have been found in the freshwater planarian *Schmidtea mediterranea*; though, the TLRs of this flatworm are TIR and LRR-only proteins rather than canonical TLRs (*Peiris & Hoyer, 2014; Forsthoefel et al. 2012*). Throughout *S. mediterranea* head regeneration, TLRs, MyD88,

TRAF, and IRAK are all upregulated, suggesting that the TLR-initiated signalling pathway is probably involved in avoiding infection throughout the regeneration process (*Peiris & Hoyer, 2014*). Therefore, further research is required to fully understand the immunological responses that TLRs trigger against infections or PAMP activation, as well as the mechanisms underlying the phylum Platyhelminthes.

Nematoda

The traditional model organism for studying nematodes is *Caenorhabditis elegans*. It has been shown that C. elegans expresses a protein that contains a TIR domain, a canonical P-type TLR (TOL-1), and other elements that are similar to those seen in TLR signalling pathways in mammals (*Forsthoefel et al. 2012; Brandt & Ringstad, 2015; Gissendanner & Kelley, 2013; Irazoqui et al. 2010; Liu & Shen, 2012; Mancuso et al. 2012; Pradel et al. 2007; Pujol et al. 2001*). However, it appears that TOL-1 does not start the NF-κB-dependent signalling pathways since *C. elegans* lacks several components of the TLR-to-NF-κB signalling pathways, including MyD88, IKK, and NF-κB. The downstream pathways that TOL-1 activates during early development are still unknown, despite the fact that prior research has demonstrated the importance of TOL-1 for C. elegans pathogen identification and early development (*Gissendanner & Kelley, 2013; Mancuso et al. 2012*).

Annelida

Davidson et al. provided initial evidence of the presence of TLRs in the genomes of annelids, such as the leech *Helobdella robusta* and the polychete worm *Capitella capitata*. The high number of TLR-like genes found in the genome of *C. capitata* is probably due to the fact that its TLR sequences are quite similar and may have resulted from recent gene duplications (*Davidson et al. 2008, Simakov et al. 2013*). A TLR known as Hm-TLR1 has been found in *Hirudo medicinalis* seems to be a chimeric mix of the cytoplasmic portion of TLR13 and the intra endosomal region of human TLR3. Microglial cells and neurons have been shown to express Hm-TLR1, and it has been proposed that Hm-TLR1 is involved in immunity

(*Schikorski et al. 2009; Cuvillier-Hot et al. 2011*). Overall, a number of studies showed that annelid TLRs are essential for neurogenesis and neuroimmunity.

Mollusca

So far, molluscan species such as *Cyclina sinensis, Biomphalaria glabrata, Chlamys farreri* and *Crassostrea gigas* have been found to contain TLR. The genome of *B. glabrata* contains 56 TLR-encoding genes, 27 encoding full TLRs, together with 2 P-type and 25 V-type TLRs (*Adema et al. 2017*). *B. glabrata* has been shown to harbour a novel snail TLR called Bg-TLR. B. glabrata becomes more susceptible to parasites when Bg-TLR is knocked down, suggesting that Bg-TLR may be important for immunological response of *B. glabrata* after infection (36). It has been shown that *C. sinensis* hemocytes include a pathogen-responsive TLR13-MyD88-NF-κB pathway, and absence of TLR13 results in the reduced expression of other adaptors in this signalling network (*Ren et al. 2016; Ren et al. 2017*). Research suggests that MyD88-dependent signalling pathway mediate the activation of downstream immunological processes in C. sinensis, particularly the antibacterial response (*Ren et al. 2016*). Further study is required for better understanding of the developmental roles of molluscan TLRs.

Arthropoda

Comparatively few studies have been conducted so far on toll-like receptors of Merostomata species. The horseshoe crab *Tachypleus tridentatus* has been shown to possess a TLR gene (tToll) that is similar in length and structure to Drosophila Toll-1 (*Inamori et al. 2010*). Interestingly, the LRRs of tToll-1 protein do not involve PAMP binding; instead, they bind to molecules that resemble Drosophila Spätzle (*Kurata et al. 2006; Coscia et al. 2011*).

Within Arthropoda, Insecta is so far the major group of hexapod invertebrates with over a million species. Insects possess the ability to develop a quick antimicrobial response upon infection (*Belinda et al. 2008; Brennan et al. 2004; Dan, 2003; Tanji & Ip, 2005; Royet et al. 2005*). There is evidence that the mammalian innate immune response and the insect antimicrobial response are similar (*Hargreaves & Medzhitov, 2005*). Research on the model

organism *D. melanogaster* has laid the groundwork for our comprehension of the basic mechanisms underlying the immune response in insects. Study has shown that D. melanogaster induction of expression of antimicrobial peptide is mediated by the Toll pathways (*Gottar et al. 2002; Tzou et al. 2002*). Toll-1, the first TLR to be identified, was found in D. melanogaster embryos in 1985. Its function was to specify the dorsal ventral polarity of the embryo. Subsequently, the genome of *D. melanogaster* was found to contain genes corresponding to other members of the Toll family (Toll-2–9), whose dual roles in immune response and embryogenesis have been gradually validated (*Hoffmann, 2003; Valanne et al. 2011*). Furthermore, a number of Drosophila TLRs perform significant roles in the preservation of tissue integrity by triggering the NF-κB-dependent apoptosis of unsuitable or mutant cells (*Ferrandon et al. 2007; Meyer et al. 2014*).

In crustacean species such as copepods, shrimps and crabs various TLRs have been found. Amid these TLRs found in shrimp species, such as *Litopenaeus vannamei, Procambarus clarkii, Penaeus monodon, Fenneropenaeus chinensis, Macrobrachium rosenbergii, and Marsupenaeus japonicus* are widely studied (*Yang et al. 2007; Arts et al. 2007; Yang et al. 2008; Wang et al. 2015; Mekata et al. 2008; Srisuk et al. 2014; Sun et al. 2017*). NF-κB is a typical downstream transcriptional component in the shrimp Toll signalling pathway, which is consistent with findings in other species (*Wang et al. 2011; Li et al. 2014; Matsuo et al. 2008*).

Echinodermata

The utmost evolved invertebrate group, echinoderms have a common evolutionary ancestor with chordates. According to reports, TLRs are essential for metazoan immunity which includes echinoderm, sea urchins and sea cucumbers (*Buckley et al. 2012*). It is notable that the purple sea urchin *S. purpuratus* has such a greatly increased innate receptor repertoire. Molecular phylogenetic tree analysis has identified 222 TLR-like genes in total (8 P-type TLRs and 214 V-type TLRs) that are present in the *S. purpuratus* genome and can be classified into seven broad categories (*Buckley et al. 2012; Hibino et al. 2006*).

TLRs in amphioxus

Amphioxus, a typical cephalochordate that is evolutionarily located in the invertebrate–vertebrate transition point, is a significant organism for studying the evolution of the TLR-associated immune system (*Li et al. 2011*). An incredibly intricate TLR system, comprising over 40 TIR adaptors and at least 48 TLRs, is encoded by the amphioxus genome (*Huang et al. 2008*). Together, the observations offer a point of reference for exploring the intricacy of the amphioxus innate immunity and suggest fresh avenues for investigating comparable vertebrate topics.

Toll-like Receptors (TLRs) in non-mammalian vertebrates

The classifications Cyclostomata, Chondrichthyes, Osteichthyes, Amphibia, Reptilia, and Aves comprise non-mammalian vertebrates. So far 28 functioning TLRs have been found in these classes in a variety of species. There have been six major subfamilies of TLRs such as TLR1, TLR3, TLR4, TLR5, TLR7, and TLR11. The large TLR1 subfamily includes TLR1, TLR2, TLR6, TLR10, TLR14, TLR15, TLR16, TLR25, TLR27, and TLR28 primarily recognises lipoproteins. While the TLR3, TLR4, and TLR5 subfamilies recognise dsRNA, LPS (but not in fish or amphibians) and bacterial flagellin respectively. TLR7, TLR8 and TLR9 are members of the TLR7 subfamily, which is involved in nucleic acid motif recognition. TLR11, TLR12, TLR13, TLR19– TLR23, and TLR26 are members of the TLR11 subfamily, which is the sixth main subfamily. Members of this family perform a variety of tasks ranging from sensing nucleic acid motifs to proteins.

Fishes

In the lowest class of vertebrates, Cyclostomata comprise two families of jawless fish that have survived: the lamprey and hagfish (*Kuraku et al. 2009*). Through polymerase chain reaction-based cloning, two TLRs (laTLR14a and laTLR14b) have been discovered in the Japanese lamprey (*Lampreta japonica*). The encoding gene for TLR14, which is interestingly a member of the TLR1 subfamily, is found in the genomes of teleosts and amphibians (*Ishii et al. 2007*).

This suggests that the existing subsets of TLRs in vertebrates evolved prior to the divergence of the jawless fish ancestor from the mammalian ancestor.

Chondrichthyes or jawed cartilaginous fish are a notable group of animals in immunological research field. They are regarded as the initial species that have developed immune responses that are adaptive. It is also interesting that the innate immune system exists at this critical stage of evolution. Based on a study of transcriptome data TLR2, TLR3, TLR6 and TLR9 have been found in the grey bamboo shark *Chiloscyllium griseum* (*Anandhakumar et al. 2012; Krishnaswamy et al. 2014*). While TLR3 of *C. griseum* is closely connected to homologs in *Rattus norvegicus* and *Canis lupus familiaris*, TLR2 of *C. griseum* is closely related to homologs in *Sus scrofa* and *Gallus gallus*. The most resemblance between TLR6 and its homologs in *Felis catus* and *Bos tarus* and between TLR9 and its homologs in *Andrias davidianus* is found.

Osteichthyes or teleost fish contain over 23,500 species and are extraordinarily diverse (*Volff, 2005*). About 21 TLRs (TLR1–TLR5, TLR5S, TLR7-TLR9, TLR13, TLR14, TLR18–TLR23, and TLR25–TLR28) have been found in a variety of teleost fish species. These TLRs include "teleost-specific" TLRs as well as orthologs of mammalian TLRs (*Quiniou et al. 2013; Boudinot et al. 2014; Zhang et al. 2013*). While teleost TLR4 appears to be structurally preserved and does not recognise LPS, unlike its mammalian counterparts, TLR1-TLR3, TLR5 and TLR7–TLR9 have structural and functional similarities with their mammalian counterparts. "Teleost-specific" TLRs include TLR5S, TLR18–TLR20, TLR23 and TLR25–TLR28. Despite their designation as "specific," these TLRs have a significant degree of structural similarity with the TLR system found in mammals (*Palti, 2011; Iliev et al. 2005*). Teleost TLR1 subfamily contain TLR1, TLR2, TLR14, TLR18, TLR25, TLR27, and TLR28.

Amphibian

There are currently at least 20 TLRs known to exist in amphibians, including TLR1, TLR2.1–TLR2.2, TLR3–TLR5, TLR6.1–TLR6.2, TLR7, TLR8.1–TLR8.2, TLR9, TLR12, TLR13, TLR14.1–TLR14.4, TLR21 and TLR22. Also, a number of soluble LRR-only TLR varieties

have been identified. In amphibians, TLR2, TLR6 and TLR8 may be duplicated and the TLR14 subfamily appears to be expanded. A putative soluble short form of TLR5 (TLRS5) is present in amphibians. It has been confirmed that TLR4 is present in the *Xenopus* genome but not CD14 or MD2 which are necessary for TLR4-mediated recognition of LPS (*Ishii et al. 2007; Boudinot et al. 2014*).

Reptilia

Reptiles hold a pivotal role in the evolution of vertebrates, owing to their distinct physiology and status as the sole poikilothermic amniotes. Nevertheless, little is known about the composition, role, and ligand specificity of TLRs in reptiles (*Zimmerman et al. 2010*). Only one species, the green anole lizard *Anolis carolinensis*, has been found in searches for reptile TLRs; they are annotated as molecules similar to mammalian TLR2, 3, 4, 5, 6, 7, and 13. The cloning, characterization, and functional analysis of *A. carolinensis* TLR5 were recently reported (*Fink et al. 2016*). The receptor or acTLR5, has a typical TLR protein structure with 22 extracellular LRRs flanked by N- and C-terminal LRR domains, an intracellular TIR domain, and a transmembrane region. From a phylogenetic perspective, acTLR5 is most separated from fish TLR5 and most similar to avian TLR5. Experiments using PAMPs to stimulate acTLR5 showed that it responded differently to bacterial flagellin (*Nie et al. 2018*).

Aves

The immunological responses of avian (birds) and mammals are essentially similar, despite their divergence approximately 300 million years ago (*Brownlie & Allan, 2011; Kaiser, 2007*). Studies on the junglefowl *G. gallus*, which is the antecedent of domestic chicken, have generated most of the knowledge about avian immunology (*Hillier et al. 2004*). Knowledge regarding the identified ligands of avian TLRs has been expanded. Ten avian Toll-like receptors (TLR1La, TLR1Lb, TLR2a, TLR2b, TLR3, TLR4, TLR5, TLR7, TLR15 and TLR21) have been identified by different studies. Six of them (TLR2a, TLR2b, TLR3, TLR4, TLR5 and TLR7) are structurally distinct orthologs of TLRs from mammal (*Brownlie & Allan, 2011; Smith et al. 2004; Yilmaz et al. 2005; Boyd et al. 2007*). Avian TLR15 a member of the

TLR1 subfamily appears to be exclusive to birds. Avian TLR21 is an ortholog of teleosts and amphibians TLR21. TLR4 and MD2 expressed in chickens. These proteins are involved in the activation of NF-κB in response to LPS stimulation, but not in the synthesis of IFN1 (*Temperley et al. 2008*). Nevertheless, the poly (I:C) activation of chicken leukocytes and ensuing upregulation of IFN1 point to the presence of the TRIF signalling pathway. As a result, the immune system of chickens may react to LPS in a TRIF-independent, MyD88-dependent manner. The reason that other than in mammals TRIF does not contribute in LPS–TLR4 signalling in chickens may be partially attributed to the lack of TICAM2 ortholog in the chicken genome. Furthermore, a large number of positively selected sites have been recorded to known ligand binding regions, representing that species-specific changes in PAMP recognition are responsible for the variations (*Keestra & van Putten, 2008; Grueber et al. 2014*).

Toll-like Receptors (TLRs) in mammals

TLRs belong to the class of pattern recognition receptors (PRRs) that sense conserved molecular patterns to trigger the innate immune response in the event of an early pathogen detection. Leucine-rich repeats (LRRs) motif, transmembrane domain, and cytoplasmic Toll/IL-1 receptor (TIR) domain are the three structural domains found in typical TLRs. While the TIR domain interacts with signal transduction adaptors to commence signalling, the LRRs motif is in charge of pathogen recognition (*Takeda et al. 2003*).

Following the initial discovery of a Toll protein in the fruit fly *Drosophila melanogaster* (*Anderson et al. 1985*), 10 human TLRs (TLR1–TLR10) and 13 mouse TLRs (TLR1–TLR13) have been identified. With a few notable exceptions, the majority of mammalian species seem to have a similar repertoire of TLR homologs (*Du et al. 2000; Tabeta et al. 2004*). For example, a mouse gene encoding the human TLR10 homolog is also found, but it seems that the gene altered by a retrovirus latter (*Basith et al. 2011*). Additionally, TLR11–TLR13 are expressed in mice but not in humans (*Mahla et al. 2013*). TLRs are able to identify molecules that are often shared by infections, referred to as pathogen associated molecular patterns (PAMPs), as well as host endogenous damage associated molecular patterns (DAMPs). Depending on the

TLR type, there are many different types of recognition. For instance, lipopolysaccharide (LPS) a constituent of Gram-negative bacteria is detected by mammalian TLR4, although bacterial 23S rRNA is recognised by murine TLR13 (*Vidya et al. 2018*).

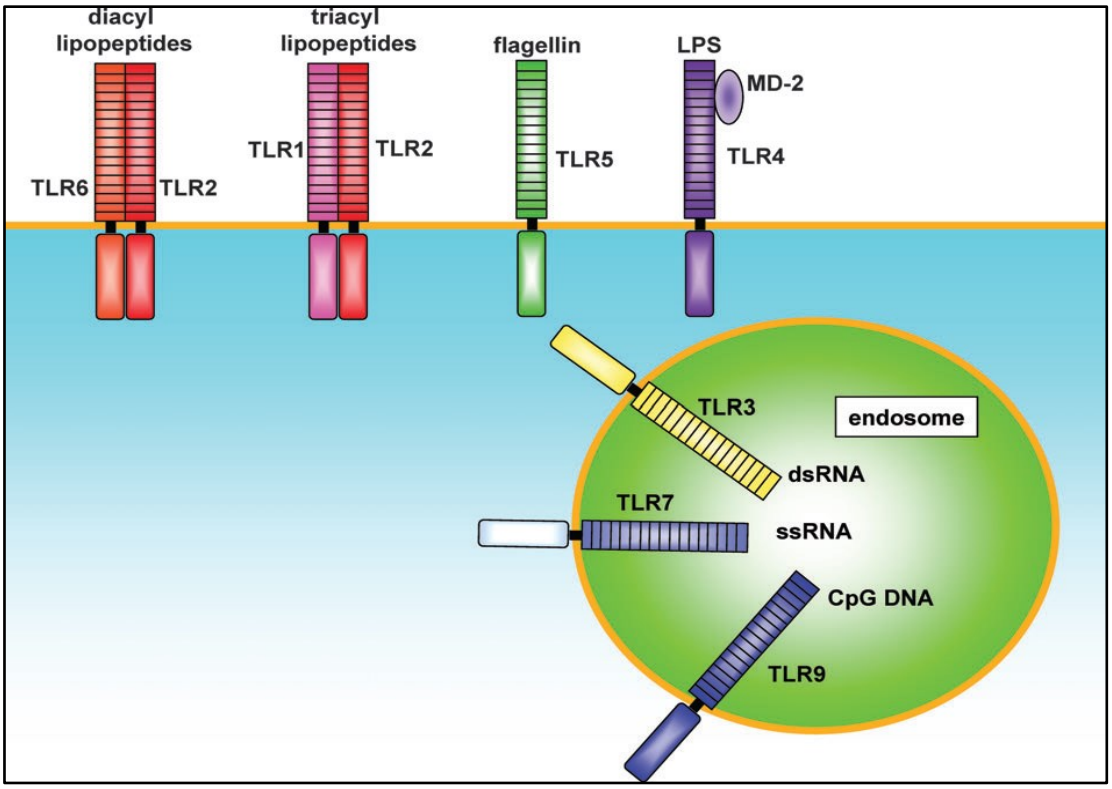


Figure 1: TLRs and their ligands. TLR2 is crucial for the identification of microbial lipopeptides. TLR1 and TLR6 work in tandem with TLR2 in order to distinguish between triacyl and diacyl lipopeptides respectively. TLR4 is receptor of LPS. TLR9 is necessary for the identification of CpG DNA. While TLR7 and TLR8 are linked to the recognition of viral-derived ssRNA, TLR3 is involved in the identification of viral dsRNA. Flagellin is recognised by TLR5. As a result, members of the TLR family are able to identify particular microbial component patterns (*Takeda and Akira 2005)*.

TLR1, TLR2 and TLR6

TLR2 recognizes wide range of microbial components. These comprise peptidoglycan and lipoteichoic acid from Gram-positive bacteria, as well as lipoproteins and lipopeptides from different pathogens (*Takeda et al. 2003*). Furthermore, it is claimed that LPS preparations from non-enterobacteria are recognised by TLR2. The quantity of acyl chains in the lipid A component of these LPS is different compared to the typical LPS of Gram-negative bacteria that TLR4 recognises and this difference leads to the differential recognition (*Netea et al.*

2002). TLR1 and TLR6 distinguish between diacyl and triacyl lipopeptides by functionally interacting with TLR2. Furthermore, TLR1 play role in identifying the outer surface lipoprotein (*Alexopoulou et al. 2002*). Additionally, TLR2 has been demonstrated to have functional interaction with other kinds of receptors including a lectin family receptor dectin-1 for the b-glucan constituent of fungal cell walls. As a result, TLR2 functions in concert with many proteins that are either physically related or unrelated to each other to recognise a broad variety of microbial products.

TLR3

Human TLR3 expression in the double-stranded RNA (dsRNA) non-responsive cell line 293 exhibits elevated activation of NF-kB responding to dsRNA. Furthermore, TLR3 deficient mice have shown reduced ability in response to dsRNA (*Alexopoulou et al. 2001*). Most viruses generate double-stranded RNA (dsRNA) while replication which triggers the synthesis of type I interferons (IFN-a/b) that have both immunostimulatory and antiviral properties. TLR3 is therefore involved in the detection of viruses and dsRNA.

TLR4

TLR4 is a crucial receptor for the recognition of LPS (*Hoshino et al. 1999*). Additionally, it has been demonstrated that TLR4 has a role in the endogenous ligand recognition, including HSP60 and HSP70 (heat shock proteins) additional domain A of fibronectins, hyaluronic acid oligosaccharides, heparan sulphate and fibrinogen. To activate TLR4, concentration of all of the endogenous ligands should to be very high. Furthermore, it has been demonstrated that the LPS contamination in the HSP70 preparation gives it the capacity to activate TLR4 (*Gao & Tsan, 2003*). Since LPS is a highly powerful immuno-activator, even minute amounts of LPS can activate TLR4, contaminating these endogenous ligand formulations. Hence, understanding of endogenous ligand recognition by TLR4 require more detailed investigation.

TLR5

The sensitivity the monomeric component of bacterial flagella is conferred by enforced expression of human TLR5 in CHO cells (*Hayashi & Smith, 2001*). Through a close physical interaction between TLR5 and flagellin it has also been demonstrated that TLR5 recognises an evolutionarily conserved region of flagellin (*Smith et al. 2004*). Intestinal epithelial cells express TLR5 on their basolateral side but not on their apical side (*Gewirtz et al. 2001*). Additionally, intestinal endothelial cells in the subepithelial compartment express TLR5 (*Maaser et al. 2004*). Furthermore, flagellin stimulates the production of inflammatory cytokines by lung epithelial cells (*Hawn et al. 2003*). These results highlight the significant role of TLR5 in mucosal surface microbial identification.

TLR7 and TLR8

Both TLR7 and TLR8 are structurally conserved and in some cases recognise the same ligand. Compounds imidazoquinoline are recognised by human TLR7 and TLR8, but not by mouse TLR8 (*Jurk et al. 2002*). It has also been demonstrated that loxoribine, a synthetic substance with antiviral and antitumor properties, is recognised by mouse TLR7 (*Lee et al. 2003; Heil et al. 2003*) Guanosine nucleoside and imidazoquinoline share a structural similarity. Consequently, it was predicted that TLR7 and human TLR8 would be able to identify the nucleic acid like structure of virus. The findings that TLR7 and human TLR8 recognise guanosine or uridine rich single stranded RNA (ssRNA) from viruses such the influenza virus, vesicular stomatitis virus and human immunodeficiency virus (*Heil et al. 2004; Diebold et al. 2004*) has demonstrated the validity of this prediction. Although the host contains a large amount of ssRNA, TLR7 and TLR8 typically do not recognise host-derived ssRNA (*Lund et al. 2004*). This could be because host-derived ssRNA is not transported to the endosome, despite the expression of TLR7 and TLR8 in the endosome (*Nie et al. 2018*).

TLR9

Analysis of TLR9-deficient mice exhibited that TLR9 is a CpG DNA receptor (*Hemmi et al. 2001*). Unmethylated CpG patterns give bacterial DNA its immunostimulatory properties. The immunostimulatory action of CpG motifs is abrogated in vertebrates due to a significant reduction in their frequency and a high degree of methylation of their cysteine residues. CpG DNA comes in minimum two types termed as A or D type and B or K type. Conventional B or K-type CpG DNA was the first to be discovered and strongly induce inflammatory cytokines like TNF-a and IL-12. A or D type CpG DNA differs structurally from ordinary CpG DNA and is more effective in stimulating plasmacytoid dendritic cells (PDC) to produce IFN-a, but less effective in stimulating IL-12 production (*Krug et al. 2001; Verthelyi et al. 2001*).

It has been demonstrated that TLR9 is necessary for the identification of both forms of CpG DNA (*Hemmi et al. 2003*). Apart from CpG DNA originating from bacteria and viruses, TLR9 is probably involved in pathogenesis of autoimmune disorders. Hence, TLR9 seems to have a role in the numerous autoimmune diseases by detecting the structure of chromatin. The mechanisms of chloroquine, that is used clinically to treat SLE and rheumatoid arthritis are not known. Chloroquine blocks TLR9 dependent signalling by inhibiting the pH-dependent maturation of endosomes act as a basic element to neutralize acidification in the vesicle (*Häcker et al. 1998*). For this, chloroquine may be an anti-inflammatory agent that inhibit TLR9 dependent immune responses.

TLR11

Recently identified TLR11 has shown its expression in epithelial cells of bladder and mediates resistance to mouse infection to uropathogenic bacteria. TLR11 deficient mice had high susceptibility to uropathogenic bacterial infections. These results suggest that mouse TLR11 mediates anti-uropathogenic bacterial response, even though the ligand is yet unknown. Studies have been suggested that humans lack a functioning TLR11 protein. These findings could suggest that the human TLR11 protein was lost to evolution because it was futile in the human context (*Zhang et al. 2004*).

Subcellular localization of Toll-like Receptors (TLRs)

Individual TLRs have different distribution within the cell. The expression of TLR1, TLR2, and TLR4 on the cell surface is shown by the positively staining cell surface with specific antibodies. Contrary, it has been shown that intracellular compartments including endosomes express TLR3, TLR7, TLR8, and TLR9 (*Heil et al. 2003; Matsumoto et al. 2003*). It has been demonstrated that endosomal maturation is necessary for TLR3, TLR7, or TLR9 mediated recognition of their ligands (*Heil et al. 2003; Diebold et al. 2004; Lund et al. 2004; Ahmad-Nejad et al. 2002*). TLR9 is drawn from the endoplasmic reticulum following non-specific absorption of CpG DNA, which is initially non-specifically trapped into endosomes by the TLR9 ligand CpG DNA (*Latz et al. 2004*). Therefore, it is can be hypothesized that during bacterial infection dendritic cells and macrophages engulf bacteria through phagocytosis.

Following bacterial degradation in phagosomes/lysosomes or endosomes/lysosomes CpG DNA is exposed, where TLR9 is expressed or recruited. When a virus infects a cell, it enters through receptor-mediated endocytosis and the viral membrane fuses with the endosomal membrane to expose the viral contents to the cytoplasm. Sometimes in endosomal compartment viral particles degradation results in the exposure of TLR ligands such dsRNA, ssRNA, and CpG DNA. After being exposed to zymosan cell surface expressed TLR2 is attracted to the phagosomal compartment of macrophages (*Underhill et al. 1999*). TLR recognition of microbial components may therefore primarily occur in the phagosomal/lysosomal or endosomal/lysosomal compartments.

Toll-like Receptors (TLRs) Signaling

The immune system "senses" risk through TLRs, which are crucial molecular sensors in order to defend the host against pathogenic microbes or endogenous threats (*Hug et al. 2018*). TLRs have been shown to play a wide range of functions including identification of self and non-self antigens, invasive pathogen detection, connecting the gap between innate and adaptive immunity and controlling the generation, proliferation, and survival of cytokines (*Vidya et al. 2018; Bhattacharyya et al. 2018; Ruysschaert & Lonez, 2015; Reuven et al. 2014*). Different

cytokines and chemokines are produced as a result of signalling pathways that are subsequently started and mediate TLR activities. TIR domain-containing adaptor-inducing IFNβ (TRIF) dependent pathways and myeloid differentiation primary response protein 88 (MyD88)-dependent pathways are the two main categories into which TLR signalling pathways are currently classified (*Akira & Takeda, 2004*).

Except TLR3, all TLRs mostly use the MyD88-dependent response. MyD88 interacts to the TIR domain of the conforming TLR by homotypic or heterotypic interactions following ligand recognition and TLR dimerization. The death domain of MyD88 then recruits IL-1 receptor-associated kinase 4 (IRAK4), which results in the creation of a Myddosome complex (*Lin et al. 2010*) and the autophosphorylation of IL-1 receptor associated kinase 1 (IRAK1). Then, by K-63-linked polyubiquitination of TAK1 and TRAF6, the protein tumour necrosis factor (TNF) receptor associated factor 6 (TRAF6) gets activated, which then activates the TAK1 or TGF-β-activated kinase (TAB) complex (*Gorjestani et al. 2012*). The subsequent process involves the phosphorylation and destruction of I kappa B alpha (IκBα) by IκB kinase (IKK). Finally, the transcription factor NF-κB translocate to the nucleus upon degradation of this inhibitor, triggering the transcription of genes that code for inflammatory cytokines (*Wang et al. 2001*).

The TRIF dependent pathway is generally thought to be exclusive to a few numbers of TLRs, including TLR3 and TLR4 in mammals. The TRIF-dependent pathway can activate transcription factors such as NF-κB, activating protein 1 (AP-1) and members of the interferon (IFN) regulatory factor (IRF) family, which together can induce the production of pro-inflammatory cytokines and/or type I IFN (IFN1) (*Hoebe et al. 2003*). The recognition of double-stranded RNA (dsRNA) activates TLR3, following the recruitment of TRIF. A branch in the signalling pathway is created when TRIF activates receptor interacting serine or threonine kinase 1 (RIPK1) and TANK-binding kinase 1 (TBK1). IRF3 is phosphorylated by the TRIF/TBK1 signalling complex, which permits the translocation to nucleus and the synthesis of IFN1.

Similar to the MyD88-dependent pathway, RIPK1 activation results in a series of signal transduction events (*Kawai & Akira, 2010*). Mammals use TLR4 as an LPS receptor. After MyD88 and MyD88-adapter-like (MAL) adaptors are recruited, the TLR4-myeloid differentiation protein 2 (MD2)-LPS complex activates early phase NF-κB and mitogen-activated protein kinase (MAPK). The TLR4-MD2-LPS complex interacts with TRAM (TRIF and TIR domain containing adapter molecule 2) adaptors once it has entered the cell by endocytosis. This TRIF-dependent pathway activates late-phase NF-κB and IRF7 in addition to inducing IFN1 production (*Shuang et al. 2015*). Ultimately, TLR signalling leads to the activation or suppression of genes that control the inflammatory response.

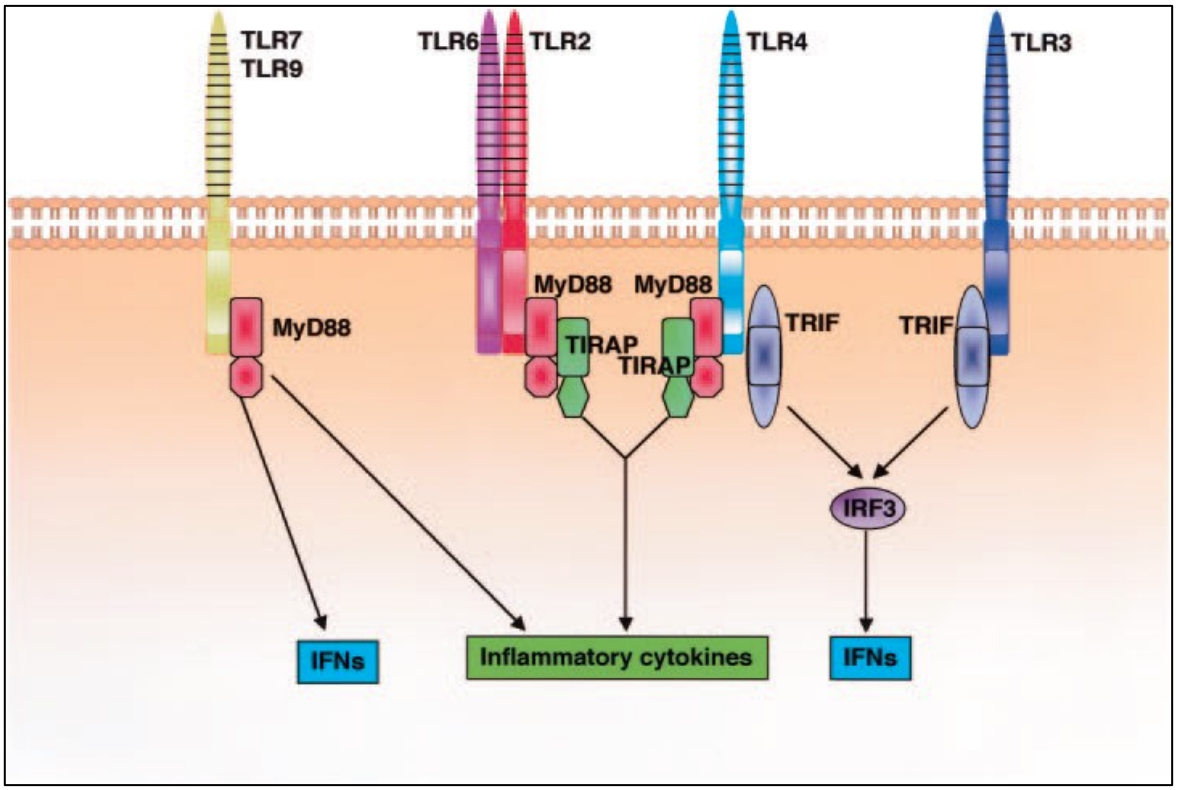


Figure 2: Adaptors involved in TLR signalling. With the exception of the TLR3 ligand, MyD88 is required for the generation of inflammatory cytokines in response to all TLR ligands. TIRAP/Mal does not participate in the MyD88-independent TLR4 signalling pathway, but it is necessary for the production of inflammatory cytokines that are dependent on TLR2 and TLR4. Both the MyD88-independent TLR4 signalling pathway and TLR3 signalling depend on TRIF. Other adaptor(s) may be involved in the induction of interferons through TLRs other than TLR7 and TLR9 (*Takeda and Akira 2005)*.

Structural Biology of Toll-like Receptors (TLRs)

Type I integral membrane receptors, TLRs have three distinct domains: a single transmembrane helix, a C-terminal cytoplasmic signalling domain and an N-terminal ligand recognition region (*Bell et al., 2003*). Because they resemble the signalling domains of members of the IL-1R family, the signalling domains of TLRs are referred to as Toll IL-1 receptor (TIR) domains. TIR domains are also present in a large number of adaptor proteins, which initiate the signalling cascade by homotypic interaction with TLR and IL-1 receptor TIR domains. Each TLR transmembrane domain has a normal stretch of 20 uncharged, primarily hydrophobic residues in it. Through their transmembrane domains, TLRs that identify PAMPs in nucleic acids interact with UNC93B, a multispan transmembrane protein that guides these TLRs to endocytic compartments. The remaining TLR paralogs pass straight to the cell surface and do not engage in interaction with UNC93B. With 550–800 amino acid residues, glycoproteins make up the N-terminal ectodomains (ECDs) of TLRs (*O'Neill & Bowie, 2007*). These ectodomains are found in endosomes or extracellular environments, where they come into contact with and identify chemicals secreted by invasive infections.

Leucine-Rich Repeats (LRRs) - the building blocks of TLRs

The LRRs usually 22–29 residues long and they contain hydrophobic residues set apart at specific intervals. TLR ECDs are made of tandem copies of such repeats (Figure 3A). Various proteins in plants, animals and microbes contain this motif, including many proteins involved in immunological recognition (*O'Neill & Bowie, 2007*). Recent review reported that all LRRs adopt a loop structure in three dimensions, starting with an extended stretch with three residues in the β strand configuration (*Bella et al. 2008*) (Figure 3B). While getting assembled into a protein numerous succeeding LRRs produce a solenoid structure where the β strands are aligned to form a hydrogen bonded parallel β sheet and the consensus hydrophobic residues point to the inside to form a stable core. The β sheet forms the concave surface of the solenoid, forcing it into a curved structure because the β strands of the LRR loops are more densely packed than the non-β parts (*Kajava, 1998*) (Figure 3C). Each LRR protein comprises four

surfaces: a concave surface, a convex surface, an ascending lateral surface made up of loops connecting the β strand to the convex surface and a descending lateral surface on the other side (*O'Neill and Bowie, 2007*).

Ribonuclease inhibitor (RI) was the first LRR protein structure to be described (*Kobe & Deisenhofer, 1995*). This protein has comparatively long LRRs with an average length of 27–29 amino acids. Each LRR contain three to four a-helix turns on its convex surface, opposite the b sheet. The 16 LRRs in RI form a "horseshoe"-shaped structure (Figure 3C). Like RI, the 19–25 LRRs that make up the TLR-ECDs also form horseshoe structures. Unlike RI, the consensus LRR of the TLRs is 24 residues long (Figure 3A), preventing the development of multi-turn helices on their convex sides. On their convex sides, the 24-residue consensus LRRs take on a variety of configuration, often containing bits of secondary structure like β strands, 3_{10} helices, and polyproline II helices (*Botos et al. 2011*).

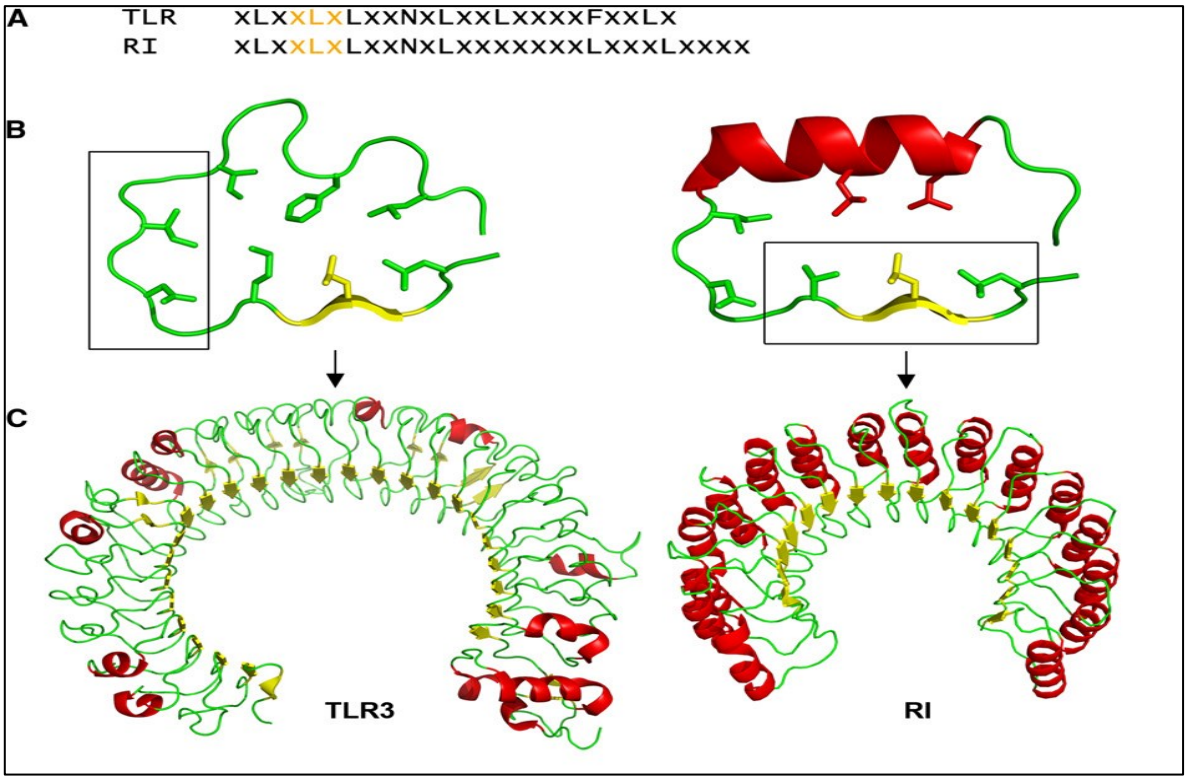


Figure 3: The Structure of Leucine-Rich Repeats. (A) LRR consensus sequences for TLR3 and ribonuclease inhibitor. (B) A LRR loop from hTLR3 and a LRR loop from RI, with the conserved residues forming a hydrophobic core. The boxed regions form the surfaces involved in ligand binding. (C) Ribbon diagram of TLR3 (2A0Z) and ribonuclease inhibitor (*Botos et al. 2011*).

TLR-ECDs are distinguished by the prevalence of LRRs that are significantly larger than the consensus 24 residues, particularly in TLRs 7, 8, and 9. These additional residues frequently form loops that protrude from the TLR-ECD horseshoe, typically on the ascending or convex side of the LRR (Figure 3B). The TLR-ECDs also have structures that cap the N and C-terminal ends, known as the LRR-NT and LRR-CT motifs (Figure 4). The LRR-NTs are disulphide-linked b-hairpins, whereas the LRR-CTs are globular structures with two helices held together by two disulphide bonds. Similar capping motifs have been found in numerous additional proteins with 24-residue LRRs (*He et al. 2003; Huizinga et al. 2002*). Most ligands bind on the concave surfaces of LRR proteins. In contrast, ligand binding is most frequently observed on the ascending lateral surface of the TLR-ECD (*Jin et al. 2007; Kang et al. 2009*) (Figure 4). This surface particularly lacks N-linked glycan and is therefore are free to interact with a ligand.

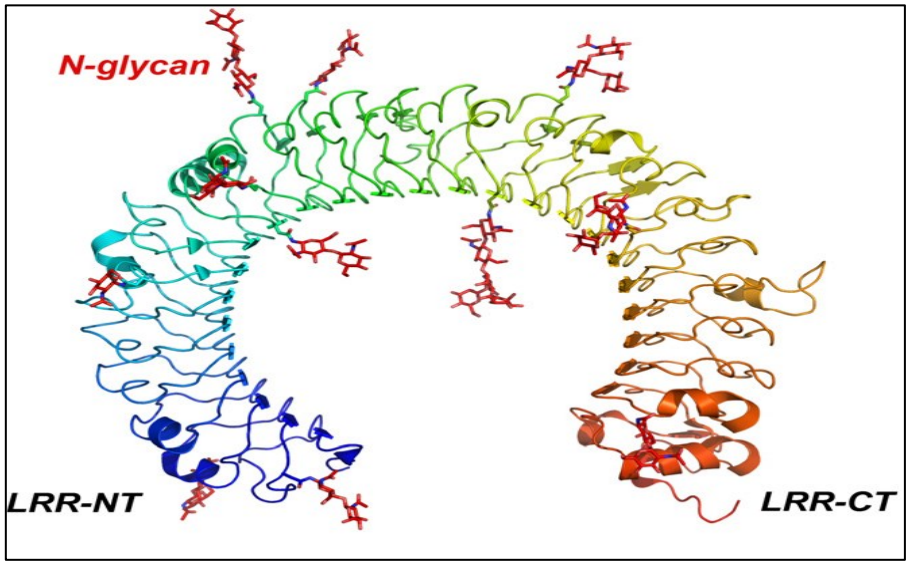


Figure 4: The Structure of a TLR-ECD (hTLR3). Top and side views of the TLR3-ECD, with the N-linked glycosyl moieties (2A0Z). The LRRs are capped by the LRR-NT and LRR-CT motifs (*Botos et al. 2011*).

TLR	Residues	LRRs[a]	N-Linked Glycosylation Sites[b]	Accession Code
1	786	19	4 (7)	Q15399
2	784	19	3 (4)	O60603
3	904	23	11 (15)	O15455
4	839	21	5 (10)	O00206
5	858	20	(7)	O60602
6	796	19	8 (9)	Q9Y2C9
7	1049	25	(14)	Q9NYK1
8	1041	25	(18)	Q9NR97
9	1032	25	(18)	Q9NR96
10	811	19	(8)	Q9BXR5

LRR, leucine-rich repeat; TLR, Toll-like receptors.
[a] The number of LRRs in the extracellular domain do not include the LRR-NT or LRR-CT motifs.
[b] Number of N-glycosylation sites observed in the crystal structure or predicted by the NetNGlyc server 1.0 in parentheses (http://www.cbs.dtu.dk/services/NetNGlyc/).

Figure 5: The main features of ten Human TLR molecules (*Botos et al. 2011*).

Structure of TLRs

Based on sequence homologies, vertebrate TLRs can be divided into six subfamilies: TLR1/ TLR2/TLR6/TLR10, TLR3, TLR4, TLR5, TLR7/TLR8/ TLR9, and TLR11/ TLR12/TLR13/ TLR21/TLR22/TLR23 (*Matsushima et al., 2007; Roach et al., 2005*). TLR paralogs are not expressed by all vertebrate species. For example, humans lack all TLR11 family members. ECDs of the ten human TLRs differ in terms of LRR counts and N-linked glycosylation. To date, the ECD structures of TLRs TLR1, TLR2, TLR3, TLR4 and TLR6 (human or mouse) have been published. All ECDs have the usual horseshoe form, the structures cannot be superimposed due to variances in curvature. Glycans are spread throughout the molecule in the known structures, with the exception of the lateral face produced by the ascending loops of LRRs. This glycan free face participates in dimerization upon ligand binding in known TLR-ligand complexes.

TLR2 is located on the plasma membrane and responds to lipid containing PAMPs such as lipoteichoic acid and di and triacylated cysteine containing lipopeptides (*Takeda et al., 2003*). It accomplishes this by creating dimeric complexes with either TLR1 or TLR6 at the plasma membrane. The TLR1/2 complex recognises tri-acylated lipopeptides like Pam_3CSK_4, while

the TLR2/6 complex recognises the di-acylated ligand, Pam_2CSK_4. According to phylogenetic studies, TLR10 belongs to the TLR-1 family (*Roach et al., 2005*).

Lipopolysaccharide (LPS), an essential component of outer membrane Gram-negative bacteria, causes a strong inflammatory response that can result in septic shock and mortality (*Beutler and Rietschel, 2003*). LPS communicates with TLR4 via the complexing coreceptor MD-2, which is bound to the lateral and concave surfaces TLR4 ECD by numerous hydrogen bonds. TLR5 is one of the few TLRs that recognises the protein PAMP bacterial flagellin (Hayashi et al., 2001). It is highly expressed in the gut, particularly in lamina propria dendritic cells (*Uematsu and Akira, 2009*), where it regulates microbiota composition (*Vijay Kumar et al., 2010*).

TLR3 recognises dsRNA, which is produced by most viruses at some stages during their life cycles and is a strong indicator of viral infection. TLR3, unlike numerous other cytoplasmic dsRNA receptors, is localised to endosomes and recognises dsRNA there. Like TLR3, members of the TLR7, TLR8, and TLR9 subfamilies are found in endosomes and recognise nucleic acid PAMPs. However, the amino acid sequences indicate that the architectures of ECD of TLR7- TLR9 differ significantly from TLR3 (*Bell et al., 2003*).

Evolution of toll-like receptor (TLR) genes

TLR diversity has been seen among species as well as within individuals, in recognition and downstream signalling pathways. This information can be valuable in determining how infectious diseases spread between species. Understanding the development of TLR genes across animals can provide us with a comprehensive understanding of changes in ligand detecting properties and host-pathogen interactions (*Miller et al. 2005*). Immunologists and evolutionary biologists are particularly interested in genetic diversity in functional immunity-associated genes like TLRs because they provide a good model for studying the selection pressure exerted by microbes on the host genome (*Quintana-Murci et al. 2013*). In response to ever-changing pathogens, these genes appear to evolve more quickly than other locations in the genome.

Evolutionary rate of a gene is expressed as the ratio of its nonsynonymous substitutions to its synonymous substitutions, and it reveals the selection constraints that act on genes. The ratio will indicate either positive or purifying selection. In comparison to mutation, selection is the dominating mechanism governing the rate of evolution of TLRs and TLRs are under strong selection for maintaining their activities (*Roach et al. 2005*). The innate immune response varies amongst mammals and TLRs change between species (*Jungi et al. 2011*). This variability is due to selective pressure on immunity-related genes, which reflect the specific conditions encountered by each species (*Zhang et al. 2010*).

Phylograms show that TLRs cluster in close species such as primates, rodents, ruminants, and cetaceans, as expected given the conserved nature of TLR sequences. Despite evolutionary restrictions, TLR evolution has resulted in variations and mutations within or between different species. A molecular tree representing all full vertebrate TLRs in GenBank has been created. The molecular tree shows six key families that encompass nearly all vertebrate TLRs (Figure 6). TLRs within a family recognise a broad class of pathogen-associated molecular patterns (PAMPs) common to that family (*Roach et al., 2005)*.

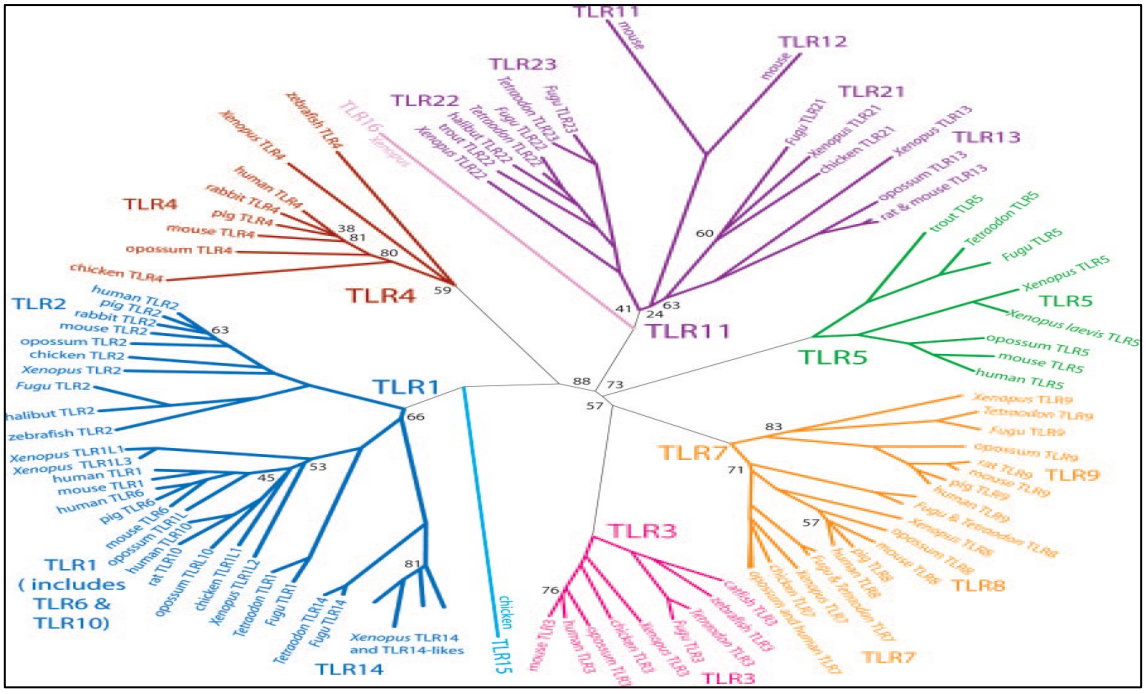


Figure 6: Molecular tree of the vertebrate TLR. Branches of each major family are shown with colors (*Roach et al., 2005)*.

The TLR2 subfamily (lipopeptide), the TLR3 family (dsRNA), the TLR4 family (LPS), the TLR5 family (flagellin), and the TLR7 to TLR 9 subfamilies (nucleic acid and heme motifs) have all been dominated by selective pressure, most likely to maintain unique PAMP recognition. TLR1, TLR2, TLR6, TLR10 and TLR14 are all members of the lipopeptide PAMP-specific TLR family. This family, like the other TLR families, has evolved through strong selection, although it has additional species-specific adaptations. The TLRs of the TLR1 family work as heterodimeric receptors, with TLR2 paired with other member of TLR1 subfamily. Because it evolved in tandem with species phylogeny, the TLR2 subfamily appears to be subject to increased selection pressure.

The TLR14 subfamily in fish might have been lost in amniotes but extended in amphibians. Since TLR14 is connected to the TLR1 subfamily, it has been hypothesised that it also interacts with TLR2. TLR15 in chicken are distant molecularly from all other TLRs. It could be resulting from the TLR1 family. Major family remained includes the TLR11-TLR13 and TLR21-TLR23 subfamilies is characterized in humans only through a pseudogene. The major divides of the TLR11 family are evidently very ancient, as most TLR11 subclades include representatives from fish and frogs, although TLR11 appears to recognise uropathogenic bacteria. The TLR16 subfamily, which is molecularly distinct from all other TLRs, may belong to the TLR11 family.

TLR11 family contain more subfamilies with respect to other family, and it has comparable diversity to the TLR1 family. Also, it includes mouse TLR11 and TLR12, the most diverse vertebrate TLRs. Therefore, the TLR11 family may face fewer purifying selection than other TLR families. The considerable diversity of TLR11, TLR12 and TLR16 could possibly ambiguous orthology for TLR21, TLR22 or TLR23. The TLR11 family has a similar number and diversity of subfamilies to the TLR1 family, which may indicate that TLR11 family members function as heterodimeric partners with each other (*Roach et al., 2005; Areal et al. 2011*).

TLRs are a class of conserved pattern recognition receptor that initiate innate and acquired immune responses. Because the TLRs play an important role in host defence, such genes

developed increasing interest in the evolutionary and population genetics literature, with variation representing a possible target of adaptive evolution. Though importance of selection that are pathogen mediated (i.e. episodic positive selection) need to be studied as these genes are not understood and not well explored in species mammals. Currently increasing bird species for which TLR sequences are available allowed investigation of the selective processes that shaped the development of the known avian TLR genes. It has been evaluated for episodic positive selection in order to find codons that have undergone purifying selection for the majority of their evolution, scattered with bursts of positive selection that may only occur in specific lineages. Genes with sequence coverage that encompassed both the extracellular leucine-rich repeat region (LRR) and intracellular domains of protein showed greater positive selection in the extracellular domain. It was reliable with theoretical estimates. These findings suggest that episodic positive selection had a significant role in the evolution of most avian TLRs, which is consistent with the loci's involvement in pathogen identification and a host-pathogen coevolution mechanism (*Grueber et al. 2014*).

The innate immune system is the first line of host defence against infections. TLRs play crucial roles in the innate immune system by recognizing molecules derived from pathogens. Studies have showed evidence that TLR-related genes have been subjected to natural selection during primate evolution. Analysis of the nucleotide sequences of 16 TLR-related genes, including TLRs (TLR1-TLR10), MYD88, TILAP, TICAM1, TICAM2, MD2 and CD14 from seven primate species. 16 TLR-related genes, included ten TLRs (TLR1–10), four genes linked to signal transduction (MYD88, TILAP, TICAM1, and TICAM2) and two genes linked to TLR4 (MD2 and CD14) in primates. MD2 and CD14 are key molecules of the LPS signaling through TLR4 (Poltorak et al. 1998; Shimazu et al. 1999; Nagai et al. 2002). The analysis of the non-synonymous/synonymous substitution ratio revealed that TLR-related genes contain both strictly conserved and rapidly evolving regions. Genomic regions of Toll/interleukin 1 receptor domains having lower frequencies of nonsynonymous substitution have undergone purifying selection. In contrast, TLR4 has a large fraction of non-synonymous changes in the extracellular domain spanning 200 amino acids, was discovered to be a likely target of positive

Darwinian selection in primate evolution. However, sequence analyses of 25 primate species, including eight hominoids, six Old World monkeys, eight New World monkeys and three prosimians found no evidence that positive Darwinian selection influenced the pattern of TLR4 sequence variations among New World monkeys and prosimians. This study revealed the molecular evolution of TLR-related genes in primates and determined that while natural selection did impact the sequence patterns of TLR-related genes during primate evolution, positive selection pressure was limited across the TLR family (*Nakajima et al. 2008*).

Studies have been conducted on how natural selection has worked on human TLRs in order to estimate the redundancy in their biological level. Sequencing of ten human TLRs in a group of 158 entities from different populations around the world, and it was discovered that intracellular TLRs activated by nucleic acids and predominantly specialised in viral recognition evolved under strong purifying selection, indicating their essential non-redundant role in host survival. Conversely, the selection restrictions on TLRs expressed on the cell surface and activated by substances other than nucleic acids have been significantly more relaxed, with larger frequencies of harmful nonsynonymous and stop mutations permitted, indicating greater redundancy. Finally, it was investigated if TLRs have undergone spatially varied selection in human populations, and it was discovered that the region comprising TLR10-TLR1-TLR6 has recently been the target of positive selection among non-Africans. Study data show that the immunological redundancy of the individual TLRs varies, indicating their unique contributions to host defence. These findings encourage the development of novel concepts for clinical and epidemiological genetics of infectious diseases (*Barreiro et al. 2009*).

Immunologists and evolutionary biologists are particularly interested in genetic variation in functional immunity-associated genes like TLRs because they provide an excellent model for studying the selection pressure exerted by microbes on the host genome (*Quintana-Murci & Clark, 2013*). In response to continuously evolving pathogens (*Lively & Dybdahl, 2000; Kuijl & Neefjes, 2009*), these genes appear to develop quicker than other locations in the genome (*Khakoo et al. 2000*; *Zelus et al. 2000; Sachidanandam et al. 2001; Downing et al. 2009*). The evolutionary rate of a gene is denoted as the ratio of its nonsynonymous substitutions to its

synonymous substitutions and it reveals the selection constraints that act on genes. The ratio will indicate either positive or purifying (stabilising) selection. In comparison to mutation, selection is the dominating force in regulating the rate of evolution of TLRs, and TLRs are subjected to intense selection to maintain their activities (*Roach et al. 2005*). The innate immune response is not the same in all animals and there is species wise variation in TLRs (*Jungi et al. 2011*). This variability is due to selective pressure on immunity-related genes, which reflect the unique conditions encountered by each species (*Zhang et al. 2010*).

For many years, TLR genes were assumed to be ideal functional candidates for increasing susceptibility or resistance to infections and inflammatory disorders. In recent years increased focus has been dedicated to understanding the precise function of these receptors. To determine the function of TLR polymorphisms in infectious disease susceptibility, relationships between various studies and populations needs to be accumulated. Various molecular phylogenetic investigations have revealed that the evolution of both cell-surface and intracellular TLRs in various species follows an almost unique paradigm. The majority of research have confirmed purifying selection as the principal force acting on TLRs. However, positive selection signatures have been identified in all TLRs from various species. The majority of the positively selected sites were located in cell-surface TLRs rather than intracellular TLRs, demonstrating the conserved characteristics of viral PAMPs recognised on intracellular TLRs versus fast escaping bacterial PAMPs detected on cell surface TLRs. Thus, viral infections are expected to have a stronger selection force on TLRs than bacterial infections. Pathogen mediated positive selection has shaped variety in mammalian TLRs. Furthermore, the selective divergence of TLRs in particular species was most likely caused by the diverse pathogenic environments that they experience. Positively selected settings are intended to improve species adaption in new environments. In other words, differences in selection limitations influence the ability of TLRs to recognise and respond to specific pathogenic profiles in their respective niches (*Priyam et al. 2018*). Various studies have suggested mostly similar trend for TLR evolution among different species, few studies on a wider range of mammalian species finds a

contradiction. It was revealed that both viral and non-viral TLRs are subject to positive selection owing to the inclusion of a broader range of species impacted by various diseases.

Although many articles have recommended a typically similar pattern for TLR evolution between diverse species, inconsistency found from the study on a large group of mammalian species. Similarity in positive selection among viral and non-viral TLRs was not aligned with preceding studies. Inclusion of greater number of species group might have affected such observations. Most of the previous studies mentioned homogeneous group of species probably get affected by a restricted number of similar viruses. Possibility of removal of non-synonymous fatal mutations by purifying selection and fixation of beneficial mutations might have caused the differences among these mammalian species (*Roach et al. 2005*). Perhaps, the extensive difference among the mammals under study, their surroundings and interaction with viruses accounted higher positively selected sites observed in viral and non-viral TLRs. Orthologs TLR share sequence and structural similarities and recognise nearly identical forms of PAMP in different species (*Keestra et al. 2007*), there are certain structural differences between TLRs and their signal transduction pathways that result in functional variability (*Bagheri & Zahmatkesh, 2018*).

Genes carry biological functions through pathways in complicated networks involving many interacting components. Studies on the effect of network design on the evolution of individual proteins aid to the understanding of the creation and evolution of signalling pathways, as well as their functional conservation. However, the relationship between network architecture and individual protein sequence evolution is still poorly understood. A network-level molecular evolution analysis was performed on the TLR signalling pathways, which is critical for innate immunity in insects and humans. It has been found that the selection constraint of genes was negatively correlated with its position along TLR signaling pathway. All genes in the TLR signalling system were highly conserved and experienced substantial purifying selection. Different nonsynonymous substitution levels determined the distribution of selective pressure throughout the pathway. The TLR signalling pathway may have existed in a common ancestor of sponges and eumetazoa, and it evolved through the TLR, IKK, IkB, and NF-kB genes, which

underwent duplication events as well as adaptor molecular enlargement, and the gene structure and conservation motif of NF-kB genes shifted throughout their evolutionary history. These findings will help us better understand the evolutionary history of the animal TLR signalling system, as well as the relationship between network design and protein sequence evolution (*Song et al., 2012*).

TLRs that initiate innate immune response have two domains: an external leucine rich repeat (LRR) and an intracellular Toll IL-receptor (TIR). LRR domains with a solenoid configuration typically evolve faster than TIR globular domains. It is critical to understand the molecular evolution and functional activities of TLRs in this context. Study of pairwise genetic distances and Ka/Ks ratios (the ratios of non-synonymous to synonymous substitution rates) between the LRR and TIR domains of vertebrate TLRs from various species (ranging from fish to primates) was performed. Among them (TLR1, TLR2, TLR3, TLR4, TLR5, TLR6, TLR7, TLR8, TLR9, TLR11/ TLR12, TLR13, TLR14, TLR21 and TLR22/ TLR23) the LRR domains evolved substantially faster than the corresponding TIR domains. The evolutionary rates of the LRR domains vary greatly across these members; LRR domains from TLR3 and TLR7 from primates to fishes have the slowest rate of evolution. In contrast, the fifteenth member, TLR10, exhibits no major alterations; its TIR domain is not well conserved (*Mikami et al. 2012*).

Despite the important of birds in vertebrate evolution, less attention has been given to their immune systems. The evolution of TLR genes has been studied in many species, but our understanding of the evolutionary properties of TLR genes in birds in the wild is restricted. Most studies focused on the structure, variation, and composition of a single gene or the analysis of selection pressure on individual genes, but neither examined the influence of the external environment or feeding habits on the evolution of avian TLR genes. The growth of avian genome data and the advancement of molecular biology in recent years have created a new opportunity for us to investigate the relationship between the adaptive evolution of birds' TLR genes and their external environment (*Velová et al. 2018*).

The phylogenetic data suggested that TLR1A and TLR1B may have differed functionally. A systematic analysis of bird TLR genes, as well as phylogenetic analyses, revealed that the TLR1 and TLR2 subfamilies diverged due to duplication. TLR1A is more closely related to TLR10 in mammals, implying that functional differentiation occurred, but not TLR2. Evolutionary study revealed that TLR genes in birds are subjected to significant purifying selection. Common positively selected codons were identified in ten avian TLR genes, with the most of sites found in the extracellular leucine-rich repeat (LRR) functional domains. The evolution of avian TLR genes was influenced by both the environment and feeding habits. Environmental stresses showed a stronger impact on TLR1B, TLR2B, TLR3 and TLR4, whereas feeding habits influenced TLR2A, TLR2B, TLR15 and TLR21. Combined with branch-site model analysis, it was discovered that habitat and feeding patterns were external variables driving the evolution of avian TLR genes, with the environment having the greatest influence. These findings revealed that TLR genes were subjected to diversified selective pressures during avian evolution, allowing them to respond differently to infections from various sources (*Yang et al. 2021, Huang et al. 2011*).

TLRs found in fish have been demonstrated to be ligand specific for TLR2, TLR3, TLR5M, TLR5S, TLR9, TLR21, and TLR22. Some research suggests that TLR2, TLR5M, TLR5S, TLR9 and TLR21 can particularly recognise PAMPs from bacteria. TLR1, TLR4, TLR14, TLR18 and TLR25 may also be bacterial sensors. TLR signalling mechanisms in fish differ from those in mammals. TLRs found in fish have direct evidence of ligand specificity. In-depth investigations need to be conducted on a constant basis to determine the ligand specificity of all TLRs in fish, particularly non-mammalian TLRs, as well as their signalling pathways. The identification of TLRs and their functions will add to the knowledge of disease resistance mechanisms in fish, as well as new insights for therapeutic intervention to modify immune response (*Fink et al. 2016; Zhang et al. 2014*).

Significance of TLRs

TLRs have an important role in innate and adaptive immunity. Their capability to detect endogenous DAMPs and exogenous PAMPs allows them to produce ligand mediated signal transduction, which is ultimately involved in the inflammatory response. In recent years, there has been a growing evidence directing to the importance of TLRs and their ligands in a variety of pathological conditions including inflammation, cancer and autoimmune disorders. Remarkably, they have a crucial role in immunotherapy and vaccination (*Vidya et al. 2018)*.

Studies have shown that TLR4 promotes injury in the liver, kidney, heart, and brain. Downregulation of TLR2, TLR4 or MyD88 in ischemia damage lowers myocardial inflammation. TLR4 has also been linked to an enhanced T cell response in burn injuries, graft inflammation, sterile damage and alloimmune responses in tissue transplantation. The overexpression of TLR2 and TLR4 on immune and other cells during sepsis has been linked to organ tissue harm. Many scientific investigations have suggested a function for TLRs in hypercholesterolemia-induced vascular damage. While it was recently established that TLR2 is substantially pro-atherogenic, TLR3 was found to be involved in the integrity protection of the of the blood vessel wall.

Response of TLR is important in tissue damage and subsequent tissue repair and regeneration, especially in the liver and intestinal epithelium. TLR2 signalling has a crucial role in wound healing. TLRs on epithelial cells detect microbial patterns and induce innate immune responses, aiding in homeostasis management. The basal layer of corneal epithelial cells expresses TLR4 and TLR5. When a break occurs in the squamous epithelium, ocular inflammation and keratitis are induced via the MyD88 dependent pathway by functioning TLR2, TLR4, and TLR9, all of which are expressed in the corneal epithelium.

Recent research has shown that endogenous TLR ligand-mediated signalling plays a key role in auto-immune diseases. The presence of bacterial DNA and peptidoglycans in the joints of people with rheumatoid arthritis (RA) and other diseases, which may increase synovial inflammation via TLR ligand-mediated signalling. TLR9 and TLR7 have also been shown to

have a role in the persistence of systemic lupus erythematosus. TLR9 detects danger signals generated by demyelinated nerves, which trigger a pathologic immune response to autoantigens in multiple sclerosis. Endogenous monosodium-urate monohydrate (MSU) crystals generated from uric acid secreted by injured cells act as DAMP, activating TLR2 and ultimately causing cartilage degradation.

TLRs have been shown to play both positive and negative functions in tumorigenesis. Though, to date, TLRs have had the opposite effect on tumour growth. TLR ligands can supress tumour growth, whilst TLR agonists can improve malignant cell survival and resistance to chemotherapy. TLRs play an important role in cancer immunotherapy. Total body irradiation (TBI) increases the activation of adaptively transplanted T lymphocytes by recognising microbial LPS by TLR4 activating innate immune system in the radiation injured gut.

TLRs play an important role in vaccinations because they act as natural adjuvants for vaccines containing attenuated live or heat-killed viruses or bacteria. TLRs play an significant role in controlling the adaptive immune response by maturing DCs, inducing the production of cytokines and co-stimulatory proteins, and reversing tolerance. As a result, as natural adjuvants in vaccines, they help DCs in better antigen presentation, resulting to a positive immune response (*Bagheri et al. 2018, Vidya et al. 2018*).

TLRs are evolutionary conserved proteins, characterization of TLRs and their ligands has contribution in understanding their function and the host defence systems against infections. To study the impact of natural selection on innate immune receptors TLRs are useful candidate molecules. Several studies were conducted and purifying selection has driven TLR evolution at least in humans. Additional research on primate species have found varying degrees of positive selection acting on their evolutionary history. These interactions may have influenced the evolution of proteins involved in direct pathogen recognition. Further research on mammal TLR genes is needed to explore for signs of positive selection.

Chapter - III

Sequence Retrieval

Sequences of TLR genes and encoding protein from mammals were retrieved from GenBank maintained by NCBI (http://www.ncbi.nlm.nih.gov/genbank/) and Ensembl maintained by EMBL-EBI (www.ensembl.org). To avoid any stochastic disparities and sample errors sequences that are error prone and redundant (partial sequences, predicted sequences, sequences having internal stop codons, non-translatable codons) were discarded (Wright, 1990). TLR nucleotide and protein sequences from different mammalian species were stored according to the TLRs. BLAST and its variants, each differentiated by the type DNA or protein of input sequence and searched database for annotation of gene or protein sequences. More inclusive database search was undertaken by using PSI-BLAST which uses an iterative pattern to search and find out distantly associated sequences. A comprehensive set of coding sequences of TLR1-TLR13 from Mammalian group constituted primary dataset for the analyses.

Multiple sequence alignment aligns many related sequences to get the best possible sequence matching. Multiple sequence alignment has the unique advantage of revealing more biological information than several pairwise alignments. As example, it enables the detection of conserved patterns of sequence and motifs across the entire family of sequence, that would otherwise be difficult to notice while comparing two sequences. A protein multiple alignment reveals several conserved and functionally important amino acid residues. Multiple sequence alignment is also required for sequence family phylogenetic analysis as well as protein secondary and tertiary structure prediction. Clustal Omega package have been developed for performing multiple sequence alignments (MSAs) to deal with large number of sequences available and the to execute big alignments rapidly and precisely.

Multivariate Analyses

Species and genes within the same genome use codons and amino acids at different frequencies. Numerous studies have been conducted on these biases in codon and amino acid usage in a range of species. Despite the fact that the genetic code is degenerate, meaning that multiple combinations of codons can produce the same protein. The mechanisms that determine non-random codon usage may also have an impact on amino acids usage in proteins. Since all codons encoding a particular amino acid may have base compositions that are either GC rich or GC poor, this can be explained by neutral processes. Furthermore, because amino acids identical functions might have varying tRNA abundances or necessitate diverse metabolic expenditures to produce, selection may be a significant factor in determining amino acid frequencies. The pattern of amino acid usage is primarily determined by the composition of the genomic bases. However, additional parameters like hydrophobicity, aromaticity, gene function, etc., have also been found to have an impact on amino acid usage (*Peden, 2000*).

Multivariate analysis (MVA) simplifies rectangular matrices in which the columns denote measurement of codon usage or amino acid usage and the rows denote specific genes. Meanwhile amino acid usage is multivariate in nature, such statistical techniques like correspondence analysis (COA). COA ordination identifies key trends data variation and distribute genes along continuous axes in according with trends. It is advantageous as it do not make any assumption of clustering the data rather distribute continuous variation correctly (*Peden, 2000*).

CodonW package analyse codon usage. It facilitates COA, a popular MVA technique for analysing codon usage. CodonW can produce a COA for codon usage, relative synonymous codon usage and amino acid usage. Additionally, codon usage analyses include investigation of optimum codons, codon and bias in dinucleotide, and base composition. CodonW examines sequences encoded using genetic codes other than the universal code (*Peden, 2000*). COA was used to explore the major trend in amino acid usage difference among the TLR genes from Mammals. For each gene, relative amino acid usage (RAAU), average hydrophobicity,

aromaticity and GC content of the TLR gene sequences were calculated employing the CodonW program.

Phylogenetic Tree

Phylogenetic tree analysis determines the ancestral relationship of a collection of sequences. Phylogeny refers to the patterns of tree branching that show evolutionary divergence. Graphical depiction of the evolutionary relationships amid biological entities such as sequences or species is presented through a phylogenetic tree. Relations among entities are apprehended by the topology or branching order and expanse of evolutionary change (branch lengths) between nodes. Root adds direction to such relationships and precisely define ancestry.

Molecular phylogenetic trees are generated through either nucleotide or protein sequences. The most important phase in the technique is to generate sequence alignment, which ascertains positional correspondence in evolution. Only the accurate alignment produces proper phylogenetic inference as aligned positions are probably related genealogically. Improper alignment causes methodical errors in the resulting tree, or sometimes entirely an erroneous tree. For this, accurate sequence alignment is essential. Multiple cutting-edge alignment programmes, such as Clustal Omega, Muscle can be used. Results of alignment from various sources should be carefully examined and linked to determine the most rational choice (*Xiong, 2006*).

Currently two major types of tree construction methods exist, with some advantages and limitations. One class of method is based on discrete characters from biological sequences of individual taxa such as maximum parsimony (MP), maximum likelihood (ML). Assumed that corresponding positional characters at in a multiple sequence alignment are homologous across all the involved sequences. Consequently, the dataset can be used to reconstruct character states of the common ancestor. Also, it is assumed that each character evolves independently hence is viewed as a separate evolutionary unit. The second category of phylogenetic methods such as Unweighted Pair Group Method Using Arithmetic Average (UPGMA), Neighbor Joining are distance based which report amount of dissimilarity among pairs of sequences estimated

using sequence alignment. Distance based approaches presume all sequences as homologous and tree branches are additive, which means that the distance amid two taxa is equal to the total of all branch lengths that connect them (*Xiong, 2006*).

Bootstrapping, a statistical procedure used to test any sampling errors in the phylogenetic tree. Repeated sampling of trees by the perturbation of dataset is done while bootstrapping. It achieves this by periodically sampling trees from marginally perturbed datasets. This allows us to analyse the robustness of the original tree. Bootstrapping is used to avoid bias in newly constructed trees caused by poor alignment or random variations in measurement of distances. The robustness of the tree constructed by generating a little modified alignment frequently with random fluctuations. Rally strong phylogenetic relationship should include sufficient features to support the relationship even if the dataset is disrupted in such a way. Or else, the noise generated during the resampling procedure is sufficient to produce alternative trees, implying that the initial topology was formed from weak phylogenetic evidence. This form of study provides a sense of the statistical confidence of the tree topology (*Xiong, 2006*).

In this study, all the TLR proteins from mammals were subjected to alignment using the Clustal Omega program (https://www.ebi.ac.uk/Tools/msa/clustalo/). The ensuing multiple sequence alignments were then used to construct the phylogenetic tree with 1000 bootstrap replicates. The latest version of MEGA software was used for Phylogenetic analysis. The Molecular Evolutionary Genetics Analysis (MEGA) software is a desktop application that allows user to compare homologous gene sequences from different species or multigene families, with a focus on inferring evolutionary relationships and patterns of DNA and protein evolution. In addition to the tools for statistical analysis of data, MEGA provides many convenient facilities for the assembly of sequence data sets from files or web-based repositories, and it includes tools for visual presentation of the results obtained in the form of interactive phylogenetic trees and evolutionary distance matrices (*Kumar et al., 2016, Kumar et al., 2018*).

Determination of the evolutionary history of genes can be done by ancestral sequence reconstruction. Aside from its use in determining the most likely evolutionary forebears of

present proteins, ancestral sequence reconstruction has proven to be an effective method for designing extremely stable proteins. Recently, various computational tools were developed that make ancestral reconstruction algorithms available to the community while leaving the most important parts of input data preparation to users. FireProtASR attempts to tackle this challenge by developing a fully automated procedure that allows even inexperienced users to acquire ancestral sequences using only a sequence query as input (*Musil et al, 2021*). FireProtASR comes with an interactive, user-friendly web interface and is freely available at https://loschmidt.chemi.muni.cz/fireprotasr/.

Evolutionary rate analysis

The neutral theory of molecular evolution states that random fixation of low fitness consequence mutations, not natural selection, is the primary cause of the diversity found within and across species. The morphology, behaviour, and physiology of species are ultimately shaped by these favourable mutations, which are infrequent at molecular level yet occur in genes and genomes. Finding molecular adaptation aids in improving comprehension of the evolutionary process. Enormous genomic data and computational resources has made it possible to the systemic analysis of genomes for positive selection study, making molecular adaptation research more fascinating than ever. Genes that encode protein, can be distinguished between synonymous or silent substitutions (nucleotide changes that do not modify the translated amino acid) and nonsynonymous or replacement substitutions. Because natural selection functions primarily at the protein level, synonymous and nonsynonymous mutations face extremely different selective forces and settle at very different rates. Thus, using the synonymous rate as a reference point, one can determine whether fixation of nonsynonymous mutations in the population is speeded or slowed by natural selection acting on the protein. A comparison of synonymous and nonsynonymous substitution rates can reflect the direction and strength of natural selection acting on the protein (*Kimura 1968; King and Jukes 1969*).

A nucleotide substitution that changes the corresponding amino acid in the protein is called a nonsynonymous substitution (denoted as Ka), whereas a nucleotide substitution that does not

change the amino acid in the protein is called a synonymous substitution (denoted as Ks). According to the neutral theory, purifying selection will eliminate nonsynonymous substitutions while tolerating synonymous ones. As a result, there will be fewer nonsynonymous than synonymous substitutions. This prediction is supported by the facts that synonymous substitutions in protein-coding genes usually exceed nonsynonymous substitutions, and the rate of evolution of functionally constrained regions of genes is slower compared to non-functionally constrained gene regions. Although, selective benefits conferred by the nonsynonymous substitution will be fixed in the population by the positive selection (*Roy et al., 2015, Roy et al., 2017*).

Calculation The ratio (ω) of rate of non-synonymous substitutions per nonsynonymous site (dN) to rate of synonymous substitutions per synonymous site (dS) indicates the impact of evolution on a gene segment. ω>1 indicates diversifying (positive) selection whereas, ω<1 signifies purifying (negative) selection (*Roy et al., 2015*). The evolutionary rates of mammalian TLRs (with reference to consensus sequence generated through Perl program) were calculated using the Codeml program included in the PAML software package (ver. 4.5) (*Nei and Gojobori, 1986; Yang, 2007*) (http://abacus.gene.ucl.ac.uk/software/paml.html) with runmode = −2 and CodonFreq = 1.

Codon-based analyses of positive selection

A gene that has an accelerated nonsynonymous substitution rate, as indicated by the nonsynonymous/synonymous rate ratio dN/dS > 1, is considered to be positively selected. This type of test is very successful at finding diversified or balancing selection because it employs excessive nonsynonymous substitutions as evidence that natural selection aided in the fixation of nonsynonymous mutations. Tests based on dN/dS may be less effective when applied to data from the same species due to lack of sequence divergences and challenges in the interpretation of the dN/dS ratio (*Kryazhimskiy and Plotkin 2008*).

Under neutrality, coding sequences are expected to have a ratio of non-synonymous substitutions (dN) over synonymous substitutions (dS) that does not significantly deviate from

1 (ω = dN/dS = 1), while significant deviations can be attributed to either positive or negative selection (ω >> 1), respectively. To investigate positive selection in individual codons of mammalian TLR sequences, the dN to dS ratios were compared using maximum likelihood (ML) frameworks, specifically the Hyphy programme implemented in the Data Monkey Web Server (http://www.datamonkey.org). Modern comparative sequencing analysis relies heavily on inferring how evolutionary forces shaped genetic diversity. Recent advances in sequence synthesis and statistical approaches enable researchers to extract more evolutionary signals from data, although at a higher processing expense. Datamonkey 2.0, a completely re-engineered web-server for analysing evolutionary signals in sequence data. We used open-source libraries to construct dynamic, robust, and scalable web applications. Datamonkey 2.0 offers curated approaches for analysing coding-sequence alignments for natural selection. It is a responsive, fully interactive, and API-enabled web application (*Weaver et al, 2018*).

The best fitted nucleotide substitution model was identified using the automatic model selection tool Data Monkey Web Server. All TLR sequences were analysed using three distinct models: single likelihood ancestor counting (SLAC) and fixed-effect likelihood (FEL). The SLAC model is based on the reconstruction of ancestral sequences and the counts of dS and dN at each codon position along the phylogenetic tree. The FEL model predicts the dN/dS ratio on a site-by-site basis, rather than assuming a priori distribution across sites. Positive selection is more strongly supported for sites found by two independent approaches. Positive selection test of individual codons of mammals TLR was performed using the Hyphy package executed in the Data Monkey Web Server that compare Ka to Ks ratio using maximum likelihood (ML) framework (*Weaver et al, 2018*).

Structural modeling

Despite the rising proficiency of different approaches to obtain protein sequences, majority of known sequences lack structural information. Protein modeling aims to predict the structure of a protein from its sequence with accuracy comparable to experimental results. This can close the structural knowledge gap in disciplines like structure-based medication design, which

would otherwise rely solely on experimentally determined structures. Furthermore, when experimental methods fail, protein modeling is the only option to gain an understanding of protein structure. Many proteins, for example, are too large for NMR study or are difficult to crystallise using X-ray diffraction methods. Homology modeling, fold recognition, and de novo structure are the available methods for protein 3D structure prediction (*Scott et al, 2014*).

Homology modeling that is also referred as template-based modelling, or comparative modeling assumes that protein three dimensional or 3D structures. Structures with similar amino acids comprise same kind of 3D structure due to structural conservation. This homology modeling process relies on two methods: sequence alignment and molecular modeling. The fundamental workflow for homology modelling starts with a given target amino acid sequence. Initially by searching the homologous sequences in known protein structure databases, alignment process begins. Coordinates of amino acids in homologous proteins with known structure are therefore used to determine corresponding amino acids coordinates of the target protein (*Muhammed and Aki-Yalcin, 2019*). Then, to reduce the unfavorable interactions among amino acid pairs molecular modeling is performed. Finally, the resulting 3D structure is examined. This homology modeling method was one of the prevalent approaches for a decade. Because of the elevated prediction speed, excellent precision for proteins having known structural homologs the homology modeling technique is very advantageous. The flaw is that it heavily relies on template structures, that means it cannot anticipate the structures of proteins for which homologs have not been discovered (*França, 2015*).

SWISS-MODEL is an automated modeling tool and it has been regularly improved since its inception and is now the most popular modeling server available on the web. The SWISS-MODEL server is intended to function with minimum user input, for example it requires only the amino acid sequence of a target protein. Because comparative modeling projects vary in complexity, some may require further user input, such as selecting a new template or adjusting the target-template alignment (*Waterhouse et al., 2018*).

The de novo modeling method searches for conformations directed by a specified energy function, which uses amino acid atomic coordinates as variables. This process generates several potential conformations, and the one with the minimum energy is chosen. The benefits of de novo modeling include the fact that it is independent of identified protein structures. It allows the prediction of protein structures without having any prior knowledge of the structure and the possibility of discovering novel structural types of protein (*Bradley et al, 2005*).

ML-based modeling is an approach for predicting the structures of target proteins using machine learning algorithms and known protein structures. Among the several ML algorithms, the most notable is deep learning (DL). In contrast to homology modeling and de novo modeling, the DL-based method is a data-driven approach that is only recently evolving. Because of the tremendous success of DL in other fields, the DL-based protein prediction strategy is projected to perform better. AlphaFold (AF) is one of several deep learning-based modeling algorithms based on the biological notion of protein structural conservation during evolution (*Yang et al, 2023*).

Molecular docking study

Protein-protein interactions are critical for cellular and immunological function, and in many situations, because the complex structure has not been empirically identified, these interactions must be modeled to gain a better understanding of their molecular foundation. The Molecular Docking approach predicts the interaction of a tiny molecule with a protein or between proteins. This allows researchers to analyse the behaviour of tiny molecules or proteins within the binding region of a target protein and gain a better understanding of the basic biochemical process driving the interaction. The methodology is structure based, requiring a 3D model with high-resolution of the target protein generated using methods such as X-ray crystallography, Nuclear Magnetic Resonance Spectroscopy, or Cryo-Electron Microscopy (*Chen et al, 2003, Agu et al, 2023*).

ZDOCK is a user-friendly protein docking server that uses rigid body docking programmes to predict the structures of protein-protein complexes and symmetric multimers. With the purpose

of offering an accessible and straightforward interface, it offers users the ability to direct the scoring and selection of output models, as well as dynamic visualisation of input structures and output docking models (*Pierce et al, 2014*). After protein-ligand docking is done, the findings are analysed to determine the most desirable candidates for future research. Binding affinity of each ligand is computed using the expected interaction energy, and the ligands are ordered accordingly. The docked structures are also examined to determine important interactions between the ligands and the protein, such as hydrogen bonds, hydrophobic interactions, and electrostatic interactions. These interactions can provide insights into the mechanisms of action of ligands and enable further optimisation of their structure (*Chen et al, 2003, Pierce et al, 2014, Agu et al, 2023*).

Biomolecular interactions between proteins regulate and control nearly every biological function in the cell. Understanding these interactions is thus an essential step in the study of biological systems. Many efforts have been made to understand the principles of protein-protein interactions. The PRODIGY web-server (https://rascar.science.uu.nl/prodigy/), an online tool for predicting the binding affinities of a protein-protein complex based on its three-dimensional structure (*Xue et al., 2016*). It is a basic yet robust binding affinity descriptor based solely on structural characteristics of protein-protein complex, particularly intermolecular interactions. PRODIGY provides binding affinity values as Gibbs free energy (ΔG, kcal/mol) or dissociation constant (Kd, M). PRODIGY measures the number of Interatomic Contacts (ICs) at a protein-protein complex interface within a 5.5 Å distance threshold and classifies them based on the polar/apolar/charged character of the interacting amino acids (*Vangone and Bonvin, 2017*).

Protein stability is one of the most critical elements determining protein function, activity, and regulation. Missense mutations can cause protein dysfunction by altering their stability and interactions with other biological components. Several investigations have found that the mutations are harmful because they reduce or enhance the stability of the corresponding protein. To measure the effects on protein stability, calculation of the changes in folding/unfolding Gibbs free energy caused by mutations is required. The computer prediction

could aid in the prioritisation of possibly functionally significant variations. PremPS, a freely available web-server (https://lilab.jysw.suda.edu.cn/research/PremPS/), forecasts the consequences of stabilising mutations with a very low bias towards anti-symmetric properties (*Chen et al, 2020*).

Protein domain identification

For protein domain identification and analysis InterPro -EMBL-EBI, PROSITE-Expasy, SMART databases were used. SMART (Simple Modular Architecture Research Tool) is a biological database that identifies and analyses protein domains within protein sequences. SMART finds protein domains in protein sequences using profile-hidden Markov models derived from multiple sequence alignments. LRR repeats of individual TLRs were identified using the web interface of SMART (http://smart.embl-heidelberg.de/) (*Schultz et al, 2000*).

The InterPro database (https://www.ebi.ac.uk/interpro/) classifies protein sequences into families, identifying functionally relevant domains and conserved regions. InterProScan is the core software that searches protein and nucleic acid sequences against InterPro signatures. Signatures are prediction models that define protein families, domains, or locations and are available from multiple databases. InterPro combines signatures indicating equivalent families, domains, or sites and includes descriptions, literature references, and Gene Ontology (GO) terms (*Paysan-Lafosse et al, 2023*).

The PROSITE database contains an array of biologically significant signatures, which are classified as patterns for short motif recognition or generalised profiles for sensitive detection of wider domains. Such databases are valuable for predicting protein function, determining family identity, and detecting remote homologues. ScanProsite offers a web interface for identifying protein matches against signatures in the PROSITE database (*Hulo et al, 2006*).

Statistical analysis t-test

Correlation coefficient between variables was calculated using the available formula in Microsoft Excel. Significance test was performed using the freely available online tools such as t-test (https://www.graphpad.com/quickcalcs/ttest1/) and one-way analysis of variance - ANOVA (https://www.socscistatistics.com/tests/anova/default2.aspx).

Chapter - IV

Natural selection on genetic diversity of TLRs

Results presented in this chapter are published in the following article:
Ghosh M, Basak S, Dutta S. Natural selection shaped the evolution of amino acid usage in mammalian toll like receptor genes. Comput Biol Chem. 2022;97:107637. doi:10.1016/j.compbiolchem.2022.107637

Background

The defense system of animal involves two type of immunity adaptive and innate immunity. Initially innate immune system produces an inflammatory response to block the growth and transmission of the pathogen during an infection. In vertebrates, in order to develop acquired immune response particularly receptors of Band T cell sense the infectious agents to produce responses that lead to its exclusion (*Janeway and Medzhitov, 2002*). Receptors associated with innate immune system are germline-encoded. They have been evolved to sense components of external pathogen also referred as pathogen-associated molecular patterns (PAMPs) which are crucial for pathogen existence or host released endogenous components in response to inflammation (*Matzinger, 1994; Yang et al. 2010; Erridge, 2010*). These receptors of innate immune system are located in serum, on cell surface, in endosomes, and in the cytoplasm (*Medzhitov, 2007*).

Being an important category of pattern recognition receptors (PRRs) the toll-like receptors (TLRs) are seen in Drosophila and mammals. Mammal TLRs play fundamental role in detection of pathogen associated patterns with the initiation of signal transduction pathways that cause genetic expression which lead to the innate and adaptive immune responses (*O'Neill, 2009, Rakoff-Nahoum & Medzhitov, 2009*). TLRs are type-I integral membrane receptors comprising an extracellular domain also known as ectodomain (ECD) containing leucine-rich repeats which facilitate the PAMPs recognition, a signal transmembrane segment, and an intracellular Toll–interleukin 1 (IL-1) receptor (TIR) domain for downstream signal transduction (*Bell et al, 2003*). In mammals there are thirteen TLRs discovered in mice (TLR1-13) and ten TLRs in humans (TLR1-10). TLR1-TLR9 is found in both mice and human, TLR10 is non-functional in mouse due to a retrovirus insertion and TLR11, TLR12 and TLR13 are not

found in human (*Takeuchi & Akira, 2010*). Depending on the subcellular distribution TLRs in humans can be classified into two categories: TLR1, TLR2, TLR4, TLR5, TLR6 and TLR10 are expressed normally on the cell surface and TLR3, TLR7, TLR8 and TLR9 are commonly found in intracellular compartments like endosomes. These human TLRs detect various PAMPs such as lipopolysaccharide (TLR4), lipopeptides (TLR2 associated TLR1 or TLR6), bacterial flagellin (TLR5), viral dsRNA (TLR3), viral or bacterial ssRNA (TLRs 7 and 8), and CpG-rich unmethylated DNA (TLR9) (*Akira et al. 2006*).

Genetic diversity in active genes associated with immune defense such as TLRs is interesting from an evolutionary perspective as these genes are an excellent model for studying the selective stress applied to the host genome by pathogen. These genes appear to evolve faster than other loci in the genome in response to pathogen that are evolving rapidly. Selection is a major factor in controlling the evolutionary rate of TLRs, mutation is also another factor and TLRs are strongly selected to maintain their functions. In different mammals innate immune response is not similar as some variation is there between different species in their TLRs. This variation is due to selective pressure on the immune system-related genes that reflect specific conditions experienced by each species (*Bagheri and Zahmatkesh, 2018*). Evolutionary genetics approaches have amplified to understand the evolutionary forces acting on the human genome that provides indispensable complement in treatment of infectious diseases. Within the perspective of infection, detecting the magnitude and pattern of environmental selection that works on the genes implicated in immune-associated procedures can deliver insight into the host defence mechanisms (*Barreiroet al. 2009*).

Amino acids and codons are used in diverse frequencies both between genes and between genes within the same genome. Degeneracy of genetic code direct the use of diverse set of codons for producing the similar protein, procedures that create non-random usage of codons are likely to influence the usage of amino acids. The possible reason behind this is the neutral processes where composition of bases of all codons that encode an amino acid might be either GC rich or GC poor (*Rao et al. 2014*). Selection also has a significant role in determining frequencies of amino acid. Often genomic base compositions play a major role on the type of amino acid

usage; other factors like hydrophobicity, gene function, level of expression etc. also influence the amino acid usage. In this study mammalian TLRs are progressively investigated to examine the effects of environmental selection on diverse set of TLRs and factors that influence selection will be explored. Natural selection on different members of TLRs family will be studied to explore their evolutionary contribution to host defense.

Methodology

Sequence retrieval and multivariate analysis on amino acid usage

Genes and their encoding protein sequences of toll-like receptors (TLR) were taken from GenBank, NCBI (http://www.ncbi.nlm.nih.gov/genbank/) and Ensembl maintained by EMBL-EBI (www.ensembl.org). By nature, amino acid usage is multivariate and need to be explored using statistical analysis like correspondence analysis (COA) (*Peden, 2000*). COA reveals major trends of variation in the dataset by arranging them along continuous axes where consecutive axis have been arranged to have diminishing effect gradually (*Roy et al. 2017*). The analyses of amino acid usage patterns of TLR genes of mammal under study were carried out using COA available in CodonW program.

Parameters like relative amino acid usage (RAAU), average hydrophobicity, GC content of genes were calculated for each TLR sequence using available option in CodonW program. Correlation coefficient between variables was calculated using the available formula in MS Excel. Significance test was performed using the freely available online tool such as t-test (https://www.graphpad.com/quickcalcs/ttest1/).

Phylogenetic analysis was performed among primate and non-primate genes of TLR. The sequences were aligned using the ClustalW program. The phylogenetic tree was constructed using Mega 7, utilizing the maximum likelihood method (*Kumar et al. 2016*).

Three dimensional structural models were generated for TLR5 protein sequences through homology modeling using SWISS-MODEL (*Waterhouse et al., 2018*). TLR5 protein structure available in Protein Data Bank (PDB) (PDB ID: 3J0A) was used as template for homology

modelling with more than 99% sequence identity and 97% query coverage in case of human (primate mammal) and 78% sequence identity and 97% query coverage in case of cattle (non-primate mammal). The structure of flagellin was truncated from crystal structure of the N-terminal fragment of zebrafish TLR5 in complex with Salmonella flagellin available in PDB (PDB ID: 3V47). As the ectodomain of the TLRs are involved in ligand recognition, the interaction study was performed on TLR5 ectodomains based on the NCBI annotation (*Savar and Bouzari, 2014; Forstnerič et al. 2016*). Molecular interaction of TLR5 protein with flagellin was performed using Z-dock software (*Pierce et al. 2014*). Then, the resulting docking data were processed and analysed considering binding energies and main interacting residues in each complex by using the PRODIGY software (*Xue et al. 2016*). Free energy of the structural complexes was calculated using PremPS server (*Chen et al. 2020*).

Estimation of evolutionary rate and mutational analysis

The impact of evolution on set of genes is indicated by the ratio (ω) i.e., ratio of non-synonymous substitution rate per non-synonymous site (Ka) to synonymous substitution rate per synonymous site (Ks). Where ω>1 point towards positive (diversifying) selection and ω< 1 signify negative (purifying) selection (*Roy & Basak, 2021*). The rate of evolution of each TLR1-TLR10 group of mammals (taking consensus sequence as reference) was estimated using the available PAL2NAL program (*Suyama & Torrents, 2006*). Residue wise evolutionary rate of TLR gene sequences were calculated using SWAKK server (*Liang et al. 2006*). This server performs a sliding 3D window analysis to calculate the ratio of non-synonymous to synonymous substitution rate (Ka/Ks) of DNA sequences that encode protein.

Positive selection test of individual codons of mammals TLR was performed using the Hyphy package executed in the Data Monkey Web Server that compare Ka to Ks ratio using maximum likelihood (ML) framework, (*Weaver et al. 2018*). The sequences of every TLR were analysed under the fixed-effect likelihood (FEL) model. This Fixed Effects Likelihood (FEL) approach uses maximum-likelihood (ML) method to deduce non-synonymous (dN) and synonymous (dS) substitution rates on the basis of per site considering a coding alignment and related

phylogeny. It is presumed in this method that selection pressure for each site remains constant throughout the phylogeny.

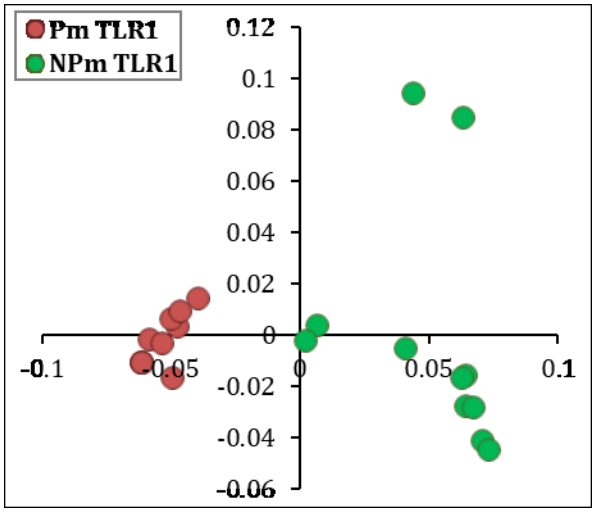
Mutational analysis was performed by using a customized script to study the mutation among the TLR sequences. Predicted consensus sequence for each TLR was used as reference sequence to identify the mutation. Consensus sequences offer promising approach in screening proteins of high stability and retain the biological activity as it predicted based on evolutionary history in which residues important for both stability and function are likely to be conserved (*Sternke et al. 2019*). Occurrences of mutation in each TLR for each species were studied across the two functional domains.

Results

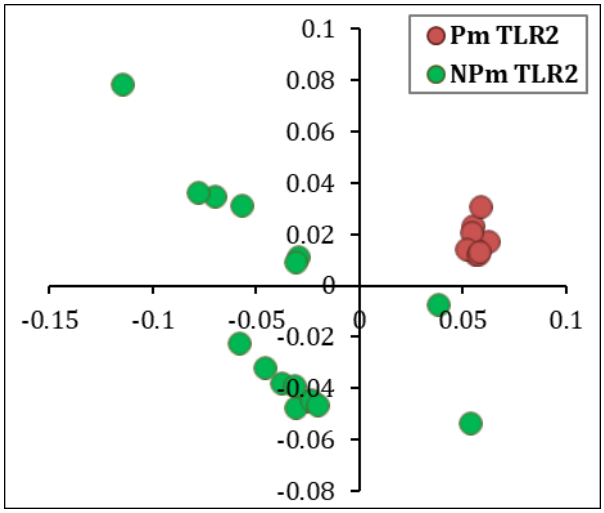
Correspondence analysis on amino acid usage of TLR genes

Correspondence analysis was performed to study the amino acid usage variation of ten different TLR genes of mammalian origin separately. The first and second major axes accounted for 54.5% and 20.1% of the total variation of amino acid usage respectively for TLR1 gene. Figure 1 shows position of genes generated during correspondence analysis on the basis of amino acid usage across the first and second major axes. Similar pattern of distribution of the amino acid usage was observed for other TLRs under study. For the ten different TLR genes these first axis always accounted the major variation which is more than 30% of the total variation of amino acid usage. It is clear from the correspondence analyses that there are two clusters. One cluster belongs to mammal which are primates and another cluster belongs to mammal other than primates. For simplicity, hereafter, TLRs from primates (Human, Gorilla, Monkey, Chimpanzee, Orangutan, Baboon etc.) will be referred to as primate mammal (Pm) TLRs and TLRs from mammal other than primates will be referred to as non-primate mammal (NPm) TLRs. Phylogenetic tree using the TLR1 genes of Pm and NPm clearly shows that Pm and NPm TLR genes are present in different branches (Figure 2). Similar pattern is observed for other TLRs. Branching pattern of phylogenetic tree follows similar trend to that of correspondence analysis.

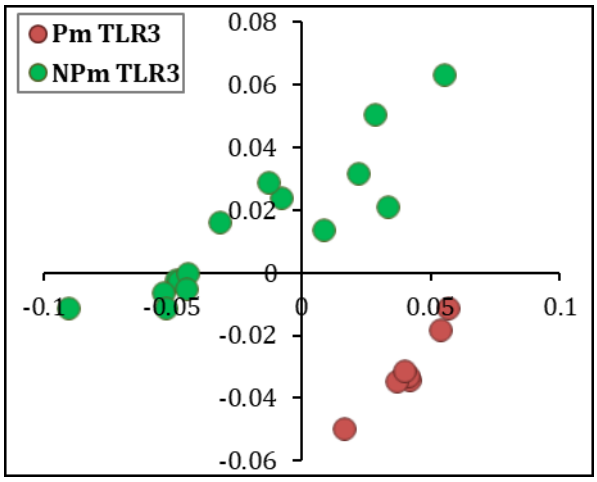
(1A)



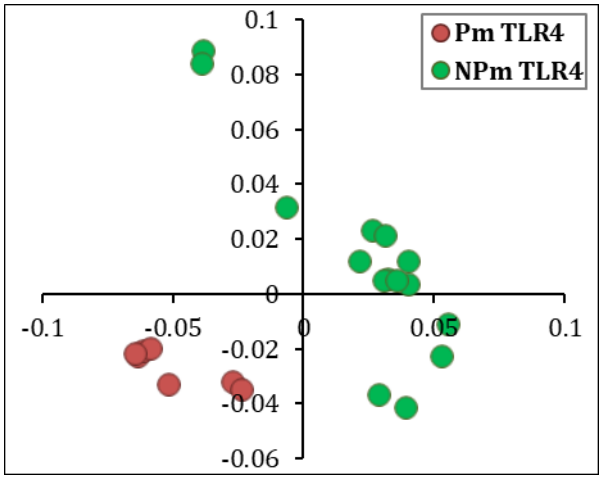
(1B)



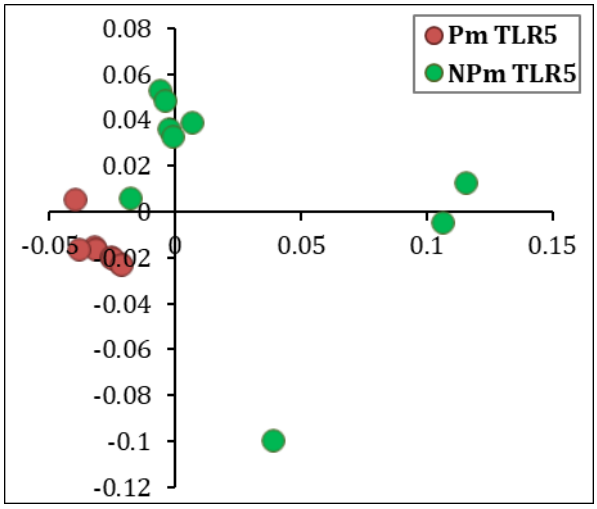
(1C)



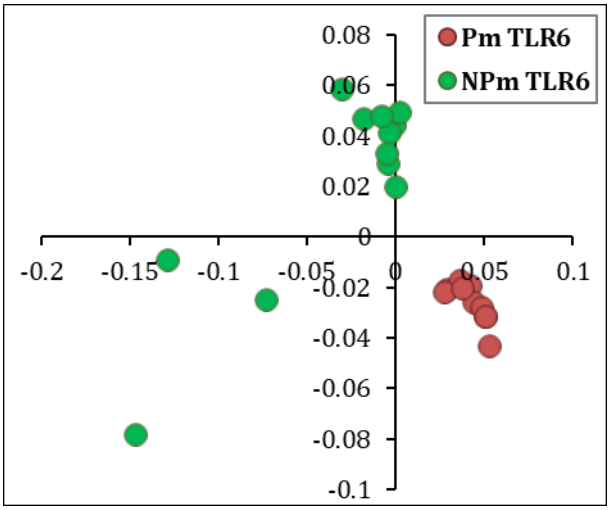
(1D)



(1E)



(1F)



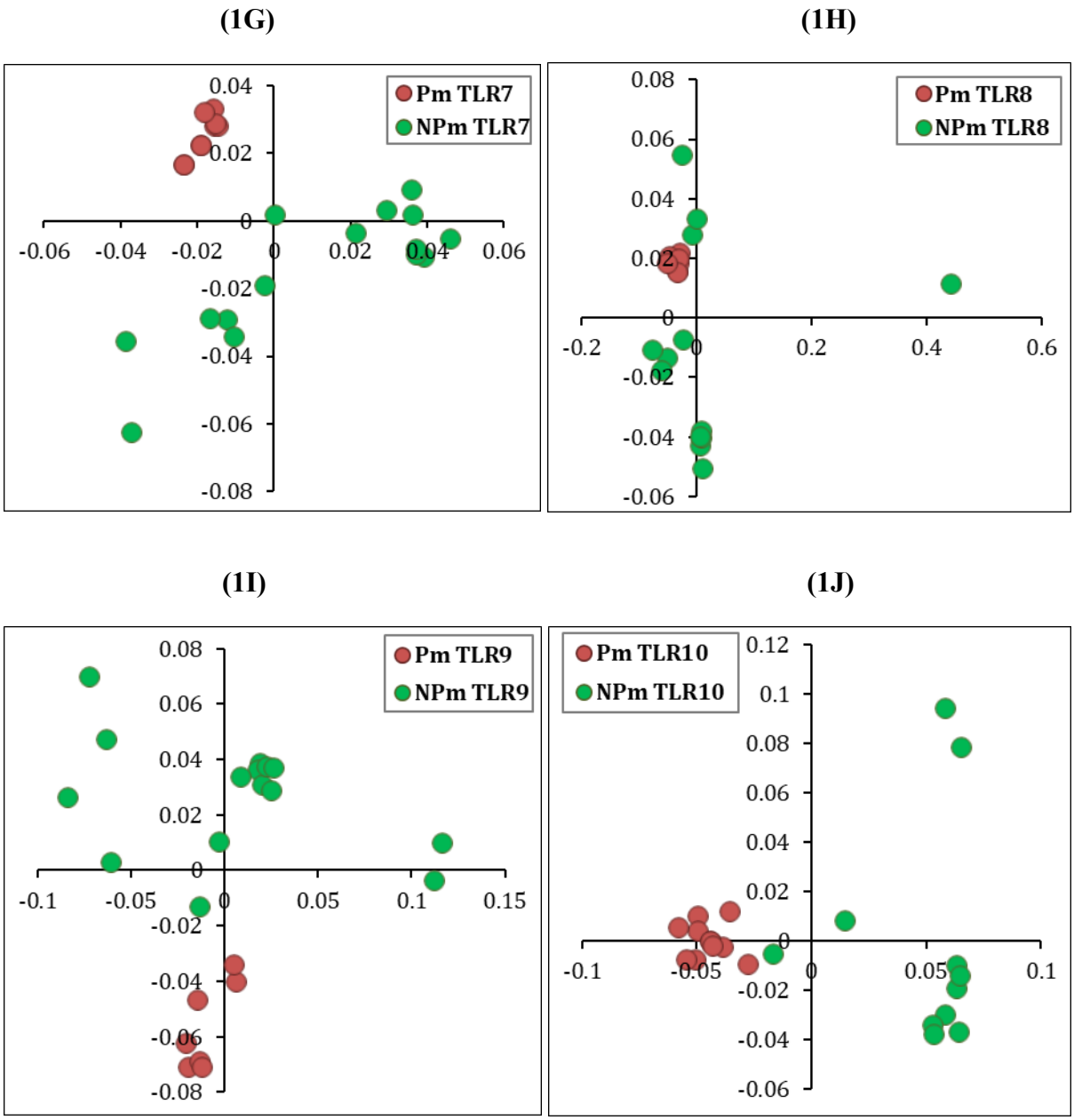
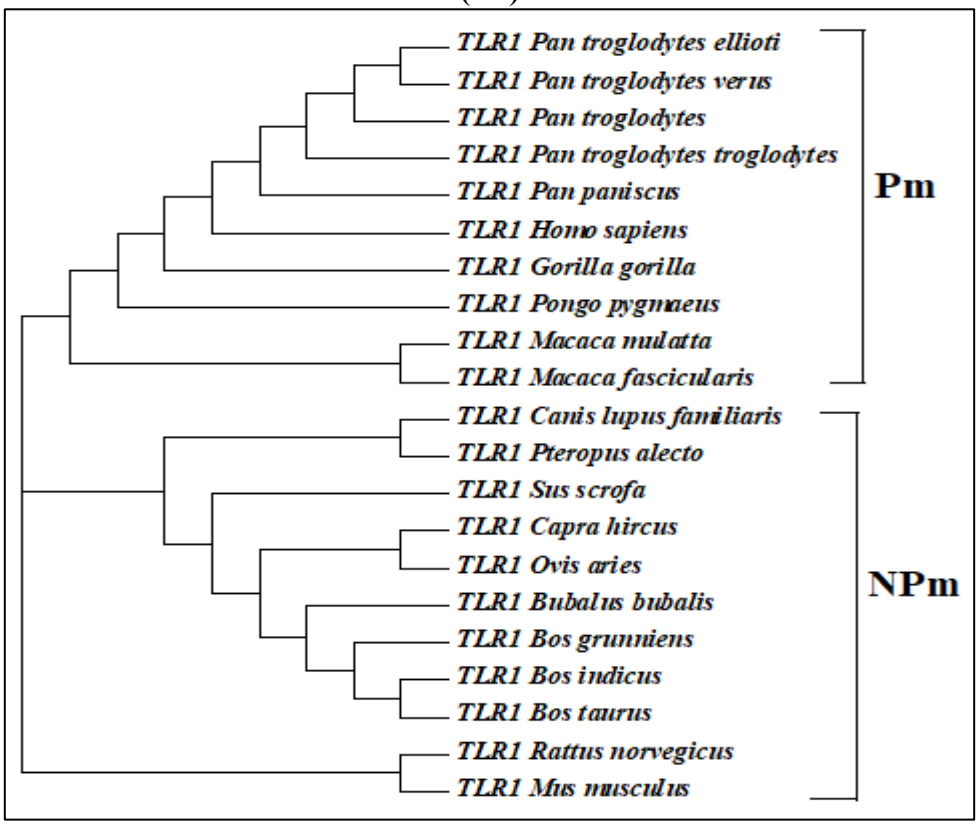
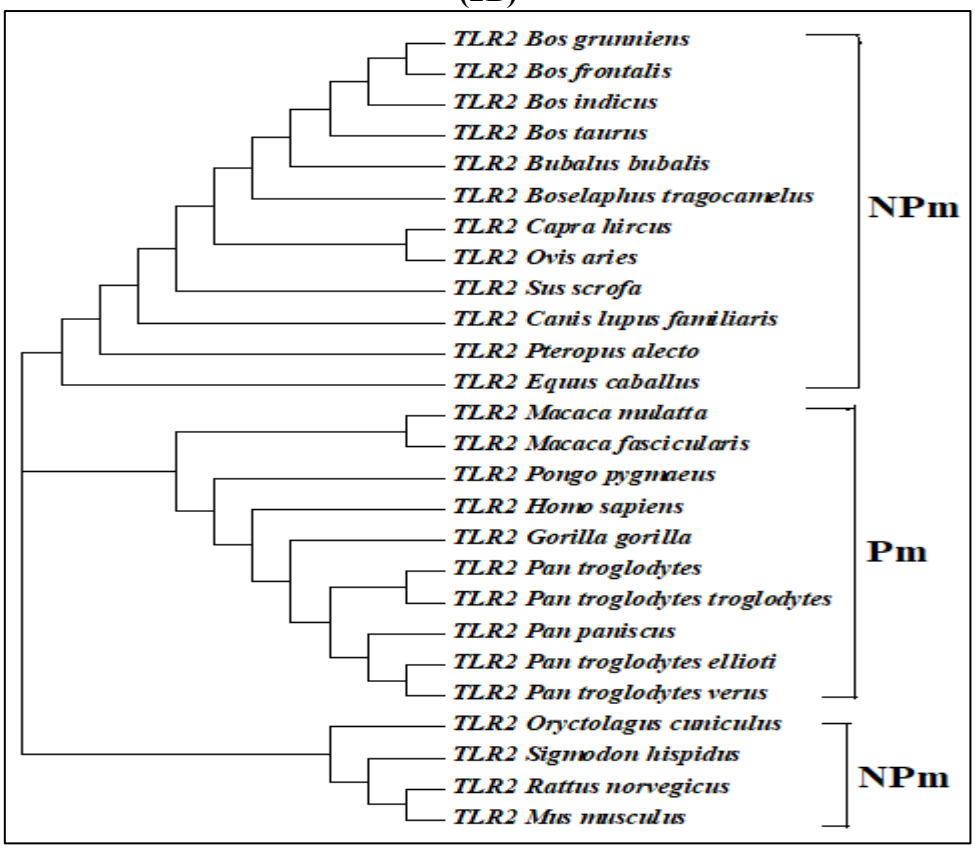


Figure 1: Distribution of TLR1-TLR10 genes along the two major axes of Correspondence analysis (COA) based on amino acid usage (AAU) data. x-axis- Axis 1 of AAU; y-axis- Axis 2 of AAU. Red coloured dots represent TLR gene sequences from Pm and green coloured dots represent TLR gene sequences from NPm. Similar pattern is observed for other TLR genes also as shown in figures 1A-1J.

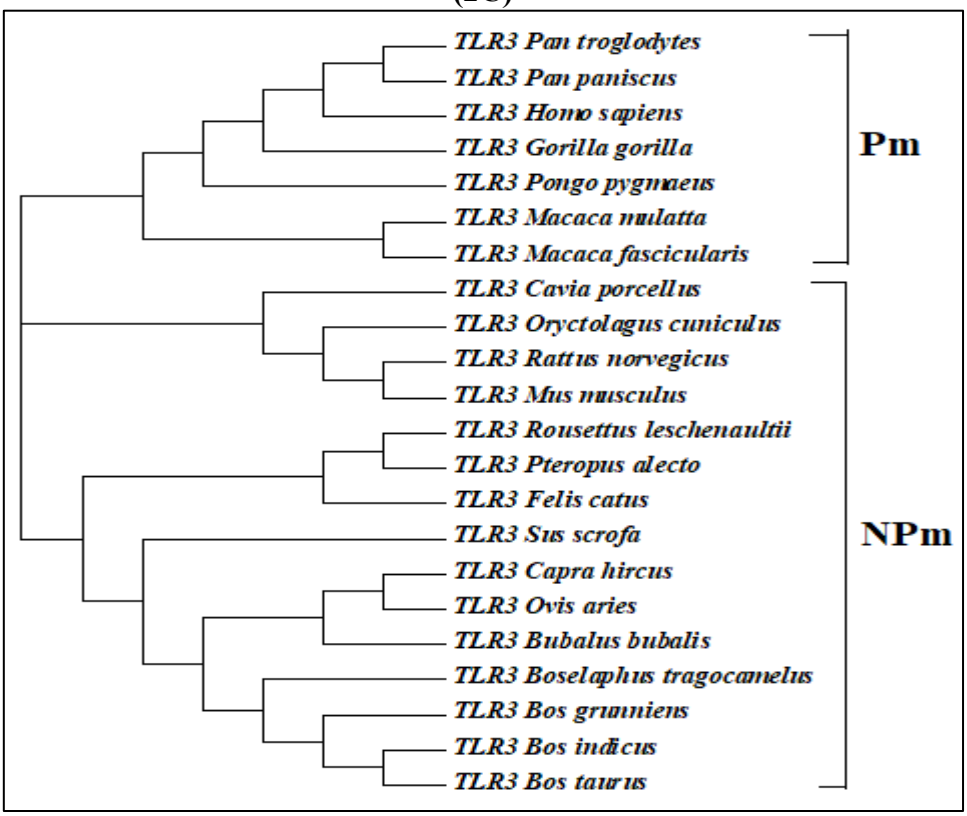
(2A)


TLR1 Pan troglodytes ellioti
TLR1 Pan troglodytes verus
TLR1 Pan troglodytes
TLR1 Pan troglodytes troglodytes
TLR1 Pan paniscus
TLR1 Homo sapiens
TLR1 Gorilla gorilla
TLR1 Pongo pygmaeus
TLR1 Macaca mulatta
TLR1 Macaca fascicularis
TLR1 Canis lupus familiaris
TLR1 Pteropus alecto
TLR1 Sus scrofa
TLR1 Capra hircus
TLR1 Ovis aries
TLR1 Bubalus bubalis
TLR1 Bos grunniens
TLR1 Bos indicus
TLR1 Bos taurus
TLR1 Rattus norvegicus
TLR1 Mus musculus
Pm
NPm

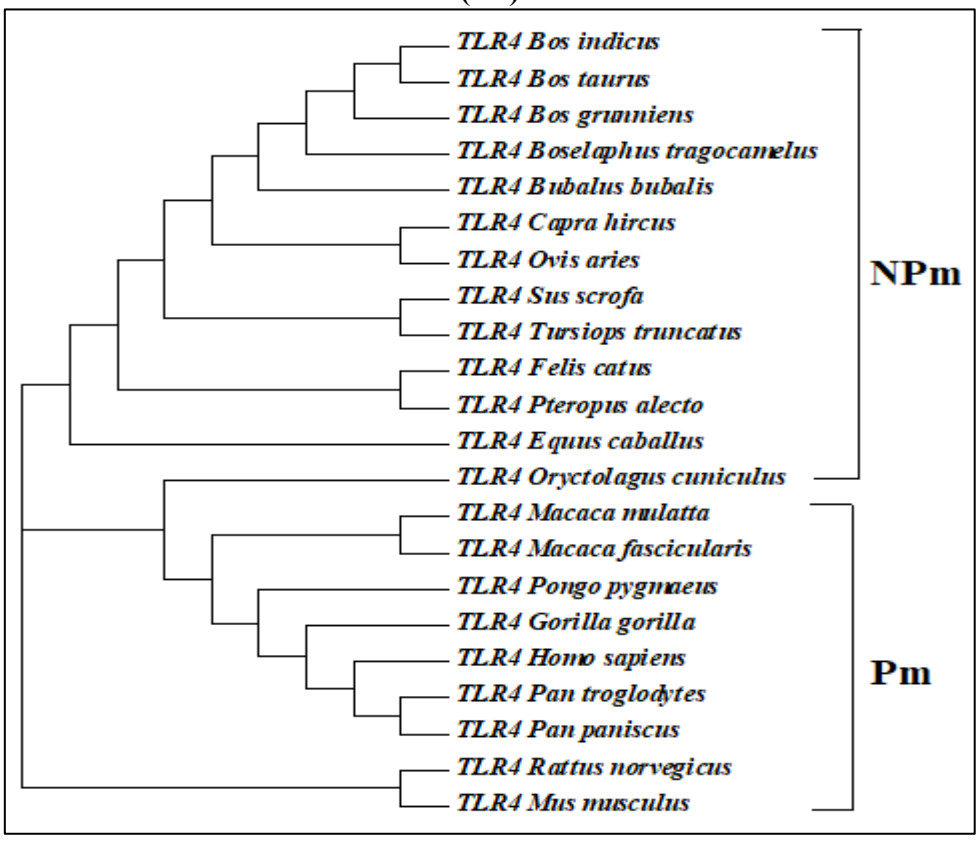
(2B)


TLR2 Bos grunniens
TLR2 Bos frontalis
TLR2 Bos indicus
TLR2 Bos taurus
TLR2 Bubalus bubalis
TLR2 Boselaphus tragocamelus
TLR2 Capra hircus
TLR2 Ovis aries
TLR2 Sus scrofa
TLR2 Canis lupus familiaris
TLR2 Pteropus alecto
TLR2 Equus caballus
TLR2 Macaca mulatta
TLR2 Macaca fascicularis
TLR2 Pongo pygmaeus
TLR2 Homo sapiens
TLR2 Gorilla gorilla
TLR2 Pan troglodytes
TLR2 Pan troglodytes troglodytes
TLR2 Pan paniscus
TLR2 Pan troglodytes ellioti
TLR2 Pan troglodytes verus
TLR2 Oryctolagus cuniculus
TLR2 Sigmodon hispidus
TLR2 Rattus norvegicus
TLR2 Mus musculus
NPm
Pm
NPm

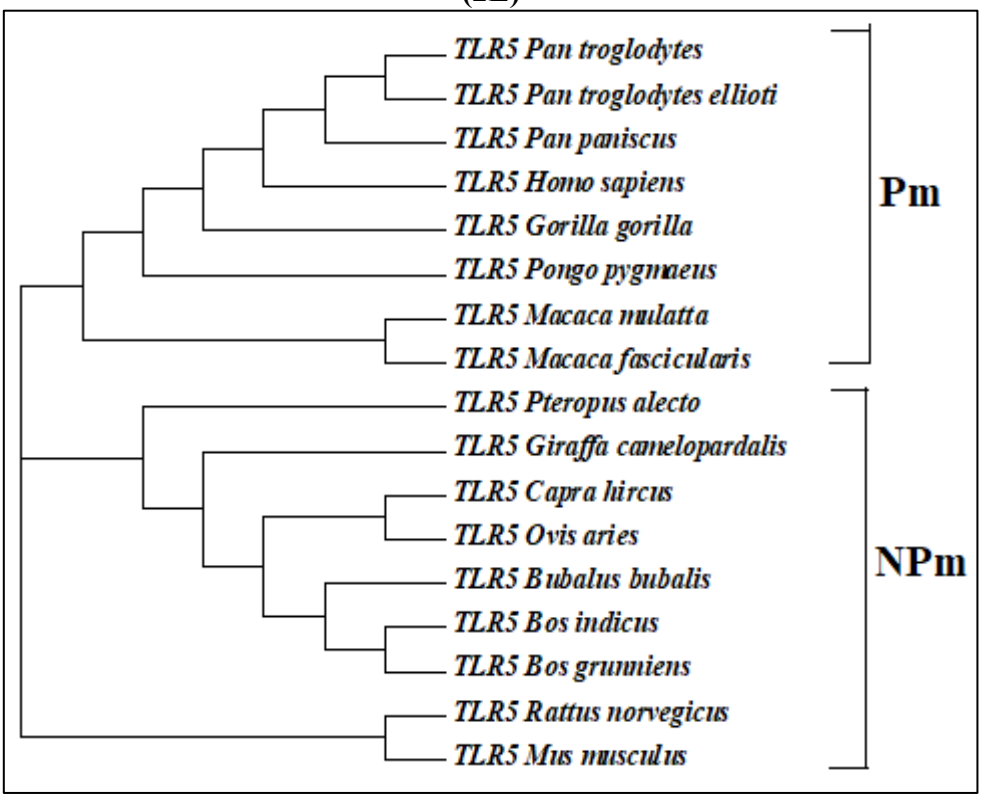
(2C)


TLR3 Pan troglodytes
TLR3 Pan paniscus
TLR3 Homo sapiens
TLR3 Gorilla gorilla
TLR3 Pongo pygmaeus
TLR3 Macaca mulatta
TLR3 Macaca fascicularis
TLR3 Cavia porcellus
TLR3 Oryctolagus cuniculus
TLR3 Rattus norvegicus
TLR3 Mus musculus
TLR3 Rousettus leschenaultii
TLR3 Pteropus alecto
TLR3 Felis catus
TLR3 Sus scrofa
TLR3 Capra hircus
TLR3 Ovis aries
TLR3 Bubalus bubalis
TLR3 Boselaphus tragocamelus
TLR3 Bos grunniens
TLR3 Bos indicus
TLR3 Bos taurus
Pm
NPm

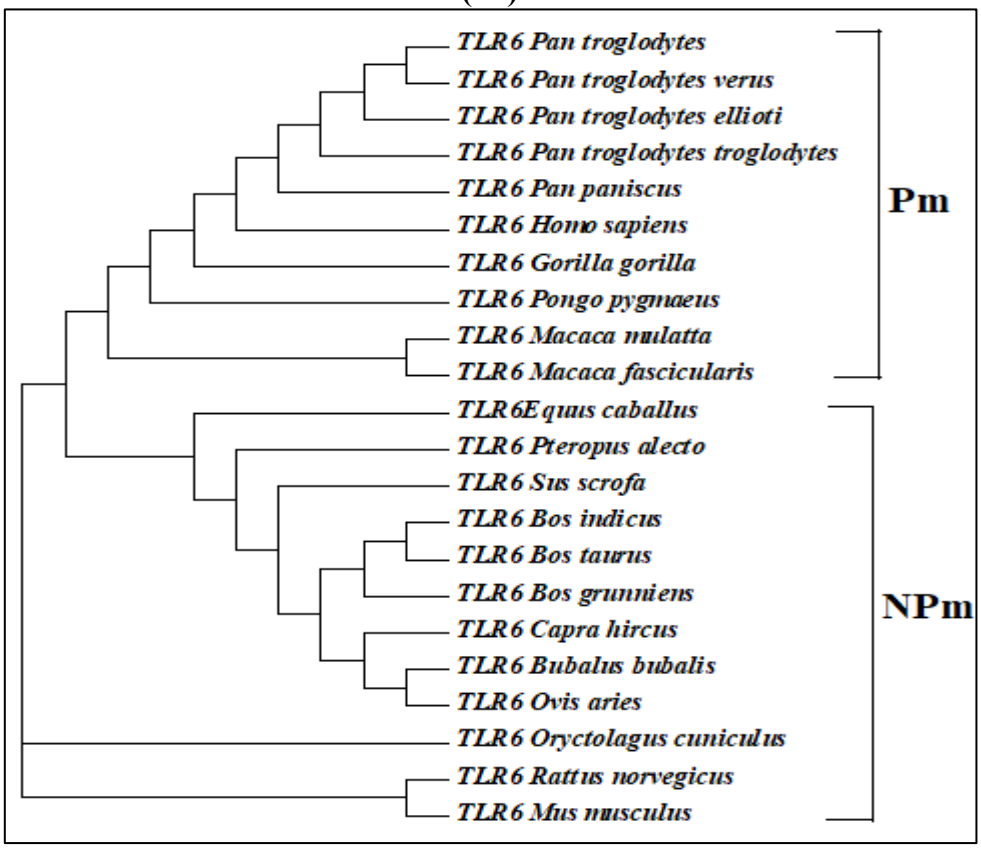
(2D)


TLR4 Bos indicus
TLR4 Bos taurus
TLR4 Bos grunniens
TLR4 Boselaphus tragocamelus
TLR4 Bubalus bubalis
TLR4 Capra hircus
TLR4 Ovis aries
TLR4 Sus scrofa
TLR4 Tursiops truncatus
TLR4 Felis catus
TLR4 Pteropus alecto
TLR4 Equus caballus
TLR4 Oryctolagus cuniculus
TLR4 Macaca mulatta
TLR4 Macaca fascicularis
TLR4 Pongo pygmaeus
TLR4 Gorilla gorilla
TLR4 Homo sapiens
TLR4 Pan troglodytes
TLR4 Pan paniscus
TLR4 Rattus norvegicus
TLR4 Mus musculus
NPm
Pm

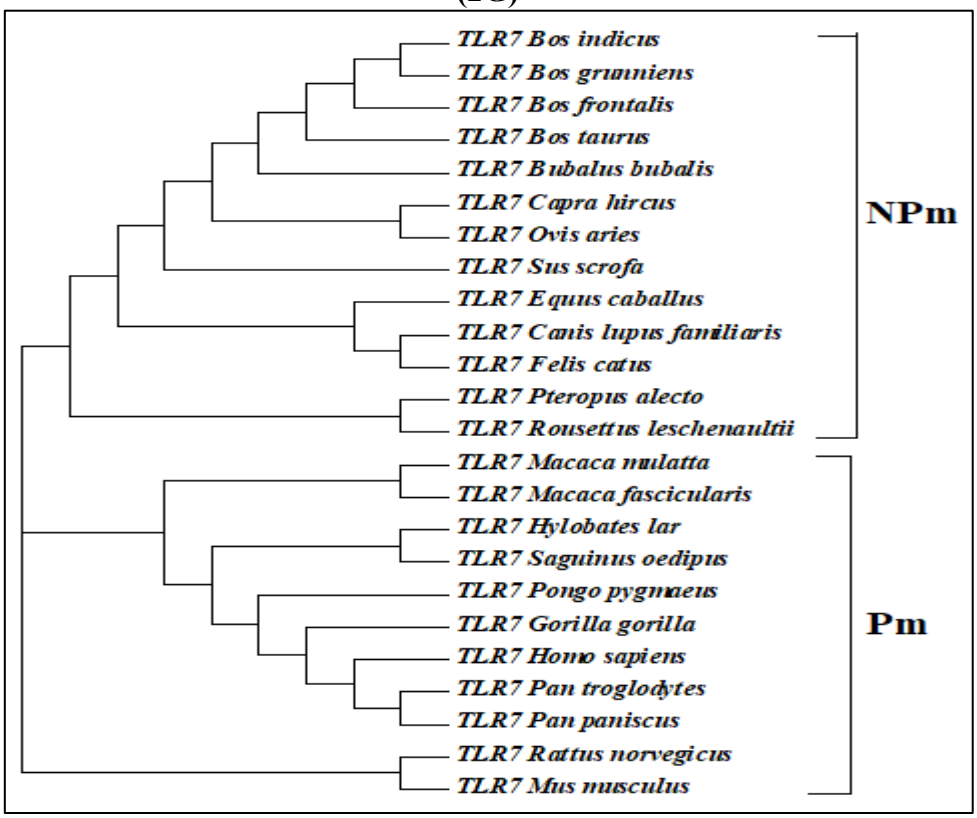
(2E)


TLR5 Pan troglodytes
TLR5 Pan troglodytes ellioti
TLR5 Pan paniscus
TLR5 Homo sapiens
TLR5 Gorilla gorilla
TLR5 Pongo pygmaeus
TLR5 Macaca mulatta
TLR5 Macaca fascicularis
TLR5 Pteropus alecto
TLR5 Giraffa camelopardalis
TLR5 Capra hircus
TLR5 Ovis aries
TLR5 Bubalus bubalis
TLR5 Bos indicus
TLR5 Bos grunniens
TLR5 Rattus norvegicus
TLR5 Mus musculus
Pm
NPm

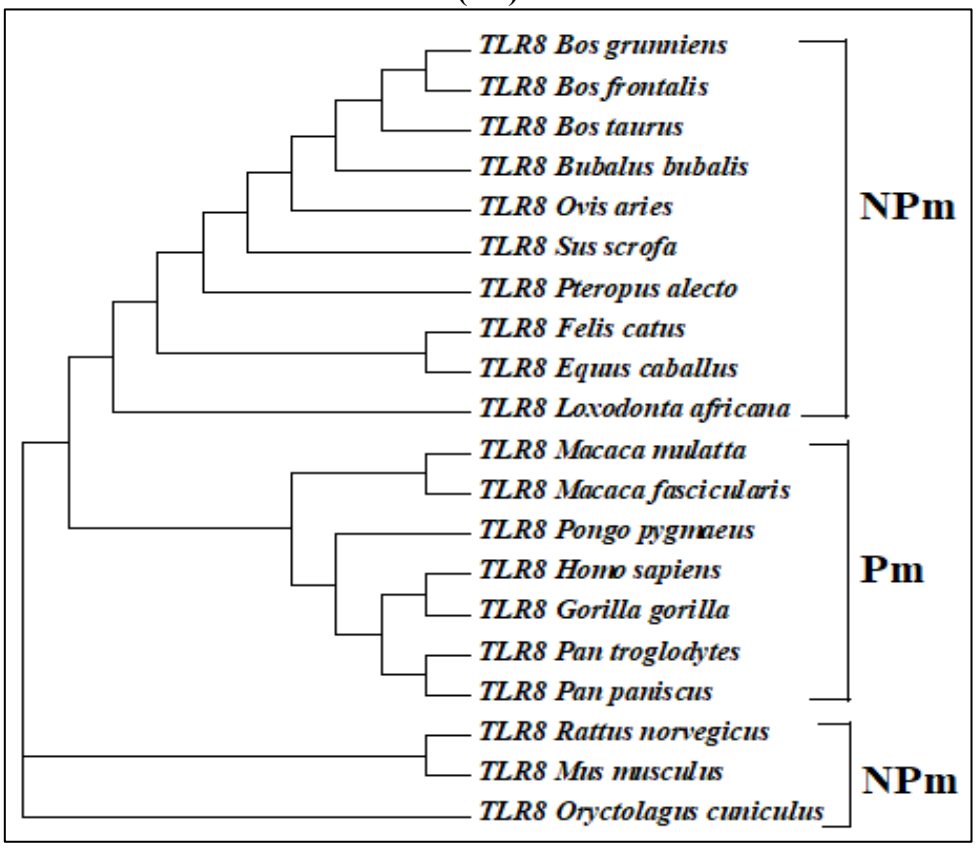
(2F)


TLR6 Pan troglodytes
TLR6 Pan troglodytes verus
TLR6 Pan troglodytes ellioti
TLR6 Pan troglodytes troglodytes
TLR6 Pan paniscus
TLR6 Homo sapiens
TLR6 Gorilla gorilla
TLR6 Pongo pygmaeus
TLR6 Macaca mulatta
TLR6 Macaca fascicularis
TLR6Equus caballus
TLR6 Pteropus alecto
TLR6 Sus scrofa
TLR6 Bos indicus
TLR6 Bos taurus
TLR6 Bos grunniens
TLR6 Capra hircus
TLR6 Bubalus bubalis
TLR6 Ovis aries
TLR6 Oryctolagus cuniculus
TLR6 Rattus norvegicus
TLR6 Mus musculus
Pm
NPm

(2G)


TLR7 Bos indicus
TLR7 Bos grunniens
TLR7 Bos frontalis
TLR7 Bos taurus
TLR7 Bubalus bubalis
TLR7 Capra hircus
TLR7 Ovis aries
TLR7 Sus scrofa
TLR7 Equus caballus
TLR7 Canis lupus familiaris
TLR7 Felis catus
TLR7 Pteropus alecto
TLR7 Rousettus leschenaultii
TLR7 Macaca mulatta
TLR7 Macaca fascicularis
TLR7 Hylobates lar
TLR7 Saguinus oedipus
TLR7 Pongo pygmaeus
TLR7 Gorilla gorilla
TLR7 Homo sapiens
TLR7 Pan troglodytes
TLR7 Pan paniscus
TLR7 Rattus norvegicus
TLR7 Mus musculus
NPm
Pm

(2H)


TLR8 Bos grunniens
TLR8 Bos frontalis
TLR8 Bos taurus
TLR8 Bubalus bubalis
TLR8 Ovis aries
TLR8 Sus scrofa
TLR8 Pteropus alecto
TLR8 Felis catus
TLR8 Equus caballus
TLR8 Loxodonta africana
TLR8 Macaca mulatta
TLR8 Macaca fascicularis
TLR8 Pongo pygmaeus
TLR8 Homo sapiens
TLR8 Gorilla gorilla
TLR8 Pan troglodytes
TLR8 Pan paniscus
TLR8 Rattus norvegicus
TLR8 Mus musculus
TLR8 Oryctolagus cuniculus
NPm
Pm
NPm

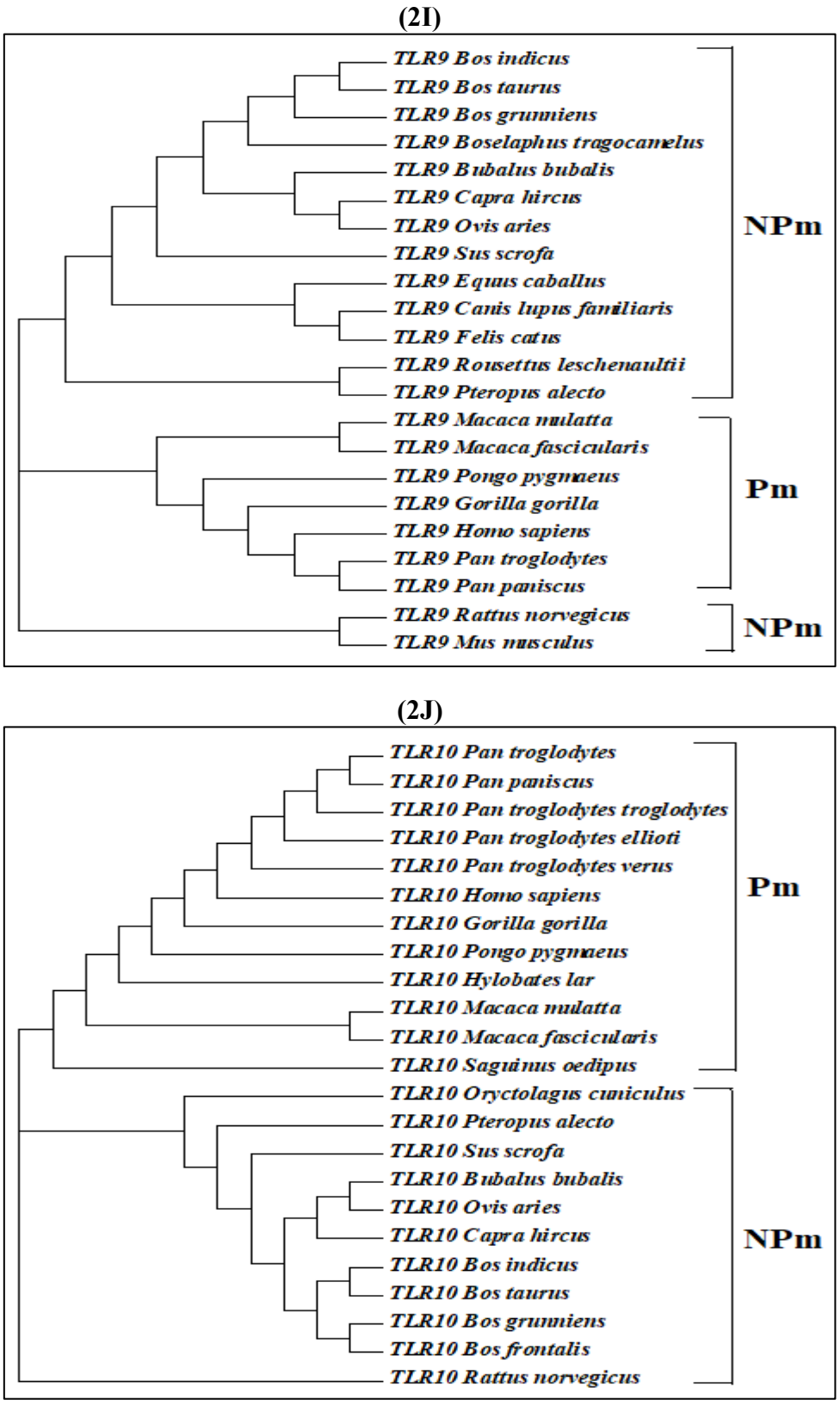


Figure 2: Phylogenetic tree of Pm and NPm genes of TLR. Similar pattern is observed for other TLRs as shown in figures 2A-2J.

Now to investigate the preference of amino acids in two different clusters we have compared the relative amino acid usage values between Pm and NPm TLR genes. Comparisons of relative amino acid usage values suggested that the twenty amino acids are differently preferred among Pm and NPm for each TLR. From the analysis it was observed that amino acids such as Phe, Met, Thr, Lys, Glu, Cys were mostly preferred in Pm TLRs whereas amino acids such as Leu, Pro, Ala, Asp, Arg, Gly were mostly preferred in NPm TLRs.

We have performed molecular docking study between TLR5 (Homo sapiens for primate and Bos indicus for non-primate) and flagellin (pathogen receptor). We have identified the preferred residues those are interacting with the flagellin and when substituted these residues with GC-rich/GC-poor, as the case may be, the stability of the TLR5-flagelin complex decreased (Figure 3).

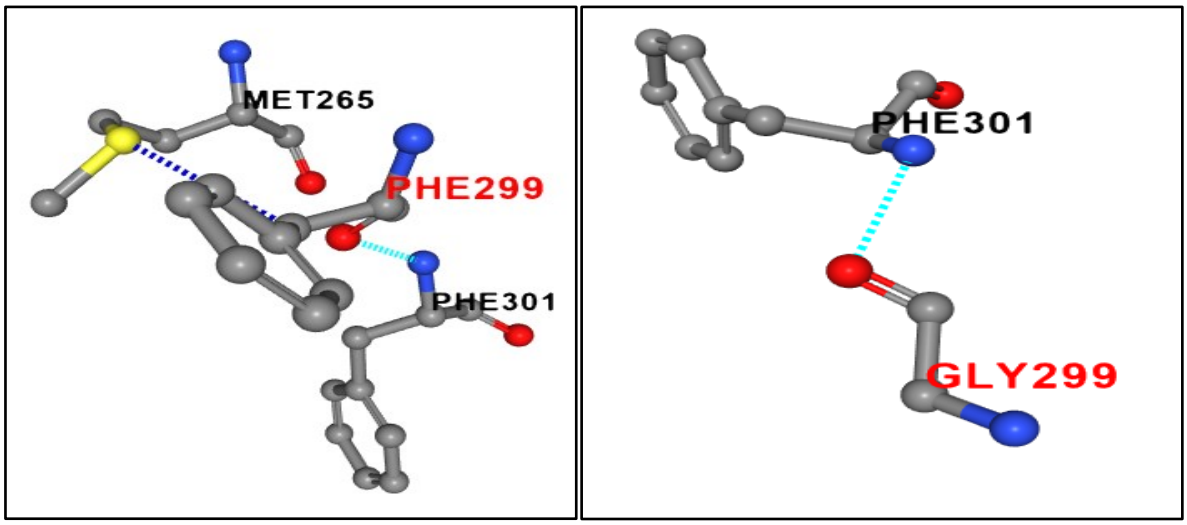
Since axis1 (horizontal axis) accounts major variation for each TLR in COA, further analysis is performed on the basis of distribution of mammal TLR genes along the horizontal axis of correspondence analysis. Significant correlation was observed between the gene position along the horizontal axis and hydrophobicity ($r=0.533$, $p<0.05$) and GC-content of the encoded proteins ($r=0.745$, $p<0.01$). Significant correlation of axis1 with GC1 ($r=0.714$, $p<0.05$), GC2 ($r=0.689$, $p<0.05$), GC3 ($r=0.668$, $p<0.05$) content of the encoded proteins were also observed.

We have compared the average GC content of TLR genes for Pm and NPm. The average GC content of TLR genes are 42.6% and 44.6% for Pm and NPm respectively. The difference of GC content of TLR genes between Pm and NPm is statistically significant ($P<0.01$). As the NPm TLR genes have higher GC content we may expect GC-rich amino acids would be preferred in NPm. Indeed, we observed that average composition of four GC-rich amino acids (Du et al., 2018) (Ala, Arg, Gly, and Pro) are higher in NPm TLR genes and the compositions of four GC-rich amino acids are positively correlated with GC contents ($r=0.836$, $p<0.001$) of the NPm TLR genes. On the other hand, we observed that average composition of AT-rich amino acids (Phe, Ile, Tyr, Asn and Lys) are higher in Pm TLR genes and their compositions are also positively correlated with AT-contents ($r=0.673$, $p<0.001$) of Pm TLR genes. All these

results support that amino acid usage have been shaped under the influence of GC-content of TLR genes.

3(A)

(i) (ii)



3(B)

(i) (ii)

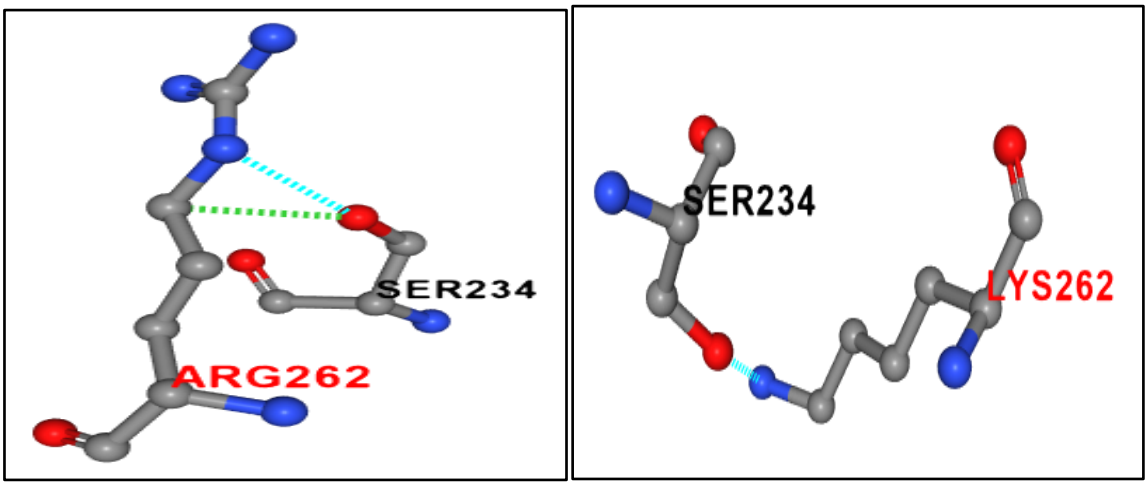


Figure 3(A): Interaction profile of a representative mutation F299G in Pm TLR5 protein indicating GC-poor to GC-rich amino acid substitution. GC-poor amino acids are preferred in Pm. The structural stability decreases when F (Phenyl alanine) is substituted by G (Glycine). (i) Wild type residue F299 having one polar interaction (sky), and one hydrophobic (blue) interaction. (ii) Mutant type residue 299G having one polar interaction (sky). **3(B):** Interaction profile of a representative mutation R2262K in NPm TLR5 protein indicating GC-rich to GC-poor amino acid substitution. GC-rich amino acids are preferred in NPm. The structural stability decreases when R (Arginine) is substituted by K (Lysine).(i) Wild type residue R262 having one polar interaction (sky) and one van der Waals (green) interactions. (ii) Mutant type residue 262K having one polar interaction (sky). Results are generated using PremPS server. ΔΔG value in both the cases is positive which indicates destabilizing mutation.

Impact of evolutionary selection pressure on TLR Genes.

We observed presence of purifying selection across all the TLR genes (both Pm and NPm) by comprehensive analysis of evolutionary rates. However, residue specific measurement of evolutionary rate shows differences of positively selected sites between Pm and NPm TLRs. Site-specific selection across the ligand binding domain also showed the same trend. These observations indicate stronger selection pressure on NPm TLR genes compared to Pm TLR genes. Positively selected sites among Pm and NPm TLRs are shown in Table 1.

The evolutionary parameters such as Non-synonymous substitution (Ka), synonymous substitution (Ks), ratio of non-synonymous and synonymous substitution (Ka/Ks) were found to differ significantly among Pm and NPm TLRs. Significant difference of these parameters was also observed across the two functional domains of Pm and NPm TLRs. These results are shown in Table 2. We have also found significant correlation of evolutionary parameters with axis1 of correspondence analysis on amino acid usage. Significant correlation of axis1 is observed with Ka in seven TLR genes, Ks in six TLR genes; Ka/Ks in five TLR genes.

Table 1: Distribution of positively selected sites among Pm and NPm TLRs.

Genes	No. of species			Total sites		Total positively selected sites		Positively selected sites in ligand binding domain		% positively selected site		% positively selected site in ligand binding domain	
	Total	Pm	NPm	Pm (length aa)	NPm (length aa)	Pm	NPm	Pm	NPm	Pm (%)	NPm (%)	Pm (%)	NPm (%)
TLR1	21	10	11	786	796	1	9	1	5	0.127	1.13	0.127	0.62
TLR2	26	10	16	784	785	0	13	0	12	0	1.65	0	1.52
TLR3	22	7	15	904	905	0	13	0	12	0	1.43	0	1.32
TLR4	22	8	14	839	844	1	32	1	28	0.119	3.79	0.119	3.31
TLR5	17	8	9	858	874	0	6	0	3	0	0.68	0	0.34
TLR6	22	10	12	796	810	0	14	0	9	0	1.72	0	1.11
TLR7	24	9	15	1049	1058	0	17	0	15	0	1.6	0	1.41
TLR8	20	7	13	1041	1091	0	20	0	18	0	1.83	0	1.64
TLR9	22	7	15	1032	1034	1	2	0	2	0.09	0.19	0	0.19
TLR10	23	12	11	811	822	0	15	0	10	0	1.82	0	1.21

Table 2: Significance test of evolutionary parameters among Pm and NPm TLR genes and across the domains. Extracellular domain of TLR (ECD), Intracellular domain of TLR (TIR) and tick mark indicates significant difference.

	Pm & NPm genes			ECD of Pm & NPm genes			TIR of Pm & NPm genes		
	Ka	**Ks**	**Ka/Ks**	**Ka**	**Ks**	**Ka/Ks**	**Ka**	**Ks**	**Ka/Ks**
TLR1	✓	✓	✓	✓	✓	✓	✓	✓	✓
TLR2	✓	✓		✓	✓	✓	✓	✓	
TLR3	✓	✓	✓	✓	✓	✓	✓		✓
TLR4			✓	✓			✓	✓	
TLR5	✓	✓	✓	✓	✓	✓	✓	✓	✓
TLR6	✓	✓	✓	✓	✓	✓	✓	✓	✓
TLR7		✓	✓	✓	✓	✓		✓	✓
TLR8					✓	✓		✓	
TLR9	✓		✓	✓		✓	✓	✓	
TLR10	✓	✓	✓	✓	✓	✓	✓	✓	✓

Correlation of evolutionary parameters with GC-content and mutational analysis.

We already observed the correlation between GC content and amino acid usage variation of TLRs through correspondence analysis. It was also found that evolutionary parameters differ significantly among Pm and NPm TLR genes. Furthermore, these evolutionary parameters such as Ka, Ks and Ka/Ks was correlated significantly with the GC content of TLR genes among mammalian species ($p<0.05$) (Table 3). Thus, GC content is playing an important role in the evolution process of amino acid sequences for most of the TLRs among Pm and NPm.

Mutations were identified for both Pm and NPm TLRs over the entire TLR sequences. But more mutations are observed in the ligand recognition domain. It endorsed that ligand recognition domain is more prone to mutation than the signaling domain. Rate of evolution (Ka/Ks) in the extracellular ligand recognition domain is more compared to intracellular signaling domain for most of the TLRs in both Pm and NPm.

Table 3: Correlation study of GC content with evolutionary parameters of TLRs.

	GC content	Ka	Correlation significant at	Ks	Correlation significant at	Ka/Ks	Correlation significant at
TLR1	0.403	0.0743	p < .01	0.1756	p < .01	0.4505	p < .05
TLR2	0.441	0.0850	p < .01	0.2391	p < .01	0.4008	p < .01
TLR3	0.403	0.0615	p < .01	0.2243	p < .01	0.2879	p < .05
TLR4	0.438	0.0994	p < .01	0.2164	p < .01	0.4829	p < .10
TLR5	0.452	0.0768	p < .01	0.2415	p < .01	0.3781	p < .01
TLR6	0.395	0.0677	p < .01	0.1883	p < .01	0.3838	p < .01
TLR7	0.410	0.0470	not significant	0.1671	not significant	0.2945	not significant
TLR8	0.418	0.1015	p < .01	0.3902	p < .01	0.4007	p < .01
TLR9	0.628	0.0685	not significant	0.4410	not significant	0.1596	not significant
TLR10	0.389	0.0607	p < .01	0.1516	p < .01	0.4020	not significant

Amino acid usage pattern of TLRs based on subcellular distribution.

Since TLRs are classified into extracellular and intracellular based on the subcellular distribution we have analyzed the amino acid usage pattern of Pm and NPm TLR genes individually. Differential amino acid usage patterns were noticed where extracellular and intracellular TLRs formed different clusters in case of Pm and NPm. In case of Pm, extracellular TLR1, TLR2, TLR6, TLR10 formed one cluster; TLR4, TLR5 were found in different clusters and intracellular TLR3, TLR7, TLR8 were present in different cluster from TLR9. In the same way, in case of NPm intracellular TLR3, TLR7, TLR8 were in different cluster and TLR9 formed another cluster. But NPm extracellular TLR1, TLR2, TLR4, TLR6, TLR10 were grouped into one cluster and TLR5 found in separate cluster. These extracellular and intracellular TLRs were distributed along the major axis shown in Figure 4. Evolutionary parameters were also checked between these two clusters of extracellular and intracellular TLRs in case of Pm and NPm respectively. The parameters Ka, Ks and Ka/Ks were found to differ significantly among these clusters. Hence, subcellular distribution is also governing the amino acid variation of TLRs for Pm and NPm independently where evolutionary selection is the most important aspect.

Discussion

The systematic study of the amino acid usage across various mammalian TLRs revealed that amino acids are used in diverse pattern among TLR genes of Pm and NPm species. In spite of similar anatomy and physiology between Pm and NPm there is disparity in amino acid usage pattern of TLRs observed in them. One key difference between these species is that primates possess a voluminous and complicated forebrain whereas non-primates possess a small brain.

Correspondence analyses established hydrophobicity and genomic GC content as the most important features causing the TLR wise variation of amino acid usage in mammal. It depicts that these factors are causing the variation in the immune response among species of a particular TLR. Significant correlation of hydrophobicity is observed among TLRs. The extracellular TLR domains are composed of leucine-rich repeats (LRR) that usually contain 22–29 length residues and have periodic hydrophobic residues positioned at discrete intervals. In three dimensions during assembling into protein multiple repeats shape as solenoid like structure, where consensus hydrophobic residues pointed inside to make a stable core of the structure (*Botos et al. 2011*). Hydrophobic residues becoming an influencing factor for amino acid usage variation of TLR genes among Pm and NPm. GC content is another influencing factor as amino acid usage of TLRs is significantly correlated with GC content. Guanine and cytosine bases proportion in the DNA molecule (GC content) being an essential qualitative aspect of genomic architecture is discussed frequently in humans and other vertebrates such as birds, mammals in relation to the evolution of the isochore structure (*Šmardaet al. 2014*).

Amino acid usage pattern study also revealed that individual Pm and NPm TLRs distribution based on subcellular location extracellular and intracellular is different. Depending on subcellular location functionality of TLRs become different due to dissimilar PAMP recognition. Cell surface expressed TLRs such as TLR1, TLR2, TLR4, TLR5, TLR6 and TLR10 mostly recognise microbial membrane components like lipoproteins, lipids; TLR3, TLR7, TLR8 and TLR9 expressed in intracellular vesicles like endoplasmic reticulum (ER), endosomes, lysosomes and endolysosomes and sense microbial nucleic acids (*Kawai and*

Akira, 2010). These factors affecting Pm and NPm TLRs which are showing distinct amino acid usage pattern between extracellular and intracellular TLRs.

Evolutionary analysis has suggested that purifying selection is the major force working on TLRs. Presence of codons that are selected positively indicates selective pressures on these immune genes lead to the most noticeable changes in the ectodomain, particularly in the variable section accountable for direct interaction with PAMPS. More mutation is observed in the extracellular domain due to the direct interaction with pathogen. Overall selective pressure within the innate immune system is stronger in non-primate mammal species compared to primate mammal species. The relation between GC contents and Ka, Ks, Ka/Ks values of TLR genes from different mammal species were observed. Correspondingly, Ka, Ks, Ka/Ks values changes with change in GC contents. The GC content is therefore consistent with the evolutionary process of amino acid sequences and contributes to the evolutionary level as a key component of amino acids between Pm and NPm TLRs. The GC content influences the emergence of proteins due to energy costs, and both the combination of bases and amino acids is involved in this process (*Du et al. 2018*).

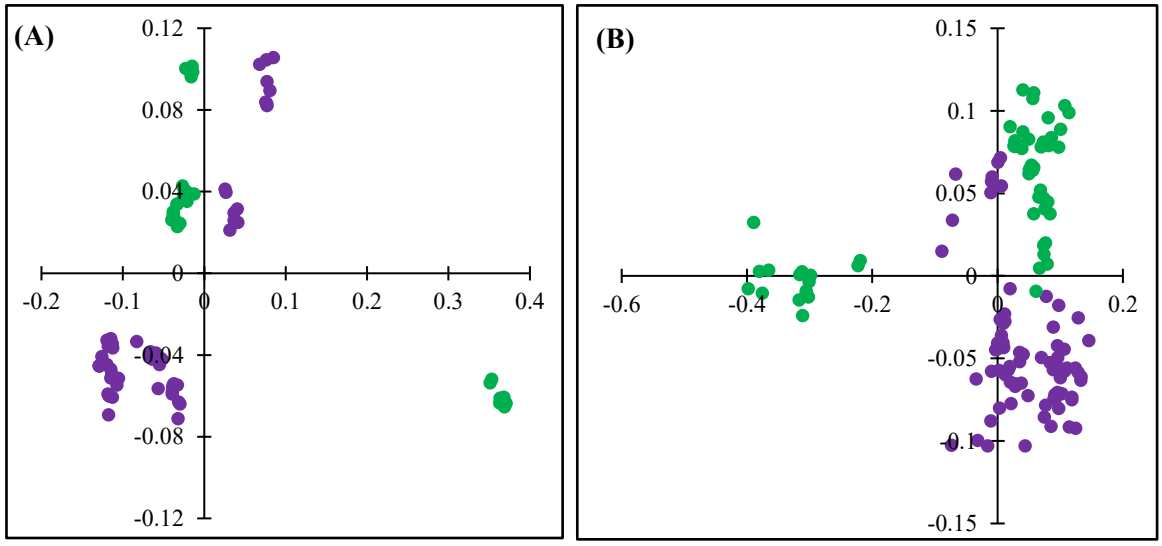


Figure 4: Distribution of TLR genes along the two major axes of Correspondence analysis (COA) based on amino acid usage (AAU) data. X-axis- Axis 1 of AAU; y-axis- Axis 2 of AAU. **(A)** TLR gene sequence of Pm, **(B)** TLR gene sequence of NPm. Violet coloured dots represent extracellular TLR gene sequences and green coloured dots represent intracellular TLR gene sequences.

Conclusion

This study reveals differential patterns of amino acid usage, evolutionary constraints of TLR genes among Pm and NPm. Amino acid composition has a significant impact on the level of TLR emergence and this is also affected by GC content. Identification of genes associated with immunity that evolves in a different way across Pm and NPm TLRs might facilitate the understanding of genetic basis for the differences in disease susceptibility (*Quach et al. 2013*). The greater extent of deviation in selection that constrain the evolution of Pm and NPm TLRs will enhance our understanding of the biological contribution of TLRs to host defence in natural setting. This study presented the divergence in the biological significance of different TLRs and offer evidences for their diverse contributions in response to host defense.

Chapter - V

Evolutionary dynamics in TLR evolution

Results presented in this chapter are published in following article:
Ghosh M, Basak S, Dutta S. Evolutionary divergence of TLR9 through ancestral sequence reconstruction. Immunogenetics. 2024;76(3):203-211. doi:10.1007/s00251-024-01338-8

Background

Toll-Like Receptors (TLRs) are considered as the primary sensors of invading microbial pathogen in the innate immune system because they detect pathogen-associated molecular patterns (PAMPs). Since the early discovery of a Toll protein in the fruit fly *Drosophila melanogaster* thirteen members of the TLR family have been identified in human (TLR1-TLR10) and mouse (TLR1-TLR13) (*Zhou et al. 2013*). It seems that most mammalian species share a similar repertoire of TLR homologs though with few exceptions (*Nie et al. 2018*). TLRs are type I integral membrane glycoproteins with a pathogen binding ectodomain (ECD) and a cytoplasmic signalling domain connected by a single transmembrane helix (*Zhou et al. 2013*). Mammalian TLR pathogen-binding ectodomains contain 19-25 extracellular leucine-rich repeats (LRRs) and a cytoplasmic toll/interleukin (IL)-1R (TIR) domain. LRRs comprising 24-29 amino acids responsible for ligand recognition and binding, while the TIR domain is responsible for downstream signalling (*Botos et al. 2011*). Surface-expressed TLRs (TLR 1, 2, 4, 5, 6 and 10) typically identify pathogen structural components, whereas endosomal TLRs (TLR 3, 7, 8, and 9) recognise nucleic acid. TLRs respond to a variety of pathogen-associated molecular patterns (PAMPs) in humans, including lipopolysaccharide (TLR4), lipopeptides (TLR2 associated with TLR1 or TLR6), bacterial flagellin (TLR5), viral dsRNA (TLR3), viral or bacterial ssRNA (TLRs 7 and 8), and CpG-rich unmethylated DNA (TLR9) (*Takeda and Akira 2005; Vidya et al. 2018*).

TLR9 is an endosomal receptor that detects bacterial DNA/CpG-containing oligodeoxynucleotides (CpG ODN). TLR9-mediated signalling is initiated within the endosome by the sequential recruitment of adaptor proteins, which in turn activates critical downstream transcription factors. Various preclinical studies showed the efficacy TLR9 agonists individually and in combination with other agents (*Karapetyan et al. 2020*).

Interaction of unmethylated CpG DNA with TLR9 activates immune responses through MyD88-dependent signaling pathway. Human trials have shown that CpG DNA can act as an adjuvant and boost the immunogenicity of the hepatitis vaccine. These findings highlight the importance of TLR ligands in triggering adaptive responses and providing new adjuvants in vaccine formulation (*Cook et al. 2004*).

Biological sequences have long been recognised as a record of evolutionary history, with accumulating mutations recording species relationships and the mechanisms driving their evolution. To avoid the recognition by the host immune system pathogens involved in recognition evolve faster. With the evolving pathogen the host receptor that recognize the pathogen also evolve to keep pace with the changes in the pathogen. These modifications in receptor can be detected as the positive selection signatures or mutation (*Areal et al. 2011*). From an evolutionary perspective, genetic variation in TLR genes linked with immunological defence is important because these genes provide a good model for investigating pathogen-induced selective stress on the host genome (*Roach et al. 2005*). In response to rapidly evolving pathogens, these genes appear to evolve quicker than other locations in the genome (*Ghosh et al. 2022*). Given enough genetic information from different species, the temporal accumulation of mutations can be used to reconstruct sequences from their common ancestors. These ancestral reconstructions serve as the foundation for many of molecular evolution approaches now a days, such as phylogenetic trees and sequence selection tests (*Muffato et al. 2023*). The Ancestral sequence reconstruction (ASR) approach begins with a multiple-sequence alignment (MSA) of the collection of relevant homolog sequences and considers evolutionary information depicted by the phylogenetic tree. It is a probabilistic strategy that investigates the deep evolutionary history of homolog sequences in order to reassemble the evolutionary trajectory of a protein. ASR can reveal sequences of long-extinct genes and organisms from which the current ones evolved, making it an important tool in evolutionary biology (*Gumulya and Gillam 2017*). Since the advent of sequencing, the reconstruction of ancestral sequences, particularly genes, has been studied extensively. Advanced methods exist to retrace the history

of sequence substitutions and leverage changes in substitution dynamics to answer specific evolutionary problems (*Merkl and Sterner 2016*).

Study of the sequence-based feature like differential amino acid usage and impact of various factors on TLRs will facilitate us to comprehend the evolutionary factors that affect innate immune genes. The evolutionary genetics approach to identify the extent of natural selection acting on these genes and the gradual changes that leads to the divergence will enhance our understanding about the mechanism of host defence mediated by TLRs.

Methodology

Data retrieval and multivariate statistical analysis

Sequences of mammalian toll-like receptor (TLR) genes and their encoding proteins representing different group of TLR such as TLR1, TLR2, TLR3, TLR4, TLR5, TLR6, TLR7, TLR8, TLR9, TLR10 were obtained from GenBank, NCBI. Toll-like receptor gene sequences were searched by using the search option available at NCBI website and mammalian species have been selected under species selection for the search operation. The output of the search operation provides coding sequence of a particular TLR. These coding sequences and their corresponding protein sequences were downloaded. TLR gene sequences from primates, rodents, artiodactyls, proboscides, perissodactyls, lagomorphs, chiropters were taken for the analysis. Sequences containing ambiguous character (other than A, T, G, C) and internal stop codons were removed from the retrieved dataset. The list of mammalian taxa chosen to investigate in this study along with their accession numbers are provided in the Supplementary Table1.

Amino acid usage is a multivariate feature by nature and studied using statistical analysis such as correspondence analysis (CoA) (*Peden, 2000*). CoA is an efficient method to explore the variation in the dataset and it reveals major tendencies of data disparities by placing them along continuous axes according to the differential trends observed, with each consecutive axis having a diminishing effect (*Roy et al. 2017*). CoA on the basis of amino acid usage (AAU) of TLR gene sequences was generated using CodonW. Estimation of physicochemical properties

like hydrophobicity, GC-content, GC3 values, effective number of codons (ENC), aromaticity of the study sequences was also performed using the CodonW program. Correlation study of the parameters were executed in Microsoft Excel. Significance test was done using the freely available web program QuickCalcs-Graphpad.

Evolutionary analysis and phylogenetic tree construction

Evolutionary selection acting on the genes under study are addressed by evolutionary rate (ω). ω is estimated as the ratio of the rate non-synonymous substitutions per non-synonymous site (Ka) and the rate of synonymous substitutions per synonymous site (Ks). $\omega > 1$ indicates positive (diversifying) selection, whereas, $\omega < 1$ indicates negative (purifying) selection. For each TLR group (Example: TLR1) their consensus nucleotide sequences (Example: TLR1_consensus) were generated. We have prepared a Perl script for generating these consensus sequences. Downloaded nucleotide sequences and the consensus sequence of each TLR groups were subjected to Clustal Omega program (Madeira et al. 2022) for the nucleotide sequence alignment. This program Clustal produces biologically meaningful multiple sequence alignments of divergent sequences. Then the evolutionary rate of the TLR genes (TLR1-TLR10) of each TLR group (Example: TLR1) were estimated relative to their consensus (Example: TLR1_consensus) sequences using Codeml program of the PAML software (ver. 4.5) with runmode = −2 and CodonFreq= 1 (*Nei and Gojobori 1986; Yang 2007*).

The protein sequences of all the mammalian TLRs were subjected to the multiple sequence alignment using Clustal Omega program (*Madeira et al. 2022*). Alignment result was saved in fasta format for further analysis. Then using that alignment construction of phylogenetic tree was done applying the maximum likelihood method with thousand bootstrap replicates in the MEGAX software (*Kumar et al. 2018*).

Reconstruction of Ancestral Protein Sequences

Common ancestral protein sequence of mammalian TLRs were predicted using FireProtASR (ancestral sequence reconstruction) v1.1 webserver with default parameter settings (*Musil et al. 2021*). Analyzing ancestral sequences in an evolutionary context to infer the ancestral

sequences at certain nodes of a tree termed as ASR. Reconstructing ancestral sequences is a well-established method for inferring the evolutionary history of genes. Along with the application in the discovery of most probable evolutionary ancestors of study protein, it has been a useful approach for the design of extremely stable proteins. This protocol enables the implementation of the automated workflow FireProt^{ASR} allowing various form of inputs and advance settings (*Khan et al. 2021*). All reconstruction methods involve a phylogenetic tree inferred from a given alignment. The quality of the tree is crucial for the reliable reconstruction. We have provided the multiple sequence alignment and the phylogenetic tree of all mammalian TLR sequences as input for our study. Upon submitting input data, the server will execute the dataset and reconstruct ancestral nodes along with their sequences.

Analysis of the ancestral sequences

We have performed sequence based and structural analysis of the identified ancestral sequences to accomplish our study. Clustal Omega program, a widely used package for carrying out multiple sequence alignment (*Madeira et al. 2022*) was used for the alignment of the ancestral protein sequences. Prediction of three-dimensional structural models of ancestral proteins were performed using AlphaFold2 (*Mirdita et al 2022*). It is an artificial intelligence system developed by DeepMind that can predict three-dimensional structures of proteins from amino acid sequences with higher accuracy (*Yang et al 2023*).

Pairwise structure alignment was performed using the structural alignment tool available in Protein Data Bank (https://www.rcsb.org/alignment). This web-based tool enables alignment of one or more structures to a particular reference structure that can be accessible from the 'Analyze' section of the menu bar. In superposed structures, RMSD is calculated between aligned pairs of the backbone C-alpha atoms. Smaller RMSD indicate better structure alignment between the two structures. TM-score (template modeling score) is a measure of topological similarity between the template and model structures. It ranges between 0 and 1, where 1 indicates a perfect match and 0 is no match between the two structures. Scores < 0.2 usually indicate that the proteins are unrelated while those >0.5 generally have the same protein fold in SCOP/CATH (*Zhang and Skolnick 2005*).

Protein domains of the ancestral sequences were annotated using ScanProsite tool (*de Castro et al. 2006*). Evolutionary parameters such as rate of non-synonymous substitutions per non-synonymous site (Ka) and rate of synonymous substitutions per synonymous site (Ks) of the ancestral sequences were analysed with respect to the root node sequence of the phylogenetic tree (*Nei and Gojobori 1986; Yang 2007*). Interaction of the ancestral protein sequences and Human_TLR9 sequence that have been used as a reference for the remaining species (*Zhou et al. 2013*) with the CpG ODN (*Areal et al. 2011*) was studied in the HDOCK. This web server enables hybrid docking algorithm of template-based modeling and free docking. The server supports protein–protein and protein–DNA/RNA docking and accepts both sequence and structure inputs for proteins. The docking scores are calculated through a knowledge-based iterative scoring function in this tool. A more negative docking score means a more possible binding model (*Yan et al. 2017*).

Results

Amino acid usage pattern of toll-like receptor genes

We used mammalian toll-like receptor (TLR1-TLR10) gene sequences to investigate the amino acid usage (AAU) pattern through correspondence analysis (CoA). Mutations are accumulated in TLR genes through various evolutionary processes. These mutations lead to the change in amino acid composition of TLRs. The CoA on the amino acid usage of mammalian TLR genes was performed to study the impact of such changes on the functionality of the encoded TLR proteins. The distribution of genes along the two major axes of the correspondence analysis is shown in Figure1. The first and second major axes accounted for 57.57% and 10.76% of the total variation of amino acid usage. A clear separation of the amino acid usage pattern of TLR9 genes with respect to other TLR (TLR1-TLR8 and TLR10) genes has been observed. Because the horizontal axis of correspondence analysis accounts for the majority of variation of the TLRs in CoA further analysis was carried out based on the distribution of mammalian TLR genes along this axis.

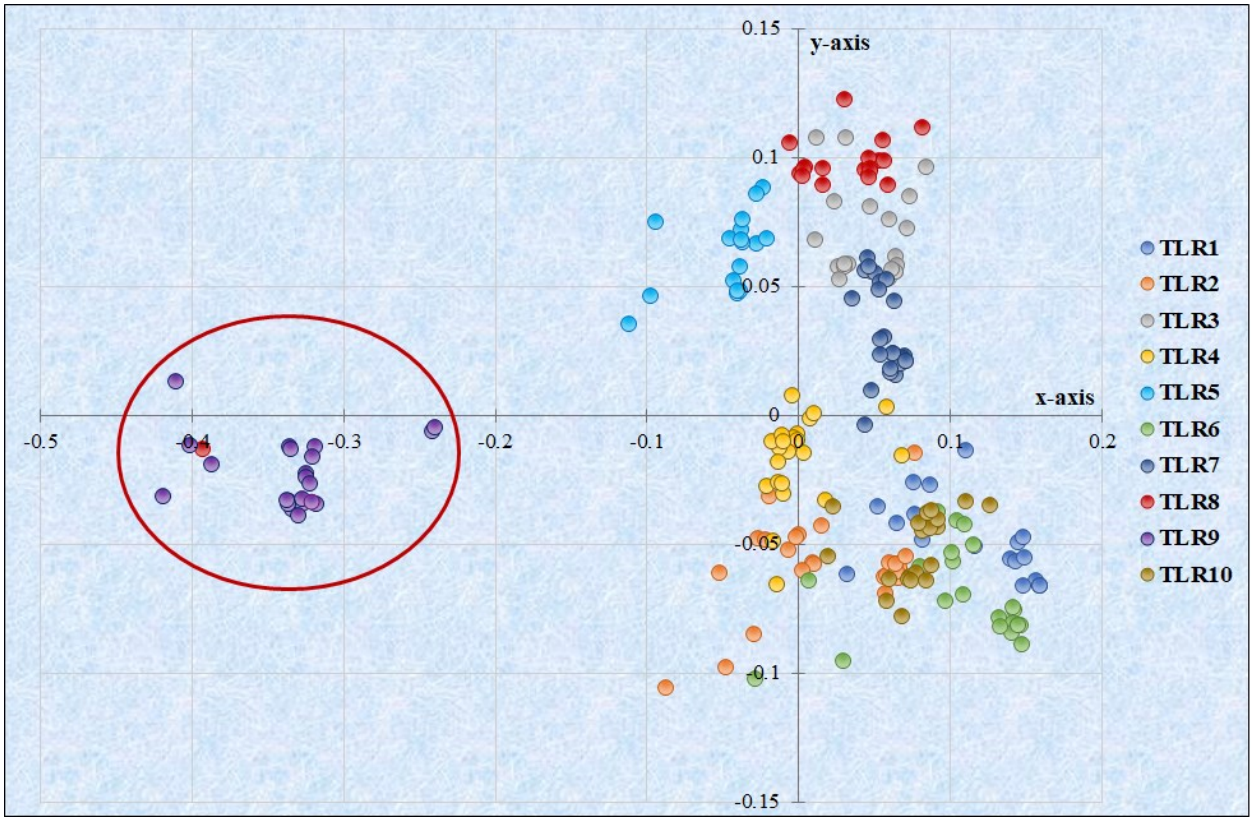


Figure 1: Distribution of mammalian toll-like receptor (TLR) genes along the two major axes of correspondence analysis (CoA) on amino acid usage. Distinct pattern of amino acid usage of TLR9 genes (violet) are marked with the red circle.

Change in amino acid usage of a gene may affect the various physicochemical properties of TLR gene. We have calculated various physicochemical parameters of TLR gene sequences to understand the factor driving this distinct amino acid usage pattern among them. The parameters such as hydrophobicity, GC-content, GC3 values, effective number of codons (ENC), aromaticity was found to differ significantly ($p < .05$) between TLR9 and other TLR (TLR1-8, TLR10) genes. Significant correlation was observed between the gene position along the horizontal axis and hydrophobicity ($r = -0.346$, $p < .01$), GC-content ($r = -0.977$, $p < .01$), GC3 values ($r = -0.96$, $p < .01$), effective number of codons (ENC) ($r = 0.825$, $p < .01$) and aromaticity ($r = 0.437$, $p < .01$) of the encoded protein. These correlation values indicate that the physicochemical parameters are contributing in the distinct amino acid usage pattern off TLR9.

Highly significant negative correlation with GC content, GC3 value indicated the influence of the codon bias. To better understand the relation between gene composition and codon bias an

ENC–GC3 scatter diagram was prepared as shown in Figure2. Such ENC–GC3 plots has been widely used to determine whether codon usage of a gene is shaped by natural selection. Significant correlation was observed between ENC and GC3 values ($r = -0.837$, $p < .01$). The solid line represents the expected curve in Figure2. TLR genes (TLR1-TLR8, TLR10) those lie on the expected curve indicate codon usage bias is only affected by mutation pressure. TLR9 genes are placed away from the expected curve, indicates that its evolution is shaped by the influence of natural selection.

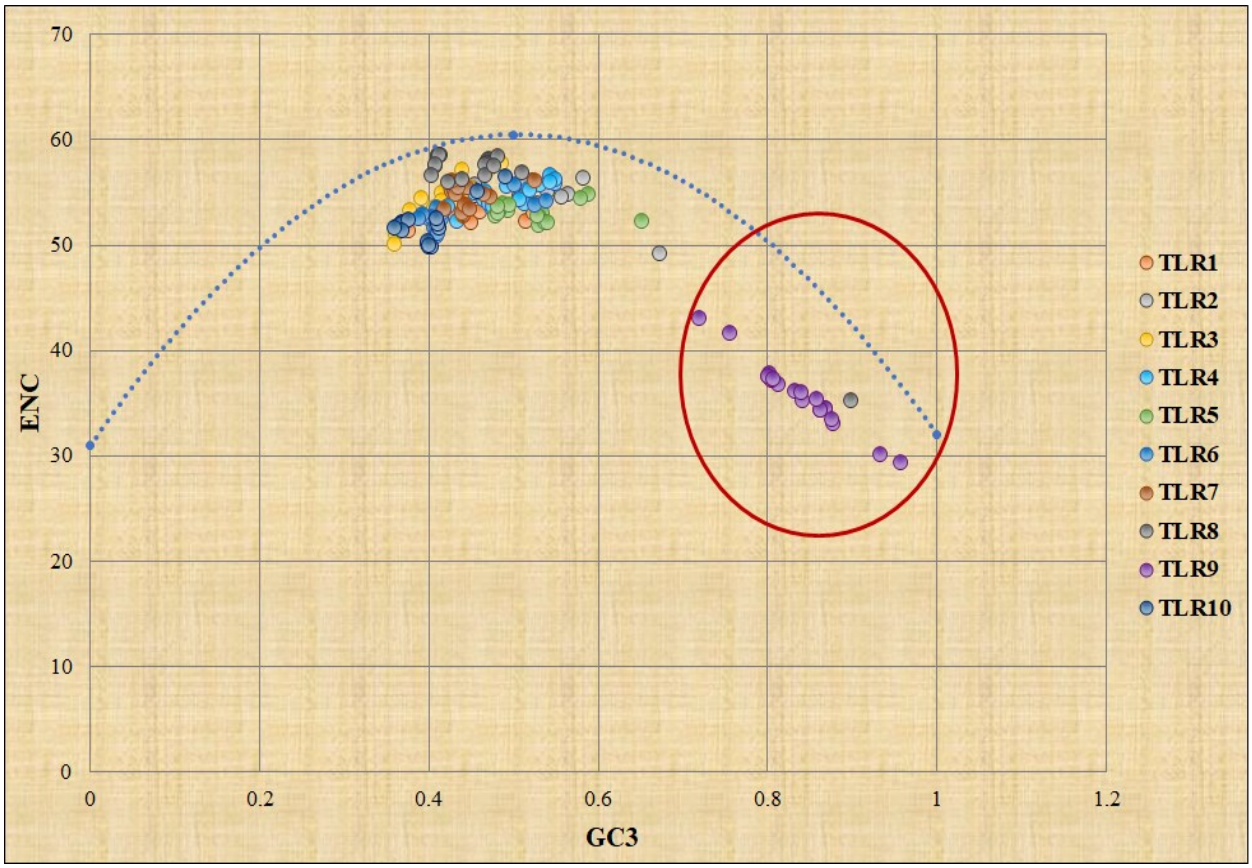


Figure 2: The plot of ENC–GC3 for mammalian toll-like receptor genes. The solid line represents the expected curve (blue). TLR genes (TLR1-TLR8, TLR10) those lie on the expected curve indicate codon usage bias is only affected by mutation pressure. TLR9 genes those are away from the expected curve indicates the influence of natural selection.

Evolutionary selection analysis

Analysis of evolutionary selection can identify specific cases of adaptation as well as general principles that guide evolution. Analysis of evolutionary processes to distinguish between neutral and adaptive changes is thus very important. To understand effect of evolutionary selection on the distinct amino acid usage pattern of TLR9, we have analyzed the evolutionary parameters such as Non-synonymous substitution (Ka), synonymous substitution (Ks), ratio of non-synonymous and synonymous substitution (Ka/Ks) of the mammalian TLR genes. Analysis of these parameters are important for the study of the dynamics of molecular evolution of TLRs. Results were compared between TLR9 and other TLR genes as we obtained the difference in amino acid usage pattern between them. We found significant difference of Ks and Ka/Ks between TLR9 and other TLRs but Ka was not statistically significant in all the cases. Average value of Ks is more and Ka/Ks is less in case of TLR9 cluster. In spite of overall purifying selection on TLR genes significant difference of non-synonymous substitution (Ka), synonymous substitution (Ks), ratio of non-synonymous and synonymous substitution (Ka/Ks) is observed. These results suggest that the evolution of TLR9 genes is highly influenced by synonymous substitution (Ks).

Ancestral sequence reconstruction

Ancestral sequence reconstruction is the calculation of ancient protein sequences on the basis of extant ones. Previous analysis suggests that TLR9 shows distinct pattern of amino acid usage and highest synonymous substitution rate with respect to other TLR genes. Thus, the ancestral sequence reconstruction through phylogenetic tree has been performed to reconstruct the evolutionary paths of the TLR protein family to study the key mechanism of the molecular evolution of TLR9. Ancestral sequence reconstruction phylogenetic tree of mammalian toll-like receptor generated from the software is shown in Supplementary Figure3. In this figure various TLR genes (For Example: TLR1, TLR2, TLR3.etc.) are marked with different colors and Nodes are assigned with Node number. All the TLR9 genes are marked in red and their ancestral Node is denoted by Node 235. Similarly, all the TLR7 and TLR8 genes are marked in orange and their ancestral Node is denoted by Node 256. TLR3 and TLR5 genes are marked

in blue and their ancestral Node is denoted by Node 299. TLR1, TLR2, TLR4, TLR6, and TLR10 genes are marked in green and their ancestral Node is denoted by Node 337. Node226 denoted the root node that leads to the evolutionary path of TLRs through Node 232, Node 233, Node 234. This entire evolutionary route of divergence of various TLRs from their common ancestor is schematically represented in Figure4. Here also the common root node is Node226. All other TLRs have been evolved from this via intermediate nodes. For example, Figure4 also depicts evolution of TLR9 from Node226 via Node235. Similarly, the evolutionary path of other TLRs from the root can be easily understood from Figure4 which is a simplified diagrammatic representation of evolutionary paths of various TLRs from root.

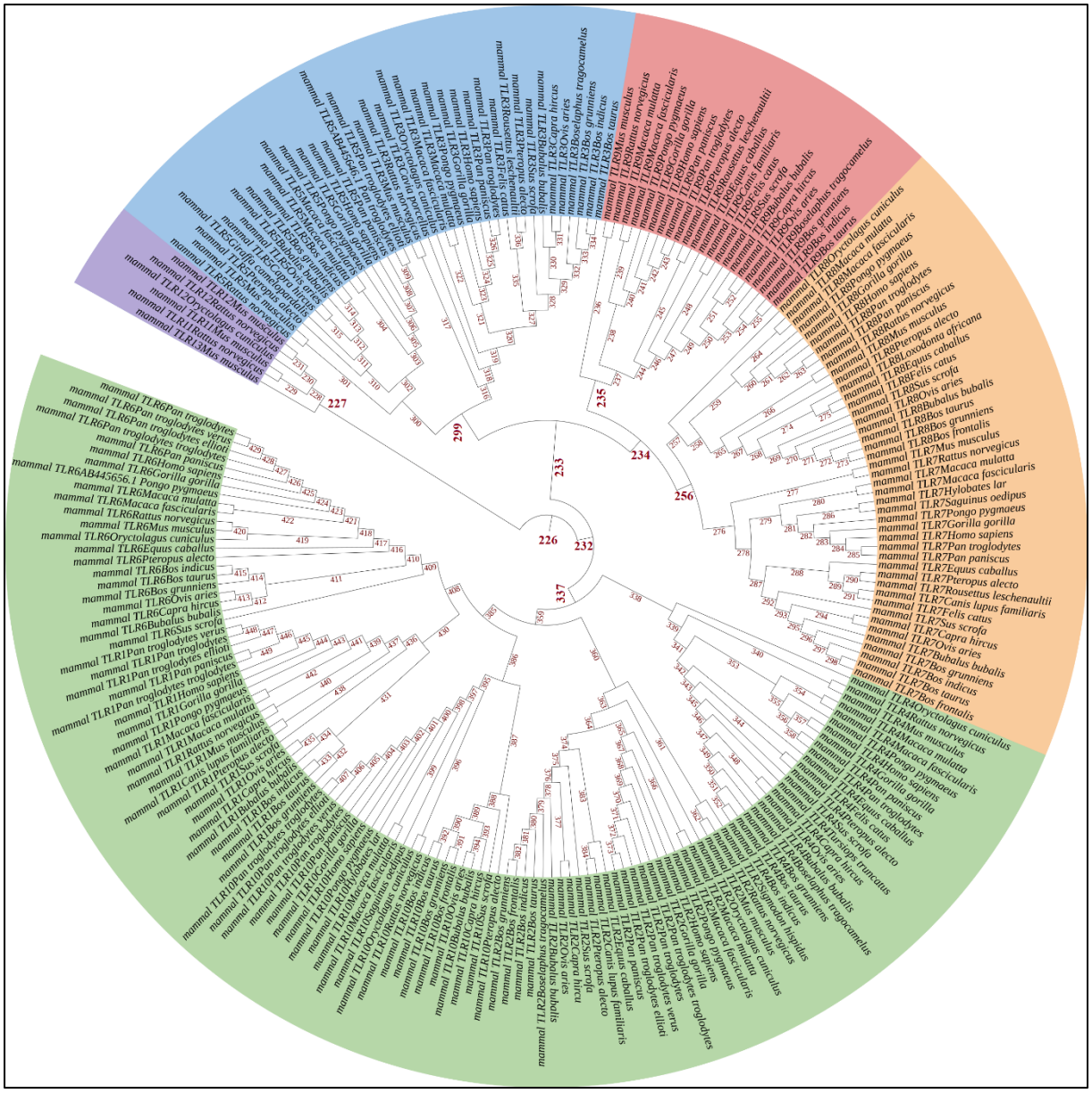


Figure 3: Phylogenetic tree of mammalian TLRs are marked with different colors and Nodes are assigned with Node number. All the TLR9 genes are marked in red and their ancestral Node is denoted by Node 235. Similarly, all the TLR7 and TLR8 genes are marked in orange and their ancestral Node is denoted by Node 256. TLR3 and TLR5 genes are marked in blue and their ancestral Node is denoted by Node 299. TLR1, TLR2, TLR4, TLR6, and TLR10 genes are marked in green and their ancestral Node is denoted by Node 337. Node226 denoted the root node that leads to the evolutionary path of TLRs through Node 232, Node 233, Node 234. This entire evolutionary route of divergence of various TLRs from their common ancestor is schematically represented in Figure4. Here also the common root node is Node226. All other TLRs have been evolved from this via intermediate nodes.

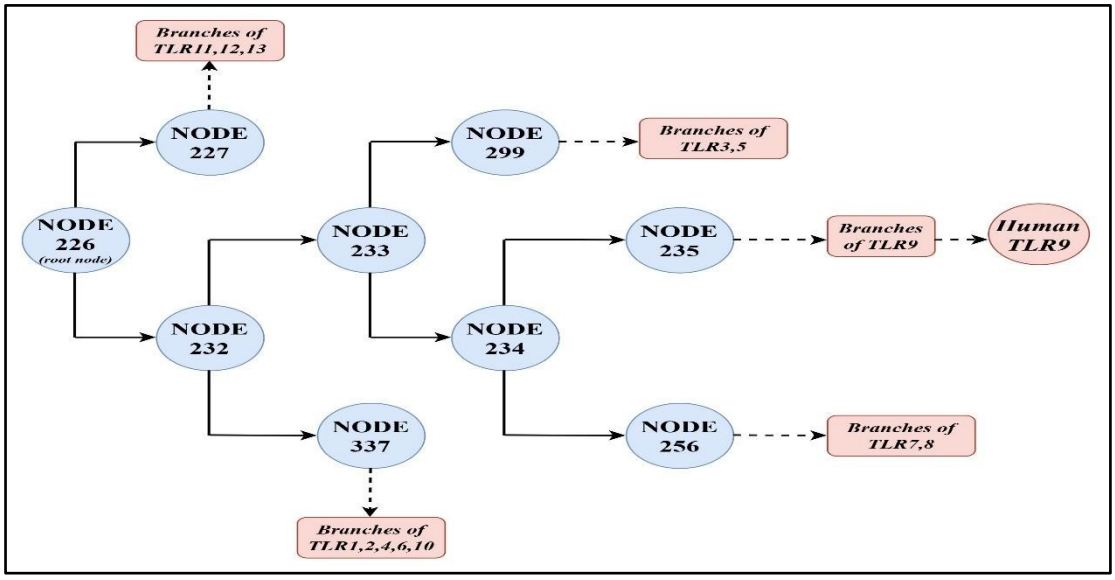


Figure 4: Simplified schematic representation of the selection of ancestral nodes from the phylogenetic tree. Node226 denotes the root node and the evolutionary pathway that leads to TLR9 follows via Node232, Node233, Node234, Node235. Node227 denotes the ancestral node of TLR11,12,13, Node337 denotes ancestral node of TLR1,2,4,6,10, Node299 denotes ancestral node of TLR3,5 and Node256 denotes ancestral node of TLR7,8.

Analysis of the ancestral sequence

We accomplished our study through sequence based and structural analysis on the selected ancestral nodes that encompasses the evolutionary path of TLR9. Sequence based analyses such as multiple sequence alignment of the ancestral sequences, analysis of the functional domains, estimation of synonymous and nonsynonymous substitution was performed in order to understand the gradual changes occurred during TLR9 evolution. Structural studies were also performed to assess the functional changes.

Multiple sequence alignment (MSA) generated a percent identity matrix of the protein sequences to provide an overview of the similarities between the sequences. The heatmap of the percent identity matrix reported from the alignment is displayed in Figure5. Higher sequence identity of TLR9 with its immediate ancestor (Node235) but lower sequence identity with the root (Node226) was observed. It suggests that the continuous changes in sequence level along the ancestral lineages lead to the distinct sequence pattern of TLR9. Prediction of domain of the selected protein sequences was done and the number of LRR in the ectodomain was calculated. The orientation of LRRs in the ancestral lineages was different compared to

Human_TLR9 and its immediate ancestral node. LRRs are the important components of the functional domains of TLRs that recognize the pathogen associated molecular pattern (PAMP). Variation in the number of LRR in the ancestors of TLR9 was observed (Figure6). It suggests that during the evolution the variations among the LRRs of the ancestral nodes contributed to the specific pattern recognition of TLR9.

To observe these differences in structural level structural models of the ancestral nodes and Human_TLR9 from the existing TLR9 group were prepared and compared through pairwise structural alignment (Supplementary File1). Root mean square deviation (RMSD) and TM-score (template modeling score) were important metric in this analysis. The RMSD values of TLR9 with the root node was higher compared to the other ancestral nodes and it gradually decreased in other nodes. These observations also showed more deviation of TLR9 from root with respect to other TLRs along the ancestral nodes in the evolution of TLR9. For all the pairwise structural alignment TM-score variation was observed but the values indicated that they are in the same protein fold.

TLR9 is a receptor for sensing bacterial DNA/CpG-containing oligodeoxynucleotides (CpG ODN) as PAMP within the endosomal compartment. Interaction study of ancestral proteins with this known ligand of Human_TLR9 was performed. It will help to understand how the present ligand is selected through evolution facilitating stronger interaction with TLR9. Interaction of Human_TLR9 and CpG ODN was also studied. Docking score of all the interactions are shown in Figure7. Highest docking score observed in case of Human_TLR9 indicated the most compatible interaction of the ligand with present TLR9. It reveals that TLR9 achieved its present conformation through the structural changes in the ancestral nodes during the course of evolution. Present TLR9 is very specific in recognizing its ligand as the ancestral nodes showed comparatively less stable interaction with this ligand.

Assessment of the evolutionary impact on the ancestral node sequences was also done by measuring the changes in non-synonymous substitution (Ka), synonymous substitution (Ks), ratio of non-synonymous and synonymous substitution (Ka/Ks) (Figure8). Gradual increase of

Ks from root to the other ancestral nodes was seen and it became extremely high in Human_TLR9. Ka value is also high in Human_TLR9 compared to the ancestral sequences. Due to high value of Ks the Ka/Ks value became very low in Human_TLR9. Influence of synonymous substitution have been shaping the TLR9 evolution compared to its ancestral nodes.

Nodes	NODE226	NODE227	NODE337	NODE232	NODE233	NODE234	NODE299	NODE256	NODE235	Human_TLR9
NODE226	100	89.75	75.86	83.83	74.79	50.92	62.53	46.73	36.46	34.81
NODE227	89.75	100	68.02	74.69	66.3	46.69	56.95	42.98	33.89	32.97
NODE337	75.86	68.02	100	85.16	73.84	49.58	60.46	45.73	35.71	34.22
NODE232	83.83	74.69	85.16	100	83.79	58.16	67.68	53.4	40.97	38.45
NODE233	74.79	66.3	73.84	83.79	100	66.49	74.95	60.56	45.49	42.73
NODE234	50.92	46.69	58.16	58.16	66.49	100	50.28	88.14	58.15	52.78
NODE299	62.53	56.95	60.46	67.68	74.95	50.28	100	45.76	36.47	35.87
NODE256	46.73	42.98	45.73	53.4	60.56	88.14	45.76	100	49.02	45.61
NODE235	36.46	33.89	35.71	40.97	45.49	58.15	36.47	49.02	100	84.09
Human_TLR9	34.81	32.97	34.22	38.45	42.73	52.78	35.87	45.61	84.09	100

Figure 5: Heatmap showing percent identity matrix of proteins obtained from multiple sequence alignment, colours correspond to the percent identity with high values (red), medium values (white) and low values (blue). Values in the box represent sequence homology in percentage. Higher sequence identity of TLR9 with its immediate ancestor (Node235) but lower sequence identity with the ancestral nodes was observed.

Figure 6: Number of LRR present in the TLR genes and the ancestral nodes are shown in the bar plot. Number of LRR in human_TLR9 is decreased from its immediate ancestor Node235.

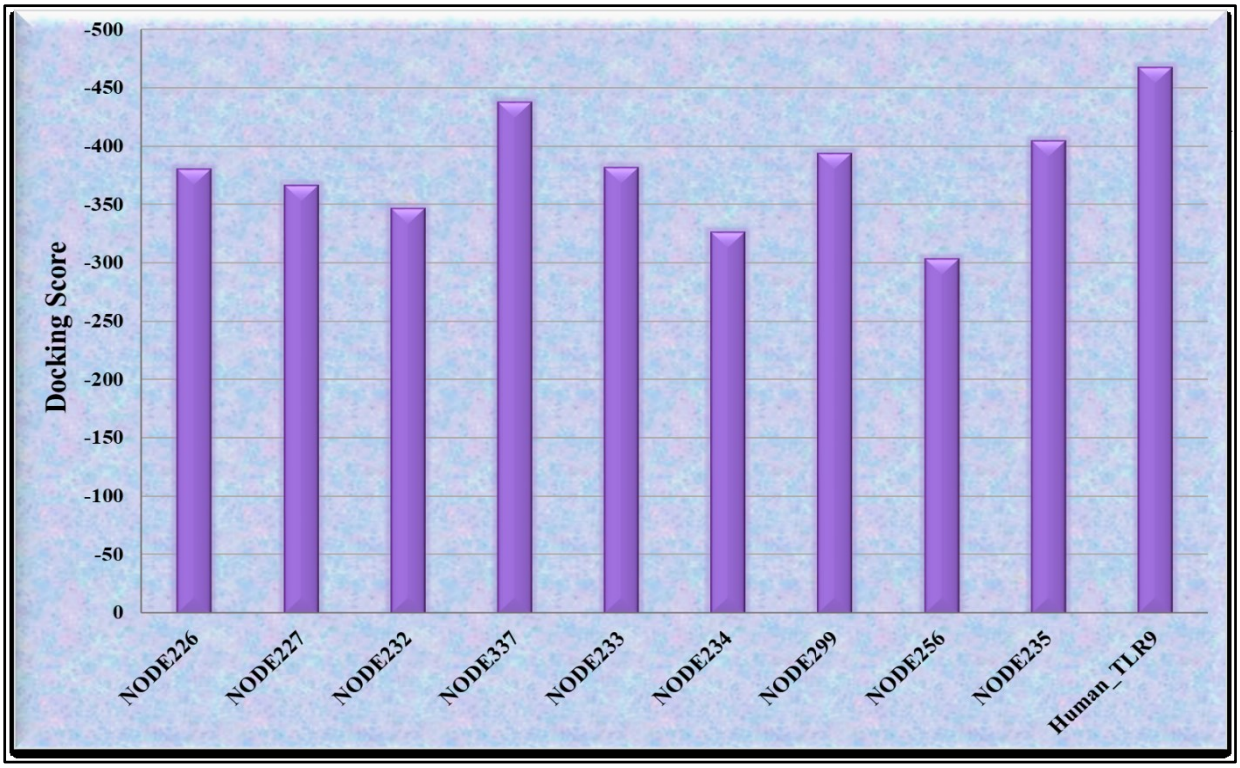


Figure 7: Docking score of the interaction analysis between selected sequences and known ligand of CpG DNA of TLR9. Highest docking score is observed in case of TLR9.

Table 1: Pairwise structural alignment results of the ancestral proteins and TLR9. All the selected ancestral sequences were compared with each other through pairwise structural sequence alignment. Two important metrics of this study are RMSD (Root mean square deviation) and TM-score (template modeling score). RMSD values of TLR9 were less when compared with the immediate ancestors but higher when compared with other ancestral nodes. TM-score values indicated that they are in the same protein fold.

Reference	Target	RMSD	TM-score	Sequence Identity	Reference	Target	RMSD	TM-score	Sequence Identity
Node226	Node227	4.23	0.5	34%	**Node299**	Node226	5.06	0.54	34%
	Node232	5.84	0.53	21%		Node227	1.44	0.89	50%
	Node233	5.76	0.49	16%		Node232	3.51	0.55	31%
	Node337	1.46	0.85	66%		Node233	4.65	0.61	37%
	Node299	5.06	0.54	34%		Node337	4.55	0.55	30%
	Node234	5.67	0.52	14%		Node234	4.09	0.61	23%
	Node256	5.67	0.52	14%		Node256	4.06	0.6	22%
	Node235	5.74	0.53	15%		Node235	3.37	0.62	24%
	TLR9	4.44	0.44	19%		TLR9	3.35	0.62	23%
Node227	Node226	4.23	0.5	34%	**Node234**	Node226	5.67	0.57	14%
	Node232	3.54	0.56	31%		Node227	4.02	0.63	24%
	Node233	4.3	0.62	40%		Node232	1.53	0.97	54%
	Node337	4.55	0.48	30%		Node233	1.66	0.87	58%
	Node299	1.44	0.89	50%		Node337	4.58	0.5	19%
	Node234	4.02	0.63	24%		Node299	4.09	0.61	23%
	Node256	3.87	0.61	24%		Node256	1.72	0.97	86%
	Node235	3.46	0.6	26%		Node235	2.07	0.88	59%
	TLR9	3.48	0.6	25%		TLR9	2.65	0.89	53%
Node232	Node226	5.84	0.53	21%	**Node256**	Node226	5.67	0.56	14%
	Node227	3.54	0.56	31%		Node227	3.87	0.61	24%
	Node233	1.76	0.77	69%		Node232	1.91	0.95	48%
	Node337	5.51	0.42	27%		Node233	1.48	0.85	53%
	Node299	3.51	0.55	31%		Node337	5.83	0.52	18%
	Node234	1.53	0.88	54%		Node299	4.06	0.6	22%
	Node256	1.91	0.87	48%		Node234	1.72	0.95	86%
	Node235	2.31	0.79	41%		Node235	2.22	0.87	49%
	TLR9	3.07	0.79	37%		TLR9	2.09	0.87	45%
Node233	Node226	5.76	0.6	16%	**Node235**	Node226	5.74	0.6	15%
	Node227	4.3	0.62	40%		Node227	3.46	0.6	26%
	Node232	1.76	0.96	69%		Node232	2.31	0.91	41%
	Node337	5.49	0.45	15%		Node233	2.18	0.83	43%
	Node299	4.65	0.61	37%		Node337	4.48	0.44	18%
	Node234	1.66	0.97	58%		Node299	3.37	0.62	24%
	Node256	1.48	0.98	53%		Node234	2.07	0.92	59%
	Node235	2.18	0.9	43%		Node256	2.22	0.92	49%
	TLR9	2.24	0.83	40%		TLR9	0.92	0.98	83%
Node337	Node226	1.46	0.85	66%	**TLR9**	Node226	4.44	0.44	19%
	Node227	4.55	0.48	30%		Node227	3.48	0.6	25%
	Node232	5.51	0.42	27%		Node232	3.07	0.79	37%
	Node233	5.49	0.45	15%		Node233	2.24	0.83	40%
	Node299	4.55	0.55	30%		Node337	4.66	0.49	14%
	Node234	4.58	0.5	19%		Node299	3.35	0.62	23%
	Node256	5.83	0.52	18%		Node234	2.65	0.89	53%
	Node235	4.48	0.44	18%		Node256	2.09	0.87	45%
	TLR9	4.66	0.49	14%		Node235	0.92	0.98	83%

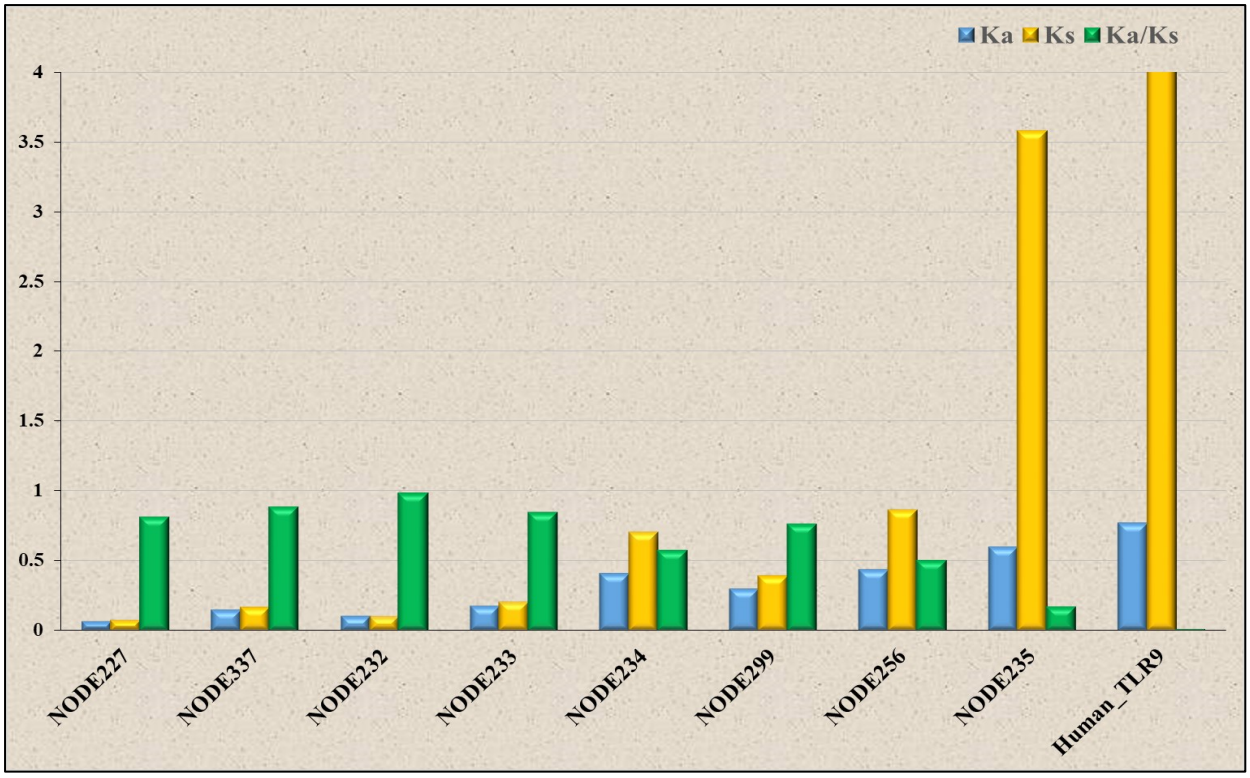


Figure 8: Synonymous (Ks) and non-synonymous (Ka) substitution rates in TLR9 and its ancestral node.

Discussion

The transmembrane pattern recognition receptor TLRs are best known for their roles in innate immunity via recognition of pathogen and initiation of signaling response. In this study, comprehensive analysis of mammalian toll-like receptor gene sequences (TLR1-TLR10) revealed that TLR9 follows a distinct pattern of evolution. Sequence based features and evolutionary constraints are found to influence the divergence of TLR9 from other TLRs. Ancestral sequence reconstruction analysis also revealed that gradual evolution of TLR genes in several ancestral lineages lead to the distinct pattern of TLR9.

Mammalian TLRs are responsible for recognition of conserved molecular pattern derived from various classes of pathogens resulting in the induction of innate immune response. Pathogen-induced selection is considered as a crucial selective mechanism driving the evolution of immune system components. We have identified various factors influencing TLR-dependent heterogeneity in amino acid usage that contribute to the differences in their immunological responses in mammals. We also found that high synonymous substitutions have shaped the

observed changes between TLR9 and other mammalian TLR genes in spite of nonsynonymous substitutions inducing the amino acid changes.

The divergence of TLR9 is demonstrated in this study through the ancestral sequence reconstruction. Analysis of the ancestral sequences also reinforced that changes occurred in the TLRs during their evolution from the ancestral lineages that mostly observed in the TLR9 and its descendants. Decrease in percent sequence identity of TLR9 from root to the ancestral nodes to the mammalian TLR9 branch of the tree depicts gradual changes happened in the sequences through accumulation of mutation. Domain-wise analysis also suggested accumulation of a greater number of mutations in the ectodomain causing variation in the number of LRR. Each TLR comprise an ectodomain with leucine-rich repeats (LRRs) that facilitate the recognition of pathogen associated molecular pattern (PAMP) and a cytoplasmic Toll/IL-1 receptor (TIR) domain that initiates downstream signaling. The mutational changes also have been influenced by gradual selection pressure on the ancestral sequences in the course of evolution. Influence of synonymous and non-synonymous substitution among the ancestral sequences is observed and the gradual selection pressure in the course of evolution leading to the distinct pattern of TLR9. Interaction study also revealed more stable interaction of the ligand with TLR9 compared to the ancestral nodes. Although decreasing docking score in other ancestral nodes indicated less stable interaction.

Conclusion

This study enables a new approach to explore the emergence of toll-like receptor through the ancestral sequence reconstruction that elucidates a distinct pattern of evolution of TLR9. It demonstrates that the evolutionary divergence of TLR9 started from the beginning and gradual accumulation of changes in the ancestral lineages leads to the distinct pattern of TLR9 compared to the other mammalian TLRs. It will elucidate the biological significance of TLR9 and provide evidence for their distinct contributions in response to host defence.

Chapter - VI

Structural and functional objectivity of TLR evolution

Results presented in this chapter are published in the articles mentioned below:

1) *Ghosh M, Basak S, Dutta S. Natural selection shaped the evolution of amino acid usage in mammalian toll like receptor genes. Comput Biol Chem. 2022;97:107637. doi:10.1016/j.compbiolchem.2022.107637*
2) *Ghosh M, Basak S, Dutta S. Evolutionary divergence of TLR9 through ancestral sequence reconstruction. Immunogenetics. 2024;76(3):203-211. doi:10.1007/s00251-024-01338-8*

Background

Plants and animals have extensive inbuilt mechanisms for recognising and responding to harmful pathogens. The innate immune system is a ubiquitous and evolutionary ancient mechanism that serves as the first line of defence of host against infections (*Janeway and Medzhitov, 2002*, *Lemaitre and Hoffmann, 2007*). In vertebrates, invertebrates, and plants, innate immunity is based on a diverse set of germline-encoded receptors known as pattern-recognition receptors (PRRs), or microbial sensors, that recognise molecular motifs shared by specific groups of microorganisms (often referred to as pathogen-associated molecular patterns or PAMPs) (*Kimbrell and Beutler, 2001*). The last decade has witnessed a lot of significant improvements in the understanding of PRR-mediated immunity, with Toll-like receptors (TLRs) being one of the largest and most studied PRR families (*Akira et al, 2001*).

The toll gene in Drosophila is the prototype of the TLR family, first discovered for its role in dorso-ventral embryo patterning (*Anderson et al, 1985*) and later demonstrated to be necessary for efficient immune responses in adult flies against fungus and Gram-positive bacteria (*Lemaitre and Hoffmann, 2007)*. Since then, homologs of the Drosophila toll have been discovered in numerous other species (*Leulier and Lemaitre, 2008*). The role of mammalian TLRs in host defence has been examined mostly in vitro through stimulation with various agonists, and knocked-out mice for one or more TLRs exhibit varying vulnerability to several experimental infections (*Qureshi and Medzhitov, 2003*). TLRs are now known to respond to a variety of pathogen-associated stimuli and transmit signalling responses necessary for the activation of innate immunity effector mechanisms and the subsequent development of adaptive immunity (*Beutler et al, 2006*).

In humans, the TLR family has ten functional members (TLR1-TLR10) (*West et al, 2006*). Human TLRs are classified according to their subcellular distribution: TLR3, TLR7, TLR8, and TLR9 are commonly found in intracellular compartments such as endosomes, whereas TLR1, TLR2, TLR4, TLR5, and TLR6 are generally expressed on the cell surface (*Akira et al, 2006*). TLRs can be further subdivided according to known agonists. Intracellular TLRs detect nucleic acid-based agonists and are particularly specialised in viral recognition, whereas cell-surface expressed TLRs detect glycolipids, lipopeptides, and flagellin, which are found in a wide range of organisms including bacteria, parasites, and fungi (*Kawai and Akira, 2006*). TLR10, which is expressed on the cell surface, is the sole orphan TLR member whose agonists and activities are currently unknown. The role of human TLRs in host defence during natural infections, as opposed to experimental infections, is only now beginning to be understood.

The evolutionary genetics method has improved our understanding of the evolutionary factors that influence the human genome, making it an essential complement to clinical and epidemiological genetics techniques (*Nielsen, 2005; Nielsen et al, 2007*). In the context of infection, determining the extent and type of natural selection acting on genes involved in immunity-related processes can provide insights into the mechanisms of host defence mediated by them, as well as distinguish between genes that are essential in host defence versus those that exhibit higher immunological redundancy (*Quintana-Murci et al, 2007*).

Methodology

Sequence retrieval and correspondence analysis

Sequences of mammalian toll-like receptor (TLR) genes and their encoding proteins representing different group of TLR such as TLR1, TLR2, TLR3, TLR4, TLR5, TLR6, TLR7, TLR8, TLR9, TLR10, TLR11, TLR12, TLR13 were obtained from GenBank, NCBI. Those sequences containing unrecognized start codon, stop codon, internal stop codons, untranslatable codons, and unrecognized character (other than a, t, g, c) have been discarded from the final dataset.

Correspondence analysis (COA) (Peden 2000) was used to investigate the major trend in amino acid usage variation among the mammalian TLRs. Since amino acid usage by its very nature is multivariate, it is necessary to analyse this data with multivariate statistical techniques i.e., COA. Correspondence analysis (COA) is an ordination technique that identifies the major trends in the variation of the data and distributes genes along continuous axes in accordance with these trends. It has the advantage of not to make any assumption that the data falls into discrete clusters and therefore represent continuous variation accurately (*Roy et al. 2017*). Parameters such ad GC content, GC3, effective number of codons (ENc), hydrophobicity, aromaticity etc. were also calculated for all the TLRs under study. These analyses were performed using the CodonW program. Correlation coefficient, statistical significance of the parameters was calculated using the tools freely available in GraphPad software.

Phylogenetic tree construction

Phylogenetic analysis provides the evolutionary relationship of a set of sequences. It involves the construction of a tree, where the nodes indicate separate evolutionary paths, and the lengths of the branches give an estimate of how distantly related the sequences represented by those branches are. Three phylogenetic trees were generated for the Mammalian TLRs using the maximum likelihood method with thousand bootstrap replicates in the MEGAX. MEGA, a comprehensive tool for performing sequence alignment and inferring phylogenetic trees was used for generating the trees (*Kumar et al. 2018*).

Evolutionary rate analysis

Evolutionary selection acting on the mammalian TLR genes are addressed by evolutionary rate (ω). ω is estimated as the ratio of the rate non-synonymous substitutions per non-synonymous site (Ka) and the rate of synonymous substitutions per synonymous site (Ks). $\omega > 1$ indicates positive (diversifying) selection, whereas, $\omega < 1$ indicates negative (purifying) selection. For each TLR group (Example: TLR1) their consensus nucleotide sequences (Example: TLR1_consensus) were generated. We have prepared a Perl script for generating these consensus sequences. Downloaded nucleotide sequences and the consensus sequence of each

TLR groups were subjected to Clustal Omega program (Madeira et al. 2022) for the nucleotide sequence alignment. Then the evolutionary rate of the TLR genes (TLR1-TLR10) of each TLR group (Example: TLR1) were estimated relative to their consensus (Example: TLR1_consensus) sequences using Codeml program of the PAML software (ver. 4.5) with runmode = −2 and CodonFreq= 1 (*Nei and Gojobori 1986; Yang 2007*).

Results

Analysis of amino acid usage pattern

Mammalian TLR genes demonstrated differential amino acid usage pattern from the correspondence analysis (COA) study. Figure1 display four different clusters based on amino acid usage pattern (marked in red circle A, B, C and D). Cluster A comprises TLR1, TLR2, TLR4, TLR6, TLR10; Cluster B comprises TLR3, TLR7, TLR8. Cluster C and Cluster D comprises TLR5 and TLR9 respectively. TLR11, TLR12, TLR13 displayed scatter distribution of genes based on amino acid usage. We found that axis1 of the COA correspond to major variation (57.57%) of amino acid usage. It is clear from Figure1 that among the four clusters TLR9 exhibit widely different amino acid usage with respect to the other three clusters along axis1. Different physico-chemical parameters such as hydrophobicity, aromaticity, GC-content, ENc were also analyzed in order to assess the factors influencing this distinct amino acid usage pattern. Significant ($p < .01$) correlation of these parameters was observed with axis1 of COA.

Impact of subcellular localization and function of TLRs were observed in clustering pattern. TLRs found on Cluster A, Cluster C are expressed extracellularly and responsible for the recognition of lipoproteins, lipopeptides, LPS etc. TLRs found on Cluster B, Cluster D are expressed intracellularly and responsible for the recognition of nucleic acid motifs.

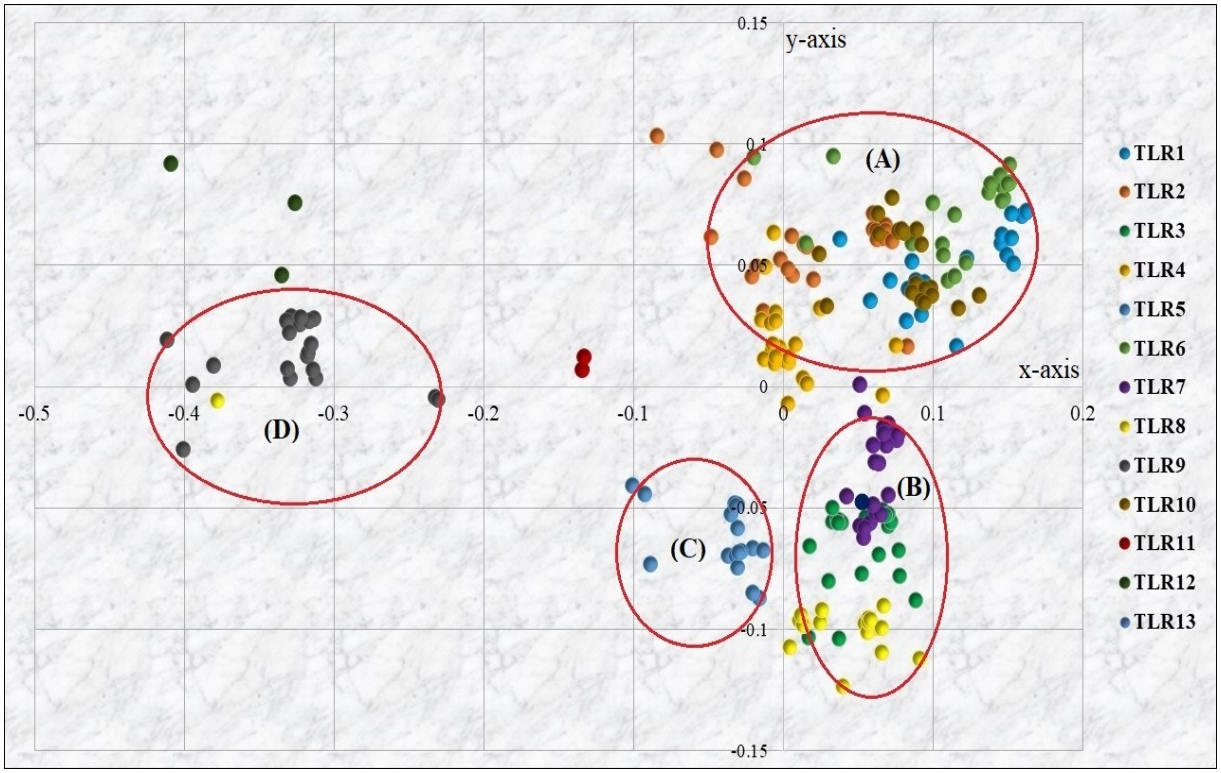


Figure 1: Distribution of mammalian toll-like receptor (TLR) genes along the two major axes of correspondence analysis (CoA) on amino acid usage. Four different clusters observed are marked with the red circle. Cluster A comprises TLR1, TLR2, TLR4, TLR6, TLR10; Cluster B comprises TLR3, TLR7, TLR8. Cluster C comprises TLR5 and Cluster D comprises TLR9.

Phylogenetic tree

Phylogenetic tree of mammalian TLR genes showing TLR wise branching pattern as displayed in Figure2. One clade contains the branches of TLR1, TLR2, TLR4, TLR6, TLR10; another clade contains the branches of TLR3, TLR7, TLR8, TLR9, TLR5 and TLR11, TLR12, TLR13 formed a separate clade. Evolutionary relationship from phylogenetic tree followed the similar trend to that of COA on amino acid usage pattern.

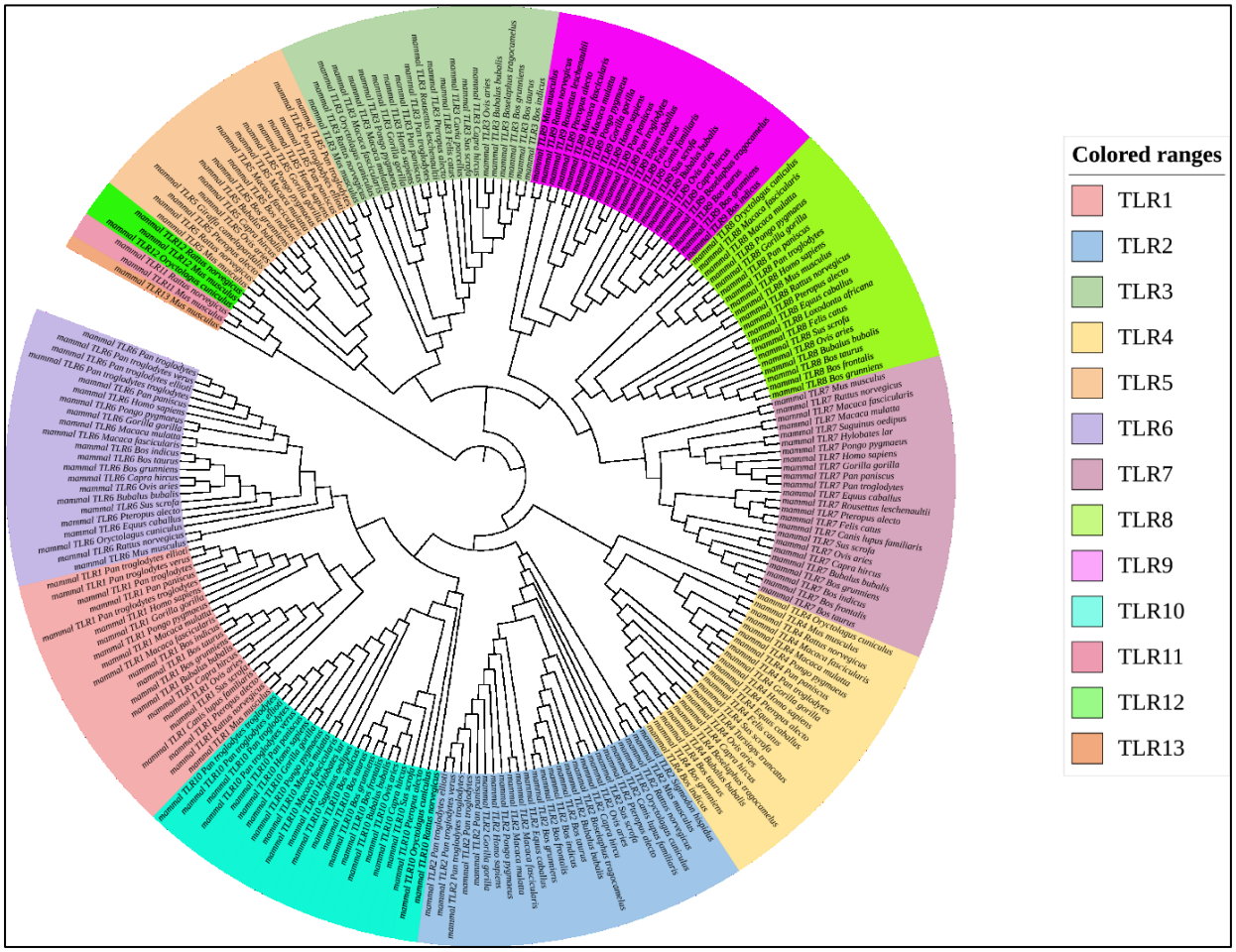


Figure 2: Phylogenetic tree of mammalian TLR genes showing TLR wise branching pattern. Individual colors represent different TLRs.

Estimation of substitution rate

Evolutionary selection acting on the mammalian TLR genes have been addressed by evolutionary rate which is estimated as the ratio of the rate of non-synonymous substitutions per non-synonymous site (Ka) and the rate of synonymous substitutions per synonymous site (Ks). Figure3 displayed overall purifying selection is observed for all the mammalian TLRs under study. Ks values indicated high number of synonymous substitutions with highest number of synonymous substitutions in TLR9. However, mammalian TLR genes exhibited several non-synonymous changes as indicated by the Ka values. Significant ($p < .01$) correlation of Ka, Ks and Ka/Ks have been observed with axis1 of correspondence analysis that accounted major variation of amino acid usage.

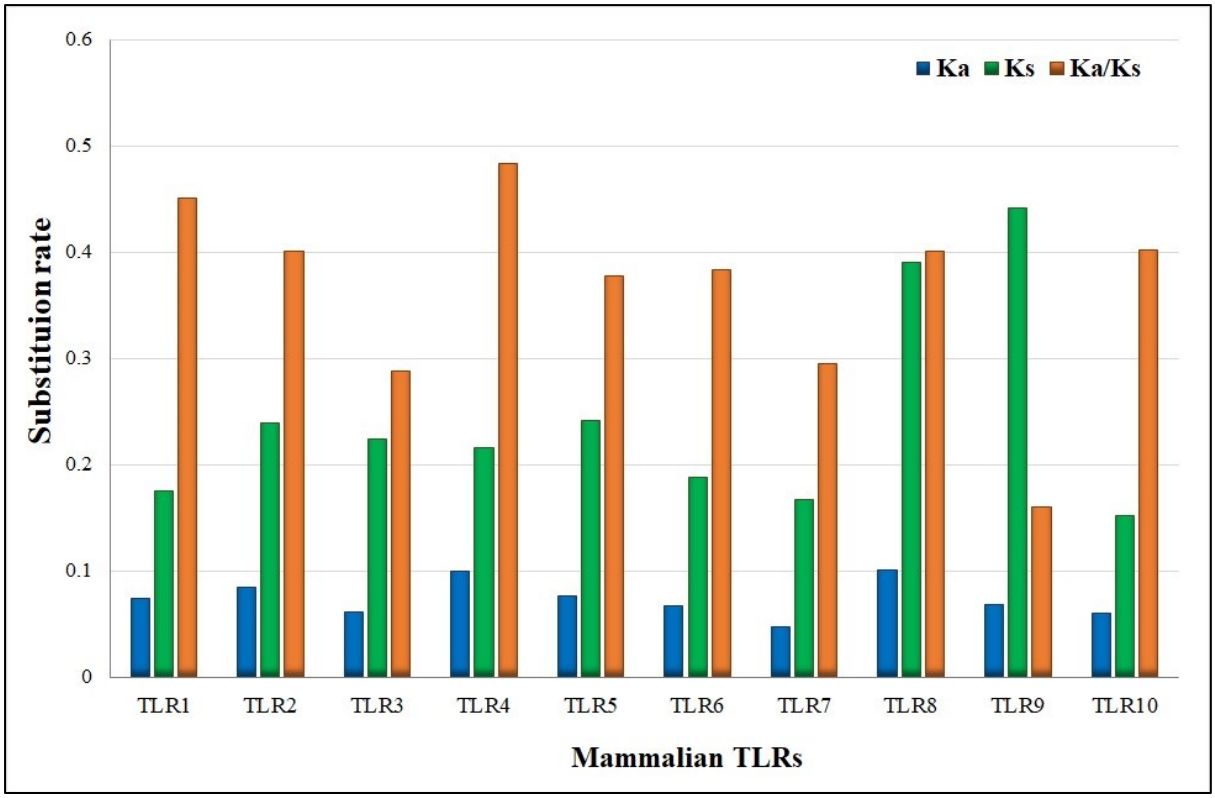


Figure 3: Bar plot showing Synonymous (Ks), non-synonymous (Ka) substitution rates and evolutionary rate (Ka/Ks) distribution of mammalian TLRs. It is clear from the plot that Ks value is higher in case of TLR9, TLR8.

Discussion

The family of vertebrate toll-like receptors (TLRs) serves as the first line of immunological defence against a variety of pathogens and is an intriguing illustration of the host-pathogen evolutionary contest. This study presents a complete comparative evolutionary genomics characterization of the vertebrate TLR family through DNA and protein level analysis. Our findings revealed the dynamic evolution of the TLRs across vertebrates with positive selection shaping adaptive evolution of host pathogen.

Amino acid usage pattern revealed distinct pattern of distribution of TLR genes among mammal and bird and dispersed pattern in fish TLRs. Clusters observed in the mammalian TLR distribution was typically influenced by their function and the subcellular localization along with the physicochemical parameters analyzed. TLR1,2,4,6,10 are surface expressed and they mostly recognize lipoproteins, lipopeptides, LPS etc., TLR5 are surface expressed and confers response to flagellin, TLR3,7,8 reside intracellularly and respond to double-stranded RNA (dsRNA), single-stranded RNA (ssRNA) and TLR9 reside intracellularly and respond to DNA. Among them TLR9 formed a distinct cluster. Scatter distribution of TLR11,12,13 have been found. Among the ten TLRs of bird the key influence of amino acid usage distribution was their function. Mammalian orthologs TLR3, TLR4, TLR5 and TLR7 recognize dsRNA, bacterial lipopolysaccharides, flagellin, ssRNA respectively. TLR1A/TLR1B and TLR2A/TLR2B arose by duplication during their evolution recognize di/triacylated lipopeptides. TLR15 is unique to birds that has evolved to perform a new function in the identification of extracellular proteases and TLR21 in birds recognises CpG DNA similarly to TLR9 in mammals showed distinct pattern. Fish TLRs having diverse function showed scattered amino acid usage pattern. Phylogenetic tree indicated the grouping of TLRs from the three taxonomic groups was analogous with their amino acid usage pattern.

Thus, the evolution of TLRs have been significantly influenced by their amino acid usage and the physico-chemical parameters of the protein. This change in amino acid usage is a result of substitution as observed from the evolutionary rate analysis. Both synonymous and non-

synonymous substitution impacted on the evolution of TLRs in mammal, bird and fish. In spite of high value of synonymous substitution there are non-synonymous substitution that has contributed to the diversity of TLRs. Positive selection is one of the distinguishing features of immune defense related genes, particularly those encoding recognition proteins, which evolve under positive selection. Positively selected sites among TLRs depicted the gradual accumulation of changes has shaped the TLR evolution. It indicated the diversity in the evolution of TLRs from various taxonomic group through accumulation of changes that lead to their distinct pattern of pathogen recognition. The location of the positively selected sites suggests that pathogens impose the utmost selective pressures that result in the alterations observed, particularly in the variable section involved for direct contact with PAMPS. This implies that they are the outcome of co-evolution.

Conclusion

This study revealed differential pattern of amino acid in the distribution of the TLRs among vertebrates particularly mammal, bird and fish. In spite of the presence of evolutionary constraints, variable rates of substitutions leading to various TLR repertoires that would have facilitated recognition and protection from a variety of diseases.

Chapter - VII

Conclusion

TLR family members recognise numerous types of pathogens and coordinate appropriate innate and adaptive immune responses. The coding sequences and functions of vertebrate TLRs are largely conserved. Similarly, TLR-mediated signalling pathways are substantially conserved. Ligand characterizations of TLRs have facilitated the understanding of the function of the TLRs and the host defense system against infections. In my thesis work an innovative approach is provided by incorporating the examination of the variation in the frequency of amino acids utilized by different TLRs of mammalian species. I have also addressed the distinct evolution of some TLRs by the molecular evolutionary approach based on ancestral reconstructions that has helped in retracing the history of sequence substitutions and leveraging changes in substitution dynamics.

My thesis work indicating that TLR genes evolved in different ways across primate and non-primate mammalian species might help to understand the genetic basis for variances in disease susceptibility with respect to host immunity. Determination of magnitude of natural selection operating on TLR genes and the progressive changes that lead to divergence have enabled better understanding of the mechanism of host defence mediated by TLRs. This work is important in integrating evolutionary genetic data into a clinical and epidemiological framework, for better understanding of the relevance of host defense genes for their survival in nature.

References

- Adema CM, Hillier LW, Jones CS, et al. Whole genome analysis of a schistosomiasis-transmitting freshwater snail [published correction appears in Nat Commun. 2017 Aug 23;8:16153. doi: 10.1038/ncomms16153]. *Nat Commun*. 2017;8:15451. Published 2017 May 16. doi:10.1038/ncomms15451
- Agu PC, Afiukwa CA, Orji OU, et al. Molecular docking as a tool for the discovery of molecular targets of nutraceuticals in diseases management. *Sci Rep*. 2023;13(1):13398. Published 2023 Aug 17. doi:10.1038/s41598-023-40160-2
- Ahmad-Nejad P, Häcker H, Rutz M, Bauer S, Vabulas RM, Wagner H. Bacterial CpG-DNA and lipopolysaccharides activate Toll-like receptors at distinct cellular compartments. *Eur J Immunol*. 2002;32(7):1958-1968. doi:10.1002/1521-4141(200207)32:7<1958::AID-IMMU1958>3.0.CO;2-U
- Akira S, Takeda K, Kaisho T. Toll-like receptors: critical proteins linking innate and acquired immunity. *Nat Immunol*. 2001;2(8):675-680. doi:10.1038/90609
- Akira S, Takeda K. Toll-like receptor signalling. *Nat Rev Immunol*. 2004;4(7):499-511. doi:10.1038/nri1391
- Akira S, Uematsu S, Takeuchi O. Pathogen recognition and innate immunity. *Cell*. 2006;124(4):783-801. doi:10.1016/j.cell.2006.02.015
- Alcaide M, Edwards SV. Molecular evolution of the toll-like receptor multigene family in birds. *Mol Biol Evol*. 2011;28(5):1703-1715. doi:10.1093/molbev/msq351
- Alexopoulou L, Holt AC, Medzhitov R, Flavell RA. Recognition of double-stranded RNA and activation of NF-kappaB by Toll-like receptor 3. *Nature*. 2001;413(6857):732-738. doi:10.1038/35099560
- Alexopoulou L, Thomas V, Schnare M, et al. Hyporesponsiveness to vaccination with Borrelia burgdorferi OspA in humans and in TLR1- and TLR2-deficient mice. *Nat Med*. 2002;8(8):878-884. doi:10.1038/nm732
- Anandhakumar C, Lavanya V, Pradheepa G, et al. Expression profile of toll-like receptor 2 mRNA in selected tissues of shark (Chiloscyllium sp.). *Fish Shellfish Immunol*. 2012;33(5):1174-1182. doi:10.1016/j.fsi.2012.09.007
- Anderson KV, Bokla L, Nüsslein-Volhard C. Establishment of dorsal-ventral polarity in the Drosophila embryo: the induction of polarity by the Toll gene product. *Cell*. 1985;42(3):791-798. doi:10.1016/0092-8674(85)90275-2
- Anderson KV, Bokla L, Nüsslein-Volhard C. Establishment of dorsal-ventral polarity in the Drosophila embryo: the induction of polarity by the Toll gene product. *Cell*. 1985;42(3):791-798. doi:10.1016/0092-8674(85)90275-2
- Areal H, Abrantes J, Esteves PJ. Signatures of positive selection in Toll-like receptor (TLR) genes in mammals. *BMC Evol Biol*. 2011;11:368. Published 2011 Dec 20. doi:10.1186/1471-2148-11-368
- Arts JA, Cornelissen FH, Cijsouw T, Hermsen T, Savelkoul HF, Stet RJ. Molecular cloning and expression of a Toll receptor in the giant tiger shrimp, Penaeus monodon. *Fish Shellfish Immunol*. 2007;23(3):504-513. doi:10.1016/j.fsi.2006.08.018
- Augustin R, Fraune S, Bosch TC. How Hydra senses and destroys microbes. *Semin Immunol*. 2010;22(1):54-58. doi:10.1016/j.smim.2009.11.002
- Ausubel FM. Are innate immune signaling pathways in plants and animals conserved?. *Nat Immunol*. 2005;6(10):973-979. doi:10.1038/ni1253
- Babik W, Dudek K, Fijarczyk A, et al. Constraint and adaptation in newt toll-like receptor genes. *Genome Biol Evol*. 2014;7(1):81-95. Published 2014 Dec 4. doi:10.1093/gbe/evu266

- Bagheri M, Zahmatkesh A. Evolution and species-specific conservation of toll-like receptors in terrestrial vertebrates. *Int Rev Immunol*. 2018;37(5):217-228. doi:10.1080/08830185.2018.1506780
- Barreiro LB, Ben-Ali M, Quach H, et al. Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet*. 2009;5(7):e1000562. doi:10.1371/journal.pgen.1000562
- Basith S, Manavalan B, Lee G, Kim SG, Choi S. Toll-like receptor modulators: a patent review (2006-2010). *Expert Opin Ther Pat*. 2011;21(6):927-944. doi:10.1517/13543776.2011.569494
- Baumgarten S, Simakov O, Esherick LY, et al. The genome of Aiptasia, a sea anemone model for coral symbiosis. *Proc Natl Acad Sci U S A*. 2015;112(38):11893-11898. doi:10.1073/pnas.1513318112
- Belinda LW, Wei WX, Hanh BT, Lei LX, Bow H, Ling DJ. SARM: a novel Toll-like receptor adaptor, is functionally conserved from arthropod to human. *Mol Immunol*. 2008;45(6):1732-1742. doi:10.1016/j.molimm.2007.09.030
- Bell JK, Mullen GE, Leifer CA, Mazzoni A, Davies DR, Segal DM. Leucine-rich repeats and pathogen recognition in Toll-like receptors. *Trends Immunol*. 2003;24(10):528-533. doi:10.1016/s1471-4906(03)00242-4
- Bella J, Hindle KL, McEwan PA, Lovell SC. The leucine-rich repeat structure. *Cell Mol Life Sci*. 2008;65(15):2307-2333. doi:10.1007/s00018-008-8019-0
- Beutler B, Jiang Z, Georgel P, et al. Genetic analysis of host resistance: Toll-like receptor signaling and immunity at large. *Annu Rev Immunol*. 2006;24:353-389. doi:10.1146/annurev.immunol.24.021605.090552
- Beutler B, Rietschel ET. Innate immune sensing and its roots: the story of endotoxin. *Nat Rev Immunol*. 2003;3(2):169-176. doi:10.1038/nri1004
- Bhattacharyya S, Varga J. Endogenous ligands of TLR4 promote unresolving tissue fibrosis: Implications for systemic sclerosis and its targeted therapy. *Immunol Lett*. 2018;195:9-17. doi:10.1016/j.imlet.2017.09.011
- Boller T, Felix G. A renaissance of elicitors: perception of microbe-associated molecular patterns and danger signals by pattern-recognition receptors. *Annu Rev Plant Biol*. 2009;60:379-406. doi:10.1146/annurev.arplant.57.032905.105346
- Bosch TC, Augustin R, Anton-Erxleben F, et al. Uncovering the evolutionary history of innate immunity: the simple metazoan Hydra uses epithelial cells for host defence. *Dev Comp Immunol*. 2009;33(4):559-569. doi:10.1016/j.dci.2008.10.004
- Botos I, Segal DM, Davies DR. The structural biology of Toll-like receptors. *Structure*. 2011;19(4):447-459. doi:10.1016/j.str.2011.02.004
- Boudinot P, Zou J, Ota T, et al. A tetrapod-like repertoire of innate immune receptors and effectors for coelacanths. *J Exp Zool B Mol Dev Evol*. 2014;322(6):415-437. doi:10.1002/jez.b.22559
- Boyd A, Philbin VJ, Smith AL. Conserved and distinct aspects of the avian Toll-like receptor (TLR) system: implications for transmission and control of bird-borne zoonoses. *Biochem Soc Trans*. 2007;35(Pt 6):1504-1507. doi:10.1042/BST0351504
- Boyd AC, Peroval MY, Hammond JA, Prickett MD, Young JR, Smith AL. TLR15 is unique to avian and reptilian lineages and recognizes a yeast-derived agonist. *J Immunol*. 2012;189(10):4930-4938. doi:10.4049/jimmunol.1101790
- Bradley P, Misura KM, Baker D. Toward high-resolution de novo structure prediction for small proteins. *Science*. 2005;309(5742):1868-1871. doi:10.1126/science.1113801
- Brandt JP, Ringstad N. Toll-like Receptor Signaling Promotes Development and Function of Sensory Neurons Required for a C. elegans Pathogen-Avoidance Behavior. *Curr Biol*. 2015;25(17):2228-2237. doi:10.1016/j.cub.2015.07.037

- Brennan CA, Anderson KV. Drosophila: the genetics of innate immune recognition and response. *Annu Rev Immunol*. 2004;22:457-483. doi:10.1146/annurev.immunol.22.012703.104626
- Brennan JJ, Gilmore TD. Evolutionary Origins of Toll-like Receptor Signaling. *Mol Biol Evol*. 2018;35(7):1576-1587. doi:10.1093/molbev/msy050
- Brownlie R, Allan B. Avian toll-like receptors. *Cell Tissue Res*. 2011;343(1):121-130. doi:10.1007/s00441-010-1026-0
- Buckley KM, Rast JP. Dynamic evolution of toll-like receptor multigene families in echinoderms. *Front Immunol*. 2012;3:136. Published 2012 Jun 5. doi:10.3389/fimmu.2012.00136
- Chen R, Li L, Weng Z. ZDOCK: an initial-stage protein-docking algorithm. *Proteins*. 2003;52(1):80-87. doi:10.1002/prot.10389
- Chen Y, Lu H, Zhang N, Zhu Z, Wang S, Li M. PremPS: Predicting the impact of missense mutations on protein stability. *PLoS Comput Biol*. 2020;16(12):e1008543. Published 2020 Dec 30. doi:10.1371/journal.pcbi.1008543
- Cook DN, Pisetsky DS, Schwartz DA. Toll-like receptors in the pathogenesis of human disease. *Nat Immunol*. 2004;5(10):975-979. doi:10.1038/ni1116
- Coscia MR, Giacomelli S, Oreste U. Toll-like receptors: an overview from invertebrates to vertebrates. *Invert Surviv J.* 2011;8(2):210–26
- Cuvillier-Hot V, Boidin-Wichlacz C, Slomianny C, Salzet M, Tasiemski A. Characterization and immune function of two intracellular sensors, HmTLR1 and HmNLR, in the injured CNS of an invertebrate. *Dev Comp Immunol*. 2011;35(2):214-226. doi:10.1016/j.dci.2010.09.011
- Davidson CR, Best NM, Francis JW, Cooper EL, Wood TC. Toll-like receptor genes (TLRs) from Capitella capitata and Helobdella robusta (Annelida). *Dev Comp Immunol*. 2008;32(6):608-612. doi:10.1016/j.dci.2007.11.004
- de Castro E, Sigrist CJ, Gattiker A, et al. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Res*. 2006;34(Web Server issue):W362-W365. doi:10.1093/nar/gkl124
- de Zoete MR, Bouwman LI, Keestra AM, van Putten JP. Cleavage and activation of a Toll-like receptor by microbial proteases. *Proc Natl Acad Sci U S A*. 2011;108(12):4968-4973. doi:10.1073/pnas.1018135108
- de Zoete MR, Bouwman LI, Keestra AM, van Putten JP. Cleavage and activation of a Toll-like receptor by microbial proteases. *Proc Natl Acad Sci U S A*. 2011;108(12):4968-4973. doi:10.1073/pnas.1018135108
- Diebold SS, Kaisho T, Hemmi H, Akira S, Reis e Sousa C. Innate antiviral responses by means of TLR7-mediated recognition of single-stranded RNA. *Science*. 2004;303(5663):1529-1531. doi:10.1126/science.1093616
- Downing T, Cormican P, O'Farrelly C, Bradley DG, Lloyd AT. Evidence of the adaptive evolution of immune genes in chicken. *BMC Res Notes*. 2009;2:254. Published 2009 Dec 15. doi:10.1186/1756-0500-2-254
- Du MZ, Zhang C, Wang H, Liu S, Wei W, Guo FB. The GC Content as a Main Factor Shaping the Amino Acid Usage During Bacterial Evolution Process. *Front Microbiol*. 2018;9:2948. Published 2018 Dec 7. doi:10.3389/fmicb.2018.02948
- Du X, Poltorak A, Wei Y, Beutler B. Three novel mammalian toll-like receptors: gene structure, expression, and evolution. *Eur Cytokine Netw*. 2000;11(3):362-371.
- Erridge C. Endogenous ligands of TLR2 and TLR4: agonists or assistants?. *J Leukoc Biol*. 2010;87(6):989-999. doi:10.1189/jlb.1209775

- Ferrandon D, Imler JL, Hetru C, Hoffmann JA. The Drosophila systemic immune response: sensing and signalling during bacterial and fungal infections. *Nat Rev Immunol*. 2007;7(11):862-874. doi:10.1038/nri2194
- Fink IR, Pietretti D, Voogdt CGP, et al. Molecular and functional characterization of Toll-like receptor (Tlr)1 and Tlr2 in common carp (Cyprinus carpio). *Fish Shellfish Immunol*. 2016;56:70-83. doi:10.1016/j.fsi.2016.06.049
- Fornarino S, Laval G, Barreiro LB, Manry J, Vasseur E, Quintana-Murci L. Evolution of the TIR domain-containing adaptors in humans: swinging between constraint and adaptation. *Mol Biol Evol*. 2011;28(11):3087-3097. doi:10.1093/molbev/msr137
- Fornůsková A, Vinkler M, Pagès M, et al. Contrasted evolutionary histories of two Toll-like receptors (Tlr4 and Tlr7) in wild rodents (MURINAE). *BMC Evol Biol*. 2013;13:194. Published 2013 Sep 12. doi:10.1186/1471-2148-13-194
- Forsthoefel DJ, James NP, Escobar DJ, et al. An RNAi screen reveals intestinal regulators of branching morphogenesis, differentiation, and stem cell proliferation in planarians. *Dev Cell*. 2012;23(4):691-704. doi:10.1016/j.devcel.2012.09.008
- Forstnerič V, Ivičak-Kocjan K, Ljubetič A, Jerala R, Benčina M. Distinctive Recognition of Flagellin by Human and Mouse Toll-Like Receptor 5. *PLoS One*. 2016;11(7):e0158894. Published 2016 Jul 8. doi:10.1371/journal.pone.0158894
- França TC. Homology modeling: an important tool for the drug discovery. *J Biomol Struct Dyn*. 2015;33(8):1780-1793. doi:10.1080/07391102.2014.971429
- Gao B, Tsan MF. Endotoxin contamination in recombinant human heat shock protein 70 (Hsp70) preparation is responsible for the induction of tumor necrosis factor alpha release by murine macrophages. *J Biol Chem*. 2003;278(1):174-179. doi:10.1074/jbc.M208742200
- Gauthier ME, Du Pasquier L, Degnan BM. The genome of the sponge Amphimedon queenslandica provides new perspectives into the origin of Toll-like and interleukin 1 receptor pathways. *Evol Dev*. 2010;12(5):519-533. doi:10.1111/j.1525-142X.2010.00436.x
- Gewirtz AT, Navas TA, Lyons S, Godowski PJ, Madara JL. Cutting edge: bacterial flagellin activates basolaterally expressed TLR5 to induce epithelial proinflammatory gene expression. *J Immunol*. 2001;167(4):1882-1885. doi:10.4049/jimmunol.167.4.1882
- Ghosh M, Basak S, Dutta S. Natural selection shaped the evolution of amino acid usage in mammalian toll like receptor genes. *Comput Biol Chem*. 2022;97:107637. doi:10.1016/j.compbiolchem.2022.107637
- Gilmore TD, Wolenski FS. NF-κB: where did it come from and why?. *Immunol Rev*. 2012;246(1):14-35. doi:10.1111/j.1600-065X.2012.01096.x
- Gissendanner CR, Kelley TD. The C. elegans gene pan-1 encodes novel transmembrane and cytoplasmic leucine-rich repeat proteins and promotes molting and the larva to adult transition. *BMC Dev Biol*. 2013;13:21. Published 2013 May 17. doi:10.1186/1471-213X-13-21
- Gorjestani S, Darnay BG, Lin X. Tumor necrosis factor receptor-associated factor 6 (TRAF6) and TGFβ-activated kinase 1 (TAK1) play essential roles in the C-type lectin receptor signaling in response to Candida albicans infection. *J Biol Chem*. 2012;287(53):44143-44150. doi:10.1074/jbc.M112.414276
- Gottar M, Gobert V, Michel T, et al. The Drosophila immune response against Gram-negative bacteria is mediated by a peptidoglycan recognition protein. *Nature*. 2002;416(6881):640-644. doi:10.1038/nature734
- Grueber CE, Wallis GP, Jamieson IG. Episodic positive selection in the evolution of avian toll-like receptor innate immunity genes. *PLoS One*. 2014;9(3):e89632. Published 2014 Mar 3. doi:10.1371/journal.pone.0089632

- Gumulya Y, Gillam EM. Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the 'retro' approach to protein engineering. *Biochem J*. 2017;474(1):1-19. doi:10.1042/BCJ20160507
- Häcker H, Mischak H, Miethke T, et al. CpG-DNA-specific activation of antigen-presenting cells requires stress kinase activity and is preceded by non-specific endocytosis and endosomal maturation. *EMBO J*. 1998;17(21):6230-6240. doi:10.1093/emboj/17.21.6230
- Hargreaves DC, Medzhitov R. Innate sensors of microbial infection. *J Clin Immunol*. 2005;25(6):503-510. doi:10.1007/s10875-005-8065-4
- Hawn TR, Verbon A, Lettinga KD, et al. A common dominant TLR5 stop codon polymorphism abolishes flagellin signaling and is associated with susceptibility to legionnaires' disease. *J Exp Med*. 2003;198(10):1563-1572. doi:10.1084/jem.20031220
- Hayashi F, Smith KD, Ozinsky A, et al. The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5. *Nature*. 2001;410(6832):1099-1103. doi:10.1038/35074106
- He XL, Bazan JF, McDermott G, et al. Structure of the Nogo receptor ectodomain: a recognition module implicated in myelin inhibition. *Neuron*. 2003;38(2):177-185. doi:10.1016/s0896-6273(03)00232-0
- Heil F, Ahmad-Nejad P, Hemmi H, et al. The Toll-like receptor 7 (TLR7)-specific stimulus loxoribine uncovers a strong relationship within the TLR7, 8 and 9 subfamily. *Eur J Immunol*. 2003;33(11):2987-2997. doi:10.1002/eji.200324238
- Heil F, Hemmi H, Hochrein H, et al. Species-specific recognition of single-stranded RNA via toll-like receptor 7 and 8. *Science*. 2004;303(5663):1526-1529. doi:10.1126/science.1093620
- Hemmi H, Kaisho T, Takeda K, Akira S. The roles of Toll-like receptor 9, MyD88, and DNA-dependent protein kinase catalytic subunit in the effects of two distinct CpG DNAs on dendritic cell subsets. *J Immunol*. 2003;170(6):3059-3064. doi:10.4049/jimmunol.170.6.3059
- Hemmi H, Takeuchi O, Kawai T, et al. A Toll-like receptor recognizes bacterial DNA [published correction appears in Nature 2001 Feb 1;409(6820):646]. *Nature*. 2000;408(6813):740-745. doi:10.1038/35047123
- Hentschel U, Piel J, Degnan SM, Taylor MW. Genomic insights into the marine sponge microbiome. *Nat Rev Microbiol*. 2012;10(9):641-654. doi:10.1038/nrmicro2839
- Hibino T, Loza-Coll M, Messier C, et al. The immune gene repertoire encoded in the purple sea urchin genome. *Dev Biol*. 2006;300(1):349-365. doi:10.1016/j.ydbio.2006.08.065
- Hoebe K, Du X, Georgel P, et al. Identification of Lps2 as a key transducer of MyD88-independent TIR signalling. *Nature*. 2003;424(6950):743-748. doi:10.1038/nature01889
- Hoffmann JA. The immune response of Drosophila. *Nature*. 2003;426(6962):33-38. doi:10.1038/nature02021
- Hoshino K, Takeuchi O, Kawai T, et al. Cutting edge: Toll-like receptor 4 (TLR4)-deficient mice are hyporesponsive to lipopolysaccharide: evidence for TLR4 as the Lps gene product. *J Immunol*. 1999;162(7):3749-3752.
- Huang S, Yuan S, Guo L, et al. Genomic analysis of the immune gene repertoire of amphioxus reveals extraordinary innate complexity and diversity. *Genome Res*. 2008;18(7):1112-1126. doi:10.1101/gr.069674.107
- Huang Y, Temperley ND, Ren L, Smith J, Li N, Burt DW. Molecular evolution of the vertebrate TLR1 gene family--a complex history of gene duplication, gene conversion, positive selection and co-evolution. *BMC Evol Biol*. 2011;11:149. Published 2011 May 28. doi:10.1186/1471-2148-11-149
- Hug H, Mohajeri MH, La Fata G. Toll-Like Receptors: Regulators of the Immune Response in the Human Gut. *Nutrients*. 2018;10(2):203. Published 2018 Feb 13. doi:10.3390/nu10020203

- Huizinga EG, Tsuji S, Romijn RA, et al. Structures of glycoprotein Ibalpha and its complex with von Willebrand factor A1 domain. *Science*. 2002;297(5584):1176-1179. doi:10.1126/science.107355
- Hulo N, Bairoch A, Bulliard V, et al. The PROSITE database. *Nucleic Acids Res*. 2006;34(Database issue):D227-D230. doi:10.1093/nar/gkj063
- Hultmark D. Drosophila immunity: paths and patterns. *Curr Opin Immunol*. 2003;15(1):12-19. doi:10.1016/s0952-7915(02)00005-5
- Hultmark D. Drosophila immunity: paths and patterns. *Curr Opin Immunol*. 2003;15(1):12-19. doi:10.1016/s0952-7915(02)00005-5
- Iliev DB, Roach JC, Mackenzie S, Planas JV, Goetz FW. Endotoxin recognition: in fish or not in fish?. *FEBS Lett*. 2005;579(29):6519-6528. doi:10.1016/j.febslet.2005.10.061
- Inamori K, Ariki S, Kawabata S. A Toll-like receptor in horseshoe crabs. *Immunol Rev*. 2004;198:106-115. doi:10.1111/j.0105-2896.2004.0131.x
- International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution [published correction appears in Nature. 2005 Feb 17;433(7027):777]. *Nature*. 2004;432(7018):695-716. doi:10.1038/nature03154
- Irazoqui JE, Urbach JM, Ausubel FM. Evolution of host innate defence: insights from Caenorhabditis elegans and primitive invertebrates. *Nat Rev Immunol*. 2010;10(1):47-58. doi:10.1038/nri2689
- Ishii A, Kawasaki M, Matsumoto M, Tochinai S, Seya T. Phylogenetic and expression analysis of amphibian Xenopus Toll-like receptors. *Immunogenetics*. 2007;59(4):281-293. doi:10.1007/s00251-007-0193-y
- Janeway CA Jr, Medzhitov R. Innate immune recognition. *Annu Rev Immunol*. 2002;20:197-216. doi:10.1146/annurev.immunol.20.083001.084359
- Jin MS, Kim SE, Heo JY, et al. Crystal structure of the TLR1-TLR2 heterodimer induced by binding of a tri-acylated lipopeptide. *Cell*. 2007;130(6):1071-1082. doi:10.1016/j.cell.2007.09.008
- Jungi TW, Farhat K, Burgener IA, Werling D. Toll-like receptors in domestic animals. *Cell Tissue Res*. 2011;343(1):107-120. doi:10.1007/s00441-010-1047-8
- Jurk M, Heil F, Vollmer J, et al. Human TLR7 or TLR8 independently confer responsiveness to the antiviral compound R-848. *Nat Immunol*. 2002;3(6):499. doi:10.1038/ni0602-499
- Kaiser P. The avian immune genome--a glass half-full or half-empty?. *Cytogenet Genome Res*. 2007;117(1-4):221-230. doi:10.1159/000103183
- Kajava AV. Structural diversity of leucine-rich repeat proteins. *J Mol Biol*. 1998;277(3):519-527. doi:10.1006/jmbi.1998.1643
- Kang JY, Nan X, Jin MS, et al. Recognition of lipopeptide patterns by Toll-like receptor 2-Toll-like receptor 6 heterodimer. *Immunity*. 2009;31(6):873-884. doi:10.1016/j.immuni.2009.09.018
- Kanwal Z, Wiegertjes GF, Veneman WJ, Meijer AH, Spaink HP. Comparative studies of Toll-like receptor signalling using zebrafish. *Dev Comp Immunol*. 2014;46(1):35-52. doi:10.1016/j.dci.2014.02.003
- Karapetyan L, Luke JJ, Davar D. Toll-Like Receptor 9 Agonists in Cancer. *Onco Targets Ther*. 2020;13:10039-10060. Published 2020 Oct 9. doi:10.2147/OTT.S247050
- Kasamatsu J, Oshiumi H, Matsumoto M, Kasahara M, Seya T. Phylogenetic and expression analysis of lamprey toll-like receptors. *Dev Comp Immunol*. 2010;34(8):855-865. doi:10.1016/j.dci.2010.03.004
- Kawai T, Akira S. Innate immune recognition of viral infection. *Nat Immunol*. 2006;7(2):131-137. doi:10.1038/ni1303
- Kawai T, Akira S. The role of pattern-recognition receptors in innate immunity: update on Toll-like receptors. *Nat Immunol*. 2010;11(5):373-384. doi:10.1038/ni.1863

- Keestra AM, de Zoete MR, Bouwman LI, van Putten JP. Chicken TLR21 is an innate CpG DNA receptor distinct from mammalian TLR9. *J Immunol*. 2010;185(1):460-467. doi:10.4049/jimmunol.0901921
- Keestra AM, de Zoete MR, van Aubel RA, van Putten JP. Functional characterization of chicken TLR5 reveals species-specific recognition of flagellin. *Mol Immunol*. 2008;45(5):1298-1307. doi:10.1016/j.molimm.2007.09.013
- Keestra AM, de Zoete MR, van Aubel RA, van Putten JP. The central leucine-rich repeat region of chicken TLR16 dictates unique ligand specificity and species-specific interaction with TLR2. *J Immunol*. 2007;178(11):7110-7119. doi:10.4049/jimmunol.178.11.7110
- Keestra AM, van Putten JP. Unique properties of the chicken TLR4/MD-2 complex: selective lipopolysaccharide activation of the MyD88-dependent pathway. *J Immunol*. 2008;181(6):4354-4362. doi:10.4049/jimmunol.181.6.4354
- Khakoo SI, Rajalingam R, Shum BP, et al. Rapid evolution of NK cell receptor systems demonstrated by comparison of chimpanzees and humans. *Immunity*. 2000;12(6):687-698. doi:10.1016/s1074-7613(00)80219-8
- Khan RT, Musil M, Stourac J, Damborsky J, Bednar D. Fully Automated Ancestral Sequence Reconstruction using FireProt^{ASR} [published correction appears in Curr Protoc. 2022 Aug;2(8):e552. doi: 10.1002/cpz1.552] [published correction appears in Curr Protoc. 2022 Aug;2(8):e551. doi: 10.1002/cpz1.551]. *Curr Protoc*. 2021;1(2):e30. doi:10.1002/cpz1.30
- Kimbrell DA, Beutler B. The evolution and genetics of innate immunity. *Nat Rev Genet*. 2001;2(4):256-267. doi:10.1038/35066006
- Kimura M. Evolutionary rate at the molecular level. *Nature*. 1968;217(5129):624-626. doi:10.1038/217624a0
- King JL, Jukes TH. Non-Darwinian evolution. *Science*. 1969;164(3881):788-798. doi:10.1126/science.164.3881.788
- Kobe B, Deisenhofer J. A structural basis of the interactions between leucine-rich repeats and protein ligands. *Nature*. 1995;374(6518):183-186. doi:10.1038/374183a0
- Krishnaswamy Gopalan T, Gururaj P, Gupta R, et al. Transcriptome profiling reveals higher vertebrate orthologous of intra-cytoplasmic pattern recognition receptors in grey bamboo shark. *PLoS One*. 2014;9(6):e100018. Published 2014 Jun 23. doi:10.1371/journal.pone.0100018
- Krug A, Rothenfusser S, Hornung V, et al. Identification of CpG oligonucleotide sequences with high induction of IFN-alpha/beta in plasmacytoid dendritic cells. *Eur J Immunol*. 2001;31(7):2154-2163. doi:10.1002/1521-4141(200107)31:7<2154::aid-immu2154>3.0.co;2-u
- Kryazhimskiy S, Plotkin JB. The population genetics of dN/dS. *PLoS Genet*. 2008;4(12):e1000304. doi:10.1371/journal.pgen.1000304
- Kucera K, Koblansky AA, Saunders LP, et al. Structure-based analysis of Toxoplasma gondii profilin: a parasite-specific motif is required for recognition by Toll-like receptor 11. *J Mol Biol*. 2010;403(4):616-629. doi:10.1016/j.jmb.2010.09.022
- Kuijl C, Neefjes J. New insight into the everlasting host-pathogen arms race. *Nat Immunol*. 2009;10(8):808-809. doi:10.1038/ni0809-808
- Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol*. 2018;35(6):1547-1549. doi:10.1093/molbev/msy096
- Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol*. 2016;33(7):1870-1874. doi:10.1093/molbev/msw054
- Kuraku S, Ota KG, Kuratani S. Jawless fishes (Cyclostomata). In: Hedges SB and Kumar S, editors. The Time Tree of Life. *Oxford*: *Oxford University Press* (2009):317–19.

- Kurata S, Ariki S, Kawabata S. Recognition of pathogens and activation of immune responses in Drosophila and horseshoe crab innate immunity. *Immunobiology*. 2006;211(4):237-249. doi:10.1016/j.imbio.2005.10.016
- Latz E, Schoenemeyer A, Visintin A, et al. TLR9 signals after translocating from the ER to CpG DNA in the lysosome. *Nat Immunol*. 2004;5(2):190-198. doi:10.1038/ni1028
- Lee J, Chuang TH, Redecke V, et al. Molecular basis for the immunostimulatory activity of guanine nucleoside analogs: activation of Toll-like receptor 7. *Proc Natl Acad Sci U S A*. 2003;100(11):6646-6651. doi:10.1073/pnas.0631696100
- Lemaitre B, Hoffmann J. The host defense of Drosophila melanogaster. *Annu Rev Immunol*. 2007;25:697-743. doi:10.1146/annurev.immunol.25.022106.141615
- Lemaitre B, Nicolas E, Michaut L, Reichhart JM, Hoffmann JA. The dorsoventral regulatory gene cassette spätzle/Toll/cactus controls the potent antifungal response in Drosophila adults. *Cell*. 1996;86(6):973-983. doi:10.1016/s0092-8674(00)80172-5
- Leulier F, Lemaitre B. Toll-like receptors--taking an evolutionary approach. *Nat Rev Genet*. 2008;9(3):165-178. doi:10.1038/nrg2303
- Li C, Chai J, Li H, et al. Pellino protein from pacific white shrimp Litopenaeus vannamei positively regulates NF-κB activation. *Dev Comp Immunol*. 2014;44(2):341-350. doi:10.1016/j.dci.2014.01.012
- Li J, Yuan S, Qi L, et al. Functional conservation and innovation of amphioxus RIP1-mediated signaling in cell fate determination. *J Immunol*. 2011;187(8):3962-3971. doi:10.4049/jimmunol.1100816
- Liang H, Zhou W, Landweber LF. SWAKK: a web server for detecting positive selection in proteins using a sliding window substitution rate analysis. *Nucleic Acids Res*. 2006;34(Web Server issue):W382-W384. doi:10.1093/nar/gkl272
- Lin SC, Lo YC, Wu H. Helical assembly in the MyD88-IRAK4-IRAK2 complex in TLR/IL-1R signalling. *Nature*. 2010;465(7300):885-890. doi:10.1038/nature09121
- Liu OW, Shen K. The transmembrane LRR protein DMA-1 promotes dendrite branching and growth in C. elegans. *Nat Neurosci*. 2011;15(1):57-63. Published 2011 Dec 4. doi:10.1038/nn.2978
- Lively CM, Dybdahl MF. Parasite adaptation to locally common host genotypes. *Nature*. 2000;405(6787):679-681. doi:10.1038/35015069
- Lund JM, Alexopoulou L, Sato A, et al. Recognition of single-stranded RNA viruses by Toll-like receptor 7. *Proc Natl Acad Sci U S A*. 2004;101(15):5598-5603. doi:10.1073/pnas.0400937101
- Maaser C, Heidemann J, von Eiff C, et al. Human intestinal microvascular endothelial cells express Toll-like receptor 5: a binding partner for bacterial flagellin. *J Immunol*. 2004;172(8):5056-5062. doi:10.4049/jimmunol.172.8.5056
- Madeira F, Pearce M, Tivey ARN, et al. Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res*. 2022;50(W1):W276-W279. doi:10.1093/nar/gkac240
- Mahla RS, Reddy MC, Prasad DV, Kumar H. Sweeten PAMPs: Role of Sugar Complexed PAMPs in Innate Immunity and Vaccine Biology. *Front Immunol*. 2013;4:248. Published 2013 Sep 2. doi:10.3389/fimmu.2013.00248
- Mancuso VP, Parry JM, Storer L, et al. Extracellular leucine-rich repeat proteins are required to organize the apical extracellular matrix and maintain epithelial junction integrity in C. elegans. *Development*. 2012;139(5):979-990. doi:10.1242/dev.075135
- Matsumoto M, Funami K, Tanabe M, et al. Subcellular localization of Toll-like receptor 3 in human dendritic cells [published correction appears in J Immunol. 2003 Nov 1;171(9):4934]. *J Immunol*. 2003;171(6):3154-3162. doi:10.4049/jimmunol.171.6.3154

- Matsuo A, Oshiumi H, Tsujita T, et al. Teleost TLR22 recognizes RNA duplex to induce IFN and protect cells from birnaviruses. *J Immunol*. 2008;181(5):3474-3485. doi:10.4049/jimmunol.181.5.3474
- Matsushima N, Tanaka T, Enkhbayar P, et al. Comparative sequence analysis of leucine-rich repeats (LRRs) within vertebrate toll-like receptors. *BMC Genomics*. 2007;8:124. Published 2007 May 21. doi:10.1186/1471-2164-8-124
- Matzinger P. Tolerance, danger, and the extended family. *Annu Rev Immunol*. 1994;12:991-1045. doi:10.1146/annurev.iy.12.040194.005015
- Medzhitov R, Preston-Hurlburt P, Janeway CA Jr. A human homologue of the Drosophila Toll protein signals activation of adaptive immunity. *Nature*. 1997;388(6640):394-397. doi:10.1038/41131
- Medzhitov R. Recognition of microorganisms and activation of the immune response. *Nature*. 2007;449(7164):819-826. doi:10.1038/nature06246
- Mekata T, Kono T, Yoshida T, Sakai M, Itami T. Identification of cDNA encoding Toll receptor, MjToll gene from kuruma shrimp, Marsupenaeus japonicus. *Fish Shellfish Immunol*. 2008;24(1):122-133. doi:10.1016/j.fsi.2007.10.006
- Merkl R, Sterner R. Ancestral protein reconstruction: techniques and applications. *Biol Chem*. 2016;397(1):1-21. doi:10.1515/hsz-2015-0158
- Meyer SN, Amoyel M, Bergantiños C, et al. An ancient defense system eliminates unfit cells from developing tissues during cell competition. *Science*. 2014;346(6214):1258236. doi:10.1126/science.1258236
- Mikami T, Miyashita H, Takatsuka S, Kuroki Y, Matsushima N. Molecular evolution of vertebrate Toll-like receptors: evolutionary rate difference between their leucine-rich repeats and their TIR domains. *Gene*. 2012;503(2):235-243. doi:10.1016/j.gene.2012.04.007
- Miller DJ, Hemmrich G, Ball EE, et al. The innate immune repertoire in cnidaria--ancestral complexity and stochastic gene loss. *Genome Biol*. 2007;8(4):R59. doi:10.1186/gb-2007-8-4-r59
- Miller SI, Ernst RK, Bader MW. LPS, TLR4 and infectious disease diversity. *Nat Rev Microbiol*. 2005;3(1):36-46. doi:10.1038/nrmicro1068
- Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. *Nat Methods*. 2022;19(6):679-682. doi:10.1038/s41592-022-01488-1
- Muffato M, Louis A, Nguyen NTT, Lucas J, Berthelot C, Roest Crollius H. Reconstruction of hundreds of reference ancestral genomes across the eukaryotic kingdom. *Nat Ecol Evol*. 2023;7(3):355-366. doi:10.1038/s41559-022-01956-z
- Muhammed MT, Aki-Yalcin E. Homology modeling in drug discovery: Overview, current applications, and future perspectives. *Chem Biol Drug Des*. 2019;93(1):12-20. doi:10.1111/cbdd.13388
- Musil M, Khan RT, Beier A, et al. FireProtASR: A Web Server for Fully Automated Ancestral Sequence Reconstruction. *Brief Bioinform*. 2021;22(4):bbaa337. doi:10.1093/bib/bbaa337
- Nagai Y, Akashi S, Nagafuku M, et al. Essential role of MD-2 in LPS responsiveness and TLR4 distribution. *Nat Immunol*. 2002;3(7):667-672. doi:10.1038/ni809
- Nakajima T, Ohtani H, Satta Y, et al. Natural selection in the TLR-related genes in the course of primate evolution. *Immunogenetics*. 2008;60(12):727-735. doi:10.1007/s00251-008-0332-0
- Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol*. 1986;3(5):418-426. doi:10.1093/oxfordjournals.molbev.a040410
- Netea MG, van Deuren M, Kullberg BJ, Cavaillon JM, Van der Meer JW. Does the shape of lipid A determine the interaction of LPS with Toll-like receptors?. *Trends Immunol*. 2002;23(3):135-139. doi:10.1016/s1471-4906(01)02169-x

- Nie L, Cai SY, Shao JZ, Chen J. Toll-Like Receptors, Associated Biological Roles, and Signaling Networks in Non-Mammals. *Front Immunol*. 2018;9:1523. Published 2018 Jul 2. doi:10.3389/fimmu.2018.01523
- Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG. Recent and ongoing selection in the human genome. *Nat Rev Genet*. 2007;8(11):857-868. doi:10.1038/nrg2187
- Nielsen R. Molecular signatures of natural selection. *Annu Rev Genet*. 2005;39:197-218. doi:10.1146/annurev.genet.39.073003.112420
- O'Neill LA, Bowie AG. The family of five: TIR-domain-containing adaptors in Toll-like receptor signalling. *Nat Rev Immunol*. 2007;7(5):353-364. doi:10.1038/nri2079
- O'Neill LA, Bryant CE, Doyle SL. Therapeutic targeting of Toll-like receptors for infectious and inflammatory diseases and cancer. *Pharmacol Rev*. 2009;61(2):177-197. doi:10.1124/pr.109.001073
- Oosting M, Cheng SC, Bolscher JM, et al. Human TLR10 is an anti-inflammatory pattern-recognition receptor. *Proc Natl Acad Sci U S A*. 2014;111(42):E4478-E4484. doi:10.1073/pnas.1410293111
- Palti Y. Toll-like receptors in bony fish: from genomics to function. *Dev Comp Immunol*. 2011;35(12):1263-1272. doi:10.1016/j.dci.2011.03.006
- Paysan-Lafosse T, Blum M, Chuguransky S, et al. InterPro in 2022. *Nucleic Acids Res*. 2023;51(D1):D418-D427. doi:10.1093/nar/gkac993
- Peden JF. Analysis of Codon Usage (Doctoral dissertation, University of Nottingham, United Kingdom), 2000
- Peiris TH, Hoyer KK, Oviedo NJ. Innate immune system and tissue regeneration in planarians: an area ripe for exploration. *Semin Immunol*. 2014;26(4):295-302. doi:10.1016/j.smim.2014.06.005
- Pierce BG, Wiehe K, Hwang H, Kim BH, Vreven T, Weng Z. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. *Bioinformatics*. 2014;30(12):1771-1773. doi:10.1093/bioinformatics/btu097
- Pila EA, Tarrabain M, Kabore AL, Hanington PC. A Novel Toll-Like Receptor (TLR) Influences Compatibility between the Gastropod Biomphalaria glabrata, and the Digenean Trematode Schistosoma mansoni. *PLoS Pathog*. 2016;12(3):e1005513. Published 2016 Mar 25. doi:10.1371/journal.ppat.1005513
- Poltorak A, He X, Smirnova I, et al. Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in Tlr4 gene. *Science*. 1998;282(5396):2085-2088. doi:10.1126/science.282.5396.2085
- Poole AZ, Weis VM. TIR-domain-containing protein repertoire of nine anthozoan species reveals coral-specific expansions and uncharacterized proteins. *Dev Comp Immunol*. 2014;46(2):480-488. doi:10.1016/j.dci.2014.06.002
- Pradel E, Zhang Y, Pujol N, Matsuyama T, Bargmann CI, Ewbank JJ. Detection and avoidance of a natural product from the pathogenic bacterium Serratia marcescens by Caenorhabditis elegans. *Proc Natl Acad Sci U S A*. 2007;104(7):2295-2300. doi:10.1073/pnas.0610281104
- Priyam M, Tripathy M, Rai U, Ghorai SM. Divergence of protein sensing (TLR 4, 5) and nucleic acid sensing (TLR 3, 7) within the reptilian lineage. *Mol Phylogenet Evol*. 2018;119:210-224. doi:10.1016/j.ympev.2017.11.018
- Pujol N, Link EM, Liu LX, et al. A reverse genetic analysis of components of the Toll signaling pathway in Caenorhabditis elegans. *Curr Biol*. 2001;11(11):809-821. doi:10.1016/s0960-9822(01)00241-x
- Putnam NH, Srivastava M, Hellsten U, et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*. 2007;317(5834):86-94. doi:10.1126/science.1139158
- Quach H, Wilson D, Laval G, et al. Different selective pressures shape the evolution of Toll-like receptors in human and African great ape populations. *Hum Mol Genet*. 2013;22(23):4829-4840. doi:10.1093/hmg/ddt335

- Quiniou SM, Boudinot P, Bengtén E. Comprehensive survey and genomic characterization of Toll-like receptors (TLRs) in channel catfish, Ictalurus punctatus: identification of novel fish TLRs. *Immunogenetics*. 2013;65(7):511-530. doi:10.1007/s00251-013-0694-9
- Quintana-Murci L, Alcaïs A, Abel L, Casanova JL. Immunology in natura: clinical, epidemiological and evolutionary genetics of infectious diseases. *Nat Immunol*. 2007;8(11):1165-1171. doi:10.1038/ni1535
- Quintana-Murci L, Clark AG. Population genetic tools for dissecting innate immunity in humans. *Nat Rev Immunol*. 2013;13(4):280-293. doi:10.1038/nri3421
- Qureshi S, Medzhitov R. Toll-like receptors and their role in experimental models of microbial infection. *Genes Immun*. 2003;4(2):87-94. doi:10.1038/sj.gene.6363937
- Raetz M, Kibardin A, Sturge CR, et al. Cooperation of TLR12 and TLR11 in the IRF8-dependent IL-12 response to Toxoplasma gondii profilin. *J Immunol*. 2013;191(9):4818-4827. doi:10.4049/jimmunol.1301301
- Rakoff-Nahoum S, Medzhitov R. Toll-like receptors and cancer. *Nat Rev Cancer*. 2009;9(1):57-63. doi:10.1038/nrc2541
- Rao Y, Wang Z, Chai X, Nie Q, Zhang X. Hydrophobicity and aromaticity are primary factors shaping variation in amino acid usage of chicken proteome. *PLoS One*. 2014;9(10):e110381. Published 2014 Oct 16. doi:10.1371/journal.pone.0110381
- Rauta PR, Samanta M, Dash HR, Nayak B, Das S. Toll-like receptors (TLRs) in aquatic animals: signaling pathways, expressions and immune responses. *Immunol Lett*. 2014;158(1-2):14-24. doi:10.1016/j.imlet.2013.11.013
- Ren Y, Ding D, Pan B, Bu W. The TLR13-MyD88-NF-κB signalling pathway of Cyclina sinensis plays vital roles in innate immune responses. *Fish Shellfish Immunol*. 2017;70:720-730. doi:10.1016/j.fsi.2017.09.060
- Reuven EM, Fink A, Shai Y. Regulation of innate immune responses by transmembrane interactions: lessons from the TLR family. *Biochim Biophys Acta*. 2014;1838(6):1586-1593. doi:10.1016/j.bbamem.2014.01.020
- Riutort M, Álvarez-Presas M, Lázaro E, Solà E, Paps J. Evolutionary history of the Tricladida and the Platyhelminthes: an up-to-date phylogenetic and systematic account. *Int J Dev Biol*. 2012;56(1-3):5-17. doi:10.1387/ijdb.113441mr
- Roach JC, Glusman G, Rowen L, et al. The evolution of vertebrate Toll-like receptors. *Proc Natl Acad Sci U S A*. 2005;102(27):9577-9582. doi:10.1073/pnas.0502272102
- Rock FL, Hardiman G, Timans JC, Kastelein RA, Bazan JF. A family of human receptors structurally related to Drosophila Toll. *Proc Natl Acad Sci U S A*. 1998;95(2):588-593. doi:10.1073/pnas.95.2.588
- Roy A, Banerjee R, Basak S. HIV Progression Depends on Codon and Amino Acid Usage Profile of Envelope Protein and Associated Host-Genetic Influence. *Front Microbiol*. 2017;8:1083. Published 2017 Jun 15. doi:10.3389/fmicb.2017.01083
- Roy A, Basak S. HIV long-term non-progressors share similar features with simian immunodeficiency virus infection of chimpanzees. *J Biomol Struct Dyn*. 2021;39(7):2447-2454. doi:10.1080/07391102.2020.1749129
- Royet J, Reichhart JM, Hoffmann JA. Sensing and signaling during infection in Drosophila. *Curr Opin Immunol*. 2005;17(1):11-17. doi:10.1016/j.coi.2004.12.002
- Ruysschaert JM, Lonez C. Role of lipid microdomains in TLR-mediated signalling. *Biochim Biophys Acta*. 2015;1848(9):1860-1867. doi:10.1016/j.bbamem.2015.03.014
- Sachidanandam R, Weissman D, Schmidt SC, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*. 2001;409(6822):928-933. doi:10.1038/35057149

- Sánchez Alvarado A. The freshwater planarian Schmidtea mediterranea: embryogenesis, stem cells and regeneration. *Curr Opin Genet Dev*. 2003;13(4):438-444. doi:10.1016/s0959-437x(03)00082-0
- Savar NS, Bouzari S. In silico study of ligand binding site of toll-like receptor 5. *Adv Biomed Res*. 2014;3:41. Published 2014 Jan 24. doi:10.4103/2277-9175.125730
- Schikorski D, Cuvillier-Hot V, Boidin-Wichlacz C, Slomianny C, Salzet M, Tasiemski A. Deciphering the immune function and regulation by a TLR of the cytokine EMAPII in the lesioned central nervous system using a leech model. *J Immunol*. 2009;183(11):7119-7128. doi:10.4049/jimmunol.0900538
- Schultz J, Copley RR, Doerks T, Ponting CP, Bork P. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res*. 2000;28(1):231-234. doi:10.1093/nar/28.1.231
- Scott W. Robinson, Avid M. Afzal, David P. Leader. Bioinformatics: Concepts, Methods, and Data. *Handbook of Pharmacogenomics and Stratified Medicine*, 2014; 13:259-287. doi:10.1016/B978-0-12-386882-4.00013-X
- Sepulcre MP, Alcaraz-Pérez F, López-Muñoz A, et al. Evolution of lipopolysaccharide (LPS) recognition and signaling: fish TLR4 does not recognize LPS and negatively regulates NF-kappaB activation. *J Immunol*. 2009;182(4):1836-1845. doi:10.4049/jimmunol.0801755
- Shimazu R, Akashi S, Ogata H, et al. MD-2, a molecule that confers lipopolysaccharide responsiveness on Toll-like receptor 4. *J Exp Med*. 1999;189(11):1777-1782. doi:10.1084/jem.189.11.1777
- Simakov O, Marletaz F, Cho SJ, et al. Insights into bilaterian evolution from three spiralian genomes. *Nature*. 2013;493(7433):526-531. doi:10.1038/nature11696
- Šmarda P, Bureš P, Horová L, et al. Ecological and evolutionary significance of genomic GC content diversity in monocots. *Proc Natl Acad Sci U S A*. 2014;111(39):E4096-E4102. doi:10.1073/pnas.1321152111
- Smith J, Speed D, Law AS, Glass EJ, Burt DW. In-silico identification of chicken immune-related genes. *Immunogenetics*. 2004;56(2):122-133. doi:10.1007/s00251-004-0669-y
- Smith KD, Andersen-Nissen E, Hayashi F, et al. Toll-like receptor 5 recognizes a conserved site on flagellin required for protofilament formation and bacterial motility [published correction appears in Nat Immunol. 2004 Apr;5(4):451]. *Nat Immunol*. 2003;4(12):1247-1253. doi:10.1038/ni1011
- Song X, Jin P, Qin S, Chen L, Ma F. The evolution and origin of animal Toll-like receptor signaling pathway revealed by network-level molecular evolutionary analyses. *PLoS One*. 2012;7(12):e51657. doi:10.1371/journal.pone.0051657
- Srisuk C, Longyant S, Senapin S, Sithigorngul P, Chaivisuthangkura P. Molecular cloning and characterization of a Toll receptor gene from Macrobrachium rosenbergii. *Fish Shellfish Immunol*. 2014;36(2):552-562. doi:10.1016/j.fsi.2013.12.025
- Srivastava M, Simakov O, Chapman J, et al. The Amphimedon queenslandica genome and the evolution of animal complexity. *Nature*. 2010;466(7307):720-726. doi:10.1038/nature09201
- Sternke M, Tripp KW, Barrick D. Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. *Proc Natl Acad Sci U S A*. 2019;116(23):11275-11284. doi:10.1073/pnas.1816707116
- Sun JJ, Xu S, He ZH, Shi XZ, Zhao XF, Wang JX. Activation of Toll Pathway Is Different between Kuruma Shrimp and *Drosophila*. *Front Immunol*. 2017;8:1151. Published 2017 Sep 20. doi:10.3389/fimmu.2017.01151
- Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*. 2006;34(Web Server issue):W609-W612. doi:10.1093/nar/gkl315

- Tabeta K, Georgel P, Janssen E, et al. Toll-like receptors 9 and 3 as essential components of innate immune defense against mouse cytomegalovirus infection. *Proc Natl Acad Sci U S A*. 2004;101(10):3516-3521. doi:10.1073/pnas.0400525101
- Takeda K, Kaisho T, Akira S. Toll-like receptors. *Annu Rev Immunol*. 2003;21:335-376. doi:10.1146/annurev.immunol.21.120601.141126
- Takeuchi O, Akira S. Pattern recognition receptors and inflammation. *Cell*. 2010;140(6):805-820. doi:10.1016/j.cell.2010.01.022
- Tanji T, Ip YT. Regulators of the Toll and Imd pathways in the Drosophila innate immune response. *Trends Immunol*. 2005;26(4):193-198. doi:10.1016/j.it.2005.02.006
- Temperley ND, Berlin S, Paton IR, Griffin DK, Burt DW. Evolution of the chicken Toll-like receptor gene family: a story of gene gain and gene loss. *BMC Genomics*. 2008;9:62. Published 2008 Feb 1. doi:10.1186/1471-2164-9-62
- Tzou P, Reichhart JM, Lemaitre B. Constitutive expression of a single antimicrobial peptide can restore wild-type resistance to infection in immunodeficient Drosophila mutants. *Proc Natl Acad Sci U S A*. 2002;99(4):2152-2157. doi:10.1073/pnas.042411999
- Uematsu S, Akira S. Immune responses of TLR5(+) lamina propria dendritic cells in enterobacterial infection. *J Gastroenterol*. 2009;44(8):803-811. doi:10.1007/s00535-009-0094-y
- Underhill DM, Ozinsky A, Hajjar AM, et al. The Toll-like receptor 2 is recruited to macrophage phagosomes and discriminates between pathogens. *Nature*. 1999;401(6755):811-815. doi:10.1038/44605
- Valanne S, Wang JH, Rämet M. The Drosophila Toll signaling pathway. *J Immunol*. 2011;186(2):649-656. doi:10.4049/jimmunol.1002302
- Vangone A, Bonvin AMJJ. PRODIGY: A Contact-based Predictor of Binding Affinity in Protein-protein Complexes. *Bio Protoc*. 2017;7(3):e2124. Published 2017 Feb 5. doi:10.21769/BioProtoc.2124
- Velová H, Gutowska-Ding MW, Burt DW, Vinkler M. Toll-Like Receptor Evolution in Birds: Gene Duplication, Pseudogenization, and Diversifying Selection. *Mol Biol Evol*. 2018;35(9):2170-2184. doi:10.1093/molbev/msy119
- Verthelyi D, Ishii KJ, Gursel M, Takeshita F, Klinman DM. Human peripheral blood cells differentially recognize and respond to two distinct CPG motifs. *J Immunol*. 2001;166(4):2372-2377. doi:10.4049/jimmunol.166.4.2372
- Vidya MK, Kumar VG, Sejian V, Bagath M, Krishnan G, Bhatta R. Toll-like receptors: Significance, ligands, signaling pathways, and functions in mammals. *Int Rev Immunol*. 2018;37(1):20-36. doi:10.1080/08830185.2017.1380200
- Vijay-Kumar M, Aitken JD, Carvalho FA, et al. Metabolic syndrome and altered gut microbiota in mice lacking Toll-like receptor 5. *Science*. 2010;328(5975):228-231. doi:10.1126/science.1179721
- Vinkler M, Bainová H, Bryja J. Protein evolution of Toll-like receptors 4, 5 and 7 within Galloanserae birds. *Genet Sel Evol*. 2014;46(1):72. Published 2014 Nov 12. doi:10.1186/s12711-014-0072-6
- Volff JN. Genome evolution and biodiversity in teleost fish. *Heredity (Edinb)*. 2005;94(3):280-294. doi:10.1038/sj.hdy.6800635
- Wang C, Deng L, Hong M, Akkaraju GR, Inoue J, Chen ZJ. TAK1 is a ubiquitin-dependent kinase of MKK and IKK. *Nature*. 2001;412(6844):346-351. doi:10.1038/35085597
- Wang PH, Gu ZH, Wan DH, et al. The shrimp NF-κB pathway is activated by white spot syndrome virus (WSSV) 449 to facilitate the expression of WSSV069 (ie1), WSSV303 and WSSV371. *PLoS One*. 2011;6(9):e24773. doi:10.1371/journal.pone.0024773

- Wang Z, Chen YH, Dai YJ, et al. A novel vertebrates Toll-like receptor counterpart regulating the anti-microbial peptides expression in the freshwater crayfish, Procambarus clarkii. *Fish Shellfish Immunol*. 2015;43(1):219-229. doi:10.1016/j.fsi.2014.12.038
- Waterhouse A, Bertoni M, Bienert S, et al. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*. 2018;46(W1):W296-W303. doi:10.1093/nar/gky427
- Weaver S, Shank SD, Spielman SJ, Li M, Muse SV, Kosakovsky Pond SL. Datamonkey 2.0: A Modern Web Application for Characterizing Selective and Other Evolutionary Processes. *Mol Biol Evol*. 2018;35(3):773-777. doi:10.1093/molbev/msx335
- Webb AE, Gerek ZN, Morgan CC, et al. Adaptive Evolution as a Predictor of Species-Specific Innate Immune Response. *Mol Biol Evol*. 2015;32(7):1717-1729. doi:10.1093/molbev/msv051
- West AP, Koblansky AA, Ghosh S. Recognition and signaling by toll-like receptors. *Annu Rev Cell Dev Biol*. 2006;22:409-437. doi:10.1146/annurev.cellbio.21.122303.115827
- Wiens M, Korzhev M, Krasko A, et al. Innate immune defense of the sponge Suberites domuncula against bacteria involves a MyD88-dependent signaling pathway. Induction of a perforin-like molecule. *J Biol Chem*. 2005;280(30):27949-27959. doi:10.1074/jbc.M504049200
- Wiens M, Korzhev M, Perovic-Ottstadt S, et al. Toll-like receptors are part of the innate immune defense system of sponges (demospongiae: Porifera). *Mol Biol Evol*. 2007;24(3):792-804. doi:10.1093/molbev/msl208
- Wlasiuk G, Nachman MW. Adaptation and constraint at Toll-like receptors in primates. *Mol Biol Evol*. 2010;27(9):2172-2186. doi:10.1093/molbev/msq104
- Wright F. The 'effective number of codons' used in a gene. *Gene*. 1990;87(1):23-29. doi:10.1016/0378-1119(90)90491-9
- Xiong J. Essential Bioinformatics. Cambridge University Press; 2006
- Xue LC, Rodrigues JP, Kastritis PL, Bonvin AM, Vangone A. PRODIGY: a web server for predicting the binding affinity of protein-protein complexes. *Bioinformatics*. 2016;32(23):3676-3678. doi:10.1093/bioinformatics/btw514
- Yan Y, Zhang D, Zhou P, Li B, Huang SY. HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. *Nucleic Acids Res*. 2017;45(W1):W365-W373. doi:10.1093/nar/gkx407
- Yang C, Zhang J, Li F, et al. A Toll receptor from Chinese shrimp Fenneropenaeus chinensis is responsive to Vibrio anguillarum infection. *Fish Shellfish Immunol*. 2008;24(5):564-574. doi:10.1016/j.fsi.2007.12.012
- Yang D, Tewary P, de la Rosa G, Wei F, Oppenheim JJ. The alarmin functions of high-mobility group proteins. *Biochim Biophys Acta*. 2010;1799(1-2):157-163. doi:10.1016/j.bbagrm.2009.11.002
- Yang J, Zhou M, Zhong Y, et al. Gene duplication and adaptive evolution of Toll-like receptor genes in birds. *Dev Comp Immunol*. 2021;119:103990. doi:10.1016/j.dci.2020.103990
- Yang LS, Yin ZX, Liao JX, et al. A Toll receptor in shrimp. *Mol Immunol*. 2007;44(8):1999-2008. doi:10.1016/j.molimm.2006.09.021
- Yang Z, Nielsen R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*. 2002;19(6):908-917. doi:10.1093/oxfordjournals.molbev.a004148
- Yang Z, Zeng X, Zhao Y, Chen R. AlphaFold2 and its applications in the fields of biology and medicine. *Signal Transduct Target Ther*. 2023;8(1):115. Published 2023 Mar 14. doi:10.1038/s41392-023-01381-z
- Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007;24(8):1586-1591. doi:10.1093/molbev/msm088

- Yilmaz A, Shen S, Adelson DL, Xavier S, Zhu JJ. Identification and sequence analysis of chicken Toll-like receptors. *Immunogenetics*. 2005;56(10):743-753. doi:10.1007/s00251-004-0740-8
- Zelus D, Robinson-Rechavi M, Delacre M, Auriault C, Laudet V. Fast evolution of interleukin-2 in mammals and positive selection in ruminants. *J Mol Evol*. 2000;51(3):234-244. doi:10.1007/s002390010085
- Zhang D, Zhang G, Hayden MS, et al. A toll-like receptor that prevents infection by uropathogenic bacteria. *Science*. 2004;303(5663):1522-1526. doi:10.1126/science.1094351
- Zhang J, Kong X, Zhou C, Li L, Nie G, Li X. Toll-like receptor recognition of bacteria in fish: ligand specificity and signal pathways. *Fish Shellfish Immunol*. 2014;41(2):380-388. doi:10.1016/j.fsi.2014.09.022
- Zhang J, Liu S, Rajendran KV, et al. Pathogen recognition receptors in channel catfish: III phylogeny and expression analysis of Toll-like receptors. *Dev Comp Immunol*. 2013;40(2):185-194. doi:10.1016/j.dci.2013.01.009
- Zhang Q, Zmasek CM, Godzik A. Domain architecture evolution of pattern-recognition receptors. *Immunogenetics*. 2010;62(5):263-272. doi:10.1007/s00251-010-0428-1
- Zhang S, Yu M, Guo Q, et al. Annexin A2 binds to endosomes and negatively regulates TLR4-triggered inflammatory responses via the TRAM-TRIF pathway. *Sci Rep*. 2015;5:15859. Published 2015 Nov 3. doi:10.1038/srep15859
- Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*. 2005;33(7):2302-2309. Published 2005 Apr 22. doi:10.1093/nar/gki524
- Zhou W, Li Y, Pan X, et al. Toll-like receptor 9 interaction with CpG ODN--an in silico analysis approach. *Theor Biol Med Model*. 2013;10:18. Published 2013 Mar 14. doi:10.1186/1742-4682-10-18
- Zimmerman LM, Vogel LA, Bowden RM. Understanding the vertebrate immune system: insights from the reptilian perspective. *J Exp Biol*. 2010;213(5):661-671. doi:10.1242/jeb.038315

Publication and Conferences

Publications:

1) Ghosh M, Basak S, Dutta S. Evolutionary divergence of TLR9 through ancestral sequence reconstruction. Immunogenetics. Published online March 5, 2024. doi:10.1007/s00251-024-01338-8 **[I.F.3.2]**
2) Ghosh M, Basak S, Dutta S. Natural selection shaped the evolution of amino acid usage in mammalian toll-like receptor genes. Comput Biol Chem. 2022 Apr;97:107637. doi:10.1016/j.compbiolchem.2022.107637 **[I.F.3.1]**

Conferences:

1) Participated and presented poster on "Evolutionary divergence of Toll-like receptor 9 (TLR9)" in the national level biotechnology conference SymBiot'23 on 30-31st October, 2023 organized by Department of Biotechnology, Manipal Institute of Technology.
2) Participated and presented poster on "In silico study of flagellin binding of mammalian toll like receptor 5 and diversification in enteric pathogen recognition" in the 16th Asian Conference on Diarrhoeal Disease and Nutrition in Kolkata, India on 11-13 November, 2022 organized by ICMR-NICED, Kolkata.

Computational Biology and Chemistry 97 (2022) 107637

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac

Natural selection shaped the evolution of amino acid usage in mammalian toll like receptor genes

Manisha Ghosh [a], Surajit Basak [a,*], Shanta Dutta [b]

[a] Division of Bioinformatics, ICMR-National Institute of Cholera and Enteric Diseases, Kolkata, India
[b] Division of Bacteriology, ICMR-National Institute of Cholera and Enteric Diseases, Kolkata, India

ARTICLE INFO

Keywords:
TLR
PAMP
GC-content
Hydrophobicity
Purifying selection
Subcellular location

Abstract: Toll-like receptors (TLRs) are important as they are able to sense diverse set of pathogens associated molecular patterns (PAMPs) as ligands. These receptors are involved in functions such as immune response, development of signaling process and cell adhesion. In the present study we are interested to analyze the influence of evolutionary selection pressure on the mutational diversity of mammalian TLR genes. We observed differential patterns of amino acid usage between primate and non-primate mammalian TLR genes. GC-content of TLR genes and hydrophobicity of the encoded proteins are the most influential factors correlated with the differential pattern of amino acid usage.The influence of the subcellular location on the amino acid usage pattern of TLRs is evident in present study. Purifying selection is uniformly present on TLR genes, positively selected sites are mostly located over the ligand binding domain. Our study clearly demonstrates that natural selection has shaped the evolution of primate and non-primate mammalian TLR genes.

1. Introduction

The defense system of animal involves two type of immunity adaptive and innate immunity. Initially innate immune system produces an inflammatory response to block the growth and transmission of the pathogen during an infection. In vertebrates, in order to develop acquired immune response particularly receptors of Band T cell sense the infectious agents to produce responses that lead to its exclusion (Janeway and Medzhitov, 2002). Receptors associated with innate immune system are germline-encoded. They have been evolved to sense components of external pathogen also referred as pathogen-associated molecular patterns (PAMPs) which are crucial for pathogen existence or host released endogenous components in response to inflammation (Matzinger, 1994; Yang et al., 2010; Erridge, 2010). These receptors of innate immune system are located in serum, on cell surface, in endosomes, and in the cytoplasm (Medzhitov, 2007).

Being an important category of pattern recognition receptors (PRRs) the toll-like receptors (TLRs) are seen in Drosophila and mammals. Mammal TLRs play fundamental role in detection of pathogen associated patterns with the initiation of signal transduction pathways that cause genetic expression which lead to the innate and adaptive immune responses (O'Neill et al., 2009; Rakoff-Nahoum and Medzhitov, 2009). TLRs are type-I integral membrane receptors comprising an extracellular domain also known as ectodomain (ECD) containing leucine-rich repeats which facilitate the PAMPs recognition, a signal transmembrane segment, and an intracellular Toll–interleukin 1 (IL-1) receptor (TIR) domain for downstream signal transduction (Bell et al., 2003). In mammals there are thirteen TLRs discovered in mice (TLR1–13) and ten TLRs in humans (TLR1–10). TLR1-TLR9 is found in both mice and human, TLR10 is non-functional in mouse due to a retrovirus insertion and TLR11, TLR12 and TLR13 are not found in human (Takeuchiand Akira, 2010). Depending on the subcellular distribution TLRs in humans can be classified into two categories: TLR1, TLR2, TLR4, TLR5, TLR6 and TLR10 are expressed normally on the cell surface and TLR3, TLR7, TLR8 and TLR9 are commonly found in intracellular compartments like endosomes. These human TLRs detect various PAMPs such as lipopolysaccharide (TLR4), lipopeptides (TLR2 associated TLR1 or TLR6), bacterial flagellin (TLR5), viral dsRNA (TLR3), viral or bacterial ssRNA (TLRs 7 and 8), and CpG-rich unmethylated DNA (TLR9) (Akira et al., 2006).

Genetic diversity in active genes associated with immune defense such as TLRs is interesting from an evolutionary perspective as these genes are an excellent model for studying the selective stress applied to the host genome by pathogen. These genes appear to evolve faster than other loci in the genome in response to pathogen that are evolving rapidly. Selection is a major factor in controlling the evolutionary rate of

* Correspondence to: Division of Bioinformatics, National Institute of Cholera and Enteric Diseases, P-33, C.I.T Road, Scheme-XM, Beliaghata, Kolkata 700010, India.
E-mail address: basak.surajit@icmr.gov.in (S. Basak).

https://doi.org/10.1016/j.compbiolchem.2022.107637
Received 2 August 2021; Received in revised form 9 November 2021; Accepted 30 January 2022
Available online 2 February 2022

TLRs, mutation is also another factor and TLRs are strongly selected to maintain their functions. In different mammals innate immune response is not similar as some variation is there between different species in their TLRs. This variation is due to selective pressure on the immune system-related genes that reflect specific conditions experienced by each species (Bagheri and Zahmatkesh, 2018). Evolutionary genetics approaches have amplified to understand the evolutionary forces acting on the human genome that provides indispensable complement in treatment of infectious diseases. Within the perspective of infection, detecting the magnitude and pattern of environmental selection that works on the genes implicated in immune-associated procedures can deliver insight into the host defense mechanisms (Barreiro et al., 2009).

Amino acids and codons are used in diverse frequencies both between genes and between genes within the same genome. Degeneracy of genetic code direct the use of diverse set of codons for producing the similar protein, procedures that create non-random usage of codons are likely to influence the usage of amino acids. The possible reason behind this is the neutral processes where composition of bases of all codons that encode an amino acid might be either GC rich or GC poor (Rao et al., 2014). Selection also has a significant role in determining frequencies of amino acid. Often genomic base compositions play a major role on the type of amino acid usage; other factors like hydrophobicity, gene function, level of expression etc. also influence the amino acid usage. In this study mammalian TLRs are progressively investigated to examine the effects of environmental selection on diverse set of TLRs and factors that influence selection will be explored. Natural selection on different members of TLRs family will be studied to explore their evolutionary contribution to host defense.

2. Materials and methods

2.1. Sequence retrieval and multivariate analysis on amino acid usage

Genes and their encoding protein sequences of toll-like receptors (TLR) were taken from GenBank, NCBI (http://www.ncbi.nlm.nih.gov/genbank/) and Ensembl maintained by EMBL-EBI (www.ensembl.org). By nature, amino acid usage is multivariate and need to be explored using statistical analysis like correspondence analysis (COA) (Peden, 2000). COA reveals major trends of variation in the dataset by arranging them along continuous axes where consecutive axis have been arranged to have diminishing effect gradually (Roy et al., 2017). The analyses of amino acid usage patterns of TLR genes of mammal under study were carried out using COA available in CodonW program.

Parameters like relative amino acid usage (RAAU), average hydrophobicity, GC content of genes were calculated for each TLR sequence using available option in CodonW program. Correlation coefficient between variables was calculated using the available formula in MS Excel. Significance test was performed using the freely available online tool such as t-test (https://www.graphpad.com/quickcalcs/ttest1/).

Phylogenetic analysis was performed among primate and non-primate genes of TLR. The sequences were aligned using the ClustalW program. The phylogenetic tree was constructed using Mega 7, utilizing the maximum likelihood method (Kumar et al., 2016).

Three dimensional structural models were generated for TLR5 protein sequences through homology modeling using SWISS-MODEL (Waterhouse et al., 2018). TLR5 protein structure available in Protein Data Bank (PDB) (PDB ID: 3J0A) was used as template for homology modeling with more than 99% sequence identity and 97% query coverage in case of human (primate mammal) and 78% sequence identity and 97% query coverage in case of cattle (non-primate mammal). The structure of flagellin was truncated from crystal structure of the N-terminal fragment of zebrafish TLR5 in complex with Salmonella flagellin available in PDB (PDB ID: 3V47). As the ectodomain of the TLRs are involved in ligand recognition, the interaction study was performed on TLR5 ectodomains based on the NCBI annotation (Savar and Bouzari, 2014; Forstnerič et al., 2016). Molecular interaction of TLR5 protein with flagellin was performed using Z-dock software (Pierce et al., 2014). Then, the resulting docking data were processed and analyzed considering binding energies and main interacting residues in each complex by using the PRODIGY software (Xue et al., 2016). Free energy of the structural complexes was calculated using PremPS server (Chen et al., 2020).

2.2. Estimation of evolutionary rate and mutational analysis

The impact of evolution on set of genes is indicated by the ratio (ω) i.e., ratio of non-synonymous substitution rate per non-synonymous site (Ka) to synonymous substitution rate per synonymous site (Ks). Where $\omega > 1$ point towards positive (diversifying) selection and $\omega < 1$ signify negative (purifying) selection (Roy and Basak, 2021). The rate of evolution of each TLR1-TLR10 group of mammals (taking consensus sequence as reference) was estimated using the available PAL2NAL program (Suyama et al., 2006). Residue wise evolutionary rate of TLR gene sequences were calculated using SWAKK server (Liang et al., 2006). This server performs a sliding 3D window analysis to calculate the ratio of non-synonymous to synonymous substitution rate (Ka/Ks) of DNA sequences that encode protein.

Positive selection test of individual codons of mammals TLR was performed using the Hyphy package executed in the Data Monkey Web Server that compare Ka to Ks ratio using maximum likelihood (ML) framework, (Weaver et al., 2018). The sequences of every TLR were analyzed under the fixed-effect likelihood (FEL) model. This Fixed Effects Likelihood (FEL) approach uses maximum-likelihood (ML) method to deduce non-synonymous (dN) and synonymous (dS) substitution rates on the basis of per site considering a coding alignment and related phylogeny. It is presumed in this method that selection pressure for each site remains constant throughout the phylogeny.

Mutational analysis was performed by using a customized script to study the mutation among the TLR sequences. Predicted consensus sequence for each TLR was used as reference sequence to identify the mutation. Consensus sequences offer promising approach in screening proteins of high stability and retain the biological activity as it predicted based on evolutionary history in which residues important for both stability and function are likely to be conserved (Sternke et al., 2019). Occurrences of mutation in each TLR for each species were studied across the two functional domains.

3. Results

3.1. Correspondence analysis on amino acid usage of TLR genes

Correspondence analysis was performed to study the amino acid usage variation of ten different TLR genes of mammalian origin separately. The first and second major axes accounted for 54.5% and 20.1% of the total variation of amino acid usage respectively for TLR1 gene. Fig. 1 shows position of genes generated during correspondence analysis on the basis of amino acid usage across the first and second major axes. Similar pattern of distribution of the amino acid usage was observed for other TLRs under study. For the ten different TLR genes these first axis always accounted the major variation which is more than 30% of the total variation of amino acid usage. It is clear from the correspondence analyses that there are two clusters. One cluster belongs to mammal which are primates and another cluster belongs to mammal other than primates. For simplicity, hereafter, TLRs from primates (Human, Gorilla, Monkey, Chimpanzee, Orangutan, Baboon etc.) will be referred to as primate mammal (Pm) TLRs and TLRs from mammal other than primates will be referred to as non-primate mammal (NPm) TLRs. Phylogenetic tree using the TLR1 genes of Pm and NPm clearly shows that Pm and NPm TLR genes are present in different branches (Fig. 2). Similar pattern is observed for other TLRs. Branching pattern of phylogenetic tree follows similar trend to that of correspondence analysis.

Now to investigate the preference of amino acids in two different

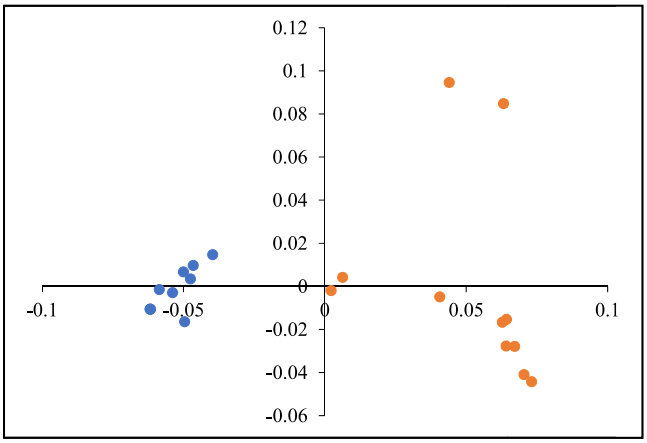


Fig. 1. Distribution of TLR1 genes along the two major axes of Correspondence analysis (COA) based on amino acid usage (AAU) data. x-axis- Axis 1 of AAU; y-axis- Axis 2 of AAU. Blue colored dots represent TLR gene sequences from Pm and orange colored dots represent TLR gene sequences from NPm. Similar pattern is observed for other TLR genes also.

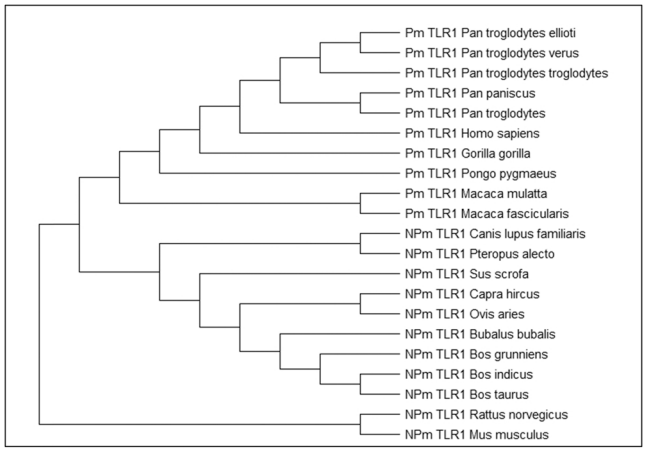


Fig. 2. Phylogenetic tree of Pm and NPm genes of TLR1. Similar pattern is observed for other TLRs.

clusters we have compared the relative amino acid usage values between Pm and NPm TLR genes. Comparisons of relative amino acid usage values suggested that the twenty amino acids are differently preferred among Pm and NPm for each TLR. From the analysis it was observed that amino acids such as Phe,Met, Thr, Lys, Glu, Cys were mostly preferred in Pm TLRs whereas amino acids such as Leu, Pro, Ala, Asp, Arg, Gly were mostly preferred in NPm TLRs.

We have performed molecular docking study between TLR5 (Homo sapiens for primate and Bos indicus for non-primate) and flagellin (pathogen receptor). We have identified the preferred residues those are interacting with the flagellin and when substituted these residues with GC-rich/GC-poor, as the case may be, the stability of the TLR5-flagellin complex decreased (Fig. 3).

Since axis1 (horizontal axis) accounts major variation for each TLR in COA, further analysis is performed on the basis of distribution of mammal TLR genes along the horizontal axis of correspondence analysis. Significant correlation was observed between the gene position along the horizontal axis and hydrophobicity ($r = 0.533$, $p < .05$) and GC-content of the encoded proteins ($r = 0.745$, $p < .01$). Significant correlation of axis1 with GC1 ($r = 0.714$, $p < .05$), GC2 ($r = 0.689$, $p < .05$), GC3 ($r = 0.668$, $p < .05$) content of the encoded proteins were also observed.

We have compared the average GC content of TLR genes for Pm and

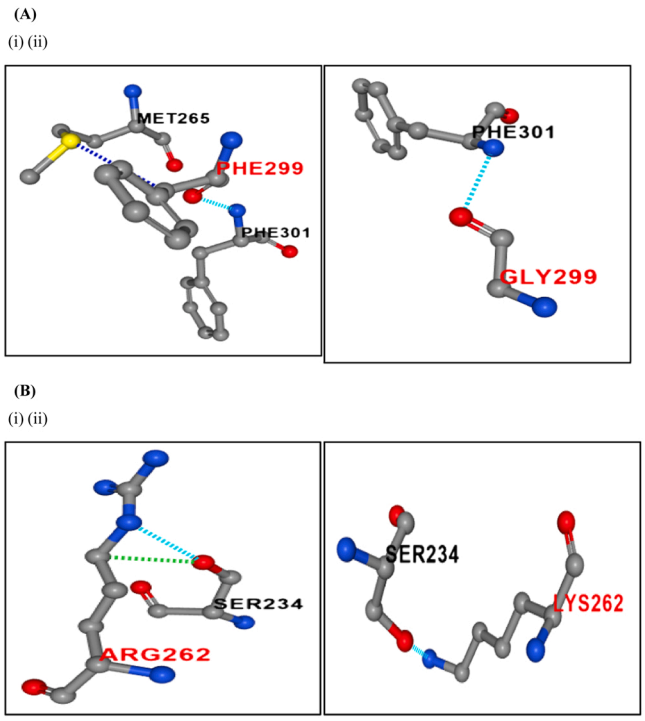


Fig. 3. (A): Interaction profile of a representative mutation F299G in Pm TLR5 protein indicating GC-poor to GC-rich amino acid substitution. GC-poor amino acids are preferred in Pm. The structural stability decreases when F (Phenyl alanine) is substituted by G (Glycine). (i) Wild type residue F299 having one polar interaction (sky), and one hydrophobic (blue) interaction. (ii) Mutant type residue 299G having one polar interaction (sky). (B): Interaction profile of a representative mutation R2262K in NPm TLR5 protein indicating GC-rich to GC-poor amino acid substitution. GC-rich amino acids are preferred in NPm. The structural stability decreases when R (Arginine) is substituted by K (Lysine).(i) Wild type residue R262 having one polar interaction (sky) and one van der Waals (green) interactions. (ii) Mutant type residue 262 K having one polar interaction (sky). Results are generated using PremPS server. ΔΔG value in both the cases is positive which indicates destabilizing mutation.

NPm. The average GC content of TLR genes are 42.6% and 44.6% for Pm and NPm respectively. The difference of GC content of TLR genes between Pm and NPm is statistically significant ($P < .01$). As the NPm TLR genes have higher GC content we may expect GC-rich amino acids would be preferred in NPm. Indeed, we observed that average composition of four GC-rich amino acids (Du et al., 2018) (Ala, Arg, Gly, and Pro) are higher in NPm TLR genes and the compositions of four GC-rich amino acids are positively correlated with GC contents ($r = 0.836$, $p < .001$) of the NPm TLR genes. On the other hand, we observed that average composition of AT-rich amino acids (Phe, Ile, Tyr, Asn and Lys) are higher in Pm TLR genes and their compositions are also positively correlated with AT-contents ($r = 0.673$, $p < .001$) of Pm TLR genes. All these results support that amino acid usage have been shaped under the influence of GC-content of TLR genes.

3.2. *Impact of evolutionary selection pressure on TLR Genes*

We observed presence of purifying selection across all the TLR genes (both Pm and NPm) by comprehensive analysis of evolutionary rates. However, residue specific measurement of evolutionary rate shows differences of positively selected sites between Pm and NPm TLRs. Site-specific selection across the ligand binding domain also showed the same trend. These observations indicate stronger selection pressure on NPm TLR genes compared to Pm TLR genes. Positively selected sites among Pm and NPm TLRs are shown in Table 1.

The evolutionary parameters such as Non-synonymous substitution

Table 1
Distribution of positively selected sites among Pm and NPm TLRs.

Genes	No. of species			Total sites		Total positively selected sites		Positively selected sites in ligand binding domain		% positively selected site		% positively selected site in ligand binding domain	
	Total	Pm	NPm	Pm (length aa)	NPm (length aa)	Pm	NPm	Pm	NPm	Pm (%)	NPm (%)	Pm (%)	NPm (%)
TLR1	21	10	11	786	796	1	9	1	5	0.127	1.13	0.127	0.62
TLR2	26	10	16	784	785	0	13	0	12	0	1.65	0	1.52
TLR3	22	7	15	904	905	0	13	0	12	0	1.43	0	1.32
TLR4	22	8	14	839	844	1	32	1	28	0.119	3.79	0.119	3.31
TLR5	17	8	9	858	874	0	6	0	3	0	0.68	0	0.34
TLR6	22	10	12	796	810	0	14	0	9	0	1.72	0	1.11
TLR7	24	9	15	1049	1058	0	17	0	15	0	1.6	0	1.41
TLR8	20	7	13	1041	1091	0	20	0	18	0	1.83	0	1.64
TLR9	22	7	15	1032	1034	1	2	0	2	0.09	0.19	0	0.19
TLR10	23	12	11	811	822	0	15	0	10	0	1.82	0	1.21

(Ka), synonymous substitution (Ks), ratio of non-synonymous and synonymous substitution (Ka/Ks) were found to differ significantly among Pm and NPm TLRs. Significant difference of these parameters was also observed across the two functional domains of Pm and NPm TLRs. These results are shown in Table 2. We have also found significant correlation of evolutionary parameters with axis1 of correspondence analysis on amino acid usage. Significant correlation of axis1 is observed with Ka in seven TLR genes, Ks in six TLR genes; Ka/Ks in five TLR genes.

3.3. Correlation of evolutionary parameters with GC-content and mutational analysis

We already observed the correlation between GC content and amino acid usage variation of TLRs through correspondence analysis. It was also found that evolutionary parameters differ significantly among Pm and NPm TLR genes. Furthermore, these evolutionary parameters such as Ka, Ks and Ka/Ks was correlated significantly with the GC content of TLR genes among mammalian species ($p < .05$) (Table 3). Thus, GC content is playing an important role in the evolution process of amino acid sequences for most of the TLRs among Pm and NPm.

Mutations were identified for both Pm and NPm TLRs over the entire TLR sequences. But more mutations are observed in the ligand recognition domain. It endorsed that ligand recognition domain is more prone to mutation than the signaling domain. Rate of evolution (Ka/Ks) in the extracellular ligand recognition domain is more compared to intracellular signaling domain for most of the TLRs in both Pm and NPm.

3.4. Amino acid usage pattern of TLRs based on subcellular distribution

Since TLRs are classified into extracellular and intracellular based on the subcellular distribution we have analyzed the amino acid usage pattern of Pm and NPm TLR genes individually. Differential amino acid usage patterns were noticed where extracellular and intracellular TLRs formed different clusters in case of Pm and NPm. In case of Pm, extracellular TLR1, TLR2, TLR6, TLR10 formed one cluster; TLR4, TLR5 were found in different clusters and intracellular TLR3, TLR7, TLR8 were present in different cluster from TLR9. In the same way, in case of NPm intracellular TLR3, TLR7, TLR8 were in different cluster and TLR9 formed another cluster. But NPm extracellular TLR1, TLR2, TLR4, TLR6, TLR10 were grouped into one cluster and TLR5 found in separate cluster. These extracellular and intracellular TLRs were distributed along the major axis shown in Fig. 4. Evolutionary parameters were also checked between these two clusters of extracellular and intracellular TLRs in case of Pm and NPm respectively. The parameters Ka, Ks and Ka/Ks were found to differ significantly among these clusters. Hence, subcellular distribution is also governing the amino acid variation of TLRs for Pm and NPm independently where evolutionary selection is the most important aspect.

Table 2
Significance test of evolutionary parameters among Pm and NPm TLR genes and across the domains.Extracellular domain of TLR (ECD), Intracellular domain of TLR (TIR) and tick mark indicates significant difference.

	Pm & NPm genes			**ECD of Pm & NPm genes**			**TIR of Pm & NPm genes**		
	Ka	Ks	Ka/Ks	Ka	Ks	Ka/Ks	Ka	Ks	Ka/Ks
TLR1	✓	✓	✓	✓	✓	✓	✓	✓	✓
TLR2	✓	✓		✓	✓	✓	✓	✓	
TLR3	✓	✓	✓	✓	✓	✓	✓		✓
TLR4			✓	✓			✓	✓	
TLR5	✓	✓	✓	✓	✓	✓	✓	✓	✓
TLR6	✓	✓	✓	✓	✓	✓	✓	✓	✓
TLR7		✓	✓	✓	✓	✓		✓	✓
TLR8					✓	✓		✓	
TLR9	✓		✓	✓		✓	✓	✓	
TLR10	✓	✓	✓	✓	✓	✓	✓	✓	✓

4. Discussion

The systematic study of the amino acid usage across various mammalian TLRs revealed that amino acids are used in diverse pattern among TLR genes of Pm and NPm species. In spite of similar anatomy and physiology between Pm and NPm there is disparity in amino acid usage pattern of TLRs observed in them. One key difference between these species is that primates possess a voluminous and complicated forebrain whereas non-primates possess a small brain.

Correspondence analyses established hydrophobicity and genomic GC content as the most important features causing the TLR wise variation of amino acid usage in mammal. It depicts that these factors are causing the variation in the immune response among species of a particular TLR. Significant correlation of hydrophobicity is observed among TLRs. The extracellular TLR domains are composed of leucine-rich repeats (LRR) that usually contain 22–29 length residues and have periodic hydrophobic residues positioned at discrete intervals. In three dimensions during assembling into protein multiple repeats shape as solenoid like structure, where consensus hydrophobic residues pointed inside to make a stable core of the structure (Botos et al., 2011). Hydrophobic residues becoming an influencing factor for amino acid usage variation of TLR genes among Pm and NPm. GC content is another influencing factor as amino acid usage of TLRs is significantly correlated with GC content. Guanine and cytosine bases proportion in the DNA molecule (GC content) being an essential qualitative aspect of genomic architecture is discussed frequently in humans and other vertebrates such as birds, mammals in relation to the evolution of the isochore structure (Šmarda et al., 2014).

Amino acid usage pattern study also revealed that individual Pm and NPm TLRs distribution based on subcellular location extracellular and intracellular is different. Depending on subcellular location functionality of TLRs become different due to dissimilar PAMP recognition. Cell

Table 3
Correlation study of GC content with evolutionary parameters of TLRs.

	GC content	Ka	Correlationsignificant at	Ks	Correlationsignificant at	Ka/Ks	Correlationsignificant at
TLR1	0.403	0.0743	$p < .01$	0.1756	$p < .01$	0.4505	$p < .05$
TLR2	0.441	0.0850	$p < .01$	0.2391	$p < .01$	0.4008	$p < .01$
TLR3	0.403	0.0615	$p < .01$	0.2243	$p < .01$	0.2879	$p < .05$
TLR4	0.438	0.0994	$p < .01$	0.2164	$p < .01$	0.4829	$p < .10$
TLR5	0.452	0.0768	$p < .01$	0.2415	$p < .01$	0.3781	$p < .01$
TLR6	0.395	0.0677	$p < .01$	0.1883	$p < .01$	0.3838	$p < .01$
TLR7	0.410	0.0470	not significant	0.1671	not significant	0.2945	not significant
TLR8	0.418	0.1015	$p < .01$	0.3902	$p < .01$	0.4007	$p < .01$
TLR9	0.628	0.0685	not significant	0.4410	not significant	0.1596	not significant
TLR10	0.389	0.0607	$p < .01$	0.1516	$p < .01$	0.4020	not significant

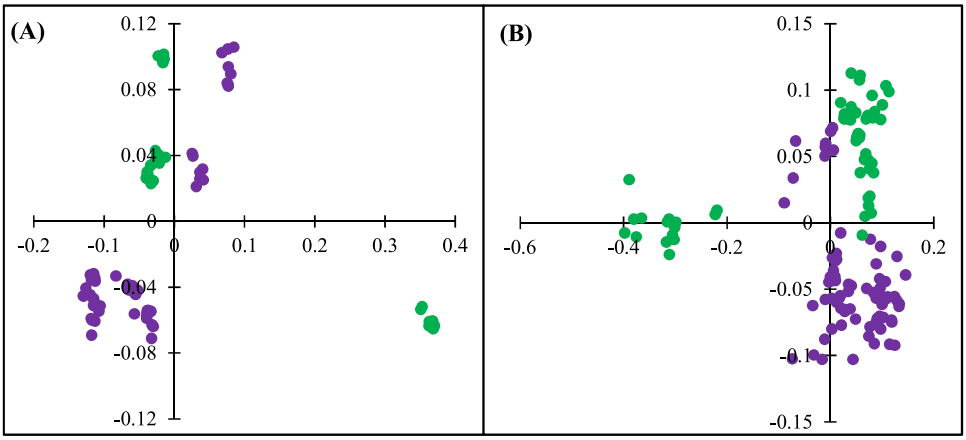


Fig. 4. Distribution of TLR genes along the two major axes of Correspondence analysis (COA) based on amino acid usage (AAU) data. X-axis- Axis 1 of AAU; y-axis- Axis 2 of AAU. (A) TLR gene sequence of Pm, (B) TLR gene sequence of NPm. Violet colored dots represent extracellular TLR gene sequences and green colored dots represent intracellular TLR gene sequences.

surface expressed TLRs such as TLR1, TLR2, TLR4, TLR5, TLR6 and TLR10 mostly recognize microbial membrane components like lipoproteins, lipids; TLR3, TLR7, TLR8 and TLR9 expressed in intracellular vesicles like endoplasmic reticulum (ER), endosomes, lysosomes and endolysosomes and sense microbial nucleic acids (Kawai and Akira, 2010). These factors affecting Pm and NPm TLRs which are showing distinct amino acid usage pattern between extracellular and intracellular TLRs.

Evolutionary analysis has suggested that purifying selection is the major force working on TLRs. Presence of codons that are selected positively indicates selective pressures on these immune genes lead to the most noticeable changes in the ectodomain, particularly in the variable section accountable for direct interaction with PAMPS. More mutation is observed in the extracellular domain due to the direct interaction with pathogen. Overall selective pressure within the innate immune system is stronger in non-primate mammal species compared to primate mammal species. The relation between GC contents and Ka, Ks, Ka/Ks values of TLR genes from different mammal species were observed. Correspondingly, Ka, Ks, Ka/Ks values changes with change in GC contents. The GC content is therefore consistent with the evolutionary process of amino acid sequences and contributes to the evolutionary level as a key component of amino acids between Pm and NPm TLRs. The GC content influences the emergence of proteins due to energy costs, and both the combination of bases and amino acids is involved in this process (Du et al., 2018).

This study reveals differential patterns of amino acid usage, evolutionary constraints of TLR genes among Pm and NPm. Amino acid composition has a significant impact on the level of TLR emergence and this is also affected by GC content. Identification of genes associated with immunity that evolves in a different way across Pm and NPm TLRs might facilitate the understanding of genetic basis for the differences in disease susceptibility (Quach et al., 2013). The greater extent of deviation in selection that constrain the evolution of Pm and NPm TLRs will enhance our understanding of the biological contribution of TLRs to host defense in natural setting. This study presented the divergence in the biological significance of different TLRs and offer evidences for their diverse contributions in response to host defense.

Authors statement

All authors have seen and approved the final version of the manuscript being submitted. Present manuscript is the authors' original work, hasn't received prior publication and isn't under consideration for publication elsewhere.

Data availability

Data will be available upon request to the Corresponding Author.

Acknowledgment

Manisha Ghosh is supported by Senior Research Fellowship by Indian Council of Medical Research (ICMR).

Conflict of interest statement

The authors declare that no conflicts of interest exist.

References

Akira, S., Uematsu, S., Takeuchi, O., 2006. Pathogen recognition and innate immunity. Cell 124 (4), 783–801.

Bagheri, M., Zahmatkesh, A., 2018. Evolution and species-specific conservation of toll-like receptors in terrestrial vertebrates. Int. Rev. Immunol. 37 (5), 217–228.
Barreiro, L.B., Ben-Ali, M., Quach, H., et al., 2009. Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. PLoS Genet. 5 (7), e1000562.
Bell, J.K., Mullen, G.E., Leifer, C.A., Mazzoni, A., Davies, D.R., Segal, D.M., 2003. Leucine-rich repeats and pathogen recognition in Toll-like receptors. Trends Immunol. 24 (10), 528–533.
Botos, I., Segal, D.M., Davies, D.R., 2011. The structural biology of Toll-like receptors. Structure 19 (4), 447–459.
Chen, Y., Lu, H., Zhang, N., Chen, Y., Zhu, Z., Wang, S., Li, M., 2020. PremPS: predicting the impact of missense mutations on protein stability. PLoS Comput. Biol. 16 (12), e1008543.
Du, M.Z., Zhang, C., Wang, H., Liu, S., Wei, W., Guo, F.B., 2018. The GC content as a main factor shaping the amino acid usage during bacterial evolution process. Front. Microbiol. 9, 2948.
Erridge, C., 2010. Endogenous ligands of TLR2 and TLR4: agonists or assistants? J. Leukoc. Biol. 87 (6), 989–999.
Forstnerič, V., Ivičak-Kocjan, K., Ljubetič, A., Jerala, R., Benčina, M., 2016. Distinctive recognition of flagellin by human and mouse toll-like receptor 5. PLoS One 11 (7), e0158894.
Janeway Jr., C.A., Medzhitov, R., 2002. Innate immune recognition. Annu. Rev. Immunol. 20, 197–216.
Kawai, T., Akira, S., 2010. The role of pattern-recognition receptors in innate immunity: update on Toll-like receptors. Nat. Immunol. 11 (5), 373–384.
Kumar, S., Stecher, G., Tamura, K., 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for bigger datasets. Mol. Biol. Evol. 33, 1870–1874.
Liang, H., Zhou, W., Landweber, L.F., 2006. SWAKK: a web server for detecting positive selection in proteins using a sliding window substitution rate analysis. Nucleic Acids Res. 34 (Web Server issue), W382–W384.
Matzinger, P., 1994. Tolerance, danger, and the extended family. Annu. Rev. Immunol. 12, 991–1045.
Medzhitov, R., 2007. Recognition of microorganisms and activation of the immune response. Nature 449, 819–826.
O'Neill, L.A., Bryant, C.E., Doyle, S.L., 2009. Therapeutic targeting of Toll-like receptors for infectious and inflammatory diseases and cancer. Pharmacol. Rev. 61 (2), 177–197.
Peden, J.F., 2000. Analysis of Codon Usage. University of Nottingham, Nottingham.
Pierce, B.G., Wiehe, K., Hwang, H., Kim, B.H., Vreven, T., Weng, Z., 2014. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. Bioinformatics 30 (12), 1771–1773.
Quach, H., Wilson, D., Laval, G., Patin, E., Manry, J., Guibert, J., Barreiro, et al., 2013. Different selective pressures shape the evolution of Toll-like receptors in human and African great ape populations. Hum. Mol. Genet. 22 (23), 4829–4840.
Rakoff-Nahoum, S., Medzhitov, R., 2009. Toll-like receptors and cancer. Nat. Rev. Cancer 9 (1), 57–63.
Rao, Y., Wang, Z., Chai, X., Nie, Q., Zhang, X., 2014. Hydrophobicity and aromaticity are primary factors shaping variation in amino acid usage of chicken proteome. PLoS One 9 (10), e110381.
Roy, A., Banerjee, R., Basak, S., 2017. HIV progression depends on codon and amino acid usage profile of envelope protein and associated host-genetic influence. Front. Microbiol. 8, 1083.
Roy, A., Basak, S., 2021. HIV long-term non-progressors share similar features with simian immunodeficiency virus infection of chimpanzees. J. Biomol. Struct. Dyn. 39 (7), 2447–2454.
Savar, N.S., Bouzari, S., 2014. In silico study of ligand binding site of toll-like receptor 5. Adv. Biomed. Res. 3, 41.
Šmarda, P., Bureš, P., Horová, L., Leitch, I.J., et al., 2014. Ecological and evolutionary significance of genomic GC content diversity in monocots. Proc. Natl. Acad. Sci. USA 111 (39), E4096–E4102.
Sternke, M., Tripp, K.W., Barrick, D., 2019. Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. Proc. Natl. Acad. Sci. USA 116 (23), 11275–11284.
Suyama, M., Torrents, D., Bork, P., 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34 (Web Server issue), W609–W612.
Takeuchi, O., Akira, S., 2010. Pattern recognition receptors and inflammation. Cell 140 (6), 805–820.
Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., de Beer, T., Rempfer, C., Bordoli, L., Lepore, R., Schwede, T., 2018. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 46 (W1), W296–W303.
Weaver, S., Shank, S.D., Spielman, S.J., Li, M., Muse, S.V., Kosakovsky Pond, S.L., 2018. Datamonkey 2.0: a modern web application for characterizing selective and other evolutionary processes. Mol. Biol. Evol. 35 (3), 773–777.
Xue, L.C., Rodrigues, J.P., Kastritis, P.L., Bonvin, A.M., Vangone, A., 2016. PRODIGY: a web server for predicting the binding affinity of protein-protein complexes. Bioinformatics 32 (23), 3676–3678.
Yang, D., Tewary, P., de la Rosa, G., Wei, F., Oppenheim, J.J., 2010. The alarmin functions of high-mobility group proteins. Biochim. Et. Biophys. Acta 1799 (1–2), 157–163.

Immunogenetics
https://doi.org/10.1007/s00251-024-01338-8

ORIGINAL ARTICLE

Evolutionary divergence of TLR9 through ancestral sequence reconstruction

Manisha Ghosh[1] · **Surajit Basak**[1] · **Shanta Dutta**[2]

Received: 1 September 2023 / Accepted: 24 February 2024

Abstract
The transmembrane pattern recognition receptor, Toll-like receptor (TLR), are best known for their roles in innate immunity via recognition of pathogen and initiation of signaling response. Mammalian TLRs recognize molecular patterns associated with pathogens and initiate innate immune response. We have studied the evolutionary diversity of mammalian TLR genes for differences in immunological response. Reconstruction of ancestral sequences is a key aspect of the molecular evolution of TLR to track changes across the TLR genes. The comprehensive analysis of mammalian TLRs revealed a distinct pattern of evolution of TLR9. Various sequence-based features such as amino acid usage, hydrophobicity, GC content, and evolutionary constraints are found to influence the divergence of TLR9 from other TLRs. Ancestral sequence reconstruction analysis also revealed that the gradual evolution of TLR genes in several ancestral lineages leads to the distinct pattern of TLR9. It demonstrates evolutionary divergence with the progressive accumulation of mutations results in the distinct pattern of TLR9.

Keywords TLR · Evolution · Phylogenetic tree · Ancestral sequence · Mutation · Diversity

Introduction

Toll-like receptors (TLRs) are considered the primary sensors of invading microbial pathogen in the innate immune system because they detect pathogen-associated molecular patterns (PAMPs). Since the early discovery of a Toll protein in the fruit fly *Drosophila melanogaster* thirteen members of the TLR family have been identified in human (TLR1-TLR10) and mouse (TLR1-TLR13) (Zhou et al. 2013). It seems that most mammalian species share a similar repertoire of TLR homologs though with few exceptions (Nie et al. 2018). TLRs are type I integral membrane glycoproteins with a pathogen-binding ectodomain (ECD) and a cytoplasmic signaling domain connected by a single transmembrane helix (Zhou et al. 2013). Mammalian TLR pathogen-binding ectodomains contain 19–25 extracellular leucine-rich repeats (LRRs) and a cytoplasmic toll/interleukin (IL)-1R (TIR) domain. LRRs comprising 24–29 amino acids are responsible for ligand recognition and binding, while the TIR domain is responsible for downstream signaling (Botos et al. 2011). Surface-expressed TLRs (TLR 1, 2, 4, 5, 6, and 10) typically identify pathogen structural components, whereas endosomal TLRs (TLR 3, 7, 8, and 9) recognize nucleic acid. TLRs respond to a variety of pathogen-associated molecular patterns (PAMPs) in humans, including lipopolysaccharide (TLR4), lipopeptides (TLR2 associated with TLR1 or TLR6), bacterial flagellin (TLR5), viral dsRNA (TLR3), viral or bacterial ssRNA (TLRs 7 and 8), and CpG-rich unmethylated DNA (TLR9) (Takeda and Akira 2005; Vidya et al. 2018).

TLR9 is an endosomal receptor that detects bacterial DNA/CpG-containing oligodeoxynucleotides (CpG ODN). TLR9-mediated signaling is initiated within the endosome by the sequential recruitment of adaptor proteins, which in turn activates critical downstream transcription factors. Various preclinical studies showed the efficacy of TLR9 agonists individually and in combination with other agents (Karapetyan et al. 2020). Interaction of unmethylated CpG DNA with TLR9 activates immune responses through the MyD88-dependent signaling pathway. Human trials have shown that CpG DNA can act as an adjuvant and boost the immunogenicity of the hepatitis vaccine. These findings highlight the importance of TLR ligands in triggering adaptive responses and providing new adjuvants in vaccine formulation (Cook et al. 2004).

Biological sequences have long been recognized as a record of evolutionary history, with accumulating mutations recording species relationships and the mechanisms driving their

✉ Surajit Basak
basaksurajit@gmail.com

[1] Division of Bioinformatics, ICMR-National Institute of Cholera and Enteric Diseases, P-33, C.I.T Road, Scheme-XM, Beliaghata Kolkata 700010, India

[2] Division of Bacteriology, ICMR-National Institute of Cholera and Enteric Diseases, Kolkata, India

Published online: 05 March 2024

evolution. To avoid the recognition by the host immune system pathogens involved in recognition evolve faster. With the evolving pathogen, the host receptor that recognizes the pathogen also evolves to keep pace with the changes in the pathogen. These modifications in receptor can be detected as the positive selection signatures or mutations (Areal et al. 2011). From an evolutionary perspective, genetic variation in TLR genes linked with immunological defence is important because these genes provide a good model for investigating pathogen-induced selective stress on the host genome (Roach et al. 2005). In response to rapidly evolving pathogens, these genes appear to evolve quicker than other locations in the genome (Ghosh et al. 2022). Given enough genetic information from different species, the temporal accumulation of mutations can be used to reconstruct sequences from their common ancestors. These ancestral reconstructions serve as the foundation for many of molecular evolution approaches nowadays, such as phylogenetic trees and sequence selection tests (Muffato et al. 2023). The ancestral sequence reconstruction (ASR) approach begins with a multiple-sequence alignment (MSA) of the collection of relevant homolog sequences and considers evolutionary information depicted by the phylogenetic tree. It is a probabilistic strategy that investigates the deep evolutionary history of homolog sequences in order to reassemble the evolutionary trajectory of a protein. ASR can reveal sequences of long-extinct genes and organisms from which the current ones evolved, making it an important tool in evolutionary biology (Gumulya and Gillam 2017). Since the advent of sequencing, the reconstruction of ancestral sequences, particularly genes, has been studied extensively. Advanced methods exist to retrace the history of sequence substitutions and leverage changes in substitution dynamics to answer specific evolutionary problems (Merkl and Sterner 2016).

The study of the sequence-based feature like differential amino acid usage and the impact of various factors on TLRs will facilitate us to comprehend the evolutionary factors that affect innate immune genes. The evolutionary genetics approach to identify the extent of natural selection acting on these genes and the gradual changes that lead to the divergence will enhance our understanding about the mechanism of host defence mediated by TLRs.

Materials and methods

Data retrieval and multivariate statistical analysis

Sequences of mammalian Toll-like receptor (TLR) genes and their encoding proteins representing different groups of TLR such as TLR1, TLR2, TLR3, TLR4, TLR5, TLR6, TLR7, TLR8, TLR9, and TLR10 were obtained from GenBank, NCBI. Toll-like receptor gene sequences were searched by using the search option available at the NCBI website and mammalian species have been selected under species selection for the search operation. The output of the search operation provides coding sequence of a particular TLR. These coding sequences and their corresponding protein sequences were downloaded. TLR gene sequences from primates, rodents, artiodactyls, proboscides, perissodactyls, lagomorphs, and chiropters were taken for the analysis. Sequences containing ambiguous character (other than A, T, G, C) and internal stop codons were removed from the retrieved dataset. The list of mammalian taxa chosen to investigate in this study along with their accession numbers is provided in the Supplementary Table 1.

Amino acid usage is a multivariate feature by nature and studied using statistical analysis such as correspondence analysis (CoA) (Peden 2000). CoA is an efficient method to explore the variation in the dataset and it reveals major tendencies of data disparities by placing them along continuous axes according to the differential trends observed, with each consecutive axis having a diminishing effect (Roy et al. 2017). CoA on the basis of amino acid usage (AAU) of TLR gene sequences was generated using CodonW. Estimation of physicochemical properties like hydrophobicity, GC-content, GC3 values, effective number of codons (ENC), and aromaticity of the study sequences was also performed using the CodonW program. The correlation study of the parameters was executed in Microsoft Excel. The significance test was done using the freely available web program QuickCalcs-Graphpad.

Evolutionary analysis and phylogenetic tree construction

Evolutionary selection acting on the genes under study is addressed by evolutionary rate (ω). ω is estimated as the ratio of the rate non-synonymous substitutions per non-synonymous site (Ka) and the rate of synonymous substitutions per synonymous site (Ks). $\omega > 1$ indicates positive (diversifying) selection, whereas, $\omega < 1$ indicates negative (purifying) selection. For each TLR group (example: TLR1) their consensus nucleotide sequences (example: TLR1_consensus) were generated. We have prepared a Perl script for generating these consensus sequences. Downloaded nucleotide sequences and the consensus sequence of each TLR group were subjected to Clustal Omega program (Madeira et al. 2022) for the nucleotide sequence alignment. This program Clustal produces biologically meaningful multiple-sequence alignments of divergent sequences. Then the evolutionary rate of the TLR genes (TLR1-TLR10) of each TLR group (example: TLR1) was estimated relative to their consensus (example: TLR1_consensus) sequences using Codeml program of the PAML software (ver. 4.5) with runmode = −2 and CodonFreq = 1 (Nei and Gojobori 1986; Yang 2007).

The protein sequences of all the mammalian TLRs were subjected to the multiple sequence alignment using the Clustal Omega program (Madeira et al. 2022). The alignment result was saved in FASTA format for further analysis. Then using that alignment, the construction of phylogenetic tree was done applying the maximum likelihood method with thousand bootstrap replicates in the MEGAX software (Kumar et al. 2018).

Reconstruction of ancestral protein sequences

The common ancestral protein sequence of mammalian TLRs was predicted using FireProtASR (ancestral sequence reconstruction) v1.1 webserver with default parameter settings (Musil et al. 2021). Analyzing ancestral sequences in an evolutionary context to infer the ancestral sequences at certain nodes of a tree is termed as ASR. Reconstructing ancestral sequences is a well-established method for inferring the evolutionary history of genes. Along with the application in the discovery of the most probable evolutionary ancestors of study protein, it has been a useful approach for the design of extremely stable proteins. This protocol enables the implementation of the automated workflow FireProt^{ASR} allowing various forms of inputs and advance settings (Khan et al. 2021). All reconstruction methods involve a phylogenetic tree inferred from a given alignment. The quality of the tree is crucial for the reliable reconstruction. We have provided the multiple sequence alignment and the phylogenetic tree of all mammalian TLR sequences as input for our study. Upon submitting input data, the server will execute the dataset and reconstruct ancestral nodes along with their sequences.

Analysis of the ancestral sequences

We have performed sequence based and structural analysis of the identified ancestral sequences to accomplish our study. The Clustal Omega program, a widely used package for carrying out multiple sequence alignment (Madeira et al. 2022), was used for the alignment of the ancestral protein sequences. The prediction of three-dimensional structural models of ancestral proteins was performed using AlphaFold2 (Mirdita et al 2022). It is an artificial intelligence system developed by DeepMind that can predict three-dimensional structures of proteins from amino acid sequences with higher accuracy (Yang et al 2023).

Pairwise structure alignment was performed using the structural alignment tool available in Protein Data Bank (https://www.rcsb.org/alignment). This web-based tool enables the alignment of one or more structures to a particular reference structure that can be accessible from the "Analyze" section of the menu bar. In superposed structures, RMSD is calculated between aligned pairs of the backbone C-alpha atoms. Smaller RMSD indicates better structure alignment between the two structures. TM-score (template modeling score) is a measure of topological similarity between the template and model structures. It ranges between 0 and 1, where 1 indicates a perfect match and 0 is no match between the two structures. Scores < 0.2 usually indicate that the proteins are unrelated while those > 0.5 generally have the same protein fold in SCOP/CATH (Zhang and Skolnick 2005).

Protein domains of the ancestral sequences were annotated using the ScanProsite tool (de Castro et al. 2006). Evolutionary parameters such as rate of non-synonymous substitutions per non-synonymous site (Ka) and rate of synonymous substitutions per synonymous site (Ks) of the ancestral sequences were analyzed with respect to the root node sequence of the phylogenetic tree (Nei and Gojobori 1986; Yang 2007). The interaction of the ancestral protein sequences and Human_TLR9 sequence that have been used as a reference for the remaining species (Zhou et al. 2013) with the CpG ODN (Areal et al. 2011) was studied in the HDOCK. This web server enables hybrid docking algorithm of template-based modeling and free docking. The server supports protein–protein and protein–DNA/RNA docking and accepts both sequence and structure inputs for proteins. The docking scores are calculated through a knowledge-based iterative scoring function in this tool. A more negative docking score means a more possible binding model (Yan et al. 2017).

Results

Amino acid usage pattern of Toll-like receptor genes

We used mammalian Toll-like receptor (TLR1-TLR10) gene sequences to investigate the amino acid usage (AAU) pattern through correspondence analysis (CoA). Mutations are accumulated in TLR genes through various evolutionary processes. These mutations lead to the change in amino acid composition of TLRs. The CoA on the amino acid usage of mammalian TLR genes was performed to study the impact of such changes on the functionality of the encoded TLR proteins. The distribution of genes along the two major axes of the correspondence analysis is shown in Fig. 1. The first and second major axes accounted for 57.57% and 10.76% of the total variation of amino acid usage. A clear separation of the amino acid usage pattern of TLR9 genes with respect to other TLR (TLR1-TLR8 and TLR10) genes has been observed. Because the horizontal axis of correspondence analysis accounts for the majority of variation of the TLRs in CoA further analysis was carried out based on the distribution of mammalian TLR genes along this axis.

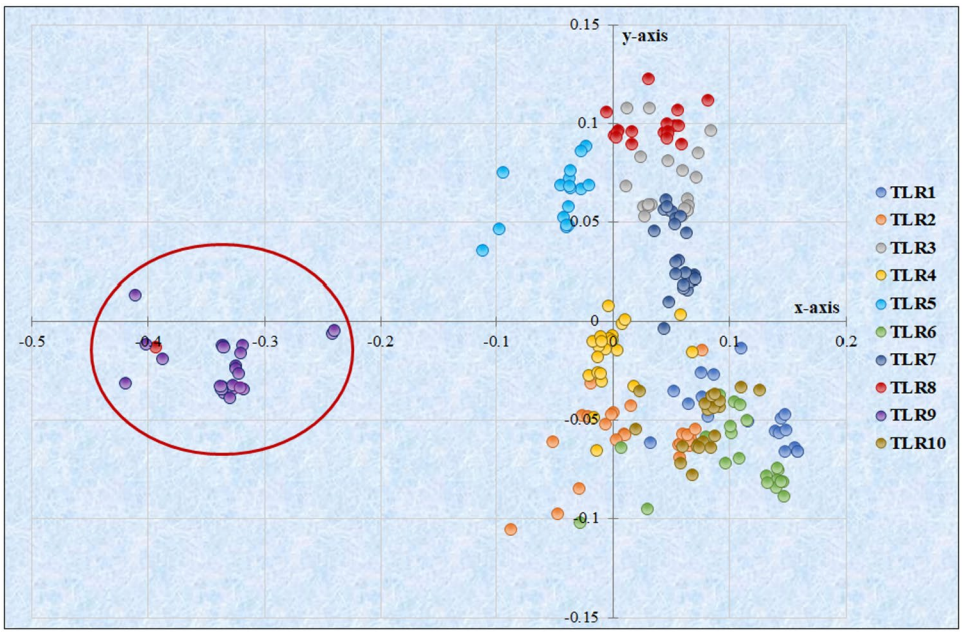


Fig. 1 Distribution of mammalian Toll-like receptor (TLR) genes along the two major axes of correspondence analysis (CoA) on amino acid usage. Distinct pattern of amino acid usage of TLR9 genes (violet) are marked with the red circle

Change in amino acid usage of a gene may affect the various physicochemical properties of TLR gene. We have calculated various physicochemical parameters of TLR gene sequences to understand the factor driving this distinct amino acid usage pattern among them. The parameters such as hydrophobicity, GC-content, GC3 values, effective number of codons (ENC), and aromaticity were found to differ significantly ($p < .05$) between TLR9 and other TLR (TLR1-8, TLR10) genes. Significant correlation was observed between the gene position along the horizontal axis and hydrophobicity ($r = -0.346$, $p < .01$), GC-content ($r = -0.977$, $p < .01$), GC3 values ($r = -0.96$, $p < .01$), effective number of codons (ENC) ($r = 0.825$, $p < .01$) and aromaticity ($r = 0.437$, $p < .01$) of the encoded protein. These correlation values indicate that the physicochemical parameters are contributing in the distinct amino acid usage pattern of TLR9.

Highly significant negative correlation with GC content, GC3 value indicated the influence of the codon bias. To better understand the relation between gene composition and codon bias, an ENC–GC3 scatter diagram was prepared as shown in Fig. 2. Such ENC–GC3 plots have been widely used to determine whether codon usage of a gene is shaped by natural selection. A significant correlation was observed between ENC and GC3 values ($r = -0.837$, $p < .01$). The solid line represents the expected curve in Fig. 2. TLR genes (TLR1-TLR8, TLR10) that lie on the expected curve indicate codon usage bias is only affected by mutation pressure. TLR9 genes are placed away from the expected curve, indicating that its evolution is shaped by the influence of natural selection.

Evolutionary selection analysis

The analysis of evolutionary selection can identify specific cases of adaptation as well as general principles that guide evolution. The analysis of evolutionary processes to distinguish between neutral and adaptive changes is thus very important. To understand the effect of evolutionary selection on the distinct amino acid usage pattern of TLR9, we have analyzed the evolutionary parameters such as non-synonymous substitution (Ka), synonymous substitution (Ks), ratio of non-synonymous and synonymous substitution (Ka/Ks) of the mammalian TLR genes. The analysis of these parameters is important for the study of the dynamics of molecular evolution of TLRs. Results were compared between TLR9 and other TLR genes as we obtained the difference in amino acid usage pattern between them. We found a significant difference of Ks and Ka/Ks between TLR9 and other TLRs, but Ka was not statistically significant in all the cases. The average value of Ks is more and Ka/Ks is less in thecase of TLR9 cluster. In spite of overall purifying selection on TLR genes, significant difference of non-synonymous substitution (Ka), synonymous substitution (Ks), ratio of non-synonymous and synonymous substitution (Ka/Ks) are observed. These results suggest that the evolution of TLR9 genes is highly influenced by synonymous substitution (Ks).

Ancestral sequence reconstruction

Ancestral sequence reconstruction is the calculation of ancient protein sequences on the basis of extant ones. The

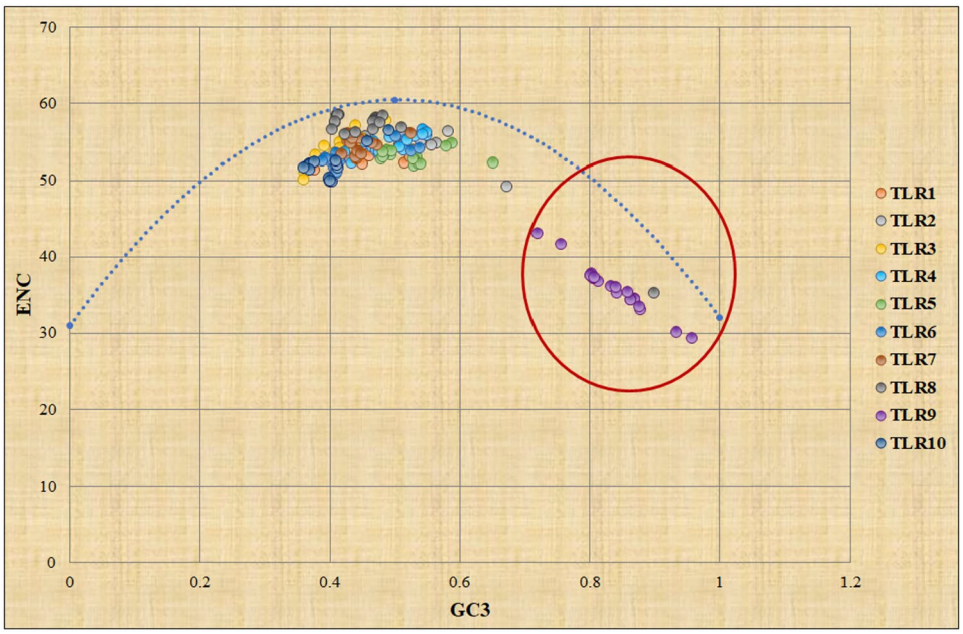


Fig. 2 The plot of ENC–GC3 for mammalian Toll-like receptor genes. The solid line represents the expected curve (blue). TLR genes (TLR1-TLR8, TLR10) those lie on the expected curve indicate codon usage bias is only affected by mutation pressure. TLR9 genes those are away from the expected curve indicate the influence of natural selection

previous analysis suggests that TLR9 shows distinct pattern of amino acid usage and the highest synonymous substitution rate with respect to other TLR genes. Thus, the ancestral sequence reconstruction through phylogenetic tree has been performed to reconstruct the evolutionary paths of the TLR protein family to study the key mechanism of the molecular evolution of TLR9. The ancestral sequence reconstruction phylogenetic tree of mammalian Toll-like receptor generated from the software is shown in Supplementary Fig. 1. In this figure, various TLR genes (for example: TLR1, TLR2, TLR3) are marked with different colors and Nodes are assigned with Node number. All the TLR9 genes are marked in red and their ancestral Node is denoted by Node 235. Similarly, all the TLR7 and TLR8 genes are marked in orange and their ancestral Node is denoted by Node 256. TLR3 and TLR5 genes are marked in blue and their ancestral Node is denoted by Node 299. TLR1, TLR2, TLR4, TLR6, and TLR10 genes are marked in green and their ancestral Node is denoted by Node 337. Node 226 denoted the root node that leads to the evolutionary path of TLRs through Node 232, Node 233, and Node 234. This entire evolutionary route of divergence of various TLRs from their common ancestor is schematically represented in Fig. 3. Here, the common root node is Node 226. All other TLRs have been evolved from this via intermediate nodes. For example, Fig. 3 also depicts the evolution of TLR9 from Node 226 via Node 235. Similarly, the evolutionary path of other TLRs from the root can be easily understood from Fig. 3 which is a simplified diagrammatic representation of evolutionary paths of various TLRs from root.

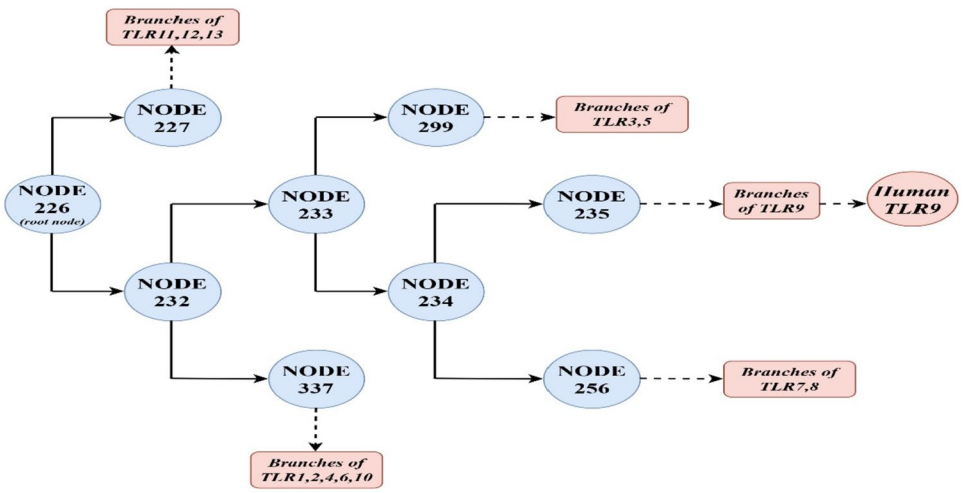


Fig. 3 Simplified schematic representation of the selection of ancestral nodes from the phylogenetic tree. Node 226 denotes the root node and the evolutionary pathway that leads to TLR9 follows via Node 232, Node 233, Node 234, and Node 235. Node 227 denotes the ancestral node of TLR11, 12, 13, Node 337 denotes ancestral node of TLR1, 2, 4, 6, 10, Node 299 denotes ancestral node of TLR3, 5, and Node 256 denotes ancestral node of TLR7, 8

Nodes	NODE226	NODE227	NODE337	NODE232	NODE233	NODE234	NODE299	NODE256	NODE235	Human_TLR9
NODE226	100	89.75	75.86	83.83	74.79	50.92	62.53	46.73	36.46	34.81
NODE227	89.75	100	68.02	74.69	66.3	46.69	56.95	42.98	33.89	32.97
NODE337	75.86	68.02	100	85.16	73.84	49.58	60.46	45.73	35.71	34.22
NODE232	83.83	74.69	85.16	100	83.79	58.16	67.68	53.4	40.97	38.45
NODE233	74.79	66.3	73.84	83.79	100	66.49	74.95	60.56	45.49	42.73
NODE234	50.92	46.69	58.16	58.16	66.49	100	50.28	88.14	58.15	52.78
NODE299	62.53	56.95	60.46	67.68	74.95	50.28	100	45.76	36.47	35.87
NODE256	46.73	42.98	45.73	53.4	60.56	88.14	45.76	100	49.02	45.61
NODE235	36.46	33.89	35.71	40.97	45.49	58.15	36.47	49.02	100	84.09
Human_TLR9	34.81	32.97	34.22	38.45	42.73	52.78	35.87	45.61	84.09	100

Fig. 4 Heatmap showing percent identity matrix of proteins obtained from multiple sequence alignment, colours correspond to the percent identity with high values (red), medium values (white) and low values (blue). Values in the box represent sequence homology in percentage. Higher sequence identity of TLR9 with its immediate ancestor (Node 235) but lower sequence identity with the ancestral nodes was observed

Analysis of the ancestral sequence

We accomplished our study through sequence based and structural analysis on the selected ancestral nodes that encompasses the evolutionary path of TLR9. Sequence-based analyses such as multiple sequence alignment of the ancestral sequences, analysis of the functional domains, estimation of synonymous, and nonsynonymous substitution were performed in order to understand the gradual changes that occurred during TLR9 evolution. Structural studies were also performed to assess the functional changes.

Multiple sequence alignment (MSA) generated a percent identity matrix of the protein sequences to provide an overview of the similarities between the sequences. The heatmap of the percent identity matrix reported from the alignment is displayed in Fig. 4. A higher sequence identity of TLR9 with its immediate ancestor (Node 235) but a lower sequence identity with the root (Node 226) was observed. It suggests that the continuous changes in sequence level along the ancestral lineages lead to the distinct sequence pattern of TLR9. The prediction of domain of the selected protein sequences was done and the number of LRR in the ectodomain was calculated. The orientation of LRRs in the ancestral lineages was different compared to Human_TLR9 and its immediate ancestral node. LRRs are the important components of the functional domains of TLRs that recognize the pathogen-associated molecular pattern (PAMP). Variation in the number of LRR in the ancestors of TLR9 was observed (Fig. 5). It suggests that during the evolution the variations among the LRRs of the ancestral nodes contributed to the specific pattern recognition of TLR9.

Fig. 5 Number of LRR present in the TLR genes and the ancestral nodes are shown in the bar plot. Number of LRR in human_TLR9 is decreased from its immediate ancestor Node235

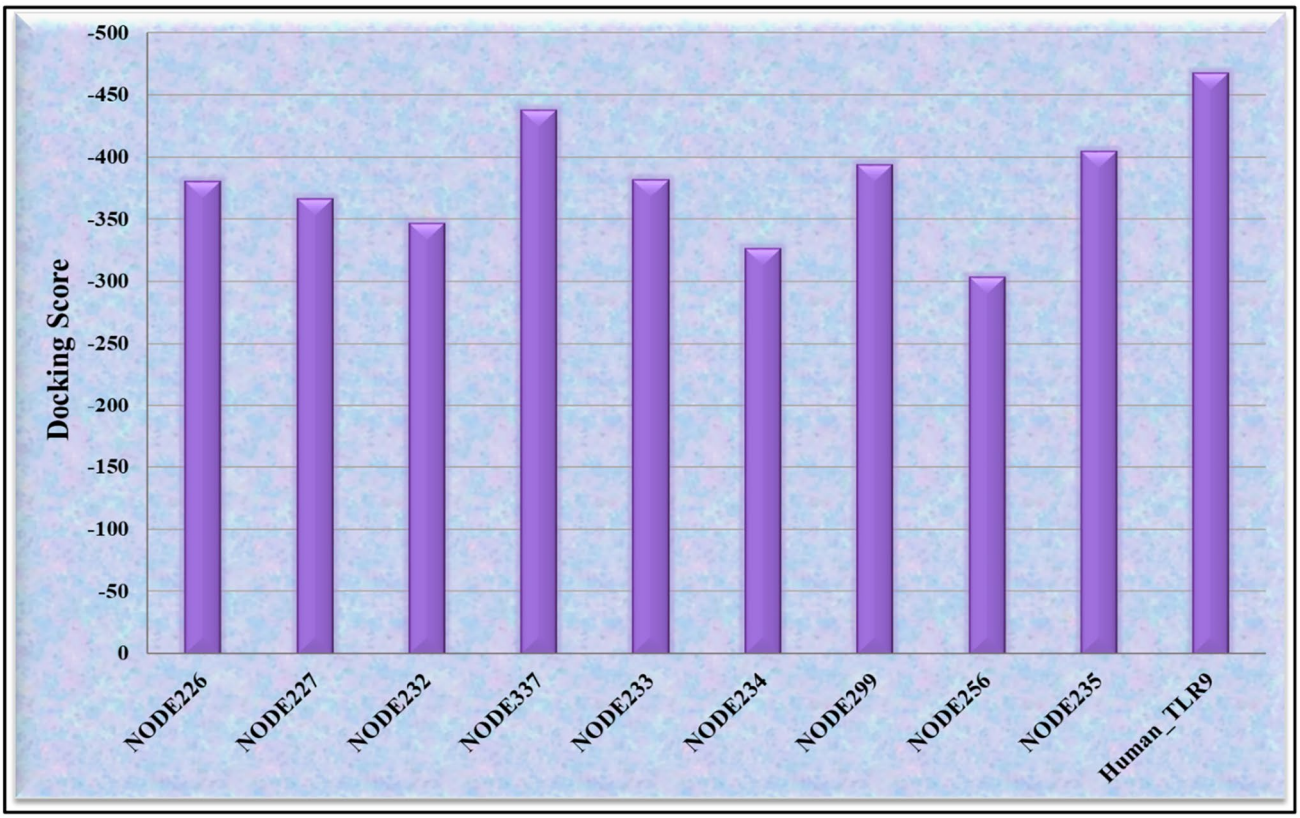


Fig. 6 Docking score of the interaction analysis between selected sequences and known ligand of CpG DNA of TLR9. The highest docking score is observed in case of TLR9

To observe these differences in structural level, structural models of the ancestral nodes and Human_TLR9 from the existing TLR9 group were prepared and compared through pairwise structural alignment (Supplementary File 1). Root mean square deviation (RMSD) and TM-score (template modeling score) were important metrics in this analysis. The RMSD values of TLR9 with the root node were higher compared to the other ancestral nodes and it gradually decreased in other nodes. These observations also showed more deviation of TLR9 from the root with respect to other TLRs along the ancestral nodes in the evolution of TLR9. For all the pairwise structural alignment, TM-score variation was observed but the values indicated that they are in the same protein fold.

TLR9 is a receptor for sensing bacterial DNA/CpG-containing oligodeoxynucleotides (CpG ODN) as PAMP within the endosomal compartment. An interaction study of ancestral proteins with this known ligand of Human_TLR9 was performed. It will help to understand how the present ligand is selected through evolution facilitating stronger interaction with TLR9. The interaction of Human_TLR9 and CpG ODN was also studied. The docking score of all the interactions is shown in Fig. 6. The highest docking score observed in the case of Human_TLR9 indicated the most compatible interaction of the ligand with the present TLR9. It reveals that TLR9 achieved its present conformation through the structural changes in the ancestral nodes during the course of evolution. Present TLR9 is very specific in recognizing its ligand as the ancestral nodes showed comparatively less stable interaction with this ligand.

The assessment of the evolutionary impact on the ancestral node sequences was also done by measuring the changes in non-synonymous substitution (Ka), synonymous substitution (Ks), ratio of non-synonymous, and synonymous substitution (Ka/Ks) (Fig. 7). Gradual increase of Ks from root to the other ancestral nodes was seen and it became extremely high in Human_TLR9. The Ka value is also high in Human_TLR9 compared to the ancestral sequences. Due to the high value of Ks, the Ka/Ks value became very low in Human_TLR9. The influence of synonymous substitution has been shaping the TLR9 evolution compared to its ancestral nodes.

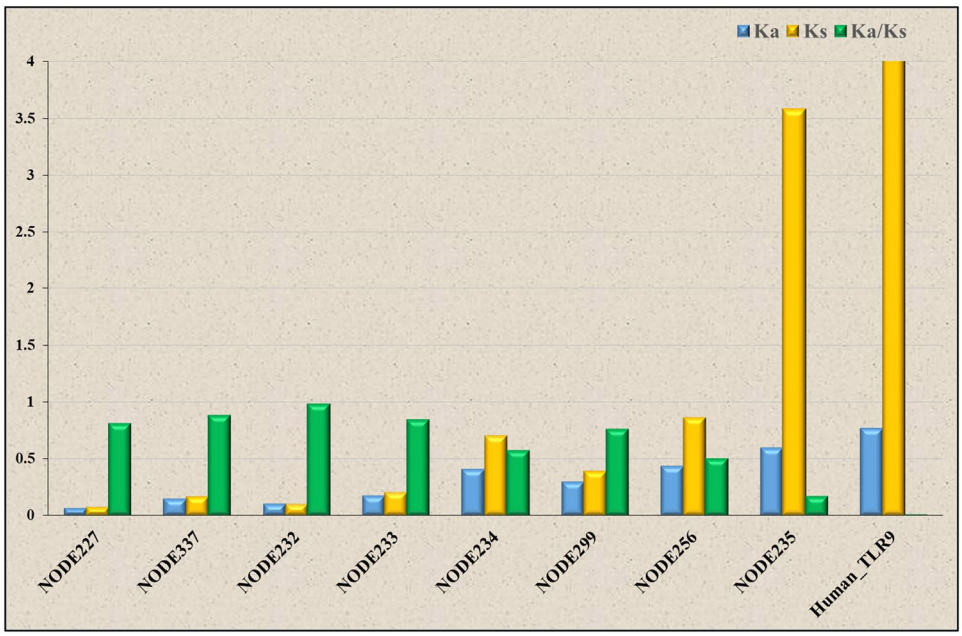


Fig. 7 Synonymous (Ks) and non-synonymous (Ka) substitution rates in TLR9 and its ancestral nodes. It is clear from the figure that both Ka as well as Ks are highest in case of TLR9

Discussion

The transmembrane pattern recognition receptor TLRs are best known for their roles in innate immunity via recognition of pathogen and initiation of signaling response. In this study, a comprehensive analysis of mammalian Toll-like receptor gene sequences (TLR1-TLR10) revealed that TLR9 follows a distinct pattern of evolution. Sequence-based features and evolutionary constraints are found to influence the divergence of TLR9 from other TLRs. Ancestral sequence reconstruction analysis also revealed that the gradual evolution of TLR genes in several ancestral lineages leads to the distinct pattern of TLR9.

Mammalian TLRs are responsible for the recognition of conserved molecular pattern derived from various classes of pathogens resulting in the induction of innate immune response. Pathogen-induced selection is considered a crucial selective mechanism driving the evolution of immune system components. We have identified various factors influencing TLR-dependent heterogeneity in amino acid usage that contribute to the differences in their immunological responses in mammals. We also found that high synonymous substitutions have shaped the observed changes between TLR9 and other mammalian TLR genes in spite of non-synonymous substitutions inducing the amino acid changes.

The divergence of TLR9 is demonstrated in this study through the ancestral sequence reconstruction. The analysis of the ancestral sequences also reinforced that changes occurred in the TLRs during their evolution from the ancestral lineages that were mostly observed in the TLR9 and its descendants. The decrease in the percent sequence identity of TLR9 from the root to the ancestral nodes to the mammalian TLR9 branch of the tree depicts gradual changes that happened in the sequences through the accumulation of mutation. The domain-wise analysis also suggested the accumulation of a greater number of mutations in the ectodomain causing variation in the number of LRR. Each TLR comprises an ectodomain with leucine-rich repeats (LRRs) that facilitate the recognition of pathogen-associated molecular pattern (PAMP) and a cytoplasmic Toll/IL-1 receptor (TIR) domain that initiates downstream signaling. The mutational changes also have been influenced by gradual selection pressure on the ancestral sequences in the course of evolution. Influence of synonymous and non-synonymous substitution among the ancestral sequences is observed and the gradual selection pressure in the course of evolution leading to the distinct pattern of TLR9. The interaction study also revealed a more stable interaction of the ligand with TLR9 compared to the ancestral nodes. Although decreasing docking score in other ancestral nodes indicated less stable interaction.

This study enables a new approach to explore the emergence of Toll-like receptor through the ancestral sequence reconstruction that elucidates a distinct pattern of evolution of TLR9. It demonstrates that the evolutionary divergence of TLR9 started from the beginning and the gradual accumulation of changes in the ancestral lineages leads to the distinct pattern of TLR9 compared to the other mammalian TLRs. It will elucidate the biological significance of TLR9 and provide evidence for their distinct contributions in response to host defence.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00251-024-01338-8.

Data Availability All sequence information is available in public databases and the accession numbers of the sequences used in the present study are provided in Supplementary Table 1.

References

Areal H, Abrantes J, Esteves PJ (2011) Signatures of positive selection in Toll-like receptor (TLR) genes in mammals. BMC Evol Biol 11:368. https://doi.org/10.1186/1471-2148-11-368

Botos I, Segal DM, Davies DR (2011) The structural biology of Toll-like receptors. Structure 19:447–459. https://doi.org/10.1016/j.str.2011.02.004

Cook DN, Pisetsky DS, Schwartz DA (2004) Toll-like receptors in the pathogenesis of human disease. Nat Immunol 5:975–979. https://doi.org/10.1038/ni1116

de Castro E, Sigrist CJ, Gattiker A et al (2006) ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res 34(W362):W365. https://doi.org/10.1093/nar/gkl124

Ghosh M, Basak S, Dutta S (2022) Natural selection shaped the evolution of amino acid usage in mammalian toll like receptor genes. Comput Biol Chem 97:107637. https://doi.org/10.1016/j.compbiolchem.2022.107637

Gumulya Y, Gillam EM (2017) Exploring the past and the future of protein evolution with ancestral sequence reconstruction: the 'retro' approach to protein engineering. Biochem J 474:1–19. https://doi.org/10.1042/BCJ20160507

Karapetyan L, Luke JJ, Davar D (2020) Toll-like receptor 9 agonists in cancer. Onco Targets Ther 13:10039–10060. https://doi.org/10.2147/OTT.S247050

Khan RT, Musil M, Stourac J (2021) Fully automated ancestral sequence reconstruction using FireProtASR. Curr Protoc 1:e30. https://doi.org/10.1002/cpz1.30

Kumar S, Stecher G, Li M et al (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35(6):1547–1549. https://doi.org/10.1093/molbev/msy096

Madeira F, Pearce M, Tivey ARN et al (2022) Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Res 50:W276–W279. https://doi.org/10.1093/nar/gkac240

Merkl R, Sterner R (2016) Ancestral protein reconstruction: techniques and applications. Biol Chem 397:1–21. https://doi.org/10.1515/hsz-2015-0158

Mirdita M, Schütze K, Moriwaki Y et al (2022) ColabFold: making protein folding accessible to all. Nat Methods 19:679–682. https://doi.org/10.1038/s41592-022-01488-1

Muffato M, Louis A, Nguyen NTT (2023) Reconstruction of hundreds of reference ancestral genomes across the eukaryotic kingdom. Nat Ecol Evol 7:355–366. https://doi.org/10.1038/s41559-022-01956-z

Musil M, Khan RT, Beier A et al (2021) FireProtASR a web server for fully automated ancestral sequence reconstruction. Brief Bioinform 22:bbaa337. https://doi.org/10.1093/bib/bbaa337

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426. https://doi.org/10.1093/oxfordjournals.molbev.a040410

Nie L, Cai SY, Shao JZ et al (2018) Toll-like receptors, associated biological roles, and signaling networks in non-mammals. Front Immunol 9:1523. https://doi.org/10.3389/fimmu.2018.01523

Peden JF (2000) Analysis of codon usage (Doctoral dissertation). University of Nottingham, United Kingdom

Roach JC, Glusman G, Rowen L et al (2005) The evolution of vertebrate Toll-like receptors. Proc Natl Acad Sci USA 102:9577–9582. https://doi.org/10.1073/pnas.0502272102

Roy A, Banerjee R, Basak S (2017) HIV progression depends on codon and amino acid usage profile of envelope protein and associated host-genetic influence. Front Microbiol 8:1083. https://doi.org/10.3389/fmicb.2017.01083

Takeda K, Akira S (2005) Toll-like receptors in innate immunity. Int Immunol 17:1–14. https://doi.org/10.1093/intimm/dxh186

Vidya MK, Kumar VG, Sejian V et al (2018) Toll-like receptors: significance, ligands, signaling pathways, and functions in mammals. Int Rev Immunol 37:20–36. https://doi.org/10.1080/08830185.2017.1380200

Yan Y, Zhang D, Zhou P et al (2017) HDOCK: a web server for protein-protein and protein-DNA/RNA docking based on a hybrid strategy. Nucleic Acids Res 45:W365–W373. https://doi.org/10.1093/nar/gkx407

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591. https://doi.org/10.1093/molbev/msm088

Yang Z, Zeng X, Zhao Y et al (2023) AlphaFold2 and its applications in the fields of biology and medicine. Signal Transduct Target Ther 8:115. https://doi.org/10.1038/s41392-023-01381-z

Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 33:2302–2309. https://doi.org/10.1093/nar/gki524

Zhou W, Li Y, Pan X, Gao Y et al (2013) Toll-like receptor 9 interaction with CpG ODN–an in silico analysis approach. Theor Biol Med Model 10:18. https://doi.org/10.1186/1742-4682-10-18