Databases and computational interaction models of Toll-like receptors

Dissertation

Fakultät für Geowissenschaften

Ludwig-Maximilians-Universität München

vorgelegt von

Jing Gong

München, den 24. Februar 2010

Erstgutachter: Prof. Dr. Wolfgang M. Heckl

Zweitgutachter: Prof. Dr. Robert W. Stark

Prüfungstermin: 26.05.2010

Abstract

Toll-like receptors (TLRs) recognize pathogen-associated molecular patterns (PAMPs) on invading organisms and are the first line of defense in innate immunity. To date, much has been learned about TLRs and their roles in autoimmune diseases are being unraveled. The autoimmune disease systemic lupus erythematosus (SLE) progresses as a consequence of the inappropriate recognition of self nucleic acids by TLRs. For the development of therapeutic approaches of SLE it is necessary to understand possible negative regulation mechanisms of TLR. Single immunoglobulin interleukin-1 receptor-related molecule (SIGIRR) is the best characterized TLR signaling inhibitor. It can interfere with the receptor complexes and attenuate the recruitment of downstream adaptors to the receptors. So far, the mechanisms of structural interactions between SIGIRR, TLRs and adaptor molecules are unknown. To develop a working hypothesis for these interactions, we constructed three-dimensional models for these single molecules based on computational predictions. Then, models of essential complexes involved in the TLR signaling and the SIGIRR inhibiting processes were yielded through protein-protein docking analysis.

With the high-throughput genome sequencing projects, a central repository for the growing amount of TLR sequence information has been created. However, subsequent annotations for these TLR sequences are incomplete. For example, the indicated numbers and positions of leucine-rich repeat (LRR) motifs contained in individual TLR ectodomains are greatly distinct or missing in established databases. In this vein, we have developed a database of TLR structural motifs called TollML (http://tollml.lrz.de). It integrates all TLR protein sequences that have been identified to date. These sequences were semi-automatically partitioned into three levels of structural motif categories. The manual motif identification procedure provided TollML with the most complete and accurate database of LRR motifs compared with other databases that contain TLR data.

LRR motifs are present not only in TLRs, but also in many other proteins. To date, more than 6,000 LRR protein sequences and more than 130 crystal structures of them have been determined. This knowledge has increased our ability to use individual LRR structures extracted from the crystal structures as building blocks to model LRR proteins with unknown structures. Because the individual LRR structures are not directly available from any protein structure database, we have developed a conformational LRR database called LRRML (http://lrrml.lrz.de). It collects three-dimensional LRR structures manually identified from all determined crystal structures of LRR-containing proteins and thus provides a source for the structural modeling and analysis of LRR proteins. With the help of TollML and LRRML, we constructed models of the human/mouse TLR5-13 ectodomains and suggested some potential receptor-ligand interaction residues based on these models.

Abstract

2

Contents

Contents

1. Introduction

The recognition of invading pathogenic organisms is critical in the proper activation of the immune system [1]. Inappropriate activation may cause immunodeficiency diseases and autoimmune diseases. The immune system consists of the antigen-unspecific innate immune system and the antigen-specific adaptive immune system. Toll-like receptors (TLRs) are responsible for the innate immunity. They recognize a wide variety of pathogen-associated molecular patterns (PAMPs) via their extracellular domains, acting to trigger intracellular signaling pathways [2]. In the signaling pathways, the association of receptors and downstream adaptors leads to the induction of inflammatory cytokines. Nevertheless, excessive production of cytokines contributes to the pathogenesis of autoimmune diseases [3]. TLR signaling must therefore be under tight negative regulation to maintain an immune balance. The single immunoglobulin interleukin-1 receptor-related molecule (SIGIRR) was shown to be involved in the negative regulation of TLR signaling [4]. Recent studies revealed that SIGIRR can inhibit the nuclear antigen-recognizing TLRs' stimulation-induced over-activation of immune responses in the body suffered by the systemic lupus erythematosus (SLE) [5]. Thus, explicit structural knowledge of molecular interactions between TLRs and SIGIRR can help to develop small drug molecules that mimic SIGIRR's functional region and bind to TLRs to block their inappropriate activation. As yet, no information has been published on the three-dimensional (3D) structure of SIGIRR, and SIGIRR's inhibiting mechanism from a structural point of view remains largely unknown. In this study, we constructed 3D structural models for the interacting domains of SIGIRR and TLRs based on homology modeling. Then, models of essential complexes involved in the TLR signaling and the SIGIRR inhibiting processes were yielded through protein-protein docking analysis. These results are expected to facilitate efforts to design further mutagenesis experiments to clarify the regulatory role of SIGIRR in innate immune responses and autoimmune diseases.

The ligation of the TLR extracellular domain starts the TLR signaling [3]. Therefore, explicit structural knowledge of receptor-ligand interactions can help to develop agonists and antagonists of TLRs that have therapeutic significance for infectious diseases. To date, several crystal structures of TLR-ligand complexes have been determined [6-9]. However, compared with the large number of known TLR protein sequences, the structures of most TLRs are still unknown. Computational methods can carry out fast and large-scale structure predictions based on these known TLR structures. The TLR extracellular domain comprises a number of intricately organized, contiguous leucine-rich repeats (LRRs). In order to provide a convenient workbench for structure predictions of TLR extracellular domains, we have developed two databases. The TollML database contains structural motif annotations for all known TLR sequences. The LRRML database contains individual 3D LRR structures

identified from all experimentally determined structures of LRR-containing proteins. None of the data in these two databases are in any other database.

This dissertation consists of seven chapters. Chapter 1 is a brief introduction. The biological background of TLRs and the bioinformatics theories commonly used in protein structure predictions are reviewed in Chapters 2-4. All computational methods used in the database construction and the TLR/SIGIRR modeling are described in Chapter 5. Chapter 6 very briefly summarizes the results published in scientific journals. The dissertation ends with a conclusion summarizing the investigations conducted (Chapter 7).

2. Immune diseases and Toll-like receptors

2.1 Immune system and immune diseases

The immune system comprises a complex network of organs, specialized tissues, cells and molecules that protects organisms from infections with pathogenic organisms and other harmful substances through triggering immune responses [1]. In general, the immune system recognizes a wide variety of invading pathogens such as viruses, bacteria, fungi and parasites. To function properly, it needs to distinguish them from the host's healthy cells and tissues. Thus, the function of immune system is a complicated balancing act based on the ability of self/non-self recognition. A malfunction of the immune system can result in diseases. On one hand, when the immune system is less active than normal, immunodeficiency occurs and results in recurring and life-threatening infections. A well-known example is the acquired immune deficiency syndrome (AIDS), which is caused by the human immunodeficiency virus (HIV) [10]. On the other hand, an overactive immune system can attack substances and tissues normally present in the host as if they were foreign organisms and result in autoimmune diseases or susceptibility to infections. SLE is a chronic autoimmune disease that primarily afflicts women of childbearing age (15-44 years) [11]. It progresses as a consequence of the production of autoantibodies to components of the host cell nucleus and attacks the body's heart, joints, skin, lungs, blood vessels, liver, kidneys and nervous system, resulting in inflammation, pain and tissue damage. SLE can be fatal. Survival for people with SLE in North America and Europe is approximately 95% at five years, 90% at 10 years and 78% at 20 years [11]. The treatment of such autoimmune diseases is typically with immunosuppression, which decreases the immune response. Currently, numerous new immunosuppressive drugs are being actively tested for SLE, such as cyclophosphamide, a nitrogen mustard alkylating agent that can decrease the immune response to various diseases and conditions [12]. Nevertheless, the use of immunosuppressive therapy in SLE carries significant risks for infection. Consequently, careful monitoring of infectious complications is warranted in patients receiving such therapies [13].

The immune system consists of two components: the innate immune system and the adaptive immune system. The innate immune system generates an immediate and a rather unspecific response to an infection or cell damage, whereas the adaptive immune system is a more sophisticated and more recently evolved antigen-specific system [14]. The adaptive immune system is present only in vertebrates, in which B and T lymphocytes utilize antigen receptors such as immunoglobulin (Ig) and T cell receptors to recognize pathogens specifically. The adaptive immune responses can also provide a long-lasting protection. They allow immunological memory, in which each pathogen is "remembered" by a signature antigen. Unlike the adaptive immune system, innate

immune systems are found in all plants and animals [15]. In fact, the innate immune system is the first line of defense against invading microorganisms. If pathogens successfully evade the innate response, vertebrates possess the adaptive immune defense that is activated by the innate response. The main players in innate immunity are phagocytes such as neutrophils, macrophages and dendritic cells. These cells express a limited number of germline-encoded pattern recognition receptors (PRRs) that specifically recognize PAMPs, which are displayed by microbes but not found in the host [16]. For viruses, nucleic acids, including double-stranded RNA (dsRNA), single-stranded RNA (ssRNA) and cytosine-phosphate-guanine (CpG) DNA, serve as PAMPs and can be recognized by multiple PRRs [17]. Following the recognition of viral nucleic acids, PRRs initiate signaling pathways that lead to the synthesis of multiple inflammatory cytokines such as type I interferon (IFN) and tumor necrosis factor-α (TNFα). In particular, type I IFN plays an essential role in the elimination of viruses. It upregulates the transcription of many IFN-inducible genes that influence protein synthesis, growth arrest and apoptosis to create an antiviral state. It also functions to enhance dendritic cell maturation, natural killer (NK) cell activation, antibody production and the differentiation of virus-specific cytotoxic T lymphocytes (CTL), which consequently facilitate adaptive immune responses in vertebrates. The best characterized family of PRRs is the TLR family.

2.2 Toll-like receptors in innate immunity

The term "Toll" was derived from a German word meaning "fantastic" or "amazing". In the early 1980s, Nüsslein-Volhard and Anderson first used "Toll" to refer to a gene that they discovered in a genetic screen of the fruit fly *Drosophila melanogaster*, the phenotype of which they thought to be "Toll" [18, 19]. Toll was initially implicated in the establishment of dorsoventral polarity in the early *Drosophila* embryo [20]. In 1996, Toll was shown by Hoffmann and co-workers to be an essential receptor for host defense against fungal infection in *Drosophila*, which has only the innate immunity [21]. One year later, Medzhitov and co-workers described a mammalian homolog of the Toll receptor (now termed TLR4), which was shown to induce the expression of genes involved in inflammatory responses [22]. Shortly after this important discovery, the progress of genome projects led to the identification of more than ten receptors in vertebrates that were direct homologs of the Toll receptor, and these became known as the Toll-like receptors (TLRs) [23]. TLRs play a key role in innate immunity. They recognize a remarkably wide variety of PAMPs, including glycolipids such as bacterial lipopolysaccharides (LPS), proteins such as bacterial flagellin and viral nucleic acids [2]. The ligation of receptors rapidly initiates intracellular signal transduction pathways to trigger the expression of genes, whose products can control innate immune responses.

All TLRs have a common domain organization, with an LRR ectodomain, a single

transmembrane helix and an intracellular Toll/interleukin-1 receptor (TIR) domain [24]. The ectodomain is a horseshoe-shaped solenoid structure consisting of 17-26 consecutive LRR motifs sandwiched between two terminal LRR modules (LRRNT and LRRCT) [25]. The sequence length of the ectodomain varies from 500 to 800 amino acids. Ectodomains are directly involved in the recognition of a variety of pathogens. The intracellular TIR domain comprises about 150 residues. It is conserved across all TLRs and is also shared by downstream signaling adaptor molecules [2]. The 20-30-amino acid-long transmembrane helix links the extracellular and intracellular portions and determines the subcellular localization of TLRs [26]. A common structure of TLRs is shown in Figure 1.
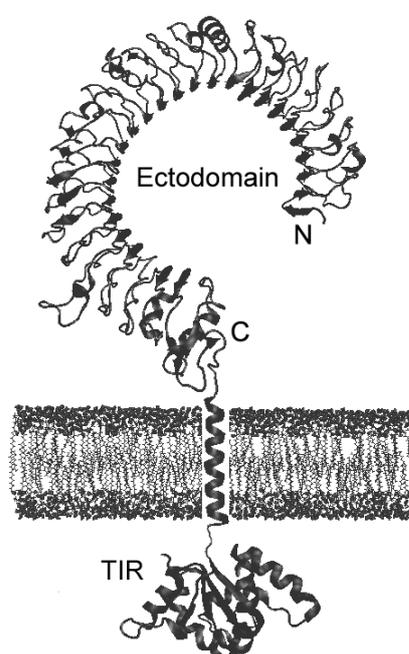
Figure 1. A typical TLR structure.

An LRR motif is defined as an array of 20-30 amino acids that is rich in the hydrophobic amino acid leucine [27]. All LRR sequences can be divided into a highly conserved segment (HCS) and a variable segment (VS). Each LRR forms a loop in which the residues of the HCS generate the concave surface of the LRR arc or horseshoe while the residues of VS form the convex surface. The HCS consists of an eleven- or twelve-residue stretch with the consensus sequence LxxLxLxxN(Cx)xL. Here, the letter L stands for Leu, Ile, Val or Phe, which form the hydrophobic core in the arc or horseshoe structure; N stands for Asn, Thr, Ser or Cys, which form hydrogen bonds between the adjacent loops, and x is any amino acid. The VS varies in length and consensus sequence; accordingly, eight types of LRRs have been proposed [27, 28] (detailed in section 6.2). Of these, only the typical type (T) and the bacterial type (S) are detected in the TLR LRRs with consensus sequences xxLxxxxLxxLxx and xxLPx(x)LPxx, respectively, where the letter P stands for Pro. As previously reported, LRRs of different types never occur simultaneously in the

same protein and probably evolved independently [27], but the T and S types are exceptions to this rule. They are often observed at the N-terminal part of one TLR protein and form the super-motif of one S-type LRR followed by two T-type LRRs (STT), which is repeated as STTSTTSTT or _TTSTTSTT [29]. It is assumed that both types evolved from a common precursor [30]. The terminal LRRNT and LRRCT modules do not match any consensus sequence of the eight types. They contain two to five cysteine residues forming disulfide bonds and stabilize the protein structure by shielding its hydrophobic core from exposure to solvents.

Mammalian TLRs comprise a large family with at least 13 members. Humans have ten members (TLR1-10), and mice have 13 (TLR1-13). Other mammalian species have between ten and 13 members [25]. TLR1-9 are conserved between the human and mouse. Although both the human and mouse have TLR10, the mouse TLR10 appears to have been damaged at some point in the past by a retrovirus. In addition, mice express TLR11, 12 and 13, none of which is represented in humans. Compared with mammals, many more TLR members ($\geq$ 20) have been identified in non-mammalian genomes [25].

According to the molecular tree of vertebrate TLRs proposed by Roach *et al*. (2005), mammalian TLR members can be divided into six major families [31]. The *TLR1* family consists of TLR1, 2, 6 and 10. This family contains fewer (20 or 21) LRRs than the other families. So far, the crystal structures of human/mouse TLR1, 2 and 6 ectodomains and human TLR1, 2 and 10 TIR domains have been determined [6, 9, 32, 33]. TLR1 and 2 heterodimerize, and the resulting dimer senses bacterial triacylated lipopeptides. TLR6 and 2 also heterodimerize and recognize bacterial diacylated lipopeptides [34, 35]. The function of TLR10 is still unknown. In the *TLR3* family, the only member (TLR3) contains 25 LRRs, as shown by the crystal structure of human/mouse TLR3, and senses synthetic and viral dsRNA [36-39]. TLR4 constitutes the *TLR4* family and contains 23 LRRs. It recognizes the Gram-negative bacterial product LPS [40]. The crystal structures of TLRs in both the *TLR1* and *TLR4* families show a feature in which irregular LRRs concentrate mainly at the central part of the ectodomains (LRR9-13) [6-9]. TLR5 in the *TLR5* family contains 22 LRRs and detects flagellin from bacteria [41]. The *TLR7* family consists of TLR7, 8 and 9 and contains 27 LRRs. A remarkable feature of the *TLR7* family is the presence of a less structured region in the middle of the ectodomain. TLR8, which is highly similar to TLR7, can sense viral ssRNA and synthetic imidazoquinolene compounds such as imiquimod [42]. TLR9 is the receptor for unmethylated CpG DNA motifs, which occur in bacterial and viral DNA [43]. TLR11, 12 and 13 are included in the *TLR11* family, which contains 24-27 LRRs. TLR11 senses parasite profilin, while the ligands of TLR12 and 13 remain unknown [44]. In contrast to other TLRs that are localized on the cell surface, the nucleic acid-recognizing TLR3 and 7-9 function in intracellular compartments such as endosomes and endolysosomes [3].

The above-mentioned crystal structures of TLR1 and 2 with triacylated lipopeptide,

TLR3 with dsRNA and TLR4 in a complex with LPS and MD2 explain the molecular mechanisms for TLR dimerization. The binding of ligand causes the C-termini of TLR ectodomains to associate, and this adds stability to the dimer. Simultaneously, the two intracellular TIR domains are brought into contact, which results in the recruitment of TIR domain-containing adaptor molecules to initiate downstream signaling [32]. Four such signaling adaptors have so far been characterized: myeloid differentiation factor-88 (MyD88), MyD88 adaptor-like protein (Mal), TIR domain-containing adaptor inducing interferon-β (TRIF) and TRIF-related adaptor protein (TRAM) (Figure 2).

MyD88 was discovered in 1990 [45]. Its function in TLR signaling was not implicated until 1998 [46, 47]. MyD88, harboring an N-terminal death domain (DD), a linker stretch and a C-terminal TIR domain, associates with the TIR domain of TLRs [48]. Upon stimulation, MyD88 recruits interleukin-1 receptor-associated protein kinase-4 (IRAK4) through the interaction of the DDs of both molecules, and it facilitates IRAK4-mediated phosphorylation of IRAK1. Activated IRAK1 then associates with tumor necrosis factor receptor-associated factor-6 (TRAF6), leading to activation of the TAK1/TAB kinase complex, which enhances activity of the IκB kinase (IKK) complex consisting of IKKα, IKKβ and IKKγ. Once activated, the IKK complex induces the phosphorylation and subsequent degradation of inhibitor nuclear factor-kappa B (IκB), which leads to nuclear translocation of transcription factor NF-κB and the induction of inflammatory cytokines such as type I IFN, TNFα, nucleosome assembly protein-1 (NAP1), interleukin-1 and interleukin-6 [49]. MyD88 is involved in the signaling pathways initiated by all TLRs, with the exception of TLR3 [50]. TLR3 exclusively uses TRIF for signaling in a MyD88-independent manner. Additionally, TLR4 can signal via both the MyD88-dependent and the TRIF-dependent pathways [2].

Mal, also known as TIR domain-containing adaptor protein (TIRAP), was the second characterized TLR adaptor protein [51]. It contains an N-terminal phosphatidylinositol 4, 5-bisphosphate (PIP2)-binding domain and a C-terminal TIR domain [52]. Mal and MyD88 play different roles in TLR2 and 4 signaling. MyD88 serves as an essential "signaling adaptor" that transmits signals from ligand-activated TLRs to downstream factors to initiate kinase-dependent signaling cascades, whereas Mal functions as a "sorting adaptor" that recruits MyD88 to the plasma membrane via its PIP2-binding domain to promote interactions between MyD88 and activated TLR2 and 4 beneath the membrane [53]. It was found that Mal's "sorting" function is only essential for the MyD88-dependent signaling pathways via TLR2 and 4, as Mal-deficient mice are not impaired in their response to TLR3, 5, 7 and 9 ligands [54]. Furthermore, binding assays by Ohnishi and co-workers suggested that MAL can simultaneously bind to TLR4 and MyD88 through its TIR domain, and there is no direct interaction between the TIR domains of TLR4 and MyD88 [55]. However, Mal was shown to be dispensable for TLR4 signaling when MyD88 is fused to the PIP2 targeting domain [53]. This observation indicated that there may be weak interactions between TIR domains of TLR4 and MyD88. An alternative Mal-independent, but MyD88-dependent,

pathway could contribute to TLR4 signaling [56].
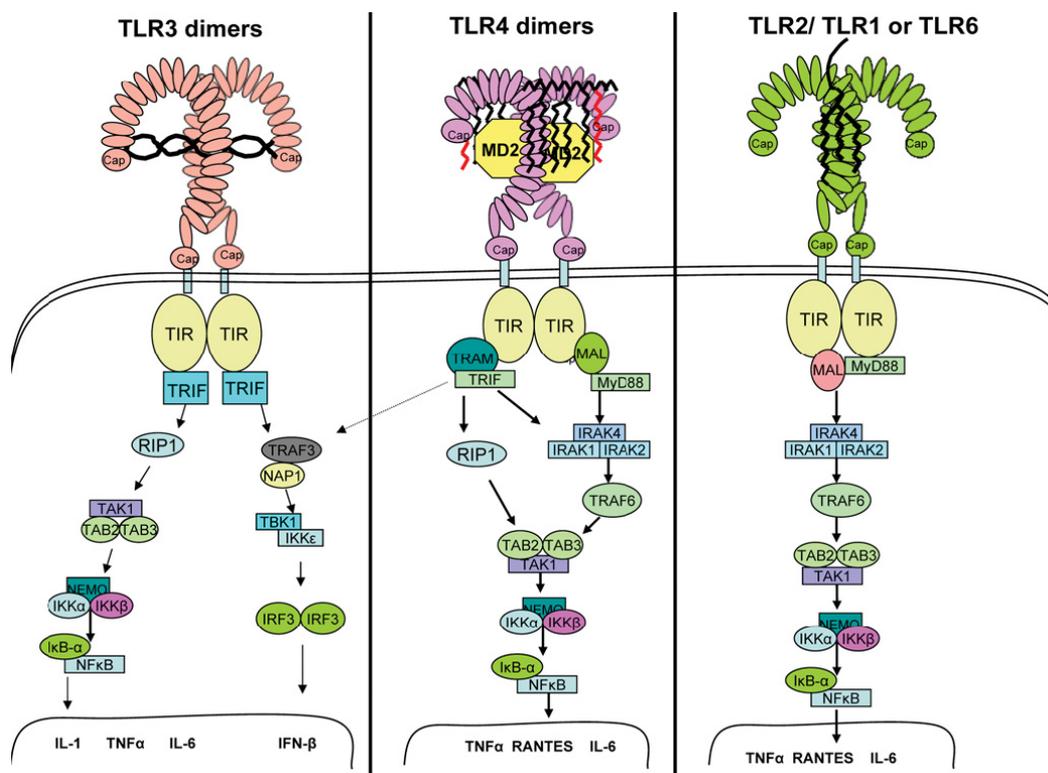
Figure 2. TLRs and their intracellular signaling pathways. The ectodomains of TLR3 (left) are monomeric in solution until they bind to dsRNA and the dimerization occurs. The dimerization is proposed to bring the two TIR domains into contact to initiate downstream signaling. Two downstream signaling pathways are subsequently triggered, leading to activation of NF-κB and IRF3. TLR4 (middle) and MD2 interact before stimulation with LPS. LPS induces dimerization of two TLR4-MD2 complexes. Dimerization also brings the TIR domains together and initiates downstream signaling through the adaptor molecules MyD88, Mal, TRIF and TRAM, which leads to the activation of the transcription factor NF-κB and the release of inflammatory cytokines. TLR2 and TLR1 (right) can heterodimerize and respond to triacylated lipopeptides, or TLR2 can bind to TLR6 and respond to diacylated lipopeptides. Both complexes utilize the adaptors MyD88 and Mal to initiate downstream signaling. Reproduced from [57] with permission from *The Biochemical Society* (http://www.biochemj.org).

TRIF is much larger than MyD88 and Mal in that it contains an extra receptor-interacting protein (RIP) homotypic interaction motif (RHIM) at the C-terminal region after its TIR domain [58]. TRIF is involved in the TLR3 and TLR4 MyD88-independent pathways. It is able to activate NF-κB through association with either RIP1 or TRAF6 (as with MyD88 and Mal). By contrast with MyD88 and Mal, TRIF is also able to associate with TRAF3 and activate the transcription factor IRF3, leading to the production of IFN-β.

TRAM, also known as TIR domain-containing adaptor molecule-2 (TICAM2) and TIR domain-containing protein (TIRP), was identified in 2003 [59]. TRAM contains an

12

N-terminal myristoylation site and a C-terminal TIR domain. It is involved in the TLR4-mediated, but not TLR3-mediated, activation of IRF3 and induction of IFN-β [60]. Thus, TRAM is essential for the TLR4-mediated, MyD88-independent, but TRIF-dependent, pathway. It acts as a "sorting adaptor" that delivers TRIF to the plasma membrane via its myristoylation site to promote interactions between TRIF and activated TLR4 [61].

In both MyD88-dependent and MyD88-independent pathways, the stimulation of TLRs triggers the induction of inflammatory cytokines. When all these cytokines are produced in excess, they induce serious systemic disorders with a high mortality rate in the host [3]. It is therefore not surprising that organisms have evolved mechanisms for modulating their TLR-mediated responses. The MyD88-dependent pathway is negatively regulated by a spliced variant of MyD88, known as MyD88 short (MyD88s). Its expression is induced in response to continuous stimulation with bacterial products or pro-inflammatory cytokines [62]. Overexpression of MyD88s results in impaired LPS-induced NF-κB activation through inhibition of IRAK4-mediated IRAK1 phosphorylation [63]. Another molecule that was shown to negatively regulate MyD88-dependent signaling is the transforming growth factor-β (TGF-β). TGF-β blocked NF-κB activation and cytokine production in response to TLR2, 4 and 5 ligands by decreasing MyD88 protein [64].

Besides MyD88s and TGF-β, the single immunoglobulin interleukin-1 receptor-related molecule (SIGIRR), also known as Toll/interleukin-1 receptor-8 (TIR8), has been found to be involved in the negative regulation of MyD88-dependent TLR signaling [4]. Overexpression of SIGIRR in Jurkat or HepG2 cells resulted in substantially reduced LPS (TLR4)-/ CpG DNA (TLR9)-induced activation of NF-κB [4, 65, 66]. Thus, SIGIRR has attracted tremendous research interest in the immunosuppressive therapy for autoimmune diseases. For example, SIGIRR can inhibit TLR7 and 9 stimulation-induced NF-κB over-activation in the body, as occurs in SLE [5]. SLE is characterized by the production of diverse autoantibodies, predominantly against nucleosomal antigens derived from the host [67]. The nucleosomal antigens are known to stimulate TLR7 and 9, which causes the activation of dendritic cells and the induction of interleukins. Homozygous deletion of the *Sigirr* gene was associated with a significant increase of interleukin production upon exposure to nucleosome immune complexes [5]. Compared with wild-type mice, *Sigirr*-deficient mice develop excessive lymphoproliferation when introduced into the context of a lupus susceptibility gene (Figure 3), as reported by Lech *et al*. (2008).
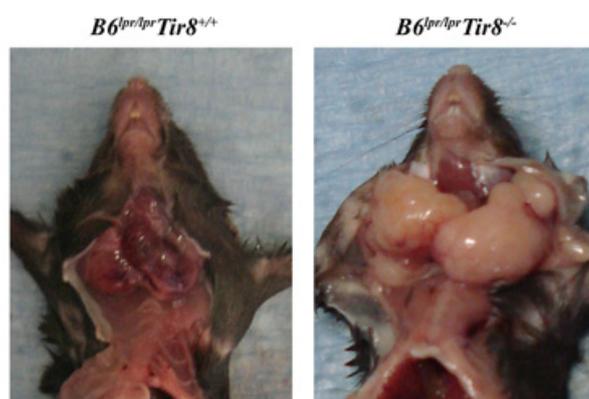
Figure 3. Lack of SIGIRR is associated with massive lymphoproliferation in B6$^{lpr/lpr}$ mice (right). Reproduced from [5] with permission from *The Journal of Experimental Medicine*.

The elucidation of the detailed molecular mechanism of SIGIRR's action is critical for understanding the regulatory role of SIGIRR in inflammatory and innate immune responses and is essential for evaluating the therapeutic potential of SIGIRR. Similar to TLRs, SIGIRR is also a membrane-bound protein consisting of an ectodomain and an intracellular TIR domain linked by a transmembrane helix. Unlike TLRs, SIGIRR's ectodomain is a small single Ig that does not support ligand-binding. Its intracellular domain cannot activate NF-κB because it lacks two essential amino acids (Ser447 and Tyr536) in its highly conserved TIR domain [68]. Moreover, the TIR domain of SIGIRR extends over that of TLRs by more than 70 amino acids at the C-terminus. To identify which part of SIGIRR is required for its interaction with TLRs, three deletion mutants of SIGIRR were generated by Qin *et al*. (2005): ΔN (lacking the extracellular Ig domain, with the deletion of amino acids 1-119), ΔC (lacking the C-terminal tail, with the deletion of amino acids 313-410) and ΔTIR (lacking the TIR domain, with the deletion of amino acids 161-313) (Figure 4A). The deletion of the intracellular TIR domain (ΔTIR) abolished the ability of SIGIRR to interact with the TLR4, MyD88, IRAK and TRAF6 upon LPS stimulation, whereas the deletion of the extracellular Ig domain (ΔN) or the C-terminus (ΔC) did not affect the interaction of SIGIRR with the TLR4 complex. Consistent with the inability of mutant ΔTIR to associate with TLR4 complex, this mutant also failed to inhibit LPS-induced NF-κB activation. These results indicated that the TIR domain of SIGIRR is necessary for both its association with TLR4 and its inhibition of LPS-induced NF-κB activation. A potential model was raised to explain these results (Figure 4B and 4C). Without SIGIRR, TLR4 dimerizes to recognize LPS and recruits MyD88 dimer, resulting in a signaling tetramer. SIGIRR can bind to TLR4 monomer only by its TIR domain, thus preventing TLR4 from dimerizing, binding ligand and recruiting MyD88 dimer. On the other hand, SIGIRR can bind to MyD88 and form a MyD88-SIGIRR heterodimer, interrupting the MyD88 homodimer formation and receptor-adapter interactions. Based on this hypothesis, we modeled 3D structures for the TIR domains of TLR4, 7, MyD88 and SIGIRR and proposed the essential molecular complexes involved in the TLR4 and 7 signaling pathways and the SIGIRR inhibiting process (detailed in section 6.3).

14
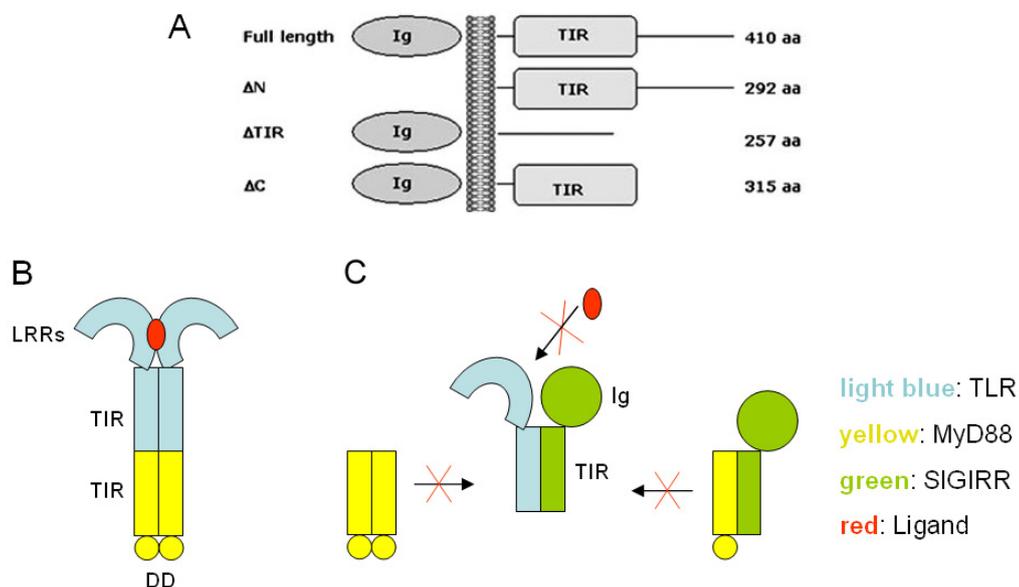
Figure 4. (A) Schematic structures of full-length SIGIRR and its deletion mutants by Qin *et al.* (2005). Protein lengths are indicated in amino acids (aa). Reproduced from [66] with permission from *The Journal of Biological Chemistry*. (B) A model of ligand-induced tetramer formation of TLR and MyD88. Without SIGIRR, TLR forms a homodimer to recognize ligands and recruit the MyD88 homodimer. (C) A model of SIGIRR inhibiting TLR signaling. The TIR domain of SIGIRR may bind to that of TLR, thus preventing TLR from homodimerization and from binding to MyD88. Simultaneously, SIGIRR may bind to the MyD88 monomer to disturb the functional MyD88 homodimer form.

3. Biological databases

3.1 Application of biological databases

A biological database is a library of life sciences information that allows for proper storage, searching and retrieval of biological data. Biological data comes in many different types depending on the research project: texts (textual information and references related to biological data), sequence data (nucleic acid and protein sequences), protein structures (3D coordinates and motifs of protein structures), links (cross-references from one database to other databases providing related resources), numerical data (gene expression data as well as other microarray data), images (2D gel and microscopic images), etc. Owing to the variety of biological data types, the biological knowledge is distributed among many different general and specialized databases. Some databases are large and contain global data collections maintained and kept up to date by the responsible organization, whereas others are small and local and may be only maintained for a limited period of time while a specific project is going on. Up to 2008, more than 1,000 biological databases had been established [69]. Most of these databases have a web interface for data search and are freely accessible. They often use keywords as a common mode of searching. Cross-references of accession numbers help to navigate from one database to another easily. Depending on the nature of information stored, these biological databases can be divided into four main categories as follows:

Nucleotide sequence databases. The International Nucleotide Sequence Database Collaboration (INSDC) consists of a joint effort to collect and disseminate databases containing DNA and RNA sequences. It involves the following computerized databases: the EMBL nucleotide sequence database [70] hosted by the European Bioinformatics Institute (EBI) at the European Molecular Biology Laboratory (EMBL), the GenBank nucleotide sequence database [71] produced at the National Center for Biotechnology Information (NCBI) (USA) and the DNA Data Bank of Japan (DDBJ) [72] located at the National Institute of Genetics of Japan. These databases have been maintained for more than 20 years.

Protein sequence databases. The world's most comprehensive resource on protein information is the universal protein resource (UniProt) [73]. The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information that is essential for modern biological research. UniProt is produced by the UniProt consortium, which consists of groups from the EBI, the Protein Information Resource (PIR) and the Swiss Institute of Bioinformatics (SIB). Each consortium member is heavily involved in protein database maintenance and annotation. EBI and SIB together produced Swiss-Prot/TrEMBL [74], while PIR produced the Protein Sequence Database

(PIR-PSD) [75]. Swiss-Prot is a curated protein sequence database that strives to provide a high level of annotation, a minimal level of redundancy and a high level of integration with other databases. TrEMBL is a computer-annotated supplement of Swiss-Prot that contains all of the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot. The PIR-PSD is a public domain protein sequence database. Moreover, in addition to storing nucleotide sequence data, NCBI stores almost all kinds of biological sequence-related data. Of these, the protein entries stored in the NCBI protein database have been compiled from a variety of sources, including Swiss-Prot, PIR, Protein Data Bank (PDB) [76] and translations from annotated coding regions in GenBank and NCBI Reference Sequence (RefSeq) [77]. Other important protein sequence databases are Pfam (a protein family database of alignments and HMMs) [78], InterPro (an integrated database of predictive protein "signatures") [79], SMART (a web-based tool and database for the study of genetically mobile domains) [80] and others.

Protein structure databases. The Protein Data Bank (PDB) [76] is the single worldwide repository for the processing and distribution of 3D structure data of large molecules of proteins, nucleic acids and complex assemblies. The data in PDB are experimentally determined by X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy and are directly submitted by biologists and biochemists from around the world. The determined structures of molecules are recorded in the PDB file format (.pdb), which is a textual file format describing the 3D coordinates of atoms and containing free-form annotations. PDB files can be displayed by various molecular visualization software tools, such as VMD [81], Jmol (www.jmol.org) and PyMOL (www.pymol.org). The Protein Structure Classification (CATH) database [82] classifies protein structures from the PDB according to a four-level hierarchy. Another protein structure classification database, called the Structural Classification of Proteins (SCOP) [83], describes structural and evolutionary relationships between all known protein structures and provides a number of links to other online resources related to protein structure and to sequence databases in general.

Specialized databases. In addition to the above-mentioned databases, there are many specialized biological databases that hold very detailed information about only one particular subject. For example, STRING: a protein-protein interaction database [84], KEGG: a metabolic pathway database [85], ArrayExpress: a microarray database [86] and Flybase: a *Drosophila* genomes database (a species-specific database) [87].

3.2 Data storage and management in databases

Biological data in the databases can be represented in heterogeneous forms. The form of flat-files structured by a field or value convention was most widely used in the past decades. Nevertheless, flat-files have a number of inevitable shortcomings [88]: i) the text format may not be described uniformly by different databases; ii) it is difficult to

represent complex (hierarchical) data; iii) the meaningful units of information are not represented well, such that it is hard to extract a subunit from a database entry; iv) the assembly of objects into bigger aggregates is difficult. Another absolutely different data storage form, XML (extensible markup language) files, can well overcome these shortcomings. XML was standardized in the 1990s. It has been accepted as an alternative data storage/transfer medium by most fundamental biological databases, such as Swiss-Prot, NCBI and PDB, since the beginning of the 21st century. XML is a set of rules for defining semantic tags that break a document into meaningful units (elements) and access the different units of the document [89]. Tags can be flexibly defined as needed via a document type definition (DTD) with a terse formal syntax. A transformation language, XML stylesheet language (XSL), can transform an XML file into another XML file or into an HTML file for web browsing. Simple examples of the XML, DTD and XSL files are shown in Figure 5.

Figure 5. Simple examples of the XML, DTD and XSL files. (A) An XML file. Tags begin with the less-than character "<" and end with the greater-than character ">". An element consists of a start tag "<…>", possibly followed by text or other complete elements, then followed by an end tag </…>. (B) A DTD file that standardizes the XML file in (A). (C) An XSL file that transforms the XML file in (A) into an HTML file. (D) A web page representation of the HTML file transformed from the XML file in (A) by the XSL file in (B).

As shown in Figure 5, an XML file is a strictly structured "collection of data." In this

sense, an XML file is definitely a database. Nevertheless, it is not a ready-to-use database yet, as a database needs a software package to control the creation, maintenance and retrieval of data. Such a software package is called a database management system (DBMS). A DBMS provides users with query languages to manipulate data and even develop database applications. XPath and XQuery (www.w3.org) are the World Wide Web Consortium-recommended and the most widely used query languages by native XML-DBMSs. Their expressions are well adapted to the structure of XML and can intelligently execute queries across the data from XML files. For example, the XPath expression *./TLRs/TLR[Name= "TLR3"]/Ligand* can directly return the ligand of the entry named "TLR3" from the XML file presented in Figure 5A. Currently, there are a number of open source native XML-DBMSs available, e.g., eXist-db (exist-db.org), 4Suite (4suite.org), myXMLDB (myxmldb.sourceforge.net) and DBDOM (dbdom.sourceforge.net). In this study, the open source native XML-DBMS eXist-db (detailed in section 5.1), which uses XPath/XQuery as query languages, was used to develop two biological databases, TollML and LRRML (detailed in sections 6.1 and 6.2).

4. Protein structure prediction

4.1 Motif identification of TLRs

Since the first TLR sequence (human TLR4) was identified in 1997 by Medzhitov and co-workers [22], only several atomic-resolution crystal structures of TLRs have been determined. They are the structures of the TIR domains of human TLR1, 2 and 10 and the ectodomains of human TLR1-4 and mouse TLR2-4 and 6 [6-9, 32, 33, 36-38]. Nevertheless, more than 2,000 TLR protein sequences have been translated by high-throughput genome sequencing projects in these years. Figure 6 shows the status of data that can be obtained using the search keywords "Toll" and "TLR" from the NCBI protein database. It is clear that the discrepancy between the rate at which novel protein sequences are discovered and the rate at which detailed structural information on proteins can be obtained from X-ray diffraction or NMR spectroscopy will persist for the foreseeable future. Thus, a comprehensive comparative analysis on the sequence level is a useful way to identify and characterize structural motifs of TLRs [30, 90] and to gain insight into how the LRR-based platform is adapted to ligand recognition. Currently, due to the variability of LRR motifs, the indicated number and positions (beginning/end of a repeating unit) of LRR motifs contained in individual TLRs are greatly distinct or missing in established databases. Protein databases that contain information about TLRs, such as Pfam, InterPro, SMART and Swiss-Prot, predict the number and positions of LRRs by various computational methods. Thus, false-negative predictions occur at high frequency. Table 1 lists the numbers of LRRs in human TLR1-4 and mouse TLR6 reported by these databases compared with those counted from the corresponding crystal structures. Moreover, there has been no collection of annotations for structural features within LRR motifs, such as HCS, VS and sequence insertions. In this vein, a specialized database that comprehensively organizes the structural motifs of TLRs is desirable. Such a database could be useful for developing pattern recognition programs, modeling structures and understanding functional mechanisms of TLRs.
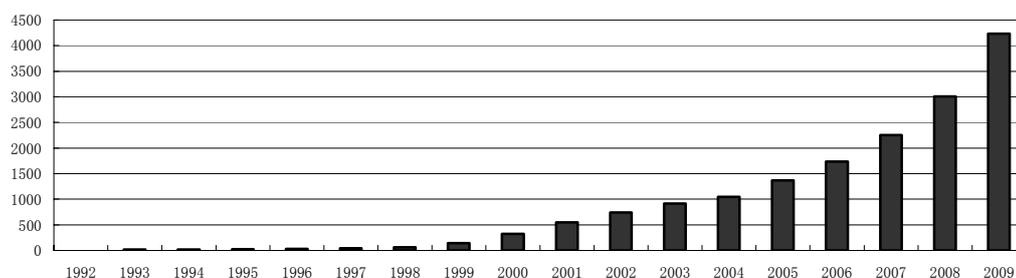
Figure 6. The growth of "Toll" and "TLR" data in the NCBI protein database (results obtained on 28 December 2009).

	TLR1	TLR2	TLR3	TLR4	TLR6
InterPro	4	4	8	7	3
Pfam	4	4	7	7	3
SMART	6	9	19	12	7
Swiss-Prot	8	14	22	21	13
Crystal Structure	21	21	25	23	21

Table 1. Comparison of the numbers of LRRs in human TLR1-4 and mouse TLR6 reported by different databases and counted from crystal structures (results obtained on 24 February 2009). The databases generally underreported many LRRs.

We have developed a TLR structural motif database called TollML (available at http://tollml.lrz.de). It is supported by the eXist-db DBMS. TollML is not only the name of the database but also stands for the Toll-like receptor markup language, aiming at a formal textual representation of TLR structural motif annotations. The TLR structural motifs are represented in three levels by the DTD of TollML: i) signal peptide, ectodomain, transmembrane domain and TIR domain of each TLR; ii) LRRs of each ectodomain; iii) HCS, VS and insertions of each LRR. These three-level structural motifs include comprehensive structural information of TLR ectodomains on the sequence level, and thus they supply an abundant and reliable source to train the LRR prediction methods of different machine-learning algorithms. We developed a position-specific weight matrix (PSWM) of LRR motifs based on the LRR and HCS partitions from TollML. A PSWM is a commonly used representation of patterns in biological sequences. It calculates scores for each alphabet symbol at each position in the pattern. Using the PSWM, the number and positions of LRRs contained in a query sequence can be predicted with both high sensitivity and high specificity. This program is called LRRFinder. It is implemented on the web page http://tollml.lrz.de. As all LRR types possess the same HCS pattern, this method can be extended to predict LRR motifs in other LRR proteins in addition to TLRs (detailed in section 6.2).

4.2 Protein structure prediction theories

Determining protein structure is critical in understanding protein function. However, the experimental determination using X-ray diffraction or NMR spectroscopy is hard work, and many proteins do not yield to these methods. The NMR spectroscopy technique determines the structure in solution and is primarily limited to relatively small proteins (usually smaller than 35 kDa). The X-ray diffraction technique determines the structure in single crystals and thus is not available for proteins that are difficult to crystallize. Moreover, both of these experimental techniques require months or years of work. It is clear that in the near future no experimental structures will be available for most of the known protein sequences. Therefore, the output of experimentally determined structures is lagging far behind the output of protein sequences. In this regard, computational prediction approaches can be used as an

alternative to bridge the gap between sequence determination and structure determination.

All current protein structure prediction methods predict the 3D structure of a protein from its amino acid sequence. A protein is a polymer of amino acids linked together by peptide bonds. Various, mostly non-covalent interactions between amino acids in the linear sequence stabilize a specific folded 3D structure (conformation) for each protein. There are four levels of protein structural organization [91]. Primary protein structure refers to the sequence of amino acids. Secondary protein structure is the local conformation primarily stabilized by hydrogen bonds between atoms of the peptide backbone. The α-helix, β-strand and turn are the most prevalent elements of secondary protein structure. Certain combinations of secondary structures give rise to different motifs, which are found in a variety of proteins and often are associated with specific functions. Tertiary protein structure results from hydrophobic interactions and disulfide bonds that stabilize the folding of the secondary structure into a compact overall arrangement or conformation. Large proteins often contain multiple independently folded regions of tertiary structures with characteristic structural and functional properties. This constitutes a quaternary protein structure. Not all proteins exhibit the quaternary structure. The 3D protein structure prediction in most cases means the prediction of a protein's tertiary structure from its primary structure.

There are three major protein structure prediction algorithms: homology modeling, threading and *ab initio* prediction. Homology modeling, also called comparative modeling, is based on the assumption that similar sequences among evolutionarily related proteins share an overall structural similarity [92]. As a consequence, a 3D model of a protein of interest (target) can be generated based on homologous protein(s) of known structure (template) that share statistically significant sequence identity (usually higher than 30%). The homology modeling procedure consists of several consecutive steps usually repeated iteratively until a satisfactory model is obtained: i) finding suitable template protein(s) related to the target; ii) aligning target and template sequences; iii) building coordinates of the 3D model based on the alignment; iv) refining and evaluating the resulting model. In the case in which homologous proteins of known structure cannot be found, threading methods are used. The amino acid sequence of an unknown structure is matched one-by-one to a database of solved structures (templates) [93]. In each case, a scoring function is used to assess the energetic compatibility of the sequence to the structure without consideration of target-template sequence identity. The structure with the best score is selected as the template for modeling. This method is also known as 3D-1D fold recognition, due to its compatibility analysis between 3D structures and linear protein sequences. Both homology modeling and threading methods use previously solved structures as one of the starting points. In contrast, the *ab initio* algorithm predicts structures based on physical principles rather than on previously solved structures [94]. Indeed, the protein amino acid sequence already contains all information needed to create a correctly folded protein [95]. This makes the *ab initio* method possible for protein structure

prediction. All the energetics involved in the protein folding process are calculated by energy functions to find the structure with the lowest free energy. However, the enumerative energy calculations are extremely time-consuming. Thus, this method can be performed only for small molecules (< 100 amino acids). At present, a variety of protein structure prediction services are available on the Internet, including molecular modeling software and model quality assessment programs (detailed in sections 5.2 and 5.3).

4.3 Structure prediction of SIGIRR

SIGIRR can negatively regulate MyD88-dependent signaling of TLRs through direct interactions of its TIR domain with that of TLR and MyD88. Understanding the inhibitory function of SIGIRR is a key issue towards the development of therapeutic approaches against autoimmune diseases caused by the production of autoantibodies to components of the cell nucleus. Thus, structural models of TIR domain interactions involved in the TLR inhibition processes will give the first insight into the understanding of SIGIRR's structure-function relationship and will provide possible theoretical frameworks for the next experiment design.

The TIR domain is conserved across all TIR domain-containing proteins. The known crystal structures of the TIR domains of human TLR1, 2 and 10 and interleukin-1 receptor accessory protein-like (IL-1RAPL) [32, 33, 96] revealed similar conformations, with a five-stranded β-sheet (βA-βE) surrounded by five α-helices (αA-αE). The β-strands and the α-helices alternate, with loop structures of varying length connecting them. The connecting region between the second β-strand (βB) and the second α-helix (αB), referred to as the BB-loop, was shown to be essential in the TIR-TIR interactions of many TIR domain-containing molecules [32, 33, 40, 97]. A single point mutation, Pro712His, in the BB-loop of mouse TLR4's TIR domain abolished the TLR4 response to LPS [40]. Mutation of the equivalent residue in human TLR2, Pro681His, disrupted signal transduction in response to stimulation by yeast and Gram-positive bacteria [32]. A heptapeptide, which mimics the BB-loop of MyD88's TIR domain, strongly interfered with the dimerization of MyD88 [97]. The crystal structure of human TLR10's TIR homodimer further proved the crucial role of the BB-loop in the TIR-TIR interactions, where the BB-loop and α-helix C of each monomer constituted the major part of the symmetric dimer interface [33].

SIGIRR can protect from lupus via suppressing TLR7 signal transduction. To determine the structural mechanism of SIGIRR's inhibitory function for TLR7, we constructed 3D models of TIR domains of human TLR4, 7, MyD88 and SIGIRR by homology modeling. We chose TLR4 as an additional example to elucidate the mechanisms involved in the negative regulation of the MyD88-dependent TLR signals by SIGIRR, because the different mechanisms of TLR4 recognizing LPS and TLR7 recognizing ssRNA may lead to different structural interactions of receptor with

SIGIRR. In order to explain how SIGIRR disturbs the MyD88-dependent TLR4 and 7 signaling, an understanding of the interaction mode of receptor and adaptor complexes without the presence of SIGIRR is indispensable. We thus performed a protein-protein docking study to construct models of TLR4-TLR4, TLR7-TLR7, MyD88-MyD88, TLR4-TLR4-MyD88-MyD88 and TLR7-TLR7-MyD88-MyD88 complexes. Previously, Dunne and co-workers modeled the TLR4-MyD88 heterodimer using TLR4 and MyD88 monomers. This monomer-to-monomer model, however, may not fully reflect the molecular interactions. Our model of the receptor dimer docking to the MyD88 dimer provided additional information for the structural interpretation. We further constructed the models of the TLR4-SIGIRR, TLR7-SIGIRR and MyD88-SIGIRR complexes that could interfere with the proper signaling complexes. All docking results indicated the importance of the BB-loop in the TIR-TIR interactions. We suggested that SIGIRR might exert its inhibitory effect through blocking the molecular interface of TLR4, TLR7 and the MyD88 adaptor mainly via its BB-loop (detailed in section 6.3).

The TIR domain of SIGIRR extends from that of other TIR domains at the C-terminal end by a 73-amino acid-long stretch (C-terminal tail). A similarity search against PDB returned no homologue structure to serve as a template. In this case, the threading method was used to determine a template structure. The finally selected template was the N-terminal domain of N-ethylmaleimide sensitive factor (PDB code: 1QCS).

4.4 Homology modeling of TLR ectodomains

The TLR ectodomain consists of consecutive LRR motifs. The structure of LRR motifs and their arrangement in repetitive stretches of variable lengths generate a versatile and highly evolvable framework for the binding of diverse proteins and non-protein ligands [90]. Up to the present, more than 130 crystal structures of LRR-containing proteins have been determined. They show that all LRR domains adopt an arc or horseshoe shape [30]. This knowledge has increased our ability to use known LRR structures as templates to model TLR ectodomains with unknown structures. In practice, due to the various repeat numbers and the distinct arrangements of LRRs in the TLR ectodomains, a proper full-length template with a sufficiently high sequence identity (≥30%, as required by homology modeling) to the TLR ectodomains is often missing. This limitation can be overcome by disassembling crystal structures into individual LRR building blocks and then reassembling the blocks to build a TLR ectodomain. In this approach, first the most similar (on the sequence level) structurally known LRR is found for each LRR in the target sequence as a local template. Second, the sequences of all local templates are combined to generate a multiple sequence alignment for the entire target sequence. Last, the templates together with the sequence alignment are input into an automatic model construction program to obtain a model as with the standard homology modeling. Such an approach requires a comprehensive database of

LRR structures to extract adequate template candidates. As yet, the individual 3D LRR structures are not directly available from the current databases, and there are only a few detailed annotations for them. Additional information such as sequence insertions and LRR types is missing. In order to consolidate this information and to provide a source for homology modeling and the structural analysis of LRR proteins, we have developed LRRML, a database and an extensible markup language description of LRR structures. It is available at http://lrrml.lrz.de. LRRs stored in LRRML were extracted from all known LRR-containing protein structures from PDB. A similarity search tool was implemented for the database. It returns the structures of similar LRRs with ranking for a query LRR sequence. After a successful test-case validation of the LRR assembly method, we constructed models of the human TLR5-10 and mouse TLR11-13 ectodomains with this method. These models can be used as the first passes for a computational simulation of ligand docking (detailed in section 6.6).

A recent study showed that the ectodomains of TLR7 and 9 are cleaved in the endolysosome to recognize ligands [98]. Only the C-terminal cleaved forms of them can recruit MyD88 on activation. This cleavage process may also happen to TLR8, because of the high homology with TLR7-9. To better understand the ligand-binding mechanisms of TLR7-9, we also developed structural models of their ectodomains in cleaved form. On the basis of the resulting models, we suggested their potential ligand-binding sites and inferred possible configurations of the receptor-ligand 2:1 complexes (detailed in section 6.5).

5. Methods

5.1 Database construction

This section describes the database management system (DBMS) supporting the TollML and LRRML databases. All programming languages used to develop the web applications of both databases are also characterized in brief.

eXist-db. eXist-db is an open source native XML-DBMS written in Java (a class-based, object-oriented programming language). The eXist-db distribution 1.1.1 was deployed inside a servlet-engine on a Linux system. eXist-db supports many web technology standards, making it an excellent platform for developing online applications. It provides storage of XML documents in hierarchical directories and efficient query processing, including keyword searches, queries on proximity of search terms or regular expressions through the query languages of XPath/XQuery.

Programming languages. All data in the two databases, TollML and LRRML, are stored in XML (www.w3.org) documents according to the syntax in the corresponding document type definition (DTD) file. The HTML display of each database entry on the web interface is directly converted from the original XML documents by the pre-designed XML Stylesheet (XSL). XPath/XQuery serve as the query languages of the databases. Complicated calculating programs for the web applications are written in Java, Perl (an interpreted, dynamic programming language), Python (an interpreted, interactive, object-oriented, extensible programming language) and R (a programming language specialized for statistical computing and graphics).

5.2 Protein structure prediction

This section describes the programs used for generating protein structural models, including protein tertiary structures and secondary structures. At present, numerous software tools for protein structure prediction have been developed. The prediction protocols from alternative software tools, however, can produce dramatically different models, even when using the exact same input target-template alignments. We then chose the most commonly accepted and long-tested software tools for our protein structure predictions.

MODELLER. MODELLER [99] is one of the most widely used computer programs for homology modeling of protein 3D structures. In the simplest case, the input is an alignment of a sequence to be modeled with the sequence(s) of template structure(s) and the atomic coordinates of the template(s). MODELLER then automatically calculates a model containing all non-hydrogen atoms. It implements homology modeling by the satisfaction of spatial restraints, by which a set of geometrical criteria

are used to create a probability density function for the location of each atom in the protein.

ModLoop. ModLoop [100] is a web tool for automated modeling of loops in protein structures. The input is the atomic coordinates of a protein structure in PDB format and the specifications of the starting and ending residues of the one or more segments to be modeled, containing no more than 20 residues in total. The output is the coordinates of the non-hydrogen atoms in the modeled segments. The modeling relies on a protocol consisting of conjugate gradient minimization and molecular dynamics simulation.

THREADER. THREADER [101] is a protein fold recognition method, whereby a query sequence is fitted (threaded) directly onto the carbon backbone coordinates of non-redundant protein structures derived from PDB. The degree of compatibility between the sequence and each proposed structure is evaluated by means of a set of empirical potentials derived from proteins of known structure. The specific aspect of this approach is that the matching of sequence to backbone coordinates is performed in full 3D space, incorporating specific pair interactions explicitly.

pGenTHREADER. pGenTHREADER [102] is a sequence profile-based structural template recognition method for protein homology modeling. It calculates sequence profiles from an input sequence and generates profile-profile alignments of the input and template sequences. The algorithm of the alignments linearly combines secondary structure specific gap-penalties, pair potentials and solvation potentials. The output is the PDB structures, which serve as candidate templates ranked by target-template similarities.

PSIPRED. PSIPRED [103] is a protein secondary structure prediction method. It performs a PSI-BLAST search [104] for a query sequence and then feeds the resulting profiles of the query through two consecutive feed-forward neural networks (a machine-learning algorithm) in order to predict secondary structure.

5.3 Protein model quality assessment

Protein model quality assessment programs (MQAPs) receive as input a 3D model/structure in PDB format and produce as output a real number representing the quality of the model. In this section, several popular and up-to-date MQAPs used in this study are listed.

ProQ. ProQ [105] is a neural network-based method to predict the quality of a protein model. It extracts structural features from an input model, such as the frequency of atom-atom contacts, and measures them either by LGscore [106] (for long target proteins) or MaxSub [107] (for short ones).

MetaMQAP. MetaMQAP [108] is a very up-to-date protein model evaluation tool. It

is a meta-predictor based on a multivariate regression model that uses scores of eight previously published model evaluation methods. MetaMQAP predicts the absolute deviation (in Ångströms) of individual Cα atoms between the model and the unknown true structure as well as global deviations (expressed as root mean square deviation).

ModFOLD. ModFOLD [109] combines scores obtained from the ModSSEA method [110], the MODCHECK method and the two ProQ methods using a neural network. It can provide (i) a single score and a P-value, which represents a quantitative measure of the confidence in a model related to the predicted quality of a single protein model; (ii) rankings for multiple models for the same protein target according to predicted model quality; (iii) predictions of the local quality (per-residue errors) within multiple models.

PROCHECK. PROCHECK [111] provides a check on the stereochemistry of a 3D protein structure or model. Its outputs comprise a number of plots, such as the Ramachandran plot [112], and a comprehensive residue-by-residue listing. These give an assessment of the overall quality of the structure as compared with well-refined structures of the same resolution and also highlight regions that may need further investigation.

5.4 Protein-protein docking

This section describes two excellent-performance rigid-body protein-protein docking tools. Rigid-body docking is known as when a conformational change does not occur within the components at any stage of complex generation.

GRAMM-X. GRAMM-X [113] uses the correlation technique Fast Fourier Transformation (FFT) for the global search of the best rigid-body conformations. During the FFT search, the protein surface representation is smoothed to account for possible conformational changes upon binding. The search results are further refined by optimization in continuous coordinates and rescoring with several knowledge-based potential terms.

ZDOCK. ZDOCK [114] is also a FFT search-based rigid-body docking method. An important feature of ZDOCK is that it employs a powerful scoring function that integrates, pairwise, the shape complementarity, desolvation and electrostatics, during the FFT search.

5.5 Protein structure visualization and analysis

The predicted protein 3D models or protein-protein interacting complexes need further examination through various visualization and analysis tools. This section describes several tools used in our studies on primary (amino acid sequence), tertiary (3D conformation) and quaternary (complex model) protein structures. Figures presenting

our models in this thesis were also rendered with these tools.

MUSCLE. MUSCLE [115] stands for MUltiple Sequence Comparison by Log-Expectation. It is a computer program for creating multiple alignments of protein sequences with high speed and accuracy, especially for large numbers of sequences. The input is a set of protein sequences, and the output is an alignment generated in a table form that exhibits the evolutionary relationship between the input sequences. In this table, the rows show the input sequences and the columns show the positions in the alignment. Gaps are inserted between the residues in the sequences so that identical or similar amino acids are aligned in successive columns.

Jalview. Jalview [116] was used to view and edit multiple sequence alignments. It allows the identification of functional residues by the comparison of subgroups of sequences arranged on a cluster tree. A number of color schemes were predefined to color alignments or groups.

VMD. VMD [81] is a molecular graphics program designed for the display and analysis of biopolymers such as proteins and nucleic acids. Molecules are displayed as one or more "representations," in which each representation embodies a particular rendering method and coloring scheme for a selected subset of atoms. It contains a set of tools for interactive problem solving in structural biology.

SPDBV. SPDBV [117] is a molecular graphics program that provides a user-friendly interface and allows the analysis of several proteins simultaneously. A very useful function of SPDBV is that the protein structures can be superimposed to deduce structural alignments and compare their active sites or any other relevant parts. Amino acid mutations, H-bonds, angles and distances between atoms can be obtained using the intuitive graphic and menu interface.

Jmol. Jmol (www.jmol.org) is a free, open source molecule viewer for biochemical structures in 3D. Its interactive web browser applet can be integrated into web pages for high-performance online 3D rendering in web-accessible databases.

SuperPose. SuperPose [118] is a web server for both pairwise and multiple protein structure superpositions using a modified quaternion eigenvalue approach. SuperPose generates sequence alignments, structure alignments, PDB coordinates and root mean square deviation statistics, as well as difference distance plots and images of the superimposed molecules.

PISA. PISA [119] is an interactive tool for the exploration of macromolecular (protein, DNA/RNA and ligand) interfaces, prediction of probable quaternary structures, database searches of structurally similar interfaces and assemblies and searches on various assembly and PDB entry parameters.

6. Results (extended abstracts of manuscripts)

6.1 Paper 1: TollML: a database of Toll-like receptor structural motifs

(Manuscript, see appendix at page 49)

Comprehensive and systematic structural motif annotations of Toll-like receptors (TLRs) on the sequence level will be useful for structural analysis and homology modeling of TLRs. Such annotations are not available in any current databases. In this regard, we have developed the TollML database, which is supported by the native extensible markup language (XML) database management system eXist-db. TollML integrated all of the TLR sequencing data from the NCBI protein database. Entries were first divided into TLR families (TLR1-23) and then were semi-automatically subdivided into three levels of structural motif categories: i) signal peptide (SP), ectodomain (ECD), transmembrane domain (TD) and Toll/interleukin-1 receptor (TIR) domain of each TLR; ii) LRRs of each ECD; iii) highly conserved segment (HCS), variable segment (VS) and insertions of each LRR. These categories can be quickly searched using an easy-to-use web interface and are dynamically displayed by graphics. Additionally, all entries have hyperlinks to various sources including NCBI, Swiss-Prot, PDB, LRRML and PubMed to provide broad external information for users. TollML also stands for Toll-like receptor markup language, a formal textual representation of TLR structural motif annotations. A database construction pipeline is shown in Figure 7.
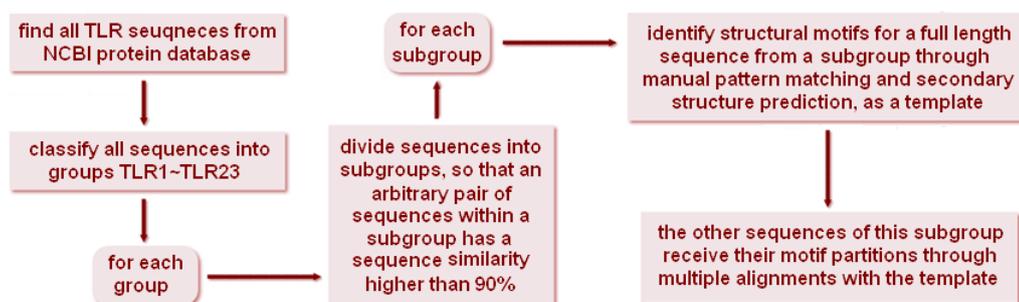
Figure 7. The TollML database construction pipeline.

The TollML release 3.1 contains 2,572 TLR entries divided into 23 families (entry distribution shown in Table 2). Of these, 2,350 entries contain an ectodomain and thus received motif annotations that correspond to the second and third levels of motif categorization. The other entries contain only a TIR domain with or without a transmembrane domain. A total of 46,720 LRR motifs were recognized from the ectodomain-containing TLRs.

TLR	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	18	19	20-23	Total
Mammalian	85	146	108	443	132	106	109	101	124	72	10	11	9	0	0	0	0	0	0	1,456
Non-mamm.	356	276	84	77	58	3	102	11	30	0	0	1	4	6	42	2	7	5	52	1,116
Total	441	422	192	520	190	109	211	112	154	72	10	12	13	6	42	2	7	5	52	2,572

Table 2. Entry distribution over TLR families for mammalian/non-mammalian groups.

TollML supplies a large and reliable source to train the LRR prediction methods of machine-learning algorithms. We developed a 20×12 position-specific weight matrix (PSWM) for LRR motif predictions using the LRR partitions (the second level of structural motifs) and the HCS partitions (the third level) from TollML. The 20 rows in the PSWM correspond to the 20 amino acids, while the twelve columns correspond to the twelve positions in the LRR HCS consensus sequence. Thus, a matrix element denotes the frequency probability of a certain amino acid occurring at a certain position in the LRR HCS. With this PSWM, each position of a query sequence will be scored to judge whether it may be the beginning position of an LRR. Finally, the number and positions of LRRs contained in the query sequence are returned. This automatic LRR prediction program is called LRRFinder. As LRR motifs of all LRR proteins possess the same HCS pattern, this method can be extended to predict LRR motifs in other LRR proteins in addition to TLRs. The TollML database and LRRFinder program are available at http://tollml.lrz.de.

6.2 Paper 2: LRRML: a conformational database and an XML description of leucine-rich repeats (LRRs)

(Manuscript, see appendix at page 57)

Leucine-rich repeats (LRRs) are present in more than 6,000 proteins. They are found in organisms ranging from viruses to eukaryotes and play an important role in protein-ligand interactions. To date, more than 100 crystal structures of LRR proteins have been determined. This knowledge has increased our ability to use individual LRR structures as building blocks to model Toll-like receptor (TLR) ectodomains or other LRR proteins. Because the individual 3D LRR structures are not directly available from the established databases and there are only a few detailed annotations for them, a conformational LRR database that is useful for homology modeling of LRR proteins is desirable. We have developed LRRML, a conformational database of LRRs. The LRRML release 0.2 contains 1,261 individual LRR structures, which were identified from 112 PDB protein structures. All LRR entries were provided with three groups of manual annotations: i) classification into eight LRR types; ii) partition into a highly conserved segment (HCS) and a variable segment (VS); iii) labeling of insertion segments (IS) longer than three amino acids according to LRR consensus sequences. In addition, an extensible markup language (XML) textual structure was defined to exchange and store the LRR conformation. A database construction pipeline is shown

in Figure 8.
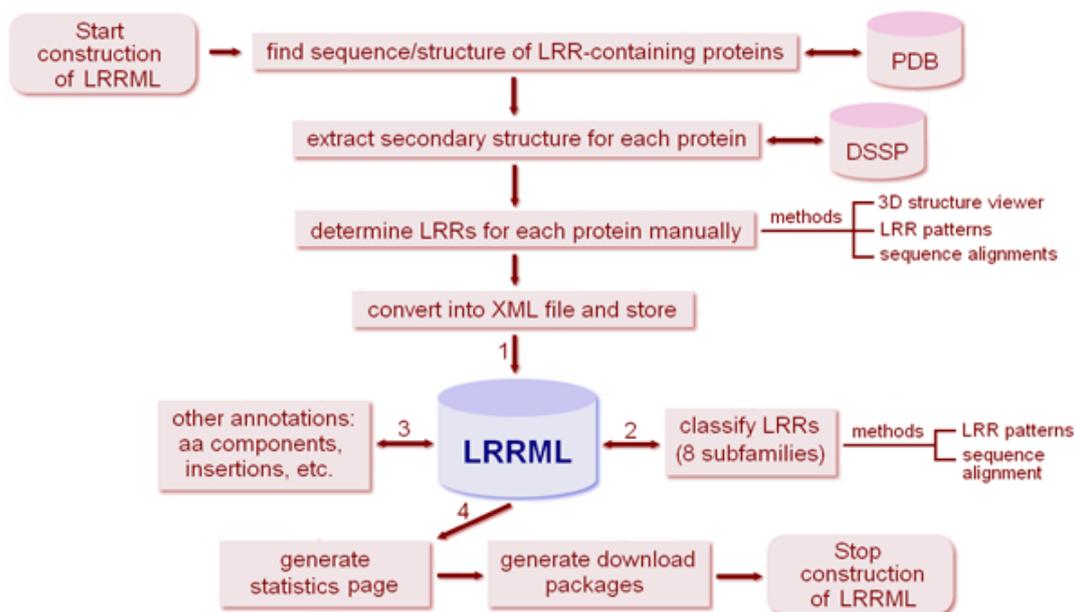
Figure 8. The LRRML database construction pipeline.

To demonstrate the capabilities of the database, we modeled the mouse TLR3 ectodomain as a test-case by combining multiple LRR templates obtained from LRRML. A comparison of the model with the corresponding crystal structure (PDB code: 3CIG) showed a very good structural agreement. In conclusion, LRRML provides a source for the homology modeling and structural analysis of LRR proteins. This database is available at http://lrrml.lrz.de.

6.3 Paper 3: Inhibition of Toll-like receptors TLR4 and 7 signaling pathways by SIGIRR: a computational approach

(Manuscript, see appendix at page 67)

Single immunoglobulin interleukin-1 receptor-related molecule (SIGIRR) has therapeutic potential towards systemic lupus erythematosus because of its inhibitory function in Toll-like receptor (TLR) signaling. So far, the mechanism of structural interactions between SIGIRR, TLRs and adaptor molecules is unclear. To develop a working hypothesis for this interaction, we constructed 3D models for the Toll/interleukin-1 receptor (TIR) domains of TLR4, 7, MyD88 and SIGIRR based on computational modeling (Figure 9). Through protein-protein docking analysis using GRAMM-X and ZDOCK (detailed in section 5.4), we developed models of essential complexes involved in the TLR4 and 7 signaling and the SIGIRR inhibiting processes. Receptor activation would trigger the formation of TLR4 and 7 TIR dimers, recruiting MyD88 TIR dimers and resulting in a signaling tetramer. Model predictions including SIGIRR revealed that SIGIRR binds to TLR4 and 7 by occupying their self-interacting

sites. On the other hand, the MyD88-SIGIRR dimer shows a resemblance to the MyD88 homodimer. That is, SIGIRR replaces a MyD88 monomer, interrupting the MyD88 homodimer formation. In all cases the BB-loop of SIGIRR plays a key role in binding.
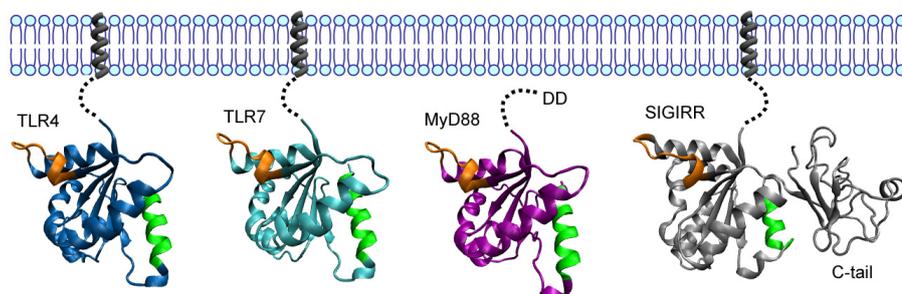
Figure 9. 3D models of TIR domains of TLR4, 7, MyD88 and SIGIRR. The BB-loops and αE regions are highlighted in orange and green respectively.

Both of the tetramers (TLR4 dimer-MyD88 dimer and TLR7 dimer-MyD88 dimer) exposed in our study demonstrated that the stimulus-induced dimerization of TIR domains creates a new negatively charged molecular pocket for the binding of the positively charged αE of the MyD88 adaptor (Figure 9). In the presence of SIGIRR, the proper shape and electric environment of the MyD88-binding pocket are completely disturbed. Remarkably, TLR4 and 7 possess a more extensive molecular interface with SIGIRR (heterodimer) than with themselves (homodimer). These observations highlighted the strong molecular affinity of SIGIRR as an inhibitor.
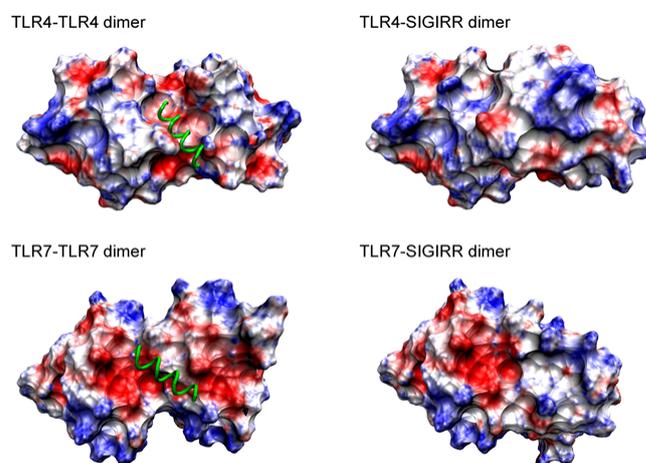
Figure 10. Surface charge distribution of TIR dimers. Both the TLR4-TLR4 and TLR7-TLR7 dimers generate a negatively charged (red) pocket adjacent to their dimer interface to hold the positively charged (blue) αE of MyD88. The incorporation of SIGIRR completely disturbed the proper shape and electric environment of the pocket. The αE is represented by a green tube, and the other part of MyD88 is omitted for better view.

In summary, we proposed a residue-detailed structural framework of SIGIRR inhibiting the TLR4 and 7 signaling pathways. These results can facilitate efforts to design further

mutagenesis experiments to clarity the regulatory role of SIGIRR in inflammatory and innate immune responses.

6.4 Paper 4: Lack of SIGIRR/TIR8 aggravates hydrocarbon oil-induced systemic lupus

(Manuscript, see appendix at page 76)

Multiple genetic factors contribute to the clinical variability of spontaneous systemic lupus erythematosus (SLE) but their role in drug-induced SLE remains largely unknown. Hydrocarbon oil-induced SLE depends on mesothelial cell apoptosis and Toll-like receptor TLR7-mediated induction of type I interferons. Hence, we hypothesized the single immunoglobulin interleukin-1 receptor-related molecule (SIGIRR, also known as TIR8), an endogenous TLR inhibitor, prevents oil-induced SLE. *Sigirr*-deficient dendritic cells expressed higher TLR7 mRNA levels and TLR7 activation resulted in increased interleukin-12 production in vitro. In vivo, lack of SIGIRR increased surface CD40 expression on spleen CD11c+ dendritic cells and MX-1, TNF, interleukin-12, BAFF and BCL-2 mRNA expression 6 months after pristane injection. Spleen cell counts of CD4-/CD8-"autoreactive" T cells and B220+ B cells were also increased in *Sigirr*-deficient mice. Serum autoantibody analysis revealed that *Sigirr*-deficiency specifically enhanced the production of rheumatoid factor (from 4 months of age) and anti-snRNP IgG (from 5 months of age) while anti-Smith IgG or anti-dsDNA IgG were independent of the *Sigirr* genotype. This effect was sufficient to significantly aggravate lupus nephritis in *Sigirr*-deficient mice. Structural model prediction identified the BB-loop of SIGIRR's intracellular TIR domain to interact with TLR7 and MyD88. BB-loop deletion was sufficient to completely abrogate SIGIRR's inhibitory effect on TLR7 signaling. Thus, SIGIRR/TIR8 protects from hydrocarbon oil-induced lupus via suppressing the TLR7-mediated activation of dendritic cells most likely via its intracellular BB-loop.

6.5 Paper 5: Homology modeling of human Toll-like receptors TLR7, 8 and 9 ligand-binding domains

(Manuscript, see appendix at page 89)

The Toll-like receptors TLR7, 8 and 9 compose a family of intracellularly localized TLRs that signal in response to pathogen-derived nucleic acids. After bacteria and viruses are internalized to the cell endosome, nucleic acids might be released to be recognized by these TLRs. So far, there are no crystallographic structures for TLR7-9. For this reason, their ligand-binding mechanisms are poorly understood. To enable first predictions of the receptor-ligand interaction sites, we developed 3D structures for the leucine-rich repeat ectodomains of human TLR7-9 based on homology modeling. The

predicted results can guide experimentalists to design site-directed mutation experiments to unravel the ligand-binding mechanisms of these receptors. To achieve a high sequence similarity between targets and templates, structural segments from all known TLR ectodomain structures (human TLR1-4 and mouse TLR3-4) were used as candidate templates for the modeling. The resulting models support previously reported essential ligand-binding residues. They also provide a basis to identify three potential receptor dimerization mechanisms. Additionally, potential ligand-binding residues are identified using combined procedures. We suggest further investigations of these residues through mutation experiments. Our modeling approach can be extended to other members of the TLR family or other repetitive proteins.

6.6 Paper 6: A leucine-rich repeat assembly approach for homology modeling of human TLR5-10 and mouse TLR11-13 ectodomains

(Manuscript, see appendix at page 98)

So far, 13 groups of mammalian Toll-like receptors (TLRs) have been identified. Most TLRs have been shown to recognize pathogen-associated molecular patterns from a wide range of invading agents and initiate both innate and adaptive immune responses. The TLR ectodomains are composed of varying numbers and types of leucine-rich repeats (LRRs). Because the crystal structures are currently missing for most TLR ligand-binding ectodomains, computational modeling enables the first predictions of their 3D structures. The determined crystal structures of TLR ectodomains can provide structural templates for homology modeling of other TLR ectodomains. However, the quality of the predicted models that are generated from full-length templates can be limited due to low sequence identity between the target and templates. To obtain better templates for modeling, we have developed a multiple LRR templates assembly approach. Individual LRR templates that are locally optimal for the target sequence are assembled into multiple templates. This method was validated through the comparison of a predicted model with the crystal structure of mouse TLR3. With this method we also constructed ectodomain models of human TLR5-10 and mouse TLR11-13 that can be used as the first passes for a computational simulation of ligand docking or to design mutation experiments. This template assembly approach can be extended to other repetitive proteins.

7. Conclusions

The topics of this thesis include the construction of databases with full-scale annotations for TLR structural motifs (TollML) and LRR conformations (LRRML) to facilitate computational modeling of the TLR structure and the computational investigation of the molecular interactions between TLRs and their inhibitor, SIGIRR.

TollML includes all known TLR sequences extracted from the NCBI protein database. Three levels of TLR structural motifs were identified and annotated by a semi-automatic procedure. The TollML release 3.1 contains 2,572 TLR entries separated into 23 families. Of these, 2,350 entries contain ectodomains, from which a total of 46,720 LRR motifs were recognized. With these LRR partitions, an automatic LRR motif prediction program named LRRFinder was developed. A five-fold cross-validation against all TollML entries indicated that the sensitivity and specificity of this program are both greater than 93%. The TollML database and the LRRFinder program are freely available at http://tollml.lrz.de. Another database is a conformational LRR database named LRRML, which contains individual 3D LRR structures with manual structural annotations. The LRRML release 0.2 contains 1,261 LRR entries identified from 112 PDB structures. A total of 548 of the 1,261 LRRs are distinct on sequence level, indicating that different molecules can share identical LRRs. This fact enhances the confidence in the computational modeling of LRR proteins using individual LRR building blocks. To simplify the homology modeling, a similarity search tool was implemented on the LRRML web page. It returns the structures of the most similar LRRs for a query LRR sequence. The LRRML database is freely available at http://lrrml.lrz.de. The TollML and LRRML stand for Toll-like receptor Markup Language and LRR Markup Language, respectively. They standardize the representation of TLR and LRR annotations for convenient information exchanges in the future. With the help of these two databases, we constructed the 3D structural models for human/mouse TLR5-13 full-length ectodomains, human TLR7-9 cleaved ectodomains and the TIR domains of human TLR4, 7, MyD88 and SIGIRR.

Through protein-protein docking analysis, we developed models of essential complexes involved in the TLR4 and 7 signaling and the SIGIRR inhibiting processes. The results could be assembled to derive a working hypothesis for the receptor signaling transductions and the SIGIRR inhibiting mode. Receptor activation would trigger the formation of TLR4- (or TLR7-) TIR dimers and the recruitment of MyD88 dimers, resulting in a signaling tetramer. SIGIRR could bind to TLR4 (or TLR7) by occupying its self-interacting sites. On the other hand, SIGIRR would replace a MyD88 monomer, interrupting the functional MyD88 homodimer. In all cases, the BB-loop of SIGIRR plays a key role in binding. Mutation experiments proved that BB-loop deletion was sufficient to completely abrogate SIGIRR's inhibitory effect on TLR7 signaling. Notably, the molecular interfaces of TLR4, 7 and MyD88 were more extensive with SIGIRR (heterodimer) than with themselves (homodimer). These

observations highlight the strong molecular affinity of SIGIRR as an inhibitor.

Based on the structural models of the cleaved TLR7-9 ectodomains, we identified potential ligand-binding residues using a combined procedure. Only highly conserved, non-negatively charged, surface residues that were positively predicted by at least two docking programs were considered as potential ligand-binding residues. Finally, we suggested three possible receptor dimerization schemes that require different minimum ligand sizes.

The TLR signaling pathways are an intricate physiological system that encompasses numerous and diverse sequential molecular interactions. The lack of structural knowledge is in part responsible for our incomplete understanding of the basis of receptor specificity and the activation mechanisms. For the receptors with solved crystal structures in a ligand-binding form, further work is required to establish how the same TLR ligand-binding domains interact with the structurally diverse ligand variants from different viruses or bacteria. Moreover, with the continual discovery of new components or regulatory functions of known components involved in TLR signaling, more and more molecular interactions need to be investigated to provide a comprehensive map of the TLR signaling network. Here, bioinformatics efforts will facilitate the understanding of all these interactions from a structural viewpoint in concert with experimental techniques. This will also be highly valuable for the future development of innovative therapies for the manipulation of infectious diseases and autoimmune diseases.

References

1.   Brusic V: **From immunoinformatics to immunomics**. *J Bioinform Comput Biol* 2003, **1**(1):179-181.

2.   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.

3.   Takeda K, Akira S: **Toll-like receptors in innate immunity**. *Int Immunol* 2005, **17**(1):1-14.

4.   Wald D, Qin J, Zhao Z, Qian Y, Naramura M, Tian L, Towne J, Sims JE, Stark GR, Li X: **SIGIRR, a negative regulator of Toll-like receptor-interleukin 1 receptor signaling**. *Nat Immunol* 2003, **4**(9):920-927.

5.   Lech M, Kulkarni OP, Pfeiffer S, Savarese E, Krug A, Garlanda C, Mantovani A, Anders HJ: **Tir8/Sigirr prevents murine lupus by suppressing the immunostimulatory effects of lupus autoantigens**. *J Exp Med* 2008, **205**(8):1879-1888.

6.   Jin MS, Kim SE, Heo JY, Lee ME, Kim HM, Paik SG, Lee H, Lee JO: **Crystal structure of the TLR1-TLR2 heterodimer induced by binding of a tri-acylated lipopeptide**. *Cell* 2007, **130**(6):1071-1082.

7.   Kim HM, Park BS, Kim JI, Kim SE, Lee J, Oh SC, Enkhbayar P, Matsushima N, Lee H, Yoo OJ *et al*: **Crystal structure of the TLR4-MD-2 complex with bound endotoxin antagonist Eritoran**. *Cell* 2007, **130**(5):906-917.

8.   Park BS, Song DH, Kim HM, Choi BS, Lee H, Lee JO: **The structural basis of lipopolysaccharide recognition by the TLR4-MD-2 complex**. *Nature* 2009, **458**(7242):1191-1195.

9.   Kang JY, Nan X, Jin MS, Youn SJ, Ryu YH, Mah S, Han SH, Lee H, Paik SG, Lee JO: **Recognition of lipopeptide patterns by Toll-like receptor 2-Toll-like receptor 6 heterodimer**. *Immunity* 2009, **31**(6):873-884.

10.   Sepkowitz KA: **AIDS--the first 20 years**. *N Engl J Med* 2001, **344**(23):1764-1772.

11.   Fauci AS, Braunwald E, Kasper DL, Hauser SL, Longo DL, Jameson JL, Loscalzo J: **Harrison's Principles of Internal Medicine**, 17 edn: McGraw-Hill Professional; 2008.

12.   Pazdur R, Wagman LD, Camphausen KA, Hoskins WJ: **Cancer Management: A Multidisciplinary Approach**, 11 edn: CMPMedica; 2008.

13.   Kang I, Park SH: **Infectious complications in SLE after immunosuppressive therapies**. *Curr Opin Rheumatol* 2003, **15**(5):528-534.

14.   Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P: **Molecular Biology of the Cell**, 5 edn: Garland Science; 2008.

15.   Litman GW, Cannon JP, Dishaw LJ: **Reconstructing immune phylogeny: new perspectives**. *Nat Rev Immunol* 2005, **5**(11):866-879.

16.   Janeway CA, Jr., Medzhitov R: **Innate immune recognition**. *Annu Rev Immunol* 2002, **20**:197-216.

17.   Kawai T, Akira S: **Innate immune recognition of viral infection**. *Nat Immunol* 2006, **7**(2):131-137.

18.   Anderson KV, Bokla L, Nusslein-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.

19.   Anderson KV, Nusslein-Volhard C: **Information for the dorsal--ventral pattern of the Drosophila embryo is stored as maternal mRNA**. *Nature* 1984, **311**(5983):223-227.

20.   Anderson KV, Jurgens G, Nusslein-Volhard C: **Establishment of dorsal-ventral polarity in the Drosophila embryo: genetic studies on the role of the Toll gene product**. *Cell* 1985, **42**(3):779-789.

21.   Lemaitre B, Nicolas E, Michaut L, Reichhart JM, Hoffmann JA: **The dorsoventral regulatory gene cassette spatzle/Toll/cactus controls the potent antifungal response in Drosophila adults**. *Cell* 1996, **86**(6):973-983.

22.   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.

23.   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.

24.   Brodsky I, Medzhitov R: **Two modes of ligand recognition by TLRs**. *Cell* 2007, **130**(6):979-981.

25.   Gong J, Wei T, Zhang N, Jamitzky F, Heckl WM, Rossle SC, Stark RW: **TollML: a database of toll-like receptor structural motifs**. *J Mol Model* 2010, **16**(7):1283-1289.

26.   Barton GM, Kagan JC, Medzhitov R: **Intracellular localization of Toll-like receptor 9 prevents recognition of self DNA but facilitates access to viral DNA**. *Nat Immunol* 2006, **7**(1):49-56.

27.   Kobe B, Kajava AV: **The leucine-rich repeat as a protein recognition motif**. *Curr Opin Struct Biol* 2001, **11**(6):725-732.

28.   Kajava AV: **Structural diversity of leucine-rich repeat proteins**. *J Mol Biol* 1998, **277**(3):519-527.

29.   Matsushima N, Ohyanagi T, Tanaka T, Kretsinger RH: **Super-motifs and evolution of tandem leucine-rich repeats within the small proteoglycans--biglycan, decorin, lumican, fibromodulin, PRELP, keratocan, osteoadherin, epiphycan, and osteoglycin**. *Proteins* 2000, **38**(2):210-225.

30.   Matsushima N, Tanaka T, Enkhbayar P, Mikami T, Taga M, Yamada K, Kuroki Y: **Comparative sequence analysis of leucine-rich repeats (LRRs) within vertebrate toll-like receptors**. *BMC Genomics* 2007, **8**:124.

31.   Roach JC, Glusman G, Rowen L, Kaur A, Purcell MK, Smith KD, Hood LE, Aderem A: **The evolution of vertebrate Toll-like receptors**. *Proc Natl Acad Sci U S A* 2005, **102**(27):9577-9582.

32.   Xu Y, Tao X, Shen B, Horng T, Medzhitov R, Manley JL, Tong L: **Structural basis for signal transduction by the Toll/interleukin-1 receptor domains**. *Nature* 2000, **408**(6808):111-115.

33.   Nyman T, Stenmark P, Flodin S, Johansson I, Hammarstrom M, Nordlund P: **The crystal structure of the human toll-like receptor 10 cytoplasmic domain reveals a putative signaling dimer**. *J Biol Chem* 2008, **283**(18):11861-11865.

34.   Takeuchi O, Sato S, Horiuchi T, Hoshino K, Takeda K, Dong Z, Modlin RL, Akira S: **Cutting edge: role of Toll-like receptor 1 in mediating immune**

response to microbial lipoproteins. *J Immunol* 2002, **169**(1):10-14.

35. Takeuchi O, Kawai T, Muhlradt PF, Morr M, Radolf JD, Zychlinsky A, Takeda K, Akira S: **Discrimination of bacterial lipoproteins by Toll-like receptor 6**. *Int Immunol* 2001, **13**(7):933-940.

36. Bell JK, Botos I, Hall PR, Askins J, Shiloach J, Segal DM, Davies DR: **The molecular structure of the Toll-like receptor 3 ligand-binding domain**. *Proc Natl Acad Sci U S A* 2005, **102**(31):10976-10980.

37. Choe J, Kelker MS, Wilson IA: **Crystal structure of human toll-like receptor 3 (TLR3) ectodomain**. *Science* 2005, **309**(5734):581-585.

38. Liu L, Botos I, Wang Y, Leonard JN, Shiloach J, Segal DM, Davies DR: **Structural basis of toll-like receptor 3 signaling with double-stranded RNA**. *Science* 2008, **320**(5874):379-381.

39. 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.

40. Poltorak A, He X, Smirnova I, Liu MY, Van Huffel C, Du X, Birdwell D, Alejos E, Silva M, Galanos C *et al*: **Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in Tlr4 gene**. *Science* 1998, **282**(5396):2085-2088.

41. Hayashi F, Smith KD, Ozinsky A, Hawn TR, Yi EC, Goodlett DR, Eng JK, Akira S, Underhill DM, Aderem A: **The innate immune response to bacterial flagellin is mediated by Toll-like receptor 5**. *Nature* 2001, **410**(6832):1099-1103.

42. Jurk M, Heil F, Vollmer J, Schetter C, Krieg AM, Wagner H, Lipford G, Bauer S: **Human TLR7 or TLR8 independently confer responsiveness to the antiviral compound R-848**. *Nat Immunol* 2002, **3**(6):499.

43. Hemmi H, Takeuchi O, Kawai T, Kaisho T, Sato S, Sanjo H, Matsumoto M, Hoshino K, Wagner H, Takeda K *et al*: **A Toll-like receptor recognizes bacterial DNA**. *Nature* 2000, **408**(6813):740-745.

44. Zhang D, Zhang G, Hayden MS, Greenblatt MB, Bussey C, Flavell RA, Ghosh S: **A toll-like receptor that prevents infection by uropathogenic bacteria**. *Science* 2004, **303**(5663):1522-1526.

45. Lord KA, Hoffman-Liebermann B, Liebermann DA: **Nucleotide sequence and expression of a cDNA encoding MyD88, a novel myeloid differentiation primary response gene induced by IL6**. *Oncogene* 1990, **5**(7):1095-1097.

46. Medzhitov R, Preston-Hurlburt P, Kopp E, Stadlen A, Chen C, Ghosh S, Janeway CA, Jr.: **MyD88 is an adaptor protein in the hToll/IL-1 receptor family signaling pathways**. *Mol Cell* 1998, **2**(2):253-258.

47. Muzio M, Natoli G, Saccani S, Levrero M, Mantovani A: **The human toll signaling pathway: divergence of nuclear factor kappaB and JNK/SAPK activation upstream of tumor necrosis factor receptor-associated factor 6 (TRAF6)**. *J Exp Med* 1998, **187**(12):2097-2101.

48. Wesche H, Henzel WJ, Shillinglaw W, Li S, Cao Z: **MyD88: an adapter that recruits IRAK to the IL-1 receptor complex**. *Immunity* 1997, **7**(6):837-847.

49. Jenkins KA, Mansell A: **TIR-containing adaptors in Toll-like receptor signalling**. *Cytokine* 2009.

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

51.   Horng T, Barton GM, Medzhitov R: **TIRAP: an adapter molecule in the Toll signaling pathway**. *Nat Immunol* 2001, **2**(9):835-841.

52.   Fitzgerald KA, Palsson-McDermott EM, Bowie AG, Jefferies CA, Mansell AS, Brady G, Brint E, Dunne A, Gray P, Harte MT *et al*: **Mal (MyD88-adapter-like) is required for Toll-like receptor-4 signal transduction**. *Nature* 2001, **413**(6851):78-83.

53.   Kagan JC, Medzhitov R: **Phosphoinositide-mediated adaptor recruitment controls Toll-like receptor signaling**. *Cell* 2006, **125**(5):943-955.

54.   Horng T, Barton GM, Flavell RA, Medzhitov R: **The adaptor molecule TIRAP provides signalling specificity for Toll-like receptors**. *Nature* 2002, **420**(6913):329-333.

55.   Ohnishi H, Tochio H, Kato Z, Orii KE, Li A, Kimura T, Hiroaki H, Kondo N, Shirakawa M: **Structural basis for the multiple interactions of the MyD88 TIR domain in TLR4 signaling**. *Proc Natl Acad Sci U S A* 2009, **106**(25):10260-10265.

56.   Monie TP, Moncrieffe MC, Gay NJ: **Structure and regulation of cytoplasmic adapter proteins involved in innate immune signaling**. *Immunol Rev* 2009, **227**(1):161-175.

57.   Carpenter S, O'Neill LA: **Recent insights into the structure of Toll-like receptors and post-translational modifications of their associated signalling proteins**. *Biochem J* 2009, **422**(1):1-10.

58.   Oshiumi H, Matsumoto M, Funami K, Akazawa T, Seya T: **TICAM-1, an adaptor molecule that participates in Toll-like receptor 3-mediated interferon-beta induction**. *Nat Immunol* 2003, **4**(2):161-167.

59.   Oshiumi H, Sasai M, Shida K, Fujita T, Matsumoto M, Seya T: **TIR-containing adapter molecule (TICAM)-2, a bridging adapter recruiting to toll-like receptor 4 TICAM-1 that induces interferon-beta**. *J Biol Chem* 2003, **278**(50):49751-49762.

60.   Yamamoto M, Sato S, Hemmi H, Uematsu S, Hoshino K, Kaisho T, Takeuchi O, Takeda K, Akira S: **TRAM is specifically involved in the Toll-like receptor 4-mediated MyD88-independent signaling pathway**. *Nat Immunol* 2003, **4**(11):1144-1150.

61.   Rowe DC, McGettrick AF, Latz E, Monks BG, Gay NJ, Yamamoto M, Akira S, O'Neill LA, Fitzgerald KA, Golenbock DT: **The myristoylation of TRIF-related adaptor molecule is essential for Toll-like receptor 4 signal transduction**. *Proc Natl Acad Sci U S A* 2006, **103**(16):6299-6304.

62.   Janssens S, Burns K, Tschopp J, Beyaert R: **Regulation of interleukin-1- and lipopolysaccharide-induced NF-kappaB activation by alternative splicing of MyD88**. *Curr Biol* 2002, **12**(6):467-471.

63.   Burns K, Janssens S, Brissoni B, Olivos N, Beyaert R, Tschopp J: **Inhibition of interleukin 1 receptor/Toll-like receptor signaling through the alternatively spliced, short form of MyD88 is due to its failure to recruit IRAK-4**. *J Exp Med* 2003, **197**(2):263-268.

64.   Naiki Y, Michelsen KS, Zhang W, Chen S, Doherty TM, Arditi M: **Transforming growth factor-beta differentially inhibits MyD88-dependent, but not TRAM- and TRIF-dependent, lipopolysaccharide-induced TLR4 signaling**. *J Biol Chem* 2005, **280**(7):5491-5495.

65.   Polentarutti N, Rol GP, Muzio M, Bosisio D, Camnasio M, Riva F, Zoja C, Benigni A, Tomasoni S, Vecchi A *et al*: **Unique pattern of expression and inhibition of IL-1 signaling by the IL-1 receptor family member**

TIR8/SIGIRR. *Eur Cytokine Netw* 2003, **14**(4):211-218.

66. Qin J, Qian Y, Yao J, Grace C, Li X: **SIGIRR inhibits interleukin-1 receptor-and toll-like receptor 4-mediated signaling through different mechanisms**. *J Biol Chem* 2005, **280**(26):25233-25241.

67. Theofilopoulos AN, Baccala R, Beutler B, Kono DH: **Type I interferons (alpha/beta) in immunity and autoimmunity**. *Annu Rev Immunol* 2005, **23**:307-336.

68. Thomassen E, Renshaw BR, Sims JE: **Identification and characterization of SIGIRR, a molecule representing a novel subtype of the IL-1R superfamily**. *Cytokine* 1999, **11**(6):389-399.

69. Galperin MY: **The Molecular Biology Database Collection: 2008 update**. *Nucleic Acids Res* 2008, **36**(Database issue):D2-4.

70. Kulikova T, Akhtar R, Aldebert P, Althorpe N, Andersson M, Baldwin A, Bates K, Bhattacharyya S, Bower L, Browne P *et al*: **EMBL Nucleotide Sequence Database in 2006**. *Nucleic Acids Res* 2007, **35**(Database issue):D16-20.

71. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S *et al*: **Database resources of the National Center for Biotechnology Information**. *Nucleic Acids Res* 2008, **36**(Database issue):D13-21.

72. Sugawara H, Ogasawara O, Okubo K, Gojobori T, Tateno Y: **DDBJ with new system and face**. *Nucleic Acids Res* 2008, **36**(Database issue):D22-24.

73. The-UniProt-Consortium: **The Universal Protein Resource (UniProt) 2009**. *Nucleic Acids Res* 2009, **37**(Database issue):D169-174.

74. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I *et al*: **The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003**. *Nucleic Acids Res* 2003, **31**(1):365-370.

75. Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J, Chen Y, Hu Z, Kourtesis P, Ledley RS, Suzek BE *et al*: **The Protein Information Resource**. *Nucleic Acids Res* 2003, **31**(1):345-347.

76. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**(1):235-242.

77. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins**. *Nucleic Acids Res* 2007, **35**(Database issue):D61-65.

78. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL *et al*: **The Pfam protein families database**. *Nucleic Acids Res* 2008, **36**(Database issue):D281-288.

79. Mulder N, Apweiler R: **InterPro and InterProScan: tools for protein sequence classification and comparison**. *Methods Mol Biol* 2007, **396**:59-70.

80. 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.

81. Humphrey W, Dalke A, Schulten K: **VMD: visual molecular dynamics**. *J Mol Graph* 1996, **14**(1):33-38, 27-38.

82. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM: **CATH--a hierarchic classification of protein domain structures**. *Structure*

1997, **5**(8):1093-1108.

83. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures**. *J Mol Biol* 1995, **247**(4):536-540.

84. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M *et al*: **STRING 8--a global view on proteins and their functional interactions in 630 organisms**. *Nucleic Acids Res* 2009, **37**(Database issue):D412-416.

85. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: **KEGG for representation and analysis of molecular networks involving diseases and drugs**. *Nucleic Acids Res* 2010, **38**(Database issue):D355-360.

86. Parkinson H, Kapushesky M, Kolesnikov N, Rustici G, Shojatalab M, Abeygunawardena N, Berube H, Dylag M, Emam I, Farne A *et al*: **ArrayExpress update--from an archive of functional genomics experiments to the atlas of gene expression**. *Nucleic Acids Res* 2009, **37**(Database issue):D868-872.

87. Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, Millburn G, Osumi-Sutherland D, Schroeder A, Seal R *et al*: **FlyBase: enhancing Drosophila Gene Ontology annotations**. *Nucleic Acids Res* 2009, **37**(Database issue):D555-559.

88. Wang L, Riethoven JJ, Robinson A: **XEMBL: distributing EMBL data in XML format**. *Bioinformatics* 2002, **18**(8):1147-1148.

89. Harold ER: **XML Bible**: IDG Books Wordwide; 1999.

90. Dolan J, Walshe K, Alsbury S, Hokamp K, O'Keeffe S, Okafuji T, Miller SF, Tear G, Mitchell KJ: **The extracellular leucine-rich repeat superfamily; a comparative survey and analysis of evolutionary relationships and expression patterns**. *BMC Genomics* 2007, **8**:320.

91. Petsko GA, Ringe D: **Protein Structure and Function**: New Science Press; 2004.

92. Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A: **Comparative protein structure modeling of genes and genomes**. *Annu Rev Biophys Biomol Struct* 2000, **29**:291-325.

93. Bowie JU, Luthy R, Eisenberg D: **A method to identify protein sequences that fold into a known three-dimensional structure**. *Science* 1991, **253**(5016):164-170.

94. Bonneau R, Baker D: **Ab initio protein structure prediction: progress and prospects**. *Annu Rev Biophys Biomol Struct* 2001, **30**:173-189.

95. Anfinsen CB: **Principles that govern the folding of protein chains**. *Science* 1973, **181**(96):223-230.

96. Khan JA, Brint EK, O'Neill LA, Tong L: **Crystal structure of the Toll/interleukin-1 receptor domain of human IL-1RAPL**. *J Biol Chem* 2004, **279**(30):31664-31670.

97. Loiarro M, Capolunghi F, Fanto N, Gallo G, Campo S, Arseni B, Carsetti R, Carminati P, De Santis R, Ruggiero V *et al*: **Pivotal Advance: Inhibition of MyD88 dimerization and recruitment of IRAK1 and IRAK4 by a novel peptidomimetic compound**. *J Leukoc Biol* 2007, **82**(4):801-810.

98. Ewald SE, Lee BL, Lau L, Wickliffe KE, Shi GP, Chapman HA, Barton GM: **The ectodomain of Toll-like receptor 9 is cleaved to generate a functional receptor**. *Nature* 2008, **456**(7222):658-662.

99. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, Shen MY, Pieper U, Sali A: **Comparative protein structure modeling using MODELLER**. *Curr Protoc Protein Sci* 2007, **Chapter 2**:Unit 2 9.

100. Fiser A, Sali A: **ModLoop: automated modeling of loops in protein structures**. *Bioinformatics* 2003, **19**(18):2500-2501.

101. Jones DT, Taylor WR, Thornton JM: **A new approach to protein fold recognition**. *Nature* 1992, **358**(6381):86-89.

102. Lobley A, Sadowski MI, Jones DT: **pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination**. *Bioinformatics* 2009, **25**(14):1761-1767.

103. Bryson K, McGuffin LJ, Marsden RL, Ward JJ, Sodhi JS, Jones DT: **Protein structure prediction servers at University College London**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W36-38.

104. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.

105. Wallner B, Elofsson A: **Can correct protein models be identified?** *Protein Sci* 2003, **12**(5):1073-1086.

106. Cristobal S, Zemla A, Fischer D, Rychlewski L, Elofsson A: **A study of quality measures for protein threading models**. *BMC Bioinformatics* 2001, **2**:5.

107. Siew N, Elofsson A, Rychlewski L, Fischer D: **MaxSub: an automated measure for the assessment of protein structure prediction quality**. *Bioinformatics* 2000, **16**(9):776-785.

108. Pawlowski M, Gajda MJ, Matlak R, Bujnicki JM: **MetaMQAP: a meta-server for the quality assessment of protein models**. *BMC Bioinformatics* 2008, **9**:403.

109. McGuffin LJ: **The ModFOLD server for the quality assessment of protein structural models**. *Bioinformatics* 2008, **24**(4):586-587.

110. McGuffin LJ: **Benchmarking consensus model quality assessment for protein fold recognition**. *BMC Bioinformatics* 2007, **8**:345.

111. Laskowski RA, MacArthur MW, Moss DS, Thornton JM: **PROCHECK: a program to check the stereochemical quality of protein structures**. *J Appl Cryst* 1993, **26**:283-291.

112. Ramachandran GN, Ramakrishnan C, Sasisekharan V: **Stereochemistry of polypeptide chain configurations**. *J Mol Biol* 1963, **7**:95-99.

113. Tovchigrechko A, Vakser IA: **GRAMM-X public web server for protein-protein docking**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W310-314.

114. Chen R, Li L, Weng Z: **ZDOCK: an initial-stage protein-docking algorithm**. *Proteins* 2003, **52**(1):80-87.

115. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**(5):1792-1797.

116. Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ: **Jalview Version 2--a multiple sequence alignment editor and analysis workbench**. *Bioinformatics* 2009, **25**(9):1189-1191.

117. Guex N, Peitsch MC: **SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling**. *Electrophoresis* 1997, **18**(15):2714-2723.

118. Maiti R, Van Domselaar GH, Zhang H, Wishart DS: **SuperPose: a simple server for sophisticated structural superposition**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W590-594.

119. Krissinel E, Henrick K: **Inference of macromolecular assemblies from crystalline state**. *J Mol Biol* 2007, **372**(3):774-797.

Acknowledgement

Acknowledgement

Appendix

Paper 1

TollML: a database of toll-like receptor structural motifs

Jing Gong, Tiandi Wei, Ferdinand Jamitzky, Wolfgang M. Heckl, Shaila C. Rössle and Robert W. Stark

TollML: a database of toll-like receptor structural motifs

Jing Gong · Tiandi Wei · Ning Zhang ·
Ferdinand Jamitzky · Wolfgang M. Heckl ·
Shaila C. Rössle · Robert W. Stark

Abstract Toll-like receptors (TLRs) play a key role in the
innate immune system. TLRs recognize pathogen-
associated molecular patterns and initiate an intracellular
kinase cascade to induce an immediate defensive response.
During recent years TLRs have become the focus of
tremendous research interest. A central repository for the
growing amount of relevant TLR sequence information has
been created. Nevertheless, structural motifs of most
sequenced TLR proteins, such as leucine-rich repeats
(LRRs), are poorly annotated in the established databases.
A database that organizes the structural motifs of TLRs
could be useful for developing pattern recognition pro-
grams, structural modeling and understanding functional
mechanisms of TLRs. We describe TollML, a database that
integrates all of the TLR sequencing data from the NCBI
protein database. Entries were first divided into TLR
families (TLR1-23) and then semi-automatically subdivided
into three levels of structural motif categories: (1) signal
peptide (SP), ectodomain (ECD), transmembrane domain
(TD) and Toll/IL-1 receptor (TIR) domain of each TLR; (2)
LRRs of each ECD; (3) highly conserved segment (HCS),
variable segment (VS) and insertions of each LRR. These
categories can be searched quickly using an easy-to-use
web interface and dynamically displayed by graphics.
Additionally, all entries have hyperlinks to various sources
including NCBI, Swiss-Prot, PDB, LRRML and PubMed in
order to provide broad external information for users. The
TollML database is available at http://tollml.lrz.de.

J. Gong · T. Wei · F. Jamitzky · W. M. Heckl · R. W. Stark
Center for Nanoscience,
Ludwig-Maximilians-Universität München,
80799 Munich, Germany

J. Gong · T. Wei (✉) · S. C. Rössle · R. W. Stark
Department of Earth and Environmental Sciences,
Ludwig-Maximilians-Universität München,
Theresienstr. 41,
80333 Munich, Germany
e-mail: tiandi@informatik.uni-muenchen.de

N. Zhang
Department of Informatics,
Ludwig-Maximilians-Universität München,
80333 Munich, Germany

F. Jamitzky
Leibniz Supercomputing Centre,
85748 Garching, Germany

W. M. Heckl
Deutsches Museum,
80538 Munich, Germany

W. M. Heckl
TUM School of Education, Technische Universität München,
80799 Munich, Germany

Keywords TollML · Toll-like receptor ·
Leucine-rich repeats · XML database · Homology modeling

Introduction

Since the *Drosophila Toll* gene was discovered in the mid-
1980s [1], genome projects have led to the identification of
13 receptors in mammalian and more than 20 receptors in
non-mammalian genomes that are homologs of *Drosophila
Toll*. These receptors have been termed collectively Toll-
like receptors (TLRs). TLRs play a key role in innate
immunity. They recognize invading microbial pathogens

and rapidly initiate intracellular signal transduction pathways to trigger expression of genes, whose products can control innate immune responses [2]. All TLRs have a common domain organization, with an extracellular ectodomain (ECD), a helical transmembrane domain (TD), and an intracellular Toll/IL-1 receptor homology (TIR) domain [3]. The ectodomain is a horseshoe-shaped solenoid structure and is directly involved in the recognition of a variety of pathogens including lipopolysaccharide, lipopeptide, cytosine-phosphate-guanine (CpG) DNA, flagellin, imidazoquinoline and dsRNA [4]. The transmembrane domain determines the subcellular localization of TLRs [5]. The TIR domain is conserved across all TLRs and IL-1 receptors, and is also shared by downstream signaling adaptor molecules. Upon receptor ligation, a TIR signaling complex is formed between the receptor and the adaptor TIR domains [6].

The TLR ectodomain contains varying numbers of leucine-rich repeat (LRR) motifs, which are arrays of 20 to 30 amino acid-long protein sequences that are enriched with the hydrophobic amino acid leucine. All LRR sequences can be divided into a highly conserved segment (HCS) and a variable segment (VS). The HCS consists of an 11 or 12 residue stretch with the consensus sequence LxxLxLxxN(Cx)xL. In this notation, the letter L represents the amino acids leucine, isoleucine, valine or phenylalanine, which form a hydrophobic core, N represents asparagine, threonine, serine or cysteine, and x is any amino acid. The variable segment can vary in both length and consensus sequence. Accordingly, several types of LRRs have been proposed [7, 8]. Of these, typical (T) type (xxLxxxxLxxLxx) and bacterial (S) type (xxLPx(x)LPxx) LRRs have been observed in TLRs [9]. All LRRs in TLRs are capped by N- and C-terminal LRRs that are usually irregular and do not match any type of LRR consensus sequences.

The atomic-detail crystal structure of the human TLR1 and TLR2 TIR domain was published in 2000 and gave the first insight into the molecular basis of TIR signaling [10]. The crystal structures of the ectodomains of human TLR1-4 and mouse TLR2-4 have also been resolved [11–16]. These structures demonstrate how the LRR-based platform is adapted to ligand recognition. Nevertheless, more than 2,000 TLR proteins have been sequenced by high-throughput genome sequencing projects. It is clear that the discrepancy between the rate at which novel protein sequences are discovered and the rate at which detailed structural information on proteins can be obtained from X-ray diffraction or nuclear magnetic resonance spectroscopy will persist for the foreseeable future. Thus, a comparative analysis at the sequence level is a useful approach to identify and characterize structural motifs of TLRs [9, 17] and to gain insight into how receptors and

ligands interact. Due to the variability of LRR motifs in TLRs, however, the indicated repeat number and positions (beginning/end of a repeating unit) for individual TLRs are quite different or missing in established databases. Currently, there is no collection of structural information for features that are contained within LRRs, such as HCS, VS and sequence insertions.

In this paper, we describe a database of TLR structural motifs called TollML. The current release (3.1) includes all known TLR sequences from the NCBI protein database [18]. Structural motifs were identified and annotated by a semi-automatic procedure that included comparison of sequences with the sequences of TLRs that have a known structure, consensus sequence matching, secondary structure prediction and multiple sequence alignments. Three levels of motif elements were generated: (1) signal peptide (SP), ectodomain (ECD), transmembrane domain (TD) and TIR domain of each TLR; (2) LRRs of each ECD; and (3) HCS, VS and insertions of each LRR. Some program application examples are presented in the last section of the paper.

Construction and content

Data extraction and pre-processing

Initial TLR sequences were extracted from the NCBI protein database. Two groups of search results were obtained using the search keys *toll** and *tlr**, where the asterisk stands for any suffix, to ensure that all TLRs were included. A manual data pre-processing step was performed before the motif identification of these sequences. We inspected the NCBI annotations of entries one-by-one to exclude TLR related molecules such as adaptors, protein kinases and transcription factors. After we performed these filtering steps, 2,572 TLR entries remained (NCBI release: 1 September 2009). We then categorized the TLRs into different families (TLR1-23) based on their original annotations. In the instances in which entries were not associated with explicit comments, we compared the sequences with well classified TLR sequences using sequence BLAST.

Motif identification

Three levels of structural motif categories were generated adapting to the structural organization of TLRs. First, each full length TLR sequence was divided into ECD, TD and TIR domain. If a sequence started with a SP, which directs the subcellular transport of a protein, the presence of a SP was also indicated. Second, the ECD of each TLR was partitioned into individual LRRs including

canonical LRRs and N/C-terminal LRRs. Third, each LRR was further divided into a highly conserved segment (HCS) and a variable segment (VS). Insertions within the VS that are longer than three residues were identified and annotated.

The procedure mentioned above was semi-automated. We first sorted the TLR sequences from each family into subgroups, so that an arbitrary pair of sequences from the same subgroup had a local sequence similarity greater than 90%. Then, the three-level structural motifs were identified manually for a selected template sequence from each subgroup. This selected template sequence was a full length sequence that had the most detailed original annotations so that the most accurate motif identifications could be performed. Finally, the other members of the subgroup received their motif partition assignments through multiple alignments with the template. More than 300 subgroups were generated for the 2,572 entries. Thus more than 300 templates were processed manually.

The manual motif identification of a template sequence combined three approaches: consensus sequence matching, secondary structure prediction and reference to original annotations or literatures. The four domains on the first level of the motif categorization (SP, ECD, TD and TIR) have characteristic sequence features, so are usually accurately divided in the original annotation of a selected template. If a template did not have clear annotations, then its sequence was compared to similar sequences with known domains to determine its domain partitions. For the second level of motif categorization, if a selected template was associated with a reliable reference such as a known crystal structure, its LRR partition was then assigned accordingly. Otherwise, we matched the LRR consensus sequence LxxLxLxxN(Cx)xL to the template sequence amino-acid-by-amino-acid and detected LRR motifs manually. In addition, protein secondary structure predictions (PredictProtein [19], NNPREDICT [20], PSIPRED [21] and SSPro [22]) helped to improve the accuracy of LRR detection because all known crystal structures of TLRs show that there is always a short β-strand (3–5 residues) beginning at approximately the third position of an LRR motif [11–15]. After an LRR was identified, consensus sequence matching was used to identify its HCS and VS motifs, as well as any insertions that were longer than three residues for the third level. Simultaneously, each LRR motif was classified into different types (detailed in Database content) according to the VS consensus sequences.

Database content

The TollML release 3.1 contains 2,572 TLR entries divided into 23 families (entry distribution shown in Table 1).

Among these, 2,350 of the sequences contain an ectodomain and thus received motif annotations that correspond to the second and third levels of motif categorization. The other entries contain only a TIR domain with or without a transmembrane domain. A total of 46,720 LRR motifs were recognized from the ectodomain containing TLRs. These LRR motifs were classified into five types: typical (T), bacterial (S), N-terminal (NT), C-terminal (CT) and irregular (I). A histogram of LRR length statistics (Fig. 1) shows the characteristic length distribution of each LRR type. The standard length of the T type LRR is 24 amino acids. A large number of T type LRRs have insertions and only some have deletions. These statistics suggest that the evolution of T type LRRs may prefer insertion over deletion. By contrast, the S type LRRs are more highly conserved. Their lengths are concentrated on 20 and 21. N-terminal LRRs vary in length and do not form a peak value in length distribution. Most C-terminal LRRs contain four cysteines that are distantly separated at the sequence level and form disulfide bonds with each other. C-terminal LRRs are generally greater than 35 amino acids in length.

Annotations for each TLR entry include:

(1) Data management information: TollML ID and access/modification date;
(2) Primary information extracted from the NCBI and related literature: FASTA sequence, biological definition, cell information, glycosylation sites and ligands;
(3) Protein family classification;
(4) Database cross links: NCBI, Swiss-Prot [23], PDB [24], LRRML [8] and PubMed [25];
(5) Three-level motif information.

Database comparison

Currently, several protein databases, such as Pfam [26], InterPro [27], SMART [28] and Swiss-Prot, contain information about TLRs. These databases predict the LRR numbers and positions for their TLR entries by various computational methods, thus resulting in a high frequency of false negative predictions. Table 2 illustrates the LRR numbers for human TLR1–10 as reported by these databases. The manual motif identification procedure discussed here provides TollML with the most complete database of LRR motifs. Although Swiss-Prot presents more accurate results than the other three databases that we investigated (Table 2), TollML has four prominent characteristics that distinguish it from Swiss-Prot:

(1) Comprehensive entry coverage. TollML covers 2,572 TLR sequences from 121 species and all sequences are provided with detailed motif annotations. Swiss-Prot

Table 1 Entry distribution over Toll-like receptor (TLR) families for mammalian/non-mammalian groups

TLR	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	18	19	20–23	Total
Mammalian	85	146	108	443	132	106	109	101	124	72	10	11	9	0	0	0	0	0	0	1,456
Non-mammalian	356	276	84	77	58	3	102	11	30	0	0	1	4	6	42	2	7	5	52	1,116
Total	441	422	192	520	190	109	211	112	154	72	10	12	13	6	42	2	7	5	52	2,572

covers 636 TLR sequences from 17 species and only 59 sequences have LRR annotations (results obtained on 24 February 2009).

(2) Structural motifs within an LRR. TollML annotates the HCS, VS and insertion for each LRR. This information is not present in any other published protein databases.

(3) Uniform LRR definition. The beginning/end positions of LRRs have been defined inconsistently across researchers due to the periodicity of LRR motifs. This variation leads to non-uniform LRR assignments in Swiss-Prot. All LRR motifs in TollML start at the beginning of the HCS and end at the end of the VS, just before the HCS of the next LRR.

(4) Accessibility of motif sequences. The amino acid sequence of any available motif is directly accessible in TollML, whereas only the full length sequence is directly accessible in Swiss-Prot.

Utility

Web application

The extensible markup language (XML) was standardized in the 1990s and is well established as a format for hierarchical biological data. TollML was designed by using

eXist [29], an XML database management system, and XPath/XQuery [30] for processing queries and web forms. The document type definition (DTD) file of TollML is provided in the electronic supplementary material (ESM; Supplementary file 1).

The entire database is browsable. When browsing, entries appear in a summary table containing ID, definition, family, species and links of motif partitions. Clicking on an entry opens an XML Stylesheet (XSLT) [30] converted HTML web page that describes the entry in detail. The original XML file can also be downloaded. The XSLT file that the program uses is provided in the ESM (Supplementary file 2). Aside from the textual view, the structural motifs of TLRs can be exhibited by three-level dynamic graphics. Figure 2 demonstrates the motif assignment for an example entry (ID: TLR_561).

On the *advanced search* page of TollML, users can search entries flexibly by inputting keywords, specifying search fields, and defining annotation contents of the output. After selected entries are returned, a *search within result* button allows for further term filtering. The resulting entries, or an arbitrarily selected subset thereof, can be sent to generate multiple sequence alignments supplied by the T-Coffee package [31]. In addition, a Wu-BLAST search tool [32] is available. A query sequence can be BLASTed against the entire database, against a certain TLR family, or against a collection of sequences marked by a user-defined label (available for registered users).
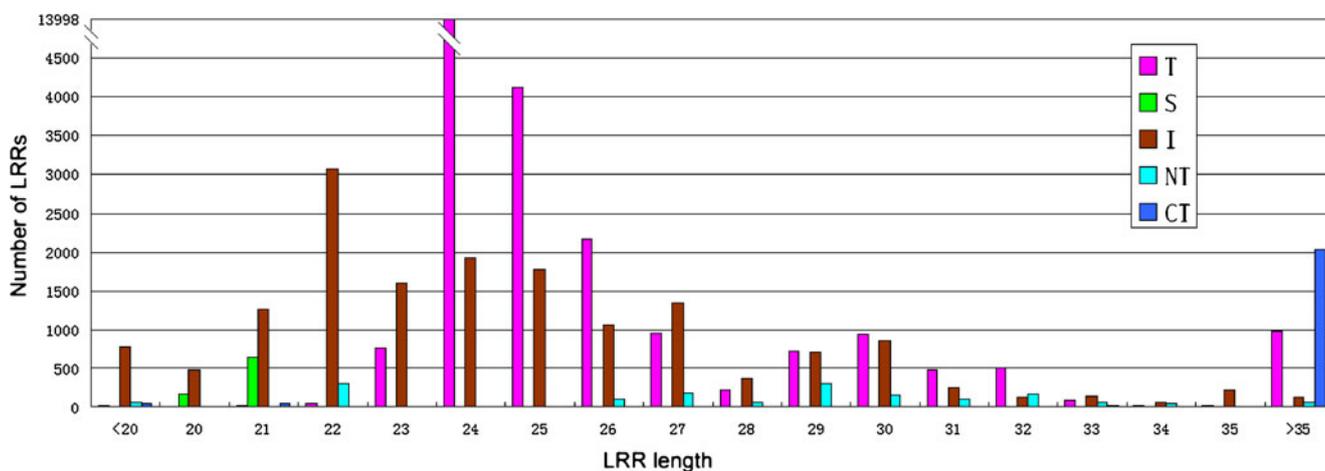
Fig. 1 Leucine-rich repeat (LRR) length distribution

Table 2 Comparison of leucine-rich repeat (LRR) numbers of human TLR1–10 in different databases (results obtained on 24 February 2009)

	TLR1	TLR2	TLR3	TLR4	TLR5	TLR6	TLR7	TLR8	TLR9	TLR10
InterPro	4	4	8	7	7	3	9	10	6	4
Pfam	4	4	7	7	8	3	8	9	6	4
SMART	6	9	19	12	10	7	16	17	19	7
Swiss-Prot	8	14	22	21	15	13	27	24	26	15
TollML	21	21	25	23	23	21	28	28	28	21

Application in LRR prediction

The automatic identification of motifs in a protein sequence is essentially a statistical pattern recognition problem. Therefore, the performance of a prediction method is thus strongly dependent on the quality and scale of the training data set. A recent program, LRRscan [17], demonstrated the feasibility of a statistics-based consensus matching algorithm applied to LRR detection. TollML supplies a large and reliable source to train LRR prediction methods of different algorithms. As an example, we developed a 20× 12 position-specific weight matrix of LRR motifs based on the LRR partitions from TollML (matrix available on the TollML webpage). A matrix element denotes the frequency probability of a certain amino acid occurring at a certain position in the LRR HCS. We have already obtained confident results through this sort of matrix scan with an appropriate cut-off score. A five-fold cross validation against all TollML entries indicates that the sensitivity and specificity are both greater than 93%. This program was implemented on the TollML webpage and has been named LRRFinder. This method can be extended to predict LRR motifs in other LRR containing proteins besides TLRs, since all LRR types possess the same HCS pattern.

Fig. 2 Online graphic display of three-level structural motifs of a database entry
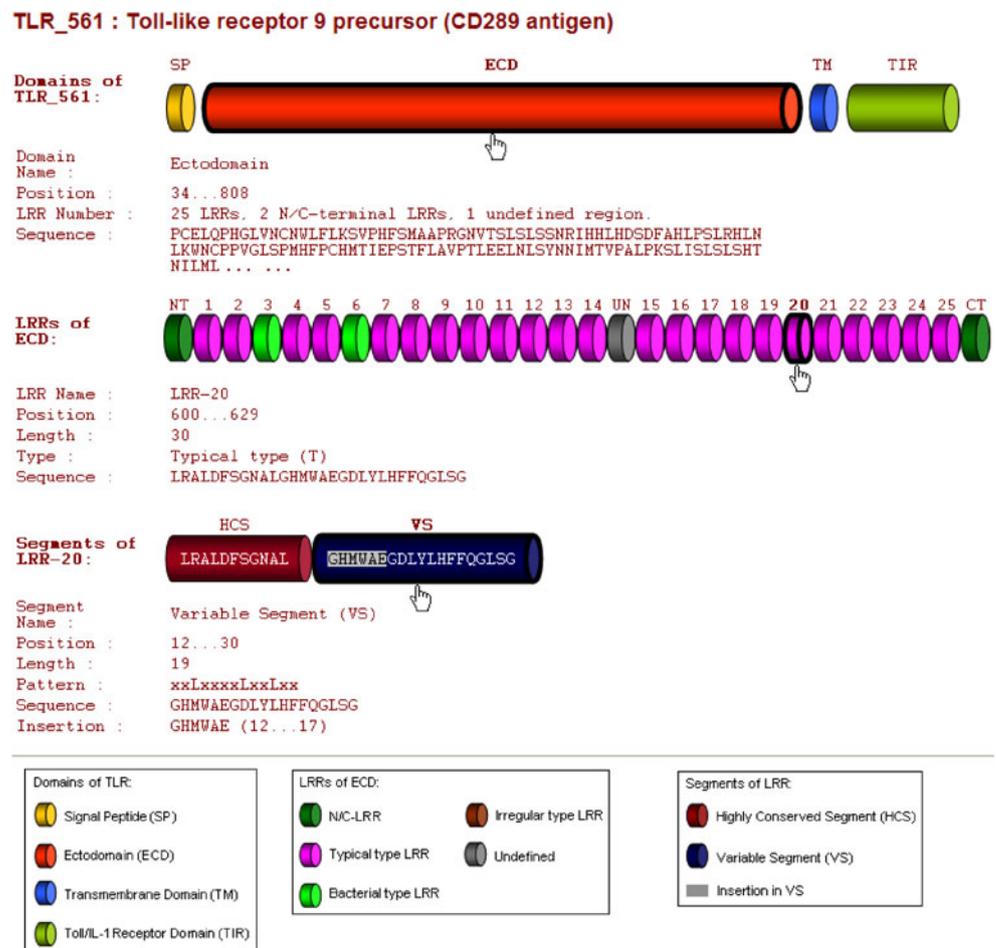
Application in homology modeling

Homology modeling is currently the most accurate computational method to predict protein structure. This system constructs a structural model for a target protein sequence from a template structure of a homologous protein. For this method to work, the target and template structures must have a sequence identity over 30%. The known crystal structures of human TLR1–4 and mouse TLR2–4 ectodomains supply valuable templates for the homology modeling of other TLR ectodomains. However, given the variability in repeat numbers and type arrangements of LRRs contained within TLRs, a full length template with a sufficient sequence identity is typically not available. This limitation can be overcome by assembling multiple LRR templates. First, all LRRs in the target sequence are identified. Then, the most similar structure-known LRR based on sequence is found for each LRR in the target sequence as a local template. Finally, all local templates are assembled to generate a multiple sequence alignment for the complete target sequence. TollML enables the direct accessibility of accurate LRR sequence partitions for the first step. The LRRML database [8] can further provide suitable LRR structural templates. This LRR template assembling approach was proven to be both feasible and significant by recent structure modeling research into human TLR7–9 [33].

Conclusions

We have developed a specialized database of TLR structural motifs called TollML. It is supported by an XML database management system and can be searched and browsed with an easy-to-use web interface. This interface is suitable for use with most graphical web browsers and has been tested on the Windows, Mac and Linux operating systems. TollML includes all TLR sequences that are published in the NCBI protein database and semi-automatically creates three levels of motif annotations. This database can help to develop motif prediction programs, to model three-dimensional structures of TLRs and to design new mutation experiments to better understand receptor–ligand or receptor–receptor interactions. We plan to update TollML every 2–3 months since the number of sequenced TLR proteins increases constantly.

Availability and requirements

This database is freely available at http://tollml.lrz.de. Any internet user can search and download data from the database, but only registered users can define and save labels for arbitrary entries.

References

1. Anderson KV, Bokla L, Nusslein-Volhard C (1985) Establishment of dorsal-ventral polarity in the Drosophila embryo: the induction of polarity by the Toll gene product. Cell 42:791–798
2. Takeda K, Akira S (2005) Toll-like receptors in innate immunity. Int Immunol 17:1–14
3. Brodsky I, Medzhitov R (2007) Two modes of ligand recognition by TLRs. Cell 130:979–981
4. Gay NJ, Gangloff M (2007) Structure and function of Toll receptors and their ligands. Annu Rev Biochem 76:141–165
5. Barton GM, Kagan JC, Medzhitov R (2006) Intracellular localization of Toll-like receptor 9 prevents recognition of self DNA but facilitates access to viral DNA. Nat Immunol 7:49–56
6. O'Neill LA, Bowie AG (2007) The family of five: TIR-domain-containing adaptors in Toll-like receptor signalling. Nat Rev Immunol 7:353–364
7. Kobe B, Kajava AV (2001) The leucine-rich repeat as a protein recognition motif. Curr Opin Struct Biol 11:725–732
8. Wei T, Gong J, Jamitzky F, Heckl WM, Stark RW, Roessle SC (2008) LRRML: a conformational database and an XML description of leucine-rich repeats (LRRs). BMC Struct Biol 8:47
9. Matsushima N, Tanaka T, Enkhbayar P, Mikami T, Taga M, Yamada K, Kuroki Y (2007) Comparative sequence analysis of leucine-rich repeats (LRRs) within vertebrate toll-like receptors. BMC Genomics 8:124
10. Xu Y, Tao X, Shen B, Horng T, Medzhitov R, Manley JL, Tong L (2000) Structural basis for signal transduction by the Toll/interleukin-1 receptor domains. Nature 408:111–115
11. Bell JK, Botos I, Hall PR, Askins J, Shiloach J, Segal DM, Davies DR (2005) The molecular structure of the Toll-like receptor 3 ligand-binding domain. Proc Natl Acad Sci USA 102:10976–10980
12. Choe J, Kelker MS, Wilson IA (2005) Crystal structure of human toll-like receptor 3 (TLR3) ectodomain. Science 309:581–585
13. Jin MS, Kim SE, Heo JY, Lee ME, Kim HM, Paik SG, Lee H, Lee JO (2007) Crystal structure of the TLR1-TLR2 heterodimer induced by binding of a tri-acylated lipopeptide. Cell 130:1071–1082
14. Kim HM, Park BS, Kim JI, Kim SE, Lee J, Oh SC, Enkhbayar P, Matsushima N, Lee H, Yoo OJ, Lee JO (2007) Crystal structure of the TLR4-MD-2 complex with bound endotoxin antagonist Eritoran. Cell 130:906–917
15. Liu L, Botos I, Wang Y, Leonard JN, Shiloach J, Segal DM, Davies DR (2008) Structural basis of toll-like receptor 3 signaling with double-stranded RNA. Science 320:379–381
16. Park BS, Song DH, Kim HM, Choi BS, Lee H, Lee JO (2009) The structural basis of lipopolysaccharide recognition by the TLR4-MD-2 complex. Nature 458:1191–1195
17. Dolan J, Walshe K, Alsbury S, Hokamp K, O'Keeffe S, Okafuji T, Miller SF, Tear G, Mitchell KJ (2007) The extracellular leucine-rich repeat superfamily; a comparative survey and analysis of evolutionary relationships and expression patterns. BMC Genomics 8:320
18. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E (2008) Database resources of the

National Center for Biotechnology Information. Nucleic Acids Res 36:D13–D21

19. Rost B, Yachdav G, Liu J (2004) The PredictProtein server. Nucleic Acids Res 32:W321–W326

20. Kneller DG, Cohen FE, Langridge R (1990) Improvements in protein secondary structure prediction by an enhanced neural network. J Mol Biol 214:171–182

21. Bryson K, McGuffin LJ, Marsden RL, Ward JJ, Sodhi JS, Jones DT (2005) Protein structure prediction servers at University College London. Nucleic Acids Res 33:W36–W38

22. Pollastri G, Przybylski D, Rost B, Baldi P (2002) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins 47:228–235

23. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N, Suzek B (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 34:D187–D191

24. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242

25. PubMed: http://www.ncbi.nlm.nih.gov/pubmed

26. Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A (2008) The Pfam protein families database. Nucleic Acids Res 36: D281–D288

27. Mulder N, Apweiler R (2007) InterPro and InterProScan: tools for protein sequence classification and comparison. Methods Mol Biol 396:59–70

28. Schultz J, Copley RR, Doerks T, Ponting CP, Bork P (2000) SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res 28:231–234

29. eXist: http://exist-db.org

30. The World Wide Web Consortium: http://www.w3.org

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

32. WU-BLAST: http://blast.wustl.edu

33. Wei T, Gong J, Jamitzky F, Heckl WM, Stark RW, Rossle SC (2009) Homology modeling of human Toll-like receptors TLR7, 8, and 9 ligand-binding domains. Protein Sci 18:1684–1691

Paper 2

LRRML: a conformational database and an XML description of leucine-rich repeats (LRRs)

Tiandi Wei, Jing Gong, Ferdinand Jamitzky, Wolfgang M. Heckl, Robert W. Stark and Shaila C. Rössle

BMC Structural Biology

LRRML: a conformational database and an XML description of leucine-rich repeats (LRRs)

Tiandi Wei[†1], Jing Gong*[†1], Ferdinand Jamitzky[1,2], Wolfgang M Heckl[1,3], Robert W Stark[1] and Shaila C Rössle[1]

Address: [1]Department of Earth and Environmental Sciences, Ludwig-Maximilians-Universität München, Theresienstr, 41, 80333 Munich, Germany, [2]Leibniz Supercomputing Centre, 85748 Garching, Germany and [3]Deutsches Museum, 80538 Munich, Germany

E-mail: Tiandi Wei - tiandi@informatik.uni-muenchen.de; Jing Gong* - gongj@informatik.uni-muenchen.de; Ferdinand Jamitzky - jamitzky@lrz.de; Wolfgang M Heckl - heckl@lmu.de; Robert W Stark - stark@lrz.uni-muenchen.de; Shaila C Rössle - shaila.roessle@lrz.uni-muenchen.de;
*Corresponding author   †Equal contributors

Abstract

Background: Leucine-rich repeats (LRRs) are present in more than 6000 proteins. They are found in organisms ranging from viruses to eukaryotes and play an important role in protein-ligand interactions. To date, more than one hundred crystal structures of LRR containing proteins have been determined. This knowledge has increased our ability to use the crystal structures as templates to model LRR proteins with unknown structures. Since the individual three-dimensional LRR structures are not directly available from the established databases and since there are only a few detailed annotations for them, a conformational LRR database useful for homology modeling of LRR proteins is desirable.

Description: We developed LRRML, a conformational database and an extensible markup language (XML) description of LRRs. The release 0.2 contains 1261 individual LRR structures, which were identified from 112 PDB structures and annotated manually. An XML structure was defined to exchange and store the LRRs. LRRML provides a source for homology modeling and structural analysis of LRR proteins. In order to demonstrate the capabilities of the database we modeled the mouse Toll-like receptor 3 (TLR3) by multiple templates homology modeling and compared the result with the crystal structure.

Conclusion: LRRML is an information source for investigators involved in both theoretical and applied research on LRR proteins. It is available at http://zeus.krist.geo.uni-muenchen.de/~lrrml.

Background

Leucine-rich repeats (LRRs) are arrays of 20 to 30 amino acid long protein segments that are unusually rich in the hydrophobic amino acid leucine. They are present in more than 6000 proteins in different organisms ranging from viruses to eukaryotes [1]. The structure of the LRRs and their arrangement in repetitive stretches of variable length generate a versatile and highly evolvable framework for the binding of manifold proteins and non-protein ligands [2]. The crystal structure of the ribonuclease inhibitor (RI) yielded the first insight into the three-dimensional molecular basis of LRRs [3]. It has a horseshoe shaped solenoid structure with parallel β-sheet lining the inner circumference and α-helices flanking its outer circumference. To date, there are over one hundred crystal structures available. All known LRR domains adopt an arc or horseshoe shape [1].

The LRR sequences can be divided into a highly conserved segment (HCS) and a variable segment (VS). The highly conserved segment consists of an 11 or 12 residue stretch with the consensus sequence LxxLxLxxN(Cx)xL. Here, the letter L stands for Leu, Ile, Val or Phe forming the hydrophobic core, N stands for Asn, Thr, Ser or Cys, and x is any amino acid. The variable segment is quite diverse in length and consensus sequence, accordingly eight classes of LRRs have been proposed [4, 5]: 'RI-like (RI)', 'Cysteine-containing (CC)', 'Bacterial (S)', 'SDS22-like (SDS22)', 'Plant-specific (PS)', 'Typical (T)', 'Treponema pallidum (Tp)' and 'CD42b-like (CD42b)'.

The discrepancy between the numbers of structure-known LRR proteins and the structure-unknown ones triggered studies focusing on the homology modeling of LRR proteins [6-8]. Homology modeling is a computational method, which is widely used to identify structural features defining molecular interactions [8-10]. The modeling results are an important input for the design of biochemical experiments. The first step of homology modeling is the selection of a structure-known protein, which serves as a template for the unknown target structure. In practice, however, it is difficult to find a complete template which has a high enough sequence identity to the target repetitive protein (single template modeling), due to different repeat numbers and varying arrangements. This limitation can be overcome by combining multiple templates. First, the most similar structure-known LRRs are found for each LRR in the target sequence as a local template. Second, all local templates are combined to generate the multiple sequence alignments for the entire target sequence. Thus, it is possible to construct a start model for further investigation, even if no adequate single template is available. Such an approach, however, requires a comprehensive database of LRRs to extract adequate template candidates. So far, the individual three-dimensional LRR structures are not directly available from the established databases and there are only a few detailed annotations for them. Additional information such as sequence insertions and types is missing. In order to consolidate this information and to provide a source for homology modeling and structural analysis of LRR proteins, we developed LRRML, a database and an extensible markup language (XML) description of LRR structures.

Construction and content

Structure-known LRR proteins were extracted from the Protein Data Bank (PDB) [11] release Sept 10, 2008. In order to ensure that all LRR proteins were found, we combined three groups of search results. First, 'leucine rich repeat', 'leucine rich repeats', 'leucine-rich repeat', 'leucine-rich repeats', 'lrr' and 'lrrs' were used as key words in the PDB quick search; second, 'SCOP classification -> Alpha and beta proteins (a/b) -> Leucine-rich repeat' was used as options in

PDB advanced search; third, 'CATH classification -> Alpha Beta -> Alpha-Beta Horseshoe -> Leucine-rich repeat' was used as options in PDB advanced search. Because of the irregularity (mutations and insertions in the sequence) of LRRs reliable identifications of LRRs contained in the LRR proteins could only be performed manually. We inspected the three-dimensional structures of the LRR proteins using molecular viewers and identified each LRR based on two criteria:

1. A LRR begins at the beginning of the highly conserved segment (HCS) and ends at the end of the variable segment (VS) (just before the HCS of the next LRR).

2. The HCS of a LRR must pose a typical conformation, i.e. a short β-sheet begins at about position 3 and a hydrophobic core is formed by the four L residues at position 1, 4, 6, and 11.

The LRRs were then manually classified according to the consensus sequences [4, 5]. In addition to the eight canonical LRR classes listed in the background section we included a new class 'other' for the N-/C-terminal LRRs and some hyper-irregular LRRs. Table 1 illustrates the consensus sequences of the eight canonical LRR classes.

During the LRR identification and classification all sequence insertions longer than 3 residues were annotated. About one tenth of entries have insertions longer than 3 residues while few entries have deletions, which suggests that the evolution of LRRs may prefer insertion to deletion.

The LRRML release 0.2 contains 1261 LRR entries from 112 PDB structures. Among them 548 LRRs are distinct on sequence level, indicating that different molecules can share identical LRRs. By superimposition, we found that they also have highly similar structures. This fact enhances the confidence in modeling LRR proteins using multiple LRR templates. A histogram of entry length distribution

Table 1: Consensus sequences of the eight canonical LRR classes [4, 5].

Classes	HCS	VS
Typical type (T)	LxxLxLxxNxL	xxLxxxxLxxLxx
Bacterial type (S)	LxxLxLxxNxL	xxLPx(x)LPxx
Ribonuclease inhibitor-like type (RI)	LxxLxLxxNxL	xxxxxxxLxxxLxxxxx
SDS22-like type (SDS22)	LxxLxLxxNxL	xxLxxLxxLxx
Cysteine-containing type (CC)	LxxLxLxxCxxL	TDxxxxxLxxxCxx
Plant-specific type (PS)	LxxLxLxxNxL	xxxLPxxLGxLxx
Treponema pallidum type (Tp)	LxxLxLPxxLxx	LxxxAFxxCxx
CD42b type (CD42b)	LxxLxLxxNxL	xxLPxxxxxxxx

L: Leu, Ile, Val, Phe; N: Asn, Thr, Ser, Cys; P: Pro; T: Thr; D: Asp; G: Gly; A: Ala; F: Phe; C: Cys; x: random residues.

(Figure 1) shows that the LRR lengths are concentrated in the interval from 20 to 29, which covers the characteristic lengths of consensus sequences of the eight canonical LRR classes. Some entries have a sequence longer than 30, because they contain large insertions. Table 2 presents the distribution of LRR entries and PDB entries over the nine classes respectively. The classification results are consistent with a previous report which showed that LRRs from different classes never occur simultaneously in the same protein and have most probably evolved independently [4]. Exceptions to this rule are the T and S types which often exist in the same protein forming the super motif 'STT' [12]. It is assumed that both evolved from a common precursor [1].

Currently, there are several protein databases containing information on LRRs, such as Pfam [13], InterPro [14], SMART [15] and Swiss-Prot [16]. These databases predict the LRR numbers and boundaries for their LRR protein entries by various computational methods, no matter whether the entries have known three-dimensional

structures or not, thereby 'false negative' occurs frequently. Table 3 lists the numbers of structure-known LRR proteins and their LRRs covered by these databases. As more detailed examples, LRR numbers of LRR proteins from different classes reported by the established databases are compared in Table 4. Additionally, the individual three-dimensional LRR structures are not directly available from these databases. In order to combine the information required for homology modeling and structural analysis, LRRML is provided with three prominent characteristics:

1. Each database entry is an individual three-dimensional LRR structure, which was identified with high accuracy.

2. Extensive annotations, such as systematic classification, secondary structures, HCS/VS partitions and sequence insertion, are provided.

3. LRRs were extracted from all structure-known LRR protein structures from PDB.
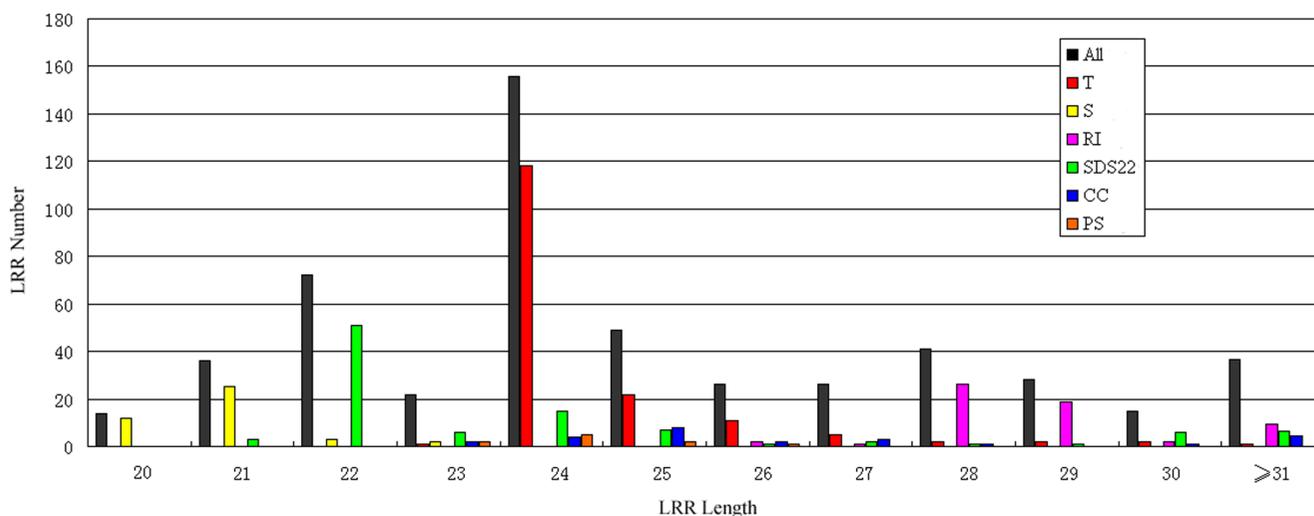
Figure 1
LRR entry length distribution. The most common entry lengths vary from 20 to 29. Each LRR class has a characteristic length distribution. Some entries have a sequence length larger than 29 due to insertions.

Table 2: Numbers of LRR and PDB entries (release 0.2) in the nine LRR classes.

	T	S	RI	SDS22	CC	PS	Tp	CD42b	Other	Total
LRR structures	272	72	151	372	184	10	0	0	200	1261
LRR entries	169	40	59	114	28	10	0	0	128	548
PDB entries		32	13	50	16	1	0	0	-	112

Up to present, no crystal structures for LRR proteins of Tp/CD42b types are determined. Different from other LRR types, the S type and T type LRRs evolved from a common precursor [1] and thus can exist in the same PDB entry simultaneously.

Table 3: Coverage of LRR proteins with PDB structures of different databases.

Databases	Numbers of LRR proteins with PDB structures	Numbers of identified LRRs
InterPro	62	325
Swiss-Prot	98	997
Pfam	48	173
SMART	84	547
LRRML	112	1261

The results were obtained on October 13, 2008.

XML description

The extensible markup language (XML) was standardized in the 90s and is well established as a format for hierarchical data. It can be queried and parsed more easily by application programs. Therefore, more and more biological databases use the XML as data saving format and database management system (DBMS) [17-19]. LRRML was designed by using eXist [20], an XML DBMS, and using XPath/XQuery [21] for processing queries and web forms. We developed a LRR markup language (LRRML) for exchanging and storing LRR structures. It consists of four blocks of information:

1. The sequence information (XML tag <l:Sequence>): amino acid sequence and sequence length.

2. The classification information (XML tag <l:Type>): class name and consensus sequences.

3. The sequence partitions (XML tag <l:Regions>): amino acid sequence, position, length and insertion of HCS and VS.

4. The corresponding PDB sources (XML tag <l:Sources>): ID, chain, LRR number and classification of the source PDB entries; serial number, position, DSSP [22] secondary structure and three-dimensional coordinates of the current LRR in these source PDB entries.

An example describing the LRR3 from PDB entry 2O6S is shown in Figure 2. The document type definition (DTD) file of LRRML is provided asAdditional file 1.

Utility

Web application

The entire database can be browsed by LRR IDs or by PDB IDs. When browsing, the entries appear in a summary table containing at first ID, type and sequence. Clicking on an ID opens an XML Stylesheet (XSLT) [21] converted HTML web page that presents the entry in detail. The original XML file and the coordinates file in PDB format can also be downloaded. The XSLT file used is provided as Additional file 2. Aside from the textual view, a LRR structure can be visualized by the online molecular viewer Jmol [23]. After loading, users can change the view settings flexibly by themselves. LRRML is provided with various search functions, including PDB ID search which returns all LRRs contained in this PDB structure, class search which returns all LRRs of this class, or length search which returns all LRRs with this sequence length. To simplify the homology modeling, the similarity search was implemented. It returns the structures of the most similar LRRs for a structure-unknown LRR. The target LRR sequence can be searched against the entire database, a certain LRR class or LRRs with a certain length. At first, a global pair wise sequence alignment with sequence identity will be generated for the target LRR and each of the LRRs in the user selected set. Then, the most similar LRRs will be returned as template candidates, ranked by sequence identity.

The DBMS provides a REST-style application programming interface (API) through HTTP, which supports GET and POST requests. A unique resource identifier (URI) 'http://zeus.krist.geo.uni-muenchen.de:8081/exist/rest/...' is treated by the server as path to a database collection. Also, request parameters can help select any required elements.

Table 4: Comparison of LRR numbers of different LRR proteins by different databases.

PDB codes	Protein functions	LRR classes	InterPro	Swiss-Prot	Pfam	SMART	LRRML
2A0Z	Immune System	T	18	22	7	20	25
1G9U	Toxin	S	7	15	1	0	15
2FT3	Structural Protein	T+S	8	8	5	9	12
1K5D	Signaling Protein	RI	2	8	0	0	11
1GWB	Glycoprotein	SDS22	6	6	4	7	8
2PIM	Signaling Protein	CC	2	16	0	6	18
1OGQ	Inhibitor	PS	7	10	2	0	10

All listed LRR numbers include N-/C-terminal LRRs. To date, only the LRRML database contains the complete set of LRRs of all LRR proteins with known structures. The results were obtained on October 13, 2008.

```xml
<l:LRR xmlns:l="http://zeus.krist.geo.uni-muenchen.de/~lrrml">
  <l:Sequence>
    <l:SLength>24</l:SLength>
    <l:SPrimary>LTQLYLGGNKLQSLPNGVFNKLTS</l:SPrimary>
  </l:Sequence>
  <l:Type>
    <l:TName>Typical type</l:TName>
    <l:TAbbr>T</l:TAbbr>
    <l:TPattern>
      <l:TPLen>24</l:TPLen>
      <l:TPSeq>LxxLxLxxNxLxxLxxxxLxxLxx</l:TPSeq>
      <l:TPHCS>LxxLxLxxNxL</l:TPHCS>
      <l:TPVS>xxLxxxxLxxLxx</l:TPVS>
    </l:TPattern>
  </l:Type>
  <l:Regions>
    <l:Region>
      <l:RName>Highly Conserved Segment</l:RName>
      <l:RNAbbr>HCS</l:RNAbbr>
      <l:RBegin>1</l:RBegin>
      <l:REnd>11</l:REnd>
      <l:RLength>11</l:RLength>
      <l:RSeq>LTQLYLGGNKL</l:RSeq>
    </l:Region>
    <l:Region>
      <l:RName>Variable Segment</l:RName>
      <l:RNAbbr>VS</l:RNAbbr>
      <l:RBegin>12</l:RBegin>
      <l:REnd>24</l:REnd>
      <l:RLength>13</l:RLength>
      <l:RSeq>QSLPNGVFNKLTS</l:RSeq>
    </l:Region>
  </l:Regions>
  <l:Sources>
    <l:Source>
      <l:PDBInfo>
        <l:PDBId>2O6S</l:PDBId>
        <l:PDBChain>A</l:PDBChain>
        <l:PDBLrrNum>7</l:PDBLrrNum>
        <l:PDBClass>Immune System</l:PDBClass>
      </l:PDBInfo>
      <l:LRRInfo>
        <l:LRRNumber>3</l:LRRNumber>
        <l:LRRBegin>77</l:LRRBegin>
        <l:LRREnd>100</l:LRREnd>
        <l:LRRSecStr>LLEEELLLLLLLLLLLLLLLLLLLL</l:LRRSecStr>
        <l:LRR3dStr>2O6S.A.7.3</l:LRR3dStr>
      </l:LRRInfo>
    </l:Source>
  </l:Sources>
</l:LRR>
```

Figure 2
The LRRML description of a LRR structure. This entry is a 24 residue long typical LRR. The first 11 residues compose its HCS and the last 13 residues compose its VS (no insertions). It is contained only in the chain A of PDB structure 2O6S (a protein involved in the immune system). It is the third one of the 7 LRRs of 2O6S, from position 77 to 100. Its secondary structure was extracted from DSSP and its three-dimensional coordinate file is available though the hyperlink on the corresponding web page.

Table 5: Sequence identities (%) of target-template LRR pairs.

	NT	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	CT	Avg
LRRML ID	406	65	212	465	151	177	110	259	8	203	64	293	270	357	65	152	259	316	152	239	239	92	80	101	173	—
PDB source	1XWD	2O6S	1G9U	1OZN	1XKU	1SQ0	2FT3	2V9S	1IO0	1H6U	2O6S	2Z64	2Z62	2Z81	2O6S	1XKU	2V9S	2Z7X	1XKU	1P8V	1P8V	2ID5	2O6Q	2ID5	1V8A	—
Identity (%)	47.60	45.83	45.83	41.67	50.00	50.00	46.15	41.67	41.38	42.31	50.00	33.33	41.67	42.86	50.00	40.00	37.50	38.46	50.00	45.83	50.00	41.67	40.00	39.29	44.12	

In the header line, 1–25 denote canonical LRRs; NT and CT denote N-/C-terminal LRRs.
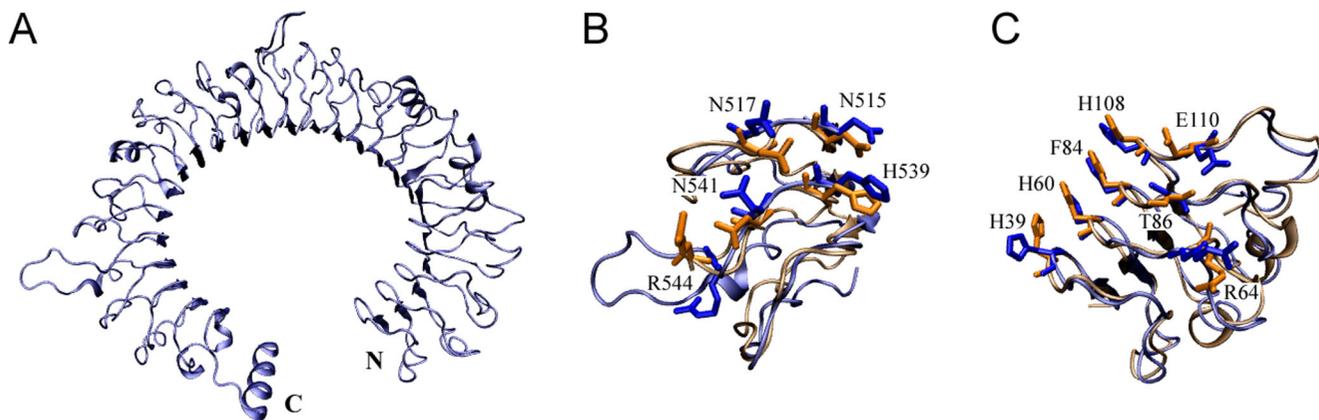
Figure 3
Comparison of model and crystal structure of mouse TLR3 ectodomain at the two ligand interaction regions. Blue: structure obtained by homology modeling; orange: crystal structure (PDB code: 3CIG). (A) The modeled backbone structure of mouse TLR3 ectodomain. (B) Model and crystal structure superimposed at the N-terminal interaction region. The root mean square deviation is 1.96 Å. (C) Superimposition at the C-terminal interaction region. The root mean square deviation is 1.9 Å. The reported interacting residues are presented with side chain and labelled with residue name and position in (B) and (C).
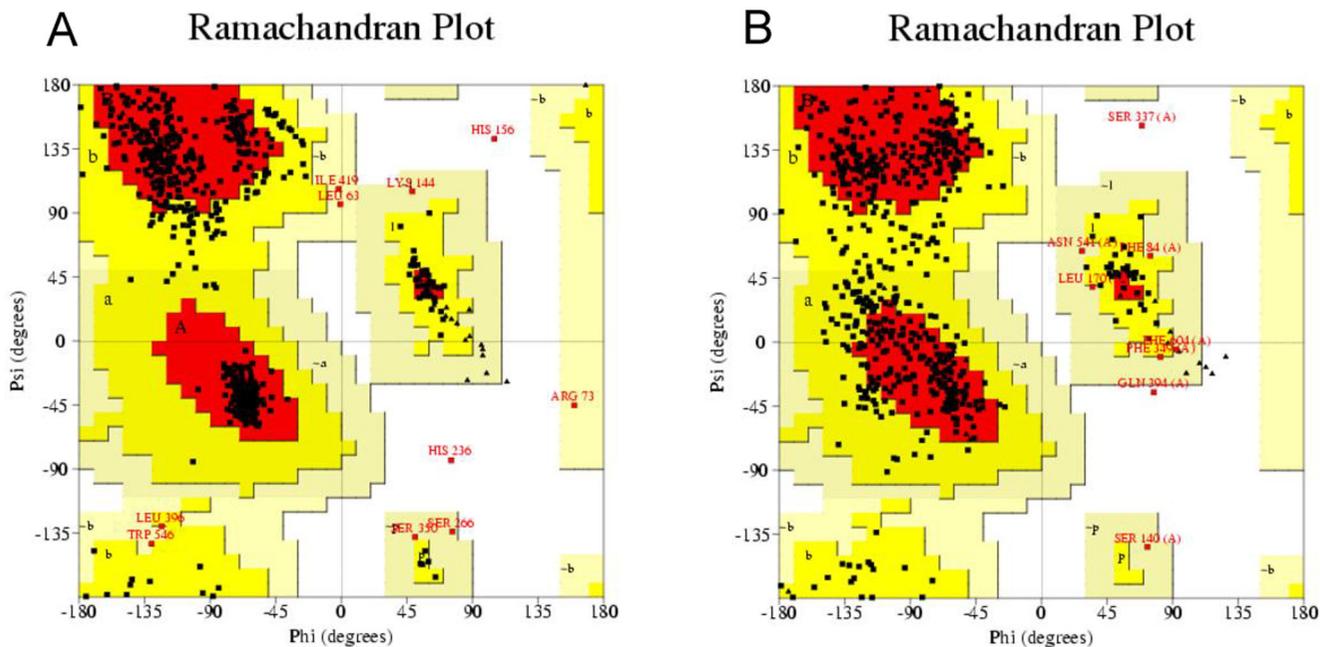
Figure 4
Ramachandran plot of model and crystal structure of mouse TLR3 ectodomain. (A) Predicted model of mouse TLR3 ectodomain. (B) Crystal structure of mouse TLR3 ectodomain. The different colored areas indicate 'disallowed' (white), 'generously allowed' (light yellow), 'additional allowed' (yellow), and 'most favored' (red) regions.

For example, '_query' executes a specified XPath/XQuery; the URL "http://zeus.krist.geo.uni-muenchen.de:8081/exist/rest/db/lrrml?_query=//LRR [.//TAbbr='S']" returns all the S type LRRs.

Application in homology modeling

LRRML was designed as a tool for template selection in homology modeling of LRR proteins. Traditionally, the template used in homology modeling is one or more full length protein structures obtained via similarity search. Nevertheless, due to the different repeat numbers and arrangements of LRRs, the sequence identity between the target and the full length template is usually not high enough for homology modeling. With LRRML the most similar structure-known LRR can be found for each LRR in the target sequence as a local template. The combination of all local templates through multiple alignments helps to achieve a high sequence identity to the target.

As test case we modeled the structure of mouse Toll-like receptor 3 (TLR3) ectodomain. We assumed that the structure of mouse TLR3 ectodomain were unknown and excluded the LRRs of mouse/human TLR3 ectodomain from LRRML. Through similarity search the optimal template for each of the 25 LRRs in mouse TLR3 was found. The sequence identity between each LRR pair (target/template LRR) is listed in Table 5. Then a 26-line multiple alignment was generated by the 25 template sequences and the target sequence as the input of MODELLER 9v3 [24]. The resulting three-dimensional model (Figure 3A) was evaluated by PROCHECK [25], with 98.2% residues falling into the most favored or allowed regions of the main chain torsion angles distribution, whereas the result of the TLR3 crystal structure (PDB code: 3CIG) was 98.6% (Figure 4). The mouse TLR3 has been shown to bind double-stranded RNA ligand with both N-terminal and C-terminal sites on the lateral side of the convex surface of TLR3 [26]. The N-terminal interaction site is composed of LRRNT and LRR1-3, and the C-terminal site is composed of LRR19-21. We superimposed the resulting model onto the crystal structure of mouse TLR3 ectodomain at the two interaction sites by using SuperPose v1.0 [27]. The root mean square deviations of the structures are 1.96 Å and 1.9 Å respectively (Figure 3B/C), indicating that the predicted model sufficiently well matched the crystal structure and was useful for prediction of ligand interaction sites. These results demonstrate that homology modeling using combined multiple templates obtained from LRRML can create valuable information to trigger further biochemical research. Interpretation of structural details, however, should be done exercising due care.

Conclusion

A specialised conformational leucine-rich repeats database called LRRML has been developed. It is supported by an XML database management system and can be searched and browsed with either an easy-to-use web interface or REST like interface. The interface is suitable for most graphical web browsers and has been tested on the Windows, Mac and Linux operating systems. LRRML contains individual three-dimensional LRR structures with manual structural annotations. It presents useful sources for homology modeling and structural analysis of LRR proteins. Since the amount of structure-determined LRR proteins constantly increases, we plan to update LRRML every 2 to 3 months.

Availability and requirements

This database is freely available at http://zeus.krist.geo.uni-muenchen.de/~lrrml.

Authors' contributions

TW and JG drafted the manuscript, extracted the data, compiled the database, wrote the code for the web interface and performed the statistical analysis. FJ, WMH, RWS and SCR conceived of the study, built the database server, participated in the database design and coordination and helped to draft the manuscript. TW and JG should be regarded as joint first authors. All authors read and approved the final manuscript.

Additional material

Additional file 1
The document type definition (DTD) file of LRRML.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1472-6807-8-47-S1.dtd]

Additional file 2
The XML Stylesheet (XSLT) of LRRML.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1472-6807-8-47-S2.xsl]

References

1. Matsushima N, Tanaka T, Enkhbayar P, Mikami T, Taga M, Yamada K and Kuroki Y: **Comparative sequence analysis of leucine-rich repeats (LRRs) within vertebrate toll-like receptors.** *BMC Genomics* 2007, **8**:124–143.
2. Dolan J, Walshe K, Alsbury S, Hokamp K, O'Keeffe S, Okafuji T, Miller SFC, Guy Tear G and Mitchell KJ: **The extracellular Leucine-Rich Repeat superfamily; a comparative survey and analysis of evolutionary relationships and expression patterns.** *BMC Genomics* 2007, **8**:320–343.

3. Kobe B and Deisenhofer J: **Crystal structure of porcine ribonuclease inhibitor, a protein with leucine-rich repeats.** *Nature* 1993, **366:**751–756.
4. Kobe B and Kajava AV: **The leucine-rich repeat as a protein recognition motif.** *Curr Opin Struct Biol* 2001, **11:**725–732.
5. Bell JK, Mullen GE, Leifer CA, Mazzoni A, Davies DR and Segal DM: **Leucine-rich repeats and pathogen recognition in Toll-like receptors.** *Trends Immunol* 2003, **24:**528–533.
6. Kajava AV: **Structural Diversity of Leucine-rich Repeat Proteins.** *J Mol Biol* 1998, **277:**519–527.
7. Stumpp MT, Forrer P, Binz HK and Plckthun A: **Designing Repeat Proteins: Modular Leucine-richRepeat Protein Libraries Based on the Mammalian Ribonuclease Inhibitor Family.** *J Mol Biol* 2003, **332:**471–487.
8. Kubarenko A, Frank M and Weber AN: **Structure-function relationships of Toll-like receptor domains through homology modelling and molecular dynamics.** *Biochem Soc Trans* 2007, **35:**1515–1518.
9. Rössle SC, Bisch PM, Lone YC, Abastado JP, Kourilsky P and Bellio M: **Mutational analysis and molecular modeling of the binding of Staphylococcus aureus enterotoxin C2 to a murine T cell receptor Vbeta10 chain.** *Eur J Immunol* 2002, **32:**2172–2178.
10. Hazai E and Bikádi Z: **Homology modeling of breast cancer resistance protein (ABCG2).** *J Struct Biol* 2008, **162:**63–74.
11. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and Bourne PE: **The Protein Data Bank.** *Nucleic Acids Res* 2000, **28:**235–242.
12. Matsushima N, Kamiya M, Suzuki N and Tanaka T: **Super-Motifs of Leucine-Rich Repeats (LRRs) Proteins.** *Genome Inform* 2000, **11:**343–345.
13. Finn RD, Tate J, Mistry J, Coggill PC, Sammut JS, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL and Bateman A: **The Pfam protein families database.** *Nucleic Acids Res* 2008, **36:**D281–288.
14. Mulder NJ and Apweiler R: **InterPro and InterProScan: tools for protein sequence classification and comparison.** *Methods Mol Biol* 2007, **396:**59–70.
15. Schultz J, Copley RR, Doerks T, Ponting CP and Bork P: **SMART: a web-based tool for the study of genetically mobile domains.** *Nucleic Acids Res* 2000, **28:**231–234.
16. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M, Martin MJ, Mazumder R, O'Donovan C, Redaschi N and Suzek B: **The Universal Protein Resource (UniProt): an expanding universe of protein information.** *Nucleic Acids Res* 2006, **34:**D187–191.
17. Heida N, Hasegawa Y, Mochizuki Y, Hirosawa K, Konagaya A and Toyoda T: **TraitMap: an XML-based genetic-map database combining multigenic loci and biomolecular networks.** *Bioinformatics* 2004, **20 Suppl 1:**i152–i160.
18. Kunz H, Derz C, Tolxdorff T and Bernarding J: **XML knowledge database of MRI-derived eye models.** *Comput Methods Programs Biomed* 2004, **73:**203–208.
19. Jiang K and Nash C: **Application of XML database technology to biological pathway datasets.** *Conf Proc IEEE Eng Med Biol Soc* 2006, **1:**4217–4220.
20. eXist-db: an open source database management system. http://exist-db.org.
21. The World Wide Web Consortium. http://www.w3.org.
22. Kabsch W and Sander C: **Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.** *Biopolymers* 1983, **22:**2577–2637.
23. Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org.
24. Fiser A, Do RK and Sali A: **Modeling of loops in protein structures.** *Protein Sci* 2000, **9:**1753–1773.
25. Laskowski RA, MacArthur MW, Moss DS and Thornton JM: **PROCHECK: a program to check the stereochemical quality of protein structures.** *J Appl Cryst* 1993, **26:**283–291.
26. Liu L, Botos I, Wang Y, Leonard JN, Shiloach J, Segal DM and Davies DR: **Structral basis of Toll-like receptor 3 signaling with double-stranded RNA.** *Science* 2008, **320:**379–381.
27. Maiti R, Van Domselaar GH, Zhang H and Wishart DS: **SuperPose: a simple server for sophisticated structural superposition.** *Nucleic Acids Res* 2004, **32:**W590–594.

Paper 3

Inhibition of the Toll-like receptors TLR4 and 7 signaling pathways by SIGIRR: a computational approach.

Jing Gong, Tiandi Wei, Robert W. Stark, Ferdinand Jamitzky, Wolfgang M. Heckl, Hans-Joachim Anders, Maciej Lech and Shaila C. Rössle

Inhibition of Toll-like receptors TLR4 and 7 signaling pathways by SIGIRR: A computational approach

Jing Gong [a,b], Tiandi Wei [a,b,*], Robert W. Stark [a,b], Ferdinand Jamitzky [a,c], Wolfgang M. Heckl [a,d,e], Hans J. Anders [f], Maciej Lech [f], Shaila C. Rössle [b]

[a] Center for Nanoscience, Ludwig-Maximilians-Universität München, 80799 Munich, Germany
[b] Department of Earth and Environmental Sciences, Ludwig-Maximilians-Universität München, 80333 Munich, Germany
[c] Leibniz Supercomputing Centre, 85748 Garching, Germany
[d] TUM School of Education, Technische Universität München, 80799 Munich, Germany
[e] Deutsches Museum, 80538 Munich, Germany
[f] Medizinische Poliklinik, Ludwig-Maximilians-Universität München, 80336 Munich, Germany

ARTICLE INFO

ABSTRACT

Toll-like receptors (TLRs) belong to the Toll-like receptor/interleukin-1 receptor (TLR/IL-1R) superfamily which is defined by a common cytoplasmic Toll/interleukin-1 receptor (TIR) domain. TLRs recognize pathogen-associated molecular patterns and initiate an intracellular kinase cascade to trigger an immediate defensive response. SIGIRR (single immunoglobulin interleukin-1 receptor-related molecule), another member of the TLR/IL-1R superfamily, acts as a negative regulator of MyD88-dependent TLR signaling. It attenuates the recruitment of MyD88 adaptors to the receptors with its intracellular TIR domain. Thus, SIGIRR is a highly important molecule for the therapy of autoimmune diseases caused by TLRs. So far, the structural mechanism of interactions between SIGIRR, TLRs and adaptor molecules is unclear. To develop a working hypothesis for this interaction, we constructed three-dimensional models for the TIR domains of TLR4, TLR7, MyD88 and SIGIRR based on computational modeling. Through protein–protein docking analysis, we developed models of essential complexes involved in the TLR4 and 7 signaling and the SIGIRR inhibiting processes. We suggest that SIGIRR may exert its inhibitory effect through blocking the molecular interface of TLR4, TLR7 and the MyD88 adaptor mainly via its BB-loop region.

© 2009 Elsevier Inc. All rights reserved.

1. Introduction

Toll-like receptors (TLRs) are essential for the innate immune system because they recognize molecules, such as single-stranded RNA or CpG DNA that are associated with pathogens. Such nuclear antigen-recognizing receptors are also important in the autoimmune disease systemic lupus erythematosus. The disease progresses as a consequence of the recognition of self nucleic acids by TLRs (Rahman and Eisenberg, 2006). For the future development of therapeutic approaches it is important to understand possible TLR inhibition mechanisms from a structural point of view.

TLRs belong to the Toll-like receptor/interleukin-1 receptor (TLR/IL-1R) superfamily, which is defined by the presence of a conserved cytoplasmic Toll/interleukin-1 receptor (TIR) domain (Bowie and O'Neill, 2000) connected to an ectodomain through a single transmembrane stretch. To date, 13 TLRs have been identi-

fied in mammals. Their ectodomains consist of 16–28 leucine-rich repeats (LRRs). These LRRs provide a variety of structural frameworks for the binding of protein and non-protein ligands including lipopolysaccharide (LPS), lipopeptide, CpG DNA, flagellin, imidazoquinoline and double-/single-stranded RNA (Gay and Gangloff, 2007). TLRs are capable of recognizing ligands in a dimer form (Latz et al., 2007; Liu et al., 2008; Park et al., 2009; Peter et al., 2009; Wei et al., 2009). Upon receptor activation, an intracellular TIR signaling complex is formed between the receptor and downstream adaptor TIR domains (O'Neill and Bowie, 2007). MyD88 (Myeloid differentiation primary response protein 88) was the first intracellular adaptor molecule characterized among all known adaptors in the TLR signaling (Takeda and Akira, 2004). It consists of an N-terminal death domain (DD) separated from its C-terminal TIR domain by a linker sequence. MyD88 also forms a dimer through DD-DD and TIR-TIR domain interactions when recruited to the receptor complex (Burns et al., 1998). MyD88 can recruit IRAK (IL-1RI-associated protein kinases) through its DD to continue signaling and, finally, to induce the nuclear factor-κB (NF-κB) leading to the expression of type I interferons. Although the MyD88-

dependent pathway is common to most TLRs, TLR3 exclusively uses TRIF (TIR domain-containing adapter inducing interferon-β) for signaling (MyD88-independent) while the TLR4 can signal via both pathways (Takeda and Akira, 2004).

SIGIRR (Single immunoglobulin interleukin-1 receptor-related molecule), also known as TIR8, was initially identified as an Ig domain-containing receptor of the TLR/IL-1R superfamily (Thomassen et al., 1999). Both the extracellular and intracellular domains of SIGIRR differ from those of other Ig domain-containing receptors, as its single extracellular Ig domain does not support ligand-binding. Its intracellular TIR domain cannot activate NF-κB because it lacks two crucial amino acids, Ser447 and Tyr536. Moreover, the TIR domain of SIGIRR extends that of the typical TLR/IL-1R superfamily member by more than 73 amino acids at the C-terminal (C-tail) (Thomassen et al., 1999). Instead, SIGIRR acts as an endogenous inhibitor for MyD88-dependent TLR and IL-1R signaling. This behavior was shown by over expression of SIGIRR in Jurkat or HepG2 cells which showed substantially reduced LPS, CpG DNA or IL-1-induced activation of NF-κB (Polentarutti et al., 2003; Qin et al., 2005; Wald et al., 2003). Thus, SIGIRR has attracted tremendous research interest because of its regulating function in cancer-related inflammation and autoimmunity (Lech et al., in press). For example, systemic lupus erythematosus is caused by TLR7-mediated induction of type I interferons. Compared with wild type mice *Sigirr*-deficient mice develop excessive lymphoproliferation when introduced into the context of a lupus susceptibility gene (Lech et al., 2008). Although the significance of SIGIRR has been widely acknowledged, its inhibition mechanism remains unclear owing to a lack of structural information.

Mutagenesis studies investigated three deletion mutants of SIGIRR (Qin et al., 2005): ΔN (lacking the extracellular Ig domain), ΔTIR (lacking the intracellular TIR domain) and ΔC (lacking the C-tail of the TIR domain with deletion of residues 313–410). The results showed that only the TIR domain (excluding the C-tail part) is necessary for SIGIRR to inhibit TLR4 signaling (Qin et al., 2005). Nevertheless, detailed structural interaction mechanisms of SIGIRR's TIR domain are still missing.

So far, the structures of the TIR domains of human TLR1, 2, 10 and IL-1RAPL have been determined by X-ray crystallography (Khan et al., 2004; Nyman et al., 2008; Xu et al., 2000). The TLR1 and 2 modules occur as monomers in solution and the packing of the molecules in the crystal lattice does not suggest a likely arrangement for a functional dimer. In contrast, the TLR10 and IL-1RAPL TIR domains were present as homodimers. Although they demonstrate different dimer conformations, a highly conserved BB-loop region plays a crucial role in both dimer interfaces. Using this information, we have constructed three-dimensional models for TIR domains of TLR4, TLR7, MyD88 and SIGIRR by homology modeling and protein threading. Models of essential molecular complexes involved in the TLR4 and 7 signaling pathways and the SIGIRR inhibiting process are proposed and compared based on results of protein–protein docking studies. We chose TLR4 as an additional example to elucidate the mechanisms involved in the negative regulation of the MyD88-dependent TLR signals by SIGIRR, because different mechanisms of TLR4 recognizing LPS and TLR7 recognizing single-stranded RNA may lead to different structural interactions of receptor with SIGIRR, which enables further insight into the molecular interaction.

2. Methods

2.1. Templates identification and sequence alignments

Amino acid sequences of the target proteins, human TLR4 (GenBank Accession No. O00206), TLR7 (Q9NYK1), MyD88 (AAC50954) and SIGIRR (CAG33619) were extracted from the NCBI protein database (Wheeler et al., 2008). Three-dimensional models of TLR4 (Asn672-Ala814), TLR7 (Cys889-Asp1036), MyD88 (Glu159-Pro296) and SIGIRR (Tyr165-Pro308, without the C-tail Arg309-Ser392) were constructed by homology modeling. Due to the homology of the target proteins, four common templates were obtained via BLAST search against the Protein Data Bank (PDB) (Berman et al., 2000). They were TLR1 (PDB code: 1FYV), TLR2 (1FYW), TLR10 (2J67) and IL-1RAPL (1T3G). Multiple sequence alignment of each target with the templates was generated with MUSCLE (Edgar, 2004) and analyzed with Jalview (Clamp et al., 2004). Because the secondary structure of the TIR domain is composed of well-organized alternating β-strands and α-helixes, we adjusted the alignments manually according to the secondary structure information to improve the alignment quality. The secondary structure of each target was predicted by PSIPRED (Bryson et al., 2005). In addition, the C-terminal tail of the TIR domain, which is unique to SIGIRR, has no homologue of known structure to serve as a template. In this case we employed the protein threading method THREADER 3.5 (Jones et al., 1995) to determine a template structure. The selected template was NSF-N (N-terminal domain of N-ethylmaleimide sensitive factor, PDB code: 1QCS).

2.2. Model construction and validation

The initial three-dimensional coordinates of the models were generated by the fully automated program MODELLER 9v3 (Fiser et al., 2000). The input files for each model were a 5-line multiple alignment file (one target and four templates) and coordinate files of the templates. During modeling, gap regions in the alignment produced 3–8 residue-long loop structures in the model, which deteriorated the model's accuracy. ModLoop (Fiser and Sali, 2003) was used to rebuild the coordinates of these loop regions. ModLoop optimizes the positions of non-hydrogen atoms of a loop (shorter than 20 residues) relying on a protocol consisting of a conjugate gradient minimization and a molecular dynamics simulation. Finally, we used the model quality assessment programs ProQ (Wallner and Elofsson, 2003), ModFOLD (McGuffin, 2008) and MetaMQAP (Pawlowski et al., 2008) to evaluate the output candidate models and select the most reliable one.

2.3. Model docking

Unrestrained pairwise model docking included eight complexes of TIR domains: TLR4-TLR4, TLR7-TLR7, MyD88-MyD88, TLR4 dimer-MyD88 dimer (tetramer), TLR7 dimer-MyD88 dimer (tetramer), TLR4-SIGIRR, TLR7-SIGIRR and MyD88-SIGIRR. We used GRAMM-X (Tovchigrechko and Vakser, 2006) and ZDOCK (Chen et al., 2003), which are widely accepted rigid-body protein–protein docking programs, to predict and assess the interactions between these complexes. Both programs rank the 10 most probable predictions out of thousands of candidates based on geometry, hydrophobicity and electrostatic complementarity of the molecular surface. We then selected the most reasonable solution from these top 10 lists in consideration of further qualifications. Briefly, these qualifications included residue conservation of the interaction sites, steric compatibility of the amino acid linker to the transmembrane helix, and knowledge from published articles (Bell et al., 2006; Loiarro et al., 2007; Nunez Miguel et al., 2007; Nyman et al., 2008; Park et al., 2009; Poltorak et al., 1998). The buried surface interaction area of dimer models were calculated with the protein interfaces, surfaces and assemblies service (PISA) at the European Bioinformatics Institute (Krissinel and Henrick, 2007).

3. Results

3.1. Molecular modeling of TIR domains

In the secondary structure-aided alignments for the homology modeling, the average target-template sequence similarity of TLR4, TLR7, MyD88 and SIGIRR was 51.7%, 50.4%, 44.5% and 42.7%, respectively (detailed in Table 1). The resulting structures exhibit a typical TIR domain conformation in which a central five-stranded parallel β-sheet (βA-βE) is surrounded by a total of five α-helixes (αA–αE) on both sides (Fig. 1A). The loops are named by the letters of the secondary structure elements that they connect. For example, the BB-loop connects β-strand B and α-helix B. The structure of NSF-N was identified as a template for SIGIRR's C-tail through protein threading. This template was first-ranked by

THREADER according to the energy Z-score (Z = 2.7: borderline significant (Jones et al., 1995)). The C-tail contains a four-stranded parallel β-sheet with an α-helix and several loop structures on one side, while the other side points to SIGIRR's TIR (Fig. 1A). These results suggest that the TIR domain and the C-tail of SIGIRR are not an integrative structure, but two interconnected individual modules. There is a 3 residue-long short linker (Leu307-Arg309) between the last secondary structure αE of SIGIRR's TIR and the first secondary structure βA of SIGIRR's C-tail. Therefore, the C-tail can only be situated next to the αE of the TIR domain. Further evaluation of the models involved analysis of geometry, stereochemistry and energy distributions of the molecules. The evaluation results (Table 2) indicate high quality for all models in terms of overall packing.

Multiple sequence alignment of TIR domains from different molecules detected seven conserved boxes in the TIR domain (Fig. 1B). Our models show that they correspond to β-strand A (βA), β-strand B (βB), BB-loop, β-strand C (βC), β-strand D (βD), β-strand E (βE) and α-helix E (αE). Functional significance can usually be observed in conserved regions. Nevertheless, the five β-strands (boxes 1, 2, 4–6) are embedded structures that form a hydrophobic core of the TIR domain and hence are not likely to interact with other molecules. Also, the αE (box 7) of SIGIRR is blocked because it is linked to the C-tail. In this vein, the BB-loop (box 3) and αE of TLR4, TLR7 and MyD88, along with the BB-loop of SIGIRR, may be

Table 1
Protein sequence similarities (%) between targets and templates.

	TLR1	TLR2	TLR10	IL-1RAPL	Avg
TLR4	53.4	57.8	51.4	44.2	51.7
TLR7	51.0	55.8	49.3	45.6	50.4
MyD88	44.5	45.3	40.6	47.4	44.5
SIGIRR	41.8	42.3	37.7	49.0	42.7
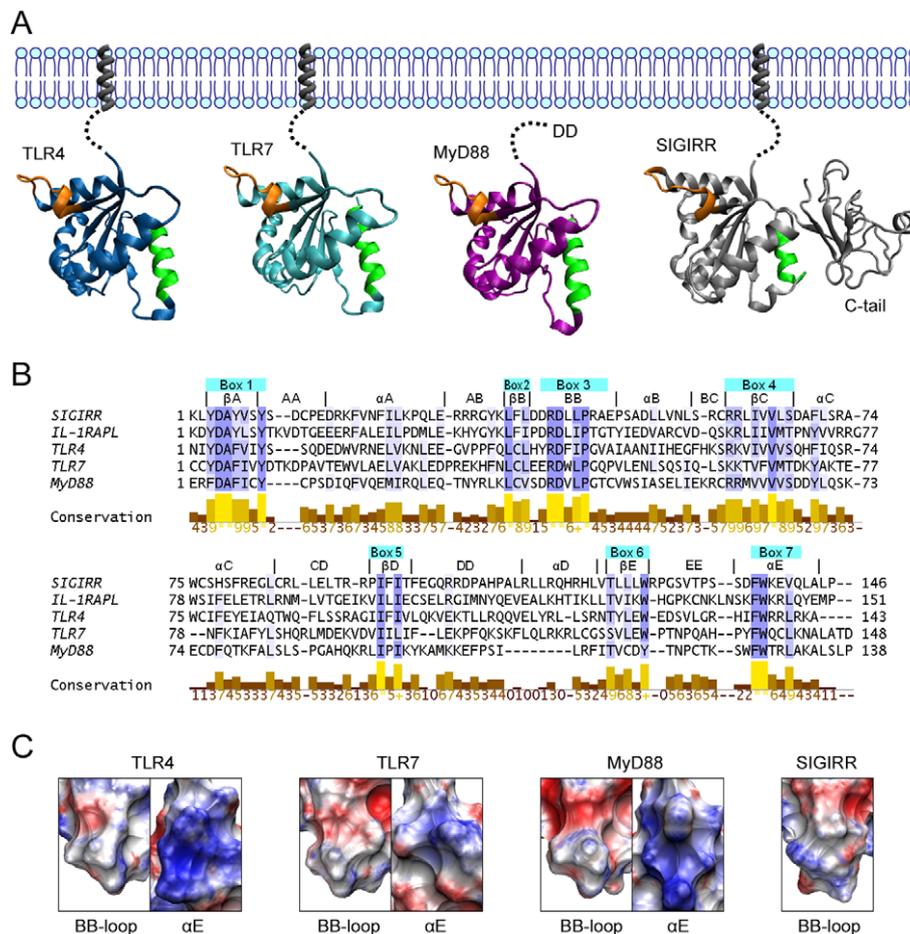
Fig. 1. Three-dimensional structures and conserved regions of TIR domains of TLR4, TLR7, MyD88 and SIGIRR. (A) The BB-loop and αE regions are highlighted in orange and green respectively. As there is a three residue-long linker between the two modules of SIGIRR's TIR, the orientation of the C-tail as shown here is speculative. (B) Multiple sequence alignment of different TIRs indicates seven conserved boxes. (C) Surface charge distribution (APBS electrostatics) of BB-loop and αE with red indicating areas of negative charge and blue indicating positive charge. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Model evaluation. ProQ_LG: >1.5 fairly good; >2.5 very good; >4 extremely good. ProQ_MS: >0.l fairly good; >0.5 very good; >0.8 extremely good. ModFOLD_Q: >0.5 medium confidence; >0.75 high confidence. ModFOLD_P: <0.05 medium confidence; <0.01 high confidence. MetaMQAP_GDT/RMSD: an ideal model has a GDT score over 59 and a RMSD around 2.0 Å.

	ProQ_LG/MS	ModFOLD_Q/P	MetaMQAP_GDT/RMSD
TLR4	4.764/0.705	0.6177/0.022	76.923/2.123 Å
TLR7	4.374/0.579	0.6199/0.022	71.791/2.138 Å
MyD88	3.966/0.628	0.5749/0.027	73.188/2.202 Å
SIGIRR	3.783/0.438	0.7589/0.010	65.068/2.737 Å
C-tail	2.018/0.300	0.7731/0.009	52.083/3.023 Å

important to ensure binding specificity achieved by different combinations of TIRs during signaling (Fig. 1A). Fig. 1C illustrates the electrostatic surface potential of these BB-loops and αEs. Accordingly, all BB-loops can be divided into two self-complementary parts. The N-terminal (upper region of BB-loops in Fig. 1C) is negatively charged, whereas the C-terminal (lower region of BB-loops in Fig. 1C) is positively charged. The αEs, by contrast, are predominantly positive.

3.2. Pairwise docking of TIR domains

The procedure of protein–protein docking is highly computationally oriented. The reliability of docking results strongly depends on the quality of docking methods. In order to verify the prediction confidence of TIR-TIR interaction of both methods GRAMM-X and ZDOCK, we unrestrainedly inputted as test case the TIR domains of human TLR10 as test case, for which the dimeric crystal structure is known. The native dimerization geometry of TLR10 was present in the top 10 solutions of both GRAMM-X and ZDOCK and was first-ranked by GRAMM-X and sixth-ranked by ZDOCK. The incompleteness of TLR10's crystal structure led to the first five incorrect predictions of ZDOCK, in which large structural gaps were involved in the dimer interfaces. It was therefore straightforward to exclude these five solutions. This test highlights the feasibility and reliability of GRAMM-X and ZDOCK applied in TIR-TIR docking.

As noted above, TIR domains are able to interact heterotypically with each other. To elucidate how SIGIRR disturbs the MyD88-dependent TLR4 and 7 signals, an understanding of the interaction mode of the TLR4 and 7 signaling complexes without the presence of SIGIRR is indispensable. We thus performed unrestrained rigid-body docking for eight TIR complexes. Each docking method returned the 10 most probable models for an input. Thus each complex received a total of 20 candidate models separated into two sets. Some models from the same set had similar conformations whereas most differed considerably from one another. There were some shared models (intersection) across both sets for each complex. These shared models were considered as more confident solutions than others. The optimal docking solution was selected for each complex from the 20 candidates based on three criteria, as follows:

1. Exclude models that do not exist in the intersection of both resulting sets.
2. Exclude models that contain a steric incompatibility of the amino acid liker to the transmembrane helix.
3. Include only those models in which the dimerization geometry is supported by reported experimental data or the dimerization interface is associated with highly conserved boxes as described in Fig. 1B.

For most complexes this three-step filtering led to a unique solution. In the case of the TLR4-TLR4 and the TLR7 dimer-

MyD88 dimer, where these three rules did not yield a unique solution, a further qualification had to be considered. The highest-ranked model by ZDOCK/GRAMM-X ranking was then accepted as the optimal model. The ZDOCK/GRAMM-X ranking and the buried surface interaction area of all optimal models are detailed in Table 3. All resulting docking models are provided in Supplementary file 1.

3.2.1. TLR4–TLR4

The signaling mechanism of TLR4 involves receptor dimerization (Park et al., 2009). After the three-step filtering two candidate models remained. Their ZDOCK/GRAMM-X rankings were 1/6 and 9/2. The first model was accepted because it was best ranked on average and ZDOCK provided a clear-cut ranking. TLR4's TIR reveals an axially symmetric dimer (Fig. 2A) with the BB-loop (involved residues: Pro714-Ala717) of one monomer protruding into a groove formed by the αC (Cys747-Ile748) and DD-loop (Gln782) of the other. The αB (Ala719) of each monomer interacts tightly with each other in the middle of both BB-loop connections. In this model, the Pro714 of one monomer and the Gln782 of the other are connected by a hydrogen bond, which supports Poltorak's (1998) conclusion that the corresponding residue Pro712 is essential to mouse TLR4's function.

3.2.2. TLR7–TLR7

TLR7's TIR forms a *face-to-tail* conformation. The BB-loop (*face*, Glu930–Pro938) of one monomer is preceded by the αE (*tail*, Tyr1024–Ala1032) of the other (Fig. 2B). The BB-loop also interacts with some other regions close to the αE, including: CD-loop (Lys982–Val983), βD (Ile986–Leu988), αD (Gly1009), DE-loop (Ser1010–Ser1011), βE (Val1012–Pro1016) and EE-loop (Thr1017 and Ala1021). Aside from this connection, the EE-loop (Thr1017–Ala1021) of the frontal monomer approximates the αA (Thr905–Glu917) of the posterior. Since the BB-loop and αE are located on opposite sides of a TIR domain, such a *face-to-tail* dimer can be extended by additional TIRs. This dimer model may be relevant for a possible oligomerization of nucleic acid-recognizing TLRs. As discussed by Bell et al. (2006), oligomers might be formed if a nucleic acid ligand is sufficiently long to aggregate several receptors.

3.2.3. MyD88–MyD88

MyD88 forms a dimer when it is incorporated into a receptor complex (Burns et al., 1998). In this model, the BB-loops (Asp195–Cys203) from both monomers were docked together in an antiparallel self-complementary packing (Fig. 2A). Additionally, both αCs (Cys233–Lys238) were brought into contact next to the BB-loop connection. The model is axially symmetric similar to the dimeric crystal structure of human TLR10 (Nyman et al., 2008). Our model is consistent with Loiarro's (2007) conclusion that a heptapeptide, which mimics the BB-loop of MyD88's TIR domain, strongly interferes with dimerization of MyD88.

Table 3
Ranking and interaction area of the selected docking models.

	ZDOCK	GRAMM-X	Interaction area (Å²)
TLR4–TLR4	1	6	639.0
TLR7–TLR7	1	3	965.3
MyD88–MyD88	1	7	737.2
TLR4 × 2-MyD88 × 2	4	1	1395.1
TLR7 × 2-MyD88 × 2	1	5	1249.3
TLR4–SIGIRR	1	6	1092.4
TLR7–SIGIRR	4	1	1055.2
MyD88–SIGIRR	2	9	818.1
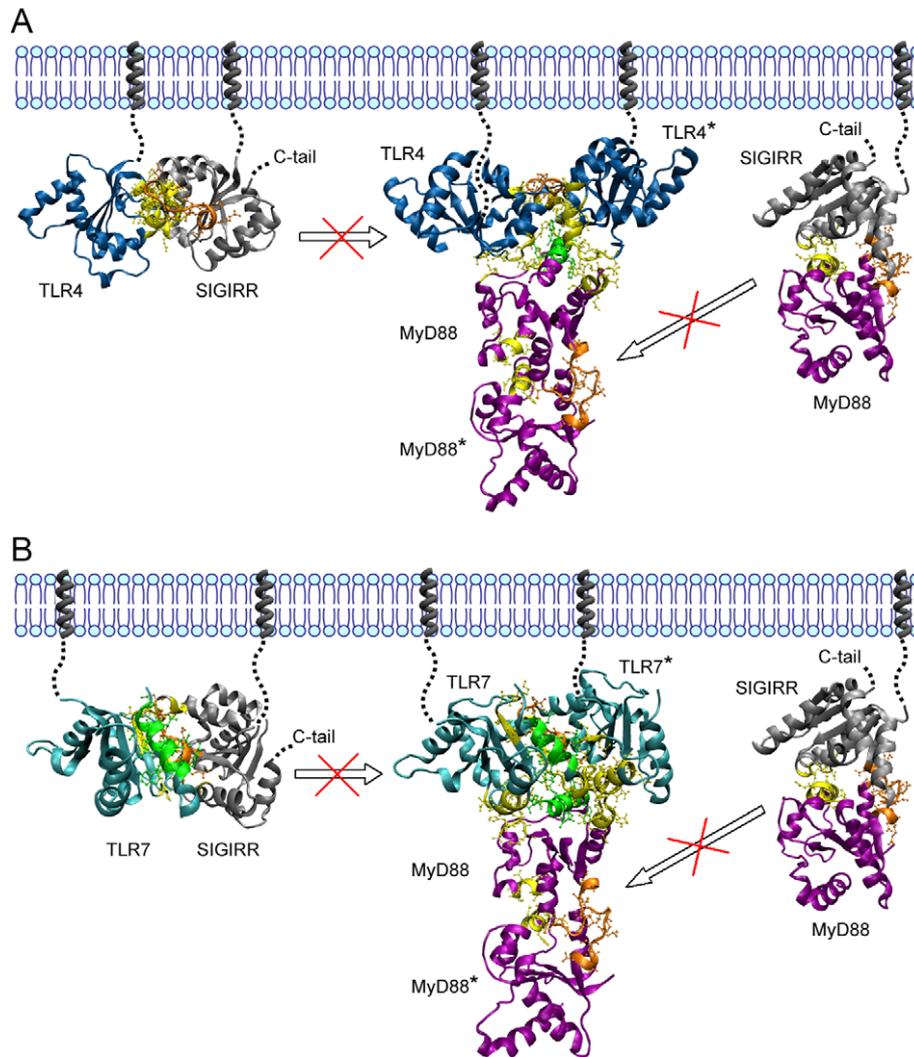
Fig. 2. Models of SIGIRR inhibiting the MyD88-dependent TLR4 and 7 signaling. Interacting regions of BB-loop and αE are labeled in orange and green respectively. Other interacting regions are labeled in yellow. All interacting residues (orange/green/yellow) are represented using CPK (Corey, Pauling & Kultun) convention. (A) Models of SIGIRR inhibiting the TLR4 signaling. (B) Models of SIGIRR inhibiting the TLR7 signaling.

3.2.4. TLR4 dimer–MyD88 dimer

The receptor dimers and MyD88 dimer described above were assembled into tetramers. The TLR4 dimer provides a negatively charged binding pocket adjacent to its interface (Fig. 3). This pocket is constituted by the αC (Gln755) of one TLR4 monomer as well as the αB (Ala719–His724) and αC (Tyr751–Thr756) of the other monomer (TLR4*). The highly conserved, positively charged αE (Cys280–Arg288) of a MyD88 monomer just fills the pocket and makes interactions with the above described residues of TLR4 (Fig. 2A). This connection is further stabilized by three surrounding links: MyD88's DE-loop (Ile271) to TLR4's CD-loop (Arg763–Ala764); MyD88's EE-loop (Asp275–Thr277) to TLR4's CD-loop (Thr756–Gln758); and MyD88's αA (Gln181–Asn186) to TLR4*'s CD-loop (Trp757–Leu760).

3.2.5. TLR7 dimer–MyD88 dimer

After the three-step filtering, two candidate models remained for this dimer. Their ZDOCK/GRAMM-X rankings were 1/5 and 4/6, respectively. The first model obtained the higher ranking in both programs and was thus accepted as optimal. Although TLR7 dimerizes in a different manner as compared to TLR4, it also generates a negatively charged αE-binding pocket for MyD88 at the corresponding location (Figs. 2B and 3). The pocket is composed of αA

(Glu906–Glu911) and EE-loop (Trp1015–Pro1019) of one TLR7 monomer, and AA-loop (Thr899–Val904) and αA (Thr905–Glu906) of the other monomer (TLR7*). This pocket connection is further stabilized by three surrounding links: TLR7's αA (Glu906) and AA-loop (Thr899–Pro902) to MyD88's CD-loop (His248–Arg251); TLR7's EE-loop (Pro1019–Ala1021) to MyD88's AB-loop (Asn186–Arg188) and βB (Leu189); and TLR7*'s αA (Glu906–Ala914) to MyD88's AB-loop (Thr185–Asn186) and αA (Gln181–Gln184). Both TLR4–MyD88 and TLR7–MyD88 tetramers show a T-shaped conformation, where the highly conserved αE of MyD88 plays a central role (Fig. 2).

3.2.6. TLR4-sigirr

SIGIRR heterodimerizes with TLR4 and acts as an inhibitor of TLR signaling (Qin et al., 2005). Our docking model exhibits an extensive interface that is composed of three patches, which indicates a strong molecular affinity (Fig. 2A). First, a consecutive stretch containing SIGIRR's BB-loop (Asp200–Glu209) and αB (Pro210–Ser211) interacts with TLR4's CD-loop (Trp757–Leu760). Second, SIGIRR's αC (Arg235–Arg243) protrudes into the groove formed by TLR4's αB (Ala719–His728) and αC (Tyr751–Gln755). Third, SIGIRR's αD (Pro268–Ala269) interacts with TLR4's BB-loop (Val716–Ala717). Fig. 3 shows that the proper MyD88's αE-binding
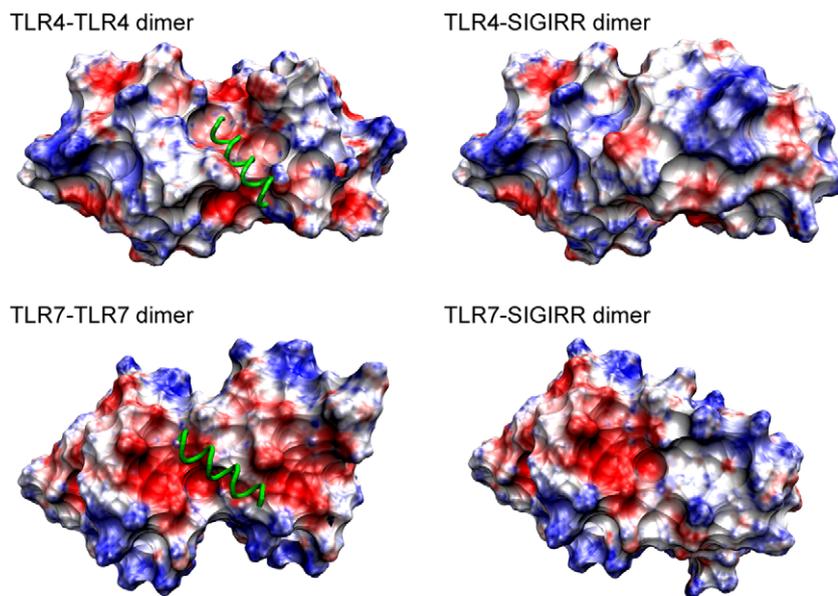
Fig. 3. Surface charge distribution of TIR dimers. Both TLR4-TLR4 and TLR7-TLR7 dimers generate a negatively charged (red) pocket adjacent to their dimer interface to hold the positively charged (blue) αE of MyD88 (charge of αE shown in Fig. 1C). The incorporation of SIGIRR completely disturbed the proper shape and electric environment of the pocket. The αE is represented by a green tube and the other part of MyD88 is omitted for better view. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

pocket presented by the TLR4–TLR4 dimer does not persist in the TLR4-SIGIRR dimer. Notably, the C-tail of SIGIRR is located on the opposite side of SIGIRR's interacting surface. Therefore, it may not participate in the dimer interface (Qin et al., 2005).

3.2.7. TLR7–SIGIRR

Similar to TLR7–TLR7, the TLR7–SIGIRR model is *face-to-tail*, with SIGIRR replacing the rear TLR7 monomer as shown in Fig. 2B. The molecular interface between TLR7 and SIGIRR is larger than that between TLR7 and TLR7. SIGIRR's BB-loop (Asp200–Ala208) together with the beginning of the adjacent αB (Pro210–Ser211) interacts with TLR7's αE (Tyr1024–Thr1035), CD-loop (Lys982–Asp984), βD (Val985–Leu988), DE-loop (Ser1010–Ser1011) and βE (Val1012–Glu1014). Simultaneously, TLR7's DD-loop (Lys993–Phe995) interacts with SIGIRR's AA-loop (Asp173) and αA (Asn182). Likewise, there is no MyD88 αE-binding pocket on this dimer (Fig. 3) and SIGIRR's C-tail does not seem to play any role (Qin et al., 2005).

3.2.8. MyD88–SIGIRR

SIGIRR interferes with the functional dimer conformation of MyD88 by heterodimerization with MyD88 (Qin et al., 2005). The molecular interface between MyD88 and SIGIRR is also quite large (Fig. 2A). SIGIRR's BB-loop (Asp201–Ala208) complements MyD88's BB-loop (Asp195–Val204) by substituting the other BB-loop in the customary MyD88 homodimer. Furthermore, SIGIRR's AA-loop (Ser172–Cys174) and αC (Arg235–Ala236) interacts with MyD88's αC (Gln229–Thr237) under the BB-loops. This model is similar to the dimeric crystal structure of human TLR10 (Nyman et al., 2008), where the BB-loop was identified as a main component of interactions. Likewise, SIGIRR's C-tail does not seem to affect the dimer (Qin et al., 2005).

4. Discussion

So far, the only crystallized dimer structure of TLR's TIR domain is the TLR10 dimer (Nyman et al., 2008), where the BB-loop and αC of each monomer constitute the major part of the symmetric dimer interface. Nunez Miguel et al. (2007) assumed that TLR4 dimerizes in a manner identical to that of TLR10 despite

having no direct evidence. However, we do not consider them to be necessarily identical, because the TIR domain has various inherent dimer conformations (Khan et al., 2004; Nyman et al., 2008; Tao et al., 2002) and TLR4 has different ligand-binding and signaling mechanisms than TLR10. Poltorak et al. (1998) reported that a single point mutation (Pro712His) of the TIR domain of murine TLR4 abolished the TLR4 response to LPS. Our human TLR4 dimer model supports their results. The corresponding residue Pro714 is located at the very tip of the BB-loop and interacts tightly with Gln782 of the other monomer. In contrast to the intensively studied TLR4, structural information about the TIR domain of TLR7 is missing. We thus propose a dimer model of the TLR7 TIR domain. The dimer interaction is maintained mainly by the BB-loop and αE, which are highly conserved among TIRs of different molecules (Fig. 1B).

Triggering of the TLR causes the adaptor protein MyD88 to be recruited to the receptor complex, which in turn promotes association with kinases IRAK4/1. Mal (MyD88-adaptor-like) is another TIR domain-containing adapter protein specifically required by the TLR2 and 4 signaling (Gray et al., 2006). A previous study indicated that Mal promotes the recruitment of MyD88 to TLR4 as a bridging factor and there is no direct interaction between MyD88 and TLR4 (Brown et al., 2006). However, Mal has been shown to be dispensable for TLR4 signaling when MyD88 is fused to a PIP2 targeting domain (Kagan and Medzhitov, 2006). Therefore, direct interactions between TIR domains of MyD88 and TLR4 may mediate signal transduction. This alternate Mal-independent pathway could contribute to signaling as discussed recently (Monie et al., 2009; Ohnishi et al., 2009). Dunne et al. (2003) modeled the TLR4-MyD88 heterodimer using TLR4 and MyD88 monomers. This monomer to monomer model, however, may not fully reflect the molecular interactions. Our model of the receptor dimer docking to the MyD88 dimer provides additional information for a structural interpretation. In particular, both tetramers (TLR4 dimer-MyD88 dimer and TLR7 dimer-MyD88 dimer) exposed in our study demonstrate that the stimulus-induced dimerization of TIR domains creates a new negatively charged molecular pocket for the binding of the positively charged αE of the MyD88 adaptor (Fig. 3). In the presence of SIGIRR, the proper shape and electric

environment of the MyD88-binding pocket are completely disturbed (Fig. 3).

The results from the pairwise docking studies presented here could be assembled to derive a working hypothesis for the TLR4 and 7 signaling transductions and the SIGIRR inhibition mode (Fig. 2). Receptor activation would trigger the formation of TLR4 and 7 TIR dimers recruiting MyD88 TIR dimers resulting in a signaling tetramer. Model predictions including SIGIRR reveal that SIGIRR binds to TLR4 and 7 by occupying their self-interacting sites. On the other hand, the MyD88-SIGIRR dimer shows a resemblance to the MyD88 homodimer. That is, SIGIRR replaces a MyD88 monomer, interrupting the MyD88 homodimer formation. In all cases the BB-loop of SIGIRR plays a key role in binding. The relative positions of all these TIR complexes to the cell/endosome membrane are difficult to expatiate because MyD88 is dissociated from the membrane, and TIR domains of TLR4, TLR7 and SIGIRR are connected to their transmembrane helix by a 20–30 amino acid-long loop stretch which endows the TIR domain with flexible depth and orientation in a cell. Remarkably, TLR4, TLR7 and MyD88 possess a more extensive molecular interface with SIGIRR (heterodimer) than with themselves (homodimer) (Table 3). Fig. 3 also shows that the spatial approximation of receptor-SIGIRR is closer than that of receptor-receptor. These observations highlight the strong molecular affinity of SIGIRR as an inhibitor. In addition, according to our model, SIGIRR's unique C-tail is distant from the active BB-loop consistent with the observation that this tail is not required for SIGIRR's inhibitory effect on TLR signaling (Qin et al., 2005).

In summary, we propose a residue-detailed structural framework of SIGIRR inhibiting the TLR4 and 7 signaling pathways. These results were obtained by computer modeling and are expected to facilitate efforts to design further site-directed mutagenesis experiments to clarity the regulatory role of SIGIRR in inflammatory and innate immune responses.

Acknowledgments

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jsb.2009.12.007.

References

Bell, J.K., Askins, J., Hall, P.R., Davies, D.R., Segal, D.M., 2006. The dsRNA binding site of human Toll-like receptor 3. Proc. Natl. Acad. Sci. USA 103, 8792–8797.
Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The protein data bank. Nucl. Acids Res. 28, 235–242.
Bowie, A., O'Neill, L.A., 2000. The interleukin-1 receptor/Toll-like receptor superfamily: signal generators for pro-inflammatory interleukins and microbial products. J. Leukoc. Biol. 67, 508–514.
Brown, V., Brown, R.A., Ozinsky, A., Hesselberth, J.R., Fields, S., 2006. Binding specificity of Toll-like receptor cytoplasmic domains. Eur. J. Immunol. 36, 742–753.
Bryson, K., McGuffin, L.J., Marsden, R.L., Ward, J.J., Sodhi, J.S., Jones, D.T., 2005. Protein structure prediction servers at University College London. Nucl. Acids Res. 33, W36–38.
Burns, K., Martinon, F., Esslinger, C., Pahl, H., Schneider, P., Bodmer, J.L., Di Marco, F., French, L., Tschopp, J., 1998. MyD88, an adapter protein involved in interleukin-1 signaling. J. Biol. Chem. 273, 12203–12209.
Chen, R., Li, L., Weng, Z., 2003. ZDOCK: an initial-stage protein-docking algorithm. Proteins 52, 80–87.
Clamp, M., Cuff, J., Searle, S.M., Barton, G.J., 2004. The Jalview Java alignment editor. Bioinformatics 20, 426–467.

Dunne, A., Ejdeback, M., Ludidi, P.L., O'Neill, L.A., Gay, N.J., 2003. Structural complementarity of Toll/interleukin-1 receptor domains in Toll-like receptors and the adaptors Mal and MyD88. J. Biol. Chem. 278, 41443–41451.
Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucl. Acids Res. 32, 1792–1797.
Fiser, A., Sali, A., 2003. ModLoop: automated modeling of loops in protein structures. Bioinformatics 19, 2500–2501.
Fiser, A., Do, R.K., Sali, A., 2000. Modeling of loops in protein structures. Protein Sci. 9, 1753–1773.
Gay, N.J., Gangloff, M., 2007. Structure and function of Toll receptors and their ligands. Annu. Rev. Biochem. 76, 141–165.
Gray, P., Dunne, A., Brikos, C., Jefferies, C.A., Doyle, S.L., O'Neill, L.A., 2006. MyD88 adapter-like (Mal) is phosphorylated by Bruton's tyrosine kinase during TLR2 and TLR4 signal transduction. J. Biol. Chem. 281, 10489–10495.
Jones, D.T., Miller, R.T., Thornton, J.M., 1995. Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. Proteins 23, 387–397.
Kagan, J.C., Medzhitov, R., 2006. Phosphoinositide-mediated adaptor recruitment controls Toll-like receptor signaling. Cell 125, 943–955.
Khan, J.A., Brint, E.K., O'Neill, L.A., Tong, L., 2004. Crystal structure of the Toll/interleukin-1 receptor domain of human IL-1RAPL. J. Biol. Chem. 279, 31664–31670.
Krissinel, E., Henrick, K., 2007. Inference of macromolecular assemblies from crystalline state. J. Mol. Biol. 372, 774–797.
Latz, E., Verma, A., Visintin, A., Gong, M., Sirois, C.M., Klein, D.C., Monks, B.G., McKnight, C.J., Lamphier, M.S., Duprex, W.P., Espevik, T., Golenbock, D.T., 2007. Ligand-induced conformational changes allosterically activate Toll-like receptor 9. Nat. Immunol. 8, 772–779.
Lech, M., Kulkarni, O.P., Pfeiffer, S., Savarese, E., Krug, A., Garlanda, C., Mantovani, A., Anders, H.J., 2008. Tir8/Sigirr prevents murine lupus by suppressing the immunostimulatory effects of lupus autoantigens. J. Exp. Med. 205, 1879–1888.
Lech, M., Skuginna, V., Kulkarni, O.P., Gong, J., Wei, T., Stark, R.W., Garlanda, C., Mantovani, A., Anders, H.J., in press. Lack of SIGIRR/TIR8 aggravates hydrocarbon oil-induced systemic lupus nephritis. J. Pathol. doi:10.1002/path.2678.
Liu, L., Botos, I., Wang, Y., Leonard, J.N., Shiloach, J., Segal, D.M., Davies, D.R., 2008. Structural basis of toll-like receptor 3 signaling with double-stranded RNA. Science 320, 379–381.
Loiarro, M., Capolunghi, F., Fanto, N., Gallo, G., Campo, S., Arseni, B., Carsetti, R., Carminati, P., De Santis, R., Ruggiero, V., Sette, C., 2007. Pivotal Advance. Inhibition of MyD88 dimerization and recruitment of IRAK1 and IRAK4 by a novel peptidomimetic compound. J. Leukoc. Biol. 82, 801–810.
McGuffin, L.J., 2008. The ModFOLD server for the quality assessment of protein structural models. Bioinformatics 24, 586–587.
Monie, T.P., Moncrieffe, M.C., Gay, N.J., 2009. Structure and regulation of cytoplasmic adapter proteins involved in innate immune signaling. Immunol. Rev. 227, 161–175.
Nunez Miguel, R., Wong, J., Westoll, J.F., Brooks, H.J., O'Neill, L.A., Gay, N.J., Bryant, C.E., Monie, T.P., 2007. A dimer of the Toll-like receptor 4 cytoplasmic domain provides a specific scaffold for the recruitment of signalling adaptor proteins. PLoS ONE 2, e788.
Nyman, T., Stenmark, P., Flodin, S., Johansson, I., Hammarstrom, M., Nordlund, P., 2008. The crystal structure of the human toll-like receptor 10 cytoplasmic domain reveals a putative signaling dimer. J. Biol. Chem. 283, 11861–11865.
O'Neill, L.A., Bowie, A.G., 2007. The family of five: TIR-domain-containing adaptors in Toll-like receptor signalling. Nat. Rev. Immunol. 7, 353–364.
Ohnishi, H., Tochio, H., Kato, Z., Orii, K.E., Li, A., Kimura, T., Hiroaki, H., Kondo, N., Shirakawa, M., 2009. Structural basis for the multiple interactions of the MyD88 TIR domain in TLR4 signaling. Proc. Natl. Acad. Sci. USA. 106, 10260–10265.
Park, B.S., Song, D.H., Kim, H.M., Choi, B.S., Lee, H., Lee, J.O., 2009. The structural basis of lipopolysaccharide recognition by the TLR4-MD-2 complex. Nature 458, 1191–1195.
Pawlowski, M., Gajda, M.J., Matlak, R., Bujnicki, J.M., 2008. MetaMQAP: a meta-server for the quality assessment of protein models. BMC Bioinform. 9, 403.
Peter, M.E., Kubarenko, A.V., Weber, A.N., Dalpke, A.H., 2009. Identification of an N-terminal recognition site in TLR9 that contributes to CpG-DNA-mediated receptor activation. J. Immunol. 182, 7690–7697.
Polentarutti, N., Rol, G.P., Muzio, M., Bosisio, D., Camnasio, M., Riva, F., Zoja, C., Benigni, A., Tomasoni, S., Vecchi, A., Garlanda, C., Mantovani, A., 2003. Unique pattern of expression and inhibition of IL-1 signaling by the IL-1 receptor family member TIR8/SIGIRR. Eur. Cytokine Netw. 14, 211–218.
Poltorak, A., He, X., Smirnova, I., Liu, M.Y., Van Huffel, C., Du, X., Birdwell, D., Alejos, E., Silva, M., Galanos, C., Freudenberg, M., Ricciardi-Castagnoli, P., Layton, B., Beutler, B., 1998. Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice. Mutations in TLR4 gene. Science 282, 2085–2088.
Qin, J., Qian, Y., Yao, J., Grace, C., Li, X., 2005. SIGIRR inhibits interleukin-1 receptor- and toll-like receptor 4-mediated signaling through different mechanisms. J. Biol. Chem. 280, 25233–25241.
Rahman, A.H., Eisenberg, R.A., 2006. The role of toll-like receptors in systemic lupus erythematosus. Springer Semin. Immunopathol. 28, 131–143.
Takeda, K., Akira, S., 2004. TLR signaling pathways. Semin. Immunol. 16, 3–9.
Tao, X., Xu, Y., Zheng, Y., Beg, A.A., Tong, L., 2002. An extensively associated dimer in the structure of the C713S mutant of the TIR domain of human TLR2. Biochem. Biophys. Res. Commun. 299, 216–221.
Thomassen, E., Renshaw, B.R., Sims, J.E., 1999. Identification and characterization of SIGIRR, a molecule representing a novel subtype of the IL-1R superfamily. Cytokine 11, 389–399.

Tovchigrechko, A., Vakser, I.A., 2006. GRAMM-X public web server for protein-protein docking. Nucl. Acids Res. 34, W310–314.

Wald, D., Qin, J., Zhao, Z., Qian, Y., Naramura, M., Tian, L., Towne, J., Sims, J.E., Stark, G.R., Li, X., 2003. SIGIRR, a negative regulator of Toll-like receptor-interleukin 1 receptor signaling. Nat. Immunol. 4, 920–927.

Wallner, B., Elofsson, A., 2003. Can correct protein models be identified? Protein Sci. 12, 1073–8106.

Wei, T., Gong, J., Jamitzky, F., Heckl, W.M., Stark, R.W., Rossle, S.C., 2009. Homology modeling of human Toll-like receptors TLR7, 8, and 9 ligand-binding domains. Protein Sci. 18, 1684–1691.

Wheeler, D.L., Barrett, T., Benson, D.A., Bryant, S.H., Canese, K., Chetvernin, V., Church, D.M., Dicuccio, M., Edgar, R., Federhen, S., Feolo, M., Geer, L.Y., Helmberg, W., Kapustin, Y., Khovayko, O., Landsman, D., Lipman, D.J., Madden, T.L., Maglott, D.R., Miller, V., Ostell, J., Pruitt, K.D., Schuler, G.D., Shumway, M., Sequeira, E., Sherry, S.T., Sirotkin, K., Souvorov, A., Starchenko, G., Tatusov, R.L., Tatusova, T.A., Wagner, L., Yaschenko, E., 2008. Database resources of the National Center for Biotechnology Information. Nucl. Acids Res. 36, D13–21.

Xu, Y., Tao, X., Shen, B., Horng, T., Medzhitov, R., Manley, J.L., Tong, L., 2000. Structural basis for signal transduction by the Toll/interleukin-1 receptor domains. Nature 408, 111–115.

Paper 4

Lack of SIGIRR/TIR8 aggravates hydrocarbon oil-induced systemic lupus

Maciej Lech, Veronika Skuginna, Onkar P. Kulkarni, Jing Gong, Tiandi Wei, Robert W. Stark, Cecilia Garlanda, Alberto Mantovani and Hans-Joachim Anders

Original Paper

Lack of SIGIRR/TIR8 aggravates hydrocarbon oil-induced lupus nephritis

Maciej Lech,[1] Veronika Skuginna,[1] Onkar P. Kulkarni,[1] Jing Gong,[2] Tiandi Wei,[2] Robert W. Stark,[2] Cecilia Garlanda,[3] Alberto Mantovani[3] and Hans-Joachim Anders[1]*

[1]*Medizinische Poliklinik, University of Munich, Germany*
[2]*Centre for NanoScience and Department for Earth and Environmental Sciences, University of Munich, Germany*
[3]*Istituto Clinico Humanitas and Fondazione Humanitas per la Ricerca, Rozzano, Italy*

*Correspondence to:
Hans-Joachim Anders,
Medizinische Poliklinik,
Universität München,
Pettenkoferstrasse 8a, 80336
Munich, Germany.
E-mail:
hjanders@med.uni-muenchen.de

No conflicts of interest were declared.

Abstract

Multiple genetic factors contribute to the clinical variability of spontaneous systemic lupus erythematosus (SLE) but their role in drug-induced SLE remain largely unknown. Hydrocarbon oil-induced SLE depends on mesothelial cell apoptosis and Toll-like receptor (TLR)-7-mediated induction of type I interferons. Hence, we hypothesized that TIR8/SIGIRR, an endogenous TLR inhibitor, prevents oil-induced SLE. *Sigirr*-deficient dendritic cells expressed higher *TLR7* mRNA levels and TLR7 activation resulted in increased IL-12 production *in vitro*. *In vivo*, lack of SIGIRR increased surface CD40 expression on spleen CD11c$^+$ dendritic cells and *MX-1, TNF, IL-12, BAFF* and *BCL-2* mRNA expression 6 months after pristane injection. Spleen cell counts of CD4$^-$/CD8$^-$ 'autoreactive' T cells and B220$^+$ B cells were also increased in *Sigirr*$^{-/-}$ mice. Serum autoantibody analysis revealed that *Sigirr* deficiency specifically enhanced the production of rheumatoid factor (from 4 months of age) and anti-snRNP IgG (from 5 months of age), while anti-Smith IgG or anti-dsDNA IgG were independent of the *Sigirr* genotype. This effect was sufficient to significantly aggravate lupus nephritis in *Sigirr*-deficient mice. Structure model prediction identified the BB loop of SIGIRR's intracellular TIR domain to interact with TLR7 and MyD88. BB loop deletion was sufficient to completely abrogate SIGIRR's inhibitory effect on TLR7 signalling. Thus, TIR8/SIGIRR protects from hydrocarbon oil-induced lupus by suppressing the TLR7-mediated activation of dendritic cells, via its intracellular BB loop. Copyright © 2009 Pathological Society of Great Britain and Ireland. Published by John Wiley & Sons, Ltd.

Introduction

Genetic and environmental factors drive both the onset and the progression of autoimmune diseases [1]. As such, a number of variants in immunoregulatory genes increase the risk for systemic lupus erythematosus (SLE) [2–4]. In mice, lack of single genes, such as *TGFβ1, DNAse1, Lyn, Fas* or *C1q*, is sufficient to cause late-onset lupus-like autoimmunity [5–9]. Mutant *Sle1* or *TLR9* can trigger lupus in C57BL/6 mice only in the presence of a second genetic factor, eg the *lpr* mutation [10,11]. Weaker modifier genes, such as IL-10 or IL-27R, enhance lupus only in a genetic context of multiple susceptibility genes, eg being provided by the specific genetic background of MRL mice [12,13].

The *TIR8* gene encodes for single immunoglobulin IL-1-related receptor (SIGIRR), a member of the Toll-like receptor (TLR)/IL-1 receptor family [14,15].

Over-expression of SIGIRR in Jurkat or HepG2 cells suppresses LPS or IL-1-induced activation of NF-κB, hence SIGIRR is an endogenous inhibitor of TLR and IL-1 signalling [14,16,17]. *Sigirr*-deficient mice develop severe immunity-mediated tissue damage upon pathogen challenge or dextran-induced damage of the intestinal epithelium [16,18–21]. Lack of SIGIRR also enhances spontaneous autoimmunity in C57BL/6$^{lpr/lpr}$ mice [22]. These mice suffer from defective Fas-induced apoptosis of autoreactive lymphocytes, which massively increases the exposure of nuclear autoantigens to the immune system [23]. Hence, the aggravated phenotype of *Sigirr*-deficient C57BL/6$^{lpr/lpr}$ mice could best be explained by a suppressive effect of Sigirr on self-RNA and -DNA-mediated activation of dendritic cells and B cells, a process known to involve TLR7 and TLR9 [10,24–27]. For example, plasmacytoid dendritic cells are the major source of type I interferons upon recognition of RNA via TLR7 [25],

whereas conventional dendritic cells produce interferon mainly upon cytosolic RNA recognition receptors [28].

Whether SIGIRR also suppresses environmentally-induced autoimmunity is unknown. To address this question would require SLE-like disease to be induced in genetically unaltered mice without a predisposing autoimmune genetic background. We therefore used the hydrocarbon oil 2,6,10,14-tetramethyl-pentadecane (pristane) to induce SLE in $Sigirr^{-/-}$ and $Sigirr^{+/+}$ mice of an identical C57BL/6 background. Intraperitoneal injection of pristane induces apoptosis of mesothelial cells, followed by granulomatous peritonitis with the formation of ectopic lymphoid tissue [29]. In this model, the persistent abundance of apoptotic peritoneal cells in the context of chronic inflammation triggers TLR7 signalling, type I interferon expression and the subsequent evolution of antinuclear antibodies, immune complex disease and lupus nephritis [30,31]. We hypothesized a role for Tir8/Sigirr in limiting pristane-induced SLE by suppressing intraperitoneal inflammation and/or autoantibody generation.

Methods

Animal studies

Sigirr-deficient mice on a F6 C57BL/6 genetic background were generated as previously described [19]. The genotype was assured by PCR in each mouse at 5 weeks of age before mice of both genotypes were intraperitoneally injected with 0.5 ml pristane (Sigma-Aldrich, Steinheim, Germany). Blood samples were taken at monthly intervals after pristane injection (at age 6 weeks) until sacrifice at 6 months after pristane injection. In a subgroup of mice, peritoneal lavage fluid was obtained at 2 and 28 days after pristane injection. All experiments were performed in accordance with the German animal care and ethics legislation and had been approved by the local government authorities.

Phenotype analysis

Flow cytometry, real-time quantitative (TaqMan) RT–PCR, and autoantibody analysis were performed as previously described [22,32]. PCR primers are listed in Table 1. Formalin-fixed tissue sections (2 μm) for periodic acid–Schiff (PAS) stains were prepared following routine protocols. The severity of kidney disease was graded by an observer blinded to the genotype of the mice, using a glomerulonephritis activity score (0–24) normally used for the assessment of human lupus nephritis [31]. Immunostaining was performed on either paraffin-embedded or frozen sections as described [26], using the following primary antibodies: anti-mouse C3c (complement, GAM/C3c/FITC, 1 : 200; Nordic Immunological Laboratories, Tilburg, The Netherlands) or

anti-mouse B220 (BD Pharmingen, Heidelberg, Germany). Negative controls included incubation with a respective isotype antibody. Semi-quantitative scoring of glomerular C3c deposits from 0 to 3+ was performed on 15 cortical glomerular sections, as described [33].

Structure and interaction model predictions

Amino acid sequences of human TLR7, MyD88 and SIGIRR (targets) were extracted from the NCBI protein database [34]. TIR domains of TLR7, MyD88 and SIGIRR (TYR165-PRO308, without the C-terminal extension) were modelled by homology modelling. Four template structures were obtained via BLAST search against the Protein Data Bank (PDB) [35]: TLR1 (PDB code 1FYV), TLR2 (1FYW), TLR10 (2J67) and IL-1RAPL (1T3G). Multiple alignments of target-template sequences and structural coordinates of templates were submitted to MODELLER 9v3 [36] to generate the three-dimensional (3D) target structures. Since the C-terminal extension of SIGIRR has no structure-known homologue, we built its structure using protein threading (THREADER 3.5 [37]). The resulting models were then evaluated by ProQ [38] and MetaMQAP [39]. The protein docking programs GRAMM-X [40] and ZDOCK [41] were used to predict the pairwise interactions between these TIR domains. Both programs can return 10 models ranked as the most probable predictions, selected from thousands of candidates, based on geometry, hydrophobicity and electrostatics complementarity of molecule surfaces.

In vitro studies

HEK 293 cells (2×10^5), stably transfected with hTLR7, were cultured overnight in 1 ml DMEM complete medium before being transiently transfected with lipofectamine 2000 (Invitrogen, Carlsbad, CA, USA) and pUNO, full-length hSIGIRR or SIGIRR mutants, all in pUNO expressing vector (Invivogen, San Diego, CA, USA) and the NF-κB luciferase reporter construct (Clontech, Mountain View, CA, USA) with a final amount of 2 μg. After 24 h the cells were stimulated with 1 μg imiquimod (Invitrogen) and luciferase activity was determined after 6 h using Promega's Dual-Glo luciferase kit. The primers were designed to anneal to the template sequences flanking the target sites, which were sequences to be deleted (ΔBB loop-SIGIRR, forward, 5′-GACTGCCCCGACCTCTTGG TGAACCTGAG-3′, reverse, 5′-CAAGAGGTCGGGG CAGTCGCTGTAGGAG-3′; ΔTIR-SIGIRR, forward, 5′-GTGGAGATACGGAAGGTGCAGTACAGGC-3′, reverse, 5′-CACCTTCCGTATCTCCACCTCCCCAT AC-3′). The pUNO-hSIGIRR plasmid was used as the template for mutagenesis. A 25 μl PCR reaction was composed of 1 μl template (200 ng), 1 μl each primer (20 pM each), 2 μl dNTP mixture, 2.5 ml 10× buffer, 0.5 μl Pfu Turbo DNA polymerase (2.5 U)

and dH$_2$O. The reaction was started with 2 min at 95 °C to pre-denature the template. This was followed by 18 cycles of 1 min at 95 °C, 1 min at 58 °C and 1 min/kbp at 68 °C. Following the PCR reaction, 1 µl DpnI (20 U; NEB, Ipswich, MA, USA) was added and the mixture was incubated at 37 °C for 1 h to degrade the original unmodified plasmid templates. After DpnI digestion, 2 µl of the mixture was used to transfect DHα-competent cells by heat shock. After a 1 h recovery in 300 µl LB medium without antibiotics, the transformed *Escherichia coli* was spread on Blas-Agar Blasticidin plates (Invivogen, San Diego, CA, USA) and incubated at 37 °C overnight. Colonies were selected and grown overnight in 3 ml LB with blasticidin.

Statistical analysis

Statistical significance was evaluated by ANOVA or by two-tailed Student's *t*-test (two group comparisons) at $p < 0.05$. Data were expressed as mean ± SEM.

Results

Sigirr modulates peritoneal cytokine production after pristane exposure

Intraperitoneal injection of pristane induces massive apoptosis of mesothelial cells and infiltrating neutrophils [29], lipogranuloma formation and interferon signalling by peritoneal macrophages [30]. Macroscopically we did not detect any difference in the peritoneal cavities of *Sigirr*$^{+/+}$ and *Sigirr*$^{-/-}$ mice 2 and 28 days after pristane injection (Figure 1A, left). Lavage fluids from these time points revealed large numbers of dead mesothelial cells and monocyte/macrophages often ingesting apoptotic neutrophils (Figure 1A, right), but cell counts were identical in *Sigirr*$^{+/+}$ and *Sigirr*$^{-/-}$ mice (not shown). Lavage fluid flow cytometry for propidium iodine and annexin V revealed comparable levels of early apoptotic (annexin V$^+$) cells and late apoptotic (annexin V/ propidium iodine$^+$) cells in *Sigirr*$^{+/+}$ and *Sigirr*$^{-/-}$ mice at all time points (not shown). Real-time RT–PCR from peritoneal lavage fluid cell mRNA showed significantly higher levels for *IL-12* and *TNFα* at 2 days and *IFNγ* and the IFNα/β-dependent gene *Mx1* 28 days after pristane injection (Figure 1B). Lavage fluid ELISA revealed significantly higher levels of TNFα but not of IL-12p40 2 days after pristane injection, while IFNα, -β and -γ were not detectable by ELISA (Figure 1C). At 28 days, TNFα and IL-12p40 levels had increased but were genotype-independent (Figure 1C).

Sigirr suppresses the activation of dendritic cells 6 months after pristane exposure

We used flow cytometry to quantify the numbers of splenic CD11c dendritic cells that stain positive for the activation marker CD40 at 6 months of age. Lack of Sigirr significantly increased the numbers of CD40$^+$ dendritic cells (Figure 2A). Lack of Sigirr was also associated with increased mRNA levels of *Mx1*, *TNFα*, *IL-4* and *IL-12* in these cells (Figure 2B). *Sigirr*-deficient spleen dendritic cells also expressed higher levels of Baff and Bcl2 (Figure 2C), which support the survival of B and/or T cells [42]. Consistent with the stronger activation of dendritic cells, *Sigirr*-deficient mice had higher IL-12p40 serum levels as compared to wild-type mice injected with pristane at 6 months of age (5.9 ± 1.6 versus 1.1 ± 0.8 ng/ml, $p < 0.0001$). Thus, Sigirr suppresses dendritic cell activation and serum IL-12p40 levels 6 months after pristane exposure.

Sigirr suppresses pristane-induced lymphoproliferation

Pristane induced mild splenomegaly at 6 months in wild-type mice. Lack of Sigirr caused a trend towards higher spleen weights, but with considerable interindividual variability (Figure 2D), but flow cytometry revealed sigificantly higher numbers of spleen cells in *Sigirr*-deficient mice treated with pristane (Figure 2E). This was due to lymph follicle hyperplasia with massive enlargement of B220$^+$ B cell areas (Figure 2F). We also observed higher numbers of CD4/CD8 double-negative 'autoreactive' T cells and CD4$^+$CD25$^+$ 'regulatory' T cells in *Sigirr*-deficient mice treated with pristane (Figure 3C, D). The latter was consistent with a significant induction of *Foxp3* mRNA levels in CD3$^+$CD4$^+$CD25$^+$ cells of *Sigirr*-deficient mice (Figure 3E). The numbers of CD3$^+$CD4$^+$ and CD3$^+$CD8$^+$ cell were not affected by the *Tir8* genotype (Figure 3A, B). However, in CD4$^+$CD25$^-$T cells, lack of Sigirr was associated with higher mRNA expression levels of the Th1 markers *T-bet* and *IFNγ* (Figure 3F) and the Th2 markers *Gata* and *IL-4* (Figure 3G). The Th17 marker *Rorγ* was significantly down-regulated, although *IL-17* mRNA levels were markedly induced in pristane-treated *Sigirr*-deficient mice (Figure 3H). Together, Sigirr suppresses the expansion of spleen B cells as well as of autoreactive and regulatory T cells after pristane injection.

Sigirr suppresses pristane-induced autoantibodies to nuclear autoantigens

We obtained serum samples at monthly intervals from pristane-treated mice of both genotypes and antibody levels were determined by ELISA. Total serum IgG levels were comparable at all time points in *Sigirr*-deficient and wild-type mice (Figure 4). Small amounts of antibodies directed against dsDNA IgG of the IgG1 isotype and against nucleosomes were detectable by ELISA but the levels did not differ between *Sigirr*-deficient and wild-type mice (Figure 4). The specificity of dsDNA autoantibodies was confirmed by *Critidiae luciliae* assay (not shown).
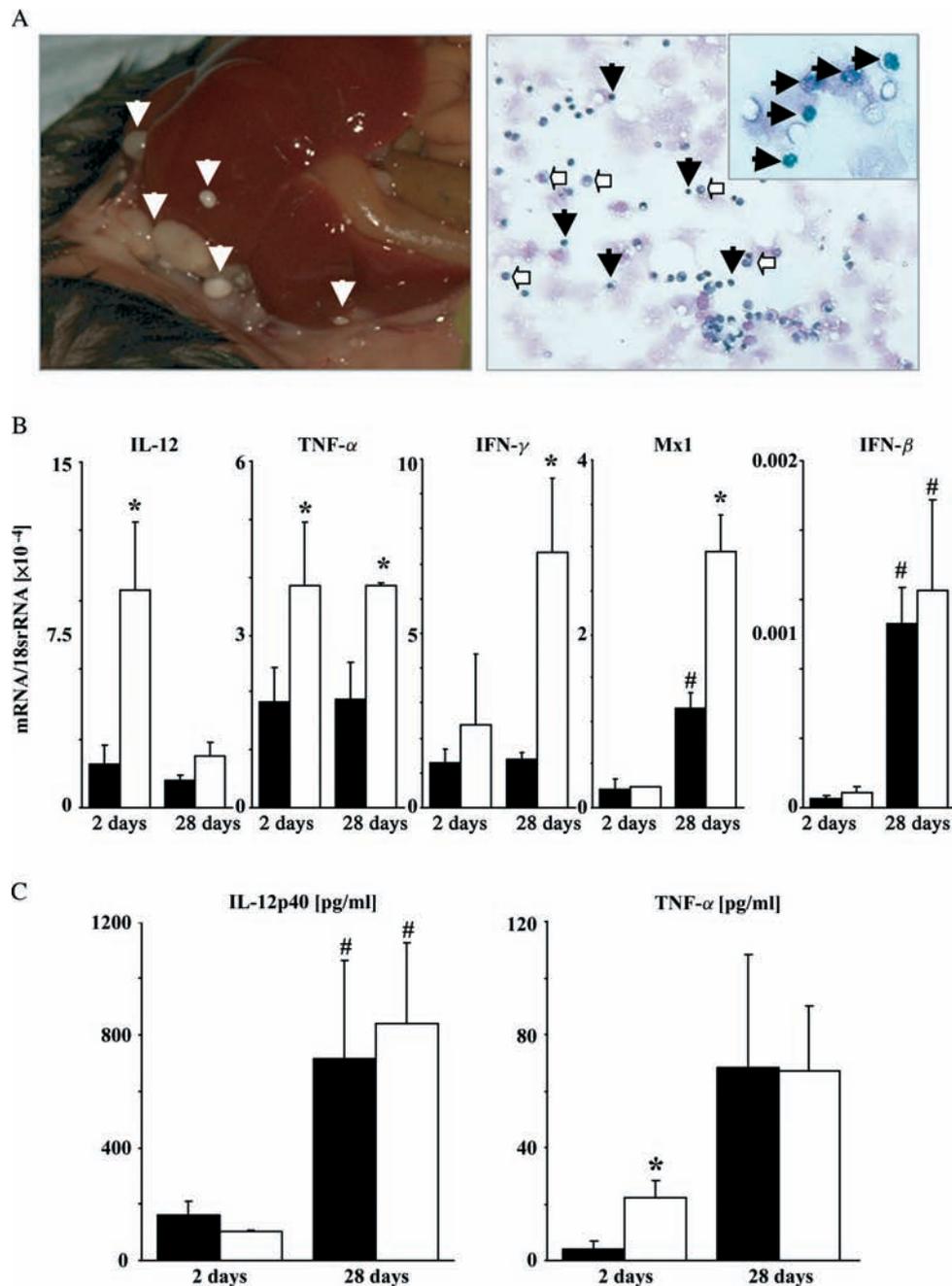
Figure 1. *Tir8/Sigirr* genotype and pristane-induced peritonitis. Pristane injection into *Sigirr*[+/+] and *Sigirr*[−/−] mice caused lipogranuloma formation in the peritoneal cavity, as indicated by white arrows in the left image of (A). The right image of (A) shows microscopic analysis of peritoneal fluids, which revealed large amounts of apoptotic cells (black arrows) inside and outside of phagocytes (white arrows and insert), independent of the *Sigirr*[−/−] genotype. (B) Real-time RT−PCR of peritoneal fluid samples taken 2 and 28 days after pristane injection was used to quantify intraperitoneal mRNA expression. Data are expressed as a ratio to respective *18s* rRNA as a reference gene. (C) Peritoneal fluid levels of TNFα and IL-12p40 were determined by ELISA. Data in (B, C) are expressed as means ± SEM of 14 mice in each group of *Sigirr*[+/+] (black bars) and *Sigirr*[−/−] mice (white bars). *$p < 0.05$ versus wild-type; #$p < 0.05$ versus 2 days

Anti-dsDNA of the IgG2a/c, IgG2b and IgG3 isotypes remained undetectable at all time points (not shown). By contrast, lack of Sigirr significantly induced the production of rheumatoid factor and anti-SnRNP IgG from 4 and 5 months after pristane exposure, respectively (Figure 4). Antibodies against the Smith antigen were produced from month 4, but the levels did not differ between the two genotypes (Figure 4). Thus, Sigirr specifically suppresses the production of rheumatoid factor and anti-SnRNP IgG but does not affect DNA autoantibody production after pristane exposure in mice.

Sigirr prevents pristane-induced lupus nephritis

Pristane does not cause major autoimmune tissue lesions in C57BL/6 mice, although mild glomerulonephritis may develop [30]. Lack of Sigirr
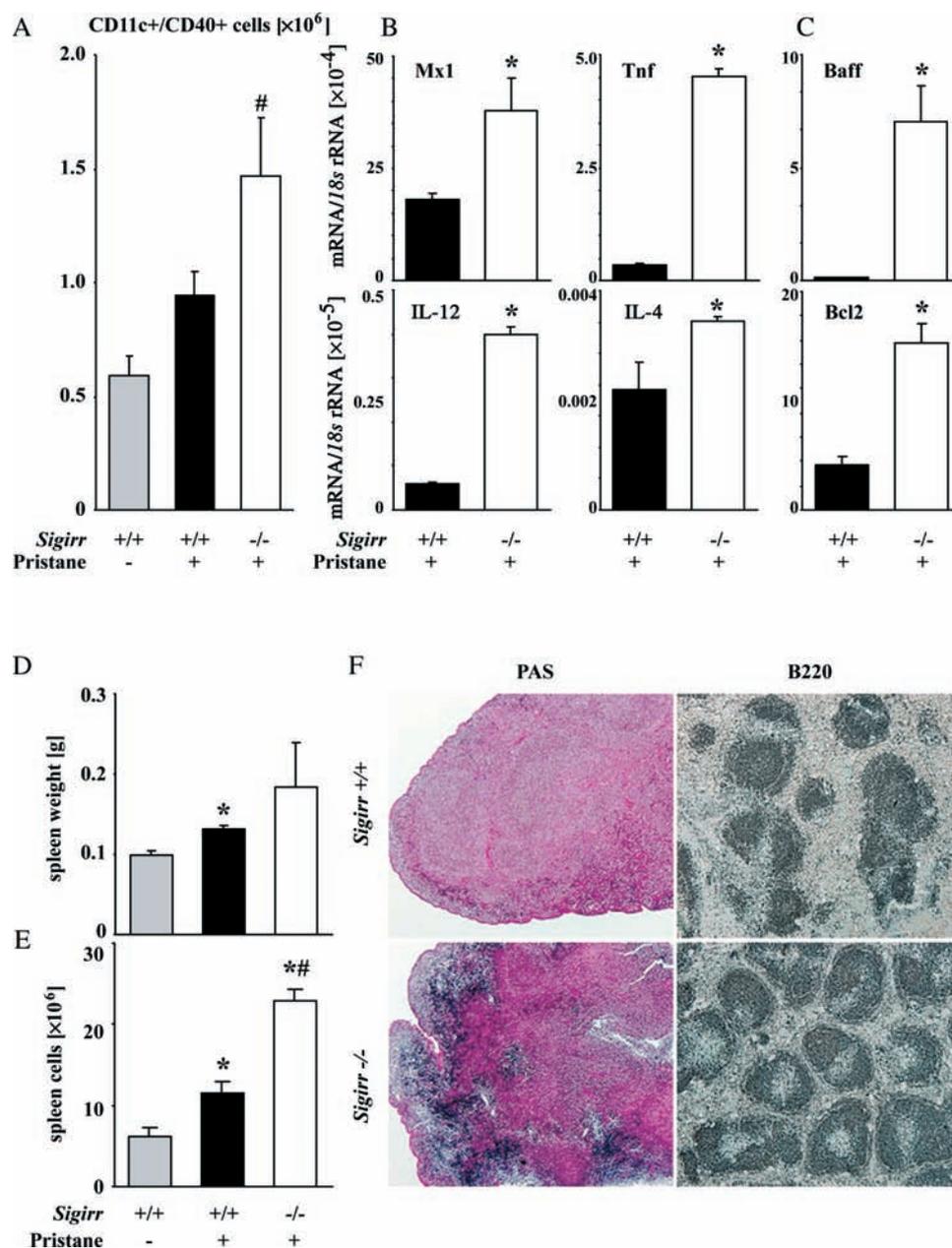
Figure 2. *Sigirr* genotype, dendritic cell activation and spleen morphology. (A) The total number of spleen CD11c$^+$ dendritic cells positive for the activation marker CD40 was quantified 6 months after pristane injection by flow cytometry. Data represent means ± SEM from 14 mice in each group. $^#p < 0.05$ versus pristane-injected wild-type mice. (B, C) RNA was isolated from spleen CD11c$^+$ cells from *Sigirr*$^{-/-}$ (white bars) and *Sigirr*$^{+/+}$ mice (black bars) 6 months after pristane injection for real-time RT−PCR analysis. Data are expressed as means of the ratio of the specific mRNA to that of 18S rRNA ± SEM. $^*p < 0.05$ versus wild-type mice. (D, E) Spleen weight (D) was determined 6 months after pristane injection in untouched wild-type mice (grey bar), pristane-injected wild-type mice (black bar) and *Sigirr*-deficient mice (white bar). Total spleen cell numbers (E) were determined by flow cytometry, as described in Methods. $^*p < 0.05$ versus untouched wild-type mice; $^#p < 0.05$ versus pristane-injected wild-type mice. (F) Representative images of spleen PAS stains and B220 immunostaining from mice at 6 months. Original magnification, ×100

was associated with diffuse mesangio-proliferative glomerulonephritis, as indicated by glomerular hypercellularity, PAS-positive matrix expansion and glomerular leukocyte infiltrates (Figure 5A). Glomerular C3c deposits were scored 1.2 ± 0.2 in *Sigirr*$^{-/-}$ mice and 0.3 ± 0.1 in wild-type mice ($p = 0.003$; Figure 5A). The composite activity score for lupus nephritis was 6.8 ± 0.6 in *Sigirr*$^{-/-}$ mice and 3.5 ± 0.3 in wild-type mice ($p = 0.0002$; Figure 5B). Albuminuria constantly increased in *Sigirr*-deficient mice and started to be significantly higher at 5 months as

compared to wild-type mice (Figure 5C). The difference was highest at 6 months. Together, Sigirr protects mice from diffuse proliferative lupus nephritis after pristane exposure.

In silico structure analysis predicts the BB loop of Sigirr's TIR domain as the interaction site with TLR7 and Myd88

Because lupus autoantigens drive SLE by ligating TLR7, we speculated that Sigirr is induced by
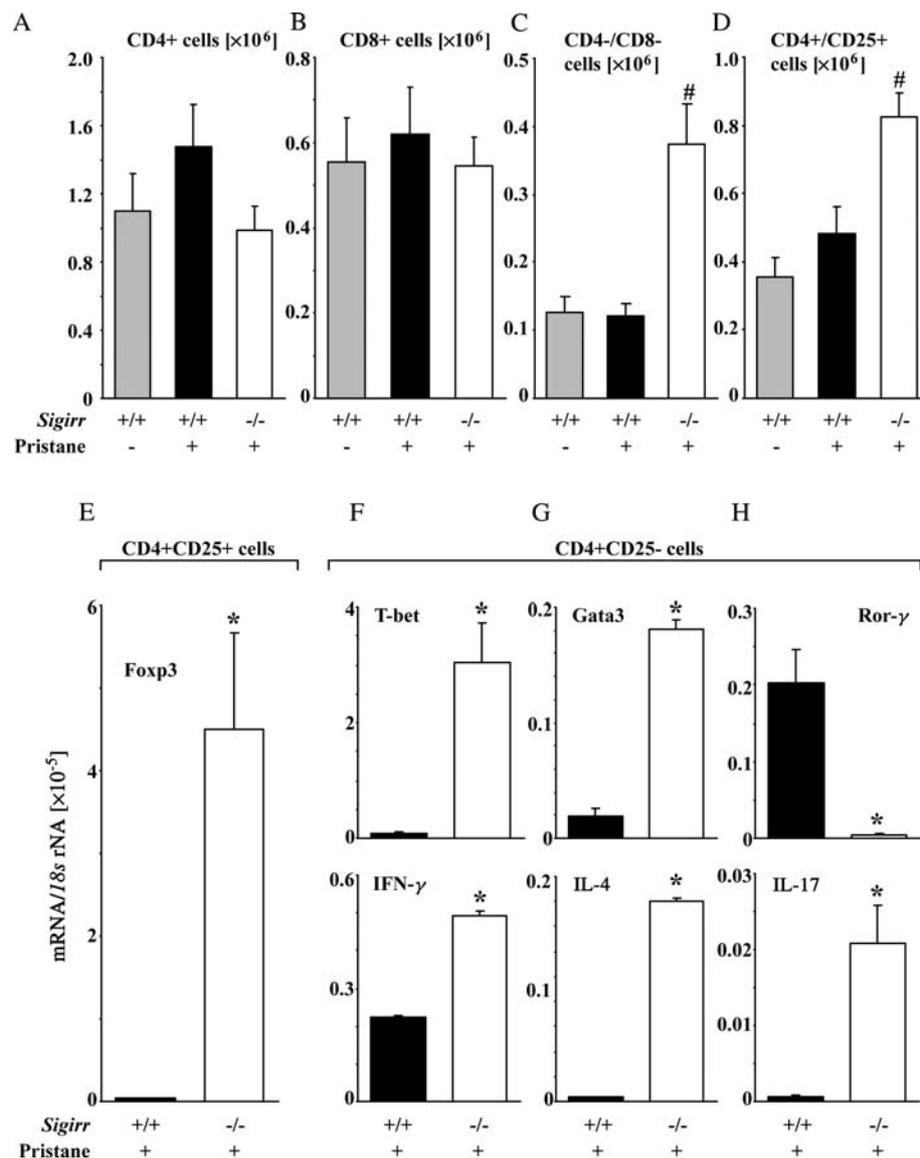
Figure 3. *Sigirr* genotype and T cell subsets. Six months after pristane injection, spleen T cell subsets were assessed by flow cytometry (A–D), as described in Methods. The histograms presents mean $\pm$ SEM of at least 14 mice in each group. (E–H) Real-time RT–PCR from CD4[+]CD25[+] cells or CD4[+]CD25[−] cells was used to quantify additional T cell markers. Data ratios to respective *18s* rRNA as a reference gene and are expressed as means $\pm$ SEM of 10 mice in each group. [#]$p < 0.05$ versus pristane-injected wild-type mice

inflammation and that it suppresses TLR7 signalling, especially in antigen-presenting cells. *Sigirr* mRNA was induced in spleen monocytes by LPS or TNFα/IFNγ, with a maximum expression level at 18 h of stimulation (Figure 6A, B). In addition, the TLR7 agonist imiquimod activated bone marrow dendritic cells to produce IL-12, a response that was five-fold higher in *Sigirr*-deficient dendritic cells (Figure 6C), perhaps also because the basal *TLR7* mRNA expression was significantly higher in *Sigirr*-deficient dendritic cells (Figure 6D). But can Sigirr directly interfere with TLR7 signalling at the structural level?

Because crystallographic structures of human TLR7, MyD88 and SIGIRR TIR domains are not available, we developed 3D structural models based on homology modelling and protein threading. The predicted structures of TLR7, MyD88 and SIGIRR TIR domains were evaluated as *extremely good* or *ideal* by several model quality assessment programs (data not shown). The BB loop (face) and the α-helix E region (neck) appeared to be conserved among all the different TIR domains (Figure 7A) and are localized on opposite regions of the SIGIRR TIR domain (Figure 7B). Then, the protein-docking softwares GRAMM-X and ZDOCK were used to predict pairwise molecular interaction sites of complexes formed by TLR7–TLR7, MyD88–MyD88, TLR7–SIGIRR and MyD88–SIGIRR. The top-ranked dimer model was selected for each complex (Figure 7C). According to these predictions, receptor activation would trigger the formation of TLR7 TIR dimers in a face-to-neck orientation, recruiting MyD88 TIR face-to-face dimers and forming a T-shaped signalling tetramer (complex B in Figure 7C). The
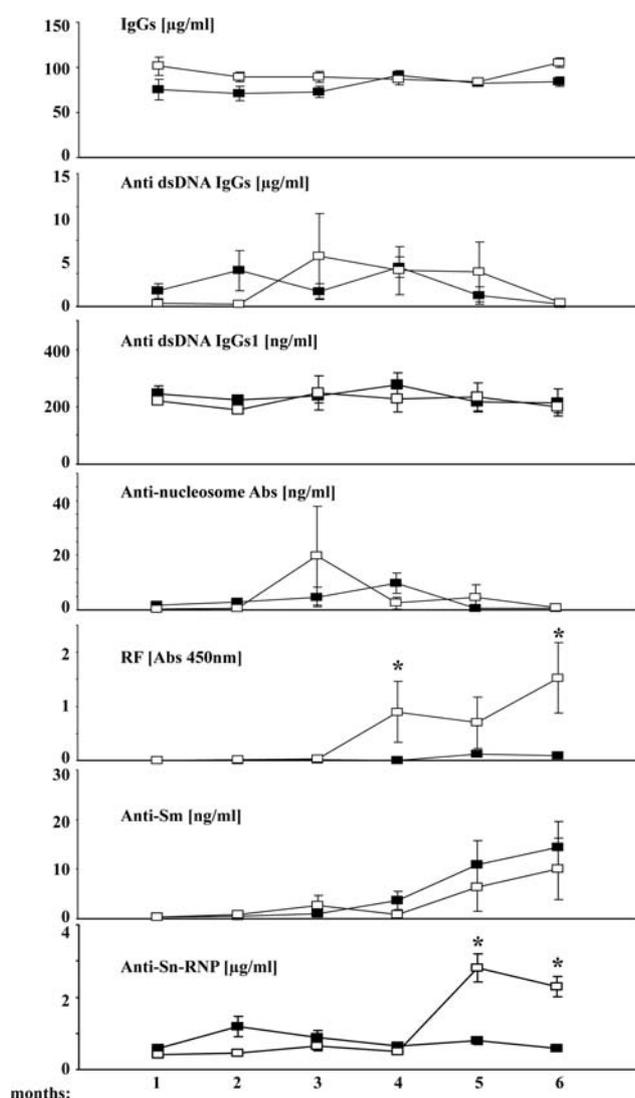
Figure 4. *Sigirr* genotype and serum immunoglobulin and lupus autoantibody levels. Wild-type mice (open squares) and *Sigirr*-deficient mice (black squares) were bled at monthly intervals after pristane injection to determine serum levels of IgG and a number of different autoantibodies, as indicated by ELISA. Data represent means $\pm$ SEM from at least 14 mice in each group. $^*p < 0.05$ versus wild-type mice at the same time point

Table 1. Primers used for RT–PCR

Gene name		Primer sequence
Baff	Forward	5′-CCTCCAAGGCATTTCCTCTT-3′
	Reverse	5′-GACTGTCTGCAGCTGATTGC-3′
Bcl2	Forward	5′-GATCCAGGATAACGGAGGCT-3′
	Reverse	5′-GGTCTTCAGAGACAGCCAGG-3′
FoxP3	Forward	5′-TTCATGCATCAGCTCTCCAC-3′
	Reverse	5′-CTGGACACCCATTCCAGACT-3′
Gata3	Forward	5′-GCCTGCGGACTCTACCATAA-3′
	Reverse	5′-AGGATGTCCCTGCTCTCCTT-3′
Ifn-γ	Forward	5′-ACAGCAAGGCGAAAAAGGAT-3′
	Reverse	5′-TGAGCTCATTGAATGCTTGG-3′
Il-4	Forward	5′-TGAACGAGGTCACAGGAGAA-3′
	Reverse	5′-CGAGCTCACTCTCTGTGGTG-3′
Il-12	Forward	5′-CTAGACAAGGGCATGCTGGT-3′
	Reverse	5′-GCTTCTCCCACAGGAGGTTT-3′
Il-17	Forward	5′-TCCAGAAGGCCCTCAGACTA-3′
	Reverse	5′-TGAGCTTCCCAGATCACAGA-3′
Mx1	Forward	5′-TCTGAGGAGAGCCAGACGAT-3′
	Reverse	5′-CTCAGGGTGTCGATGAGGTC-3′
Ror-γ	Forward	5′-ACAGAGACACCACCGGACAT-3′
	Reverse	5′-GGTGATAACCCCGTAGTGGA-3′
Tbet	Forward	5′-TCAACCAGCACCAGACAGAG-3′
	Reverse	5′-ATCCTGTAATGGCTTGTGGG-3′
Tnf-α	Forward	5′-CCACCACGCTCTTCTGTCTAC-3′
	Reverse	5′-AGGGTCTGGGCCATAGAACT-3′
18s RNA	Forward	5′-GCAATTATTCCCCATGAACG-3′
	Reverse	5′-AGGGCCTCACTAAACCATCC-3′

A Sigirr TIR mutant lacking the BB loop can no longer block TLR7 signalling

To verify the functional role of the BB loop for TLR7 signalling, we over-expressed full-length TLR7 in HEK293 cells, together with full-length Sigirr or various Sigirr mutants. Full-length Sigirr potently suppressed NF-κB reporter gene expression 6 h after stimulation with the TLR7 agonist imiquimod (Figure 7D). By contrast, lack of the TIR domain or the BB loop only completely abrogated this inhibitory effect on TLR7 signalling. Together, Sigirr TIR's BB loop mediates the inhibitory effect of Sigirr on TLR7 signalling.

Discussion

Lack of Sigirr clearly aggravated pristane-induced autoimmune tissue injury. In the kidney, *Sigirr*-deficiency was associated with diffuse proliferative lupus nephritis and significant albuminuria as compared to wild-type C57BL/6 mice, which revealed only minor glomerular abnormalities. The role of Sigirr in inhibiting pristane-induced autoimmunity was clearly documented by increased dendritic cell activation, increased numbers of CD4/CD8 double-negative T cells and B cells in spleen as well as increased serum

stimulus-induced dimerization of TIR domains creates a molecular pocket for the binding of α-helix E of the MyD88 adaptor. Model predictions, including SIGIRR's TIR, revealed that SIGIRR interacts with TLR7 by occupying TLR7 TIR's α-helix E region with its BB loop, which should interrupt TLR7 homodimer formation (complex A in Figure 7C). In addition, SIGIRR's BB loop was predicted to interact with the BB loop of MyD88, which should interrupt MyD88 homodimer formation (complex C in Figure 7C). According to our model, SIGIRR TIR's unique C-terminal extension (ca. 100 amino acids; Figure 7B) is located distant from the BB loop, consistent with the observation that this extension is not required for SIGIRR's inhibitory effect on TLR signalling (16).

Figure 5. *Sigirr* genotype and lupus nephritis 6 months after pristane injection. (A) Renal sections were stained with PAS and immunostaining was performed for complement factor C3c (DAPI stains cell nuclei blue), as indicated. Original magnification, $\times 200$. Images are representative for 10 mice in each group. (B) The lupus nephritis activity index (range 0–24) was assessed on renal PAS-stained sections at 6 months after pristane injection, as described in Methods. (C) Urinary albumin : creatinine ratio was determined from urine samples taken from at least 14 wild-type mice (open squares) and *Sigirr*-deficient mice (black squares) at monthly intervals after pristane injection. [#]$p < 0.05$ versus pristane-injected wild-type mice

levels of IL-12 and selected autoantibodies in *Sigirr*-deficient mice. In this regard, the data from pristane-induced lupus matches our previous data obtained from *Sigirr*-deficient C57BL/6[lpr/lpr] mice with spontaneous autoimmunity [22]. However, in C57BL/6[lpr/lpr] mice Sigirr had a global suppressive effect on the evolution of hypergammaglobulinaemia and autoantibodies of multiple specificities as early as at 2 months

of age [22]. By contrast, in pristane-induced lupus, lack of Sigirr massively increased the production of rheumatic factor and RNA autoantibodies, but not before 4–5 months after pristane exposure. This was most obvious for anti-Sm IgG and anti-U1snRNP, because the Sm (Smith) antigen is the protein component and U1snRNP the RNA component of the U1snRNP ribonucleoprotein complex [43,44].
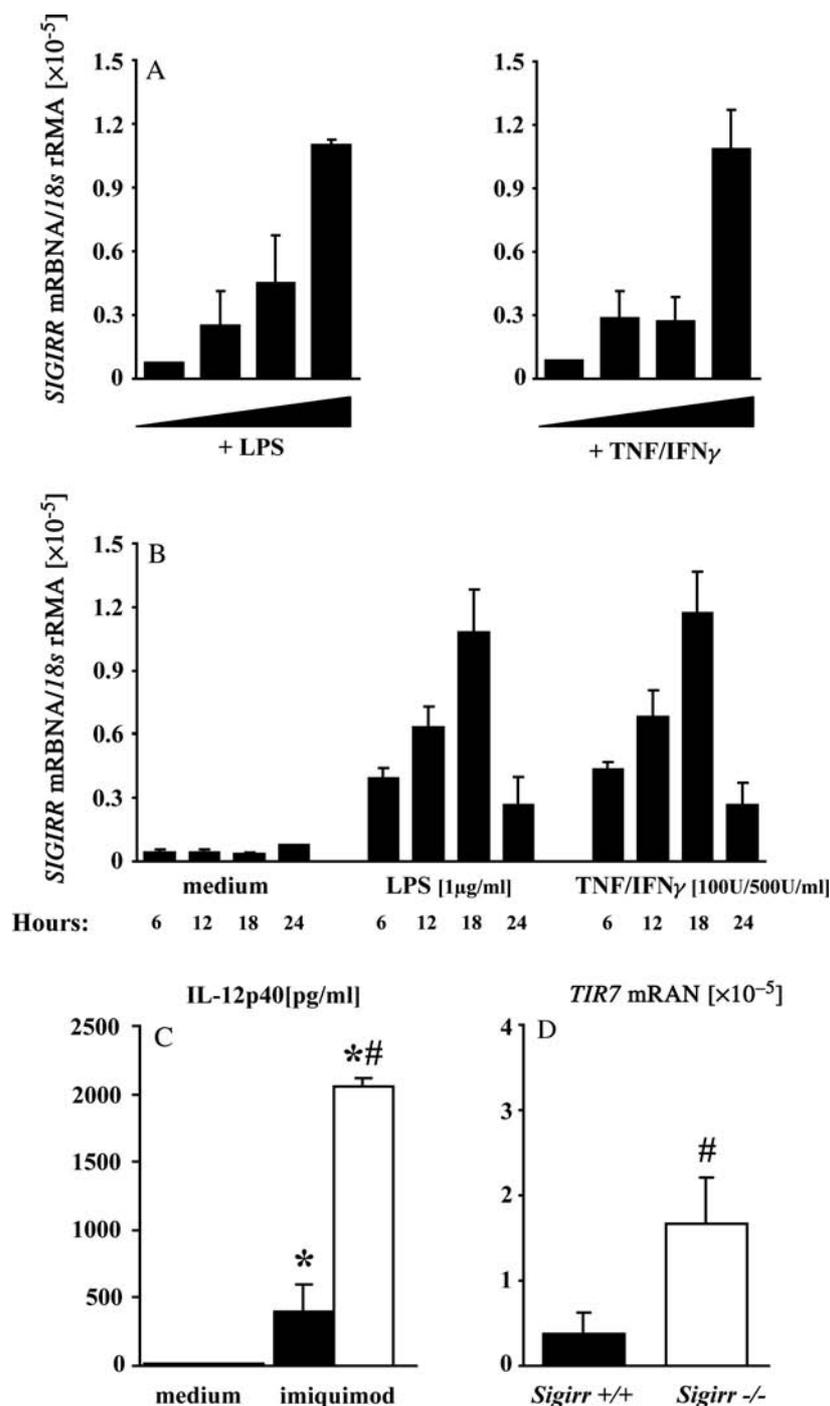
Figure 6. Sigirr suppresses TLR7 signalling in dendritic cells. (A) *Sigirr* mRNA expression was determined by real-time RT–PCR on RNA samples from spleen monocytes 12 h after stimulation with various doses of LPS or TNFα/IFNγ. (B) Sigirr induction was quantified in spleen monocytes at various time intervals after stimulation with a given dose of LPS or TNFα/IFNγ. Data represent mean ratios of Sigirr/*18s* rRNA ± SEM from three independent experiments. (C) Bone marrow dendritic cells from wild-type mice (black bars) and *Sigirr*-deficient mice (white bars) were stimulated with imiquimod, as described in Methods. IL-12p40 was determined after 24 h in cell culture supernatants. (D) Basal *TLR7* mRNA expression was determined in the same cells by real-time RT–PCR and is illustrated as a ratio to the respective *18s* rRNA expression. Data represent means ± SEM from three independent experiments. *$p < 0.05$ versus medium; #$p < 0.05$ versus wild-type mice

Because *Sigirr*-deficient mice did not display a broader spectrum of autoantibodies than wild-type mice, Sigirr does not seem to directly promote loss-of-tolerance or epitope spreading. Obviously, Sigirr rather specifically fosters the expansion of IgG and RNA autoreactive lymphocyte clones that produce the necessary components for pathogenic RNA immune complexes. In turn, such RNA immune complexes are known trigger TLR7 activation and type I interferon signalling, a positive amplification loop [24,45–46]. In fact, pristane-induced lupus is driven by TLR7 signalling [30,31] and, most interestingly, lack of *TLR7* selectively impaired the production of pristane-induced snRNP antibodies [30,31]. We therefore

Figure 7. Models of SIGIRR inhibiting the TLR7 signalling pathway. (A) Amino acid sequence homologies of the intracellular TIR domains of human TLR7, MyD88 and SIGIRR. (B) The 3D structural model of SIGIRR's intracellular TIR domain was predicted as described in Methods. The BB loop is indicated in orange; the α-helix E region is indicated in green. Note that the SIGIRR-specific C-terminal extension links at the α-helix E. (C) Complex B, TLR7 homodimers (dark blue), linked by BB loop (orange)$-\alpha$-helix E (green) interaction, bind MyD88 homodimers (light blue), forming a T-shaped conformation. MyD88 homodimers are formed by BB loop interactions (orange). The TLR7–MyD88 complex may not form when SIGIRR recruits to this complex as follows: the predicted TLR7–SIGIRR interaction (complex A) should affect TLR7 homodimer formation, and SIGIRR–MyD88 interaction (complex C) may affect MyD88 homodimer formation (SIGIRR interfering sites indicated by open arrows). (D) HEK293 cells, constitutively expressing TLR7, were seeded at a concentration of 2×10^5 cells/well in a 12-well plate, and cultured overnight in 1 ml DMEM complete medium. The next day the cells were transfected with *TIR8* or *TIR8* mutants or pUNO control vector and the NF-κB luciferase reporter construct, as described in Methods. Cells were stimulated with 1 μg imiquimod and luciferase activity was determined after 6 h. The values are percentages of the imiquimod-stimulated cells expressing empty pUNO vector. Data represent means $\pm$ SEM of at least two independent experiments. *$p < 0.05$ versus pUNO control

hypothesized that Sigirr may directly inhibit TLR7 signalling.

Previous studies suggested that Sigirr can inhibit TLR4 (but not TLR3 or TLR5) signalling via interaction at the level of their respective intracellular TIR domains, which inhibits the necessary recruitment of its adaptor Myd88 [17]. Our structure-based prediction model of putative TIR–TIR interaction sites identified the BB loop of SIGIRR's intracellular TIR domain as the most likely interaction site with TLR7, as well

as with MyD88. Deletion of the BB loop completely abrogated SIGIRR's inhibitory effect on TLR7 signalling, as did the deletion of the entire TIR domain. These data are in line with the report of Qin *et al* [17] and further prove that SIGIRR also inhibits TLR7 via this mechanism. In addition, Sigirr suppresses *TLR7* mRNA expression, which adds to its suppressive effect on TLR7 signalling.

In pristane-induced lupus, the immunoregulatory function of Sigirr clearly localizes to the central lymphoid organs, as indicated by the impact of *Sigirr* deficiency on lymphocyte numbers and activation states and autoantibody production.

We conclude that genes that regulate autoantigen-driven dendritic cell activation determine environmentally triggered autoimmunity, and that *Tir8/Sigirr* loss-of-function mutations represent a novel genetic risk factor for hydrocarbon oil-induced autoimmunity in mice.

Acknowledgements

References

1. Goodnow CC. Multistep pathogenesis of autoimmune disease. *Cell* 2007;**130:**25–35.
2. Rahman A, Isenberg DA. Systemic lupus erythematosus. *N Engl J Med* 2008;**358:**929–939.
3. Harley JB, Alarcon-Riquelme ME, Criswell LA, Jacob CO, Kimberly RP, Moser KL, *et al*. Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci. *Nat Genet* 2008;**40:**204–210.
4. Tsokos GC, Nambiar MP, Tenbrock K, Juang YT. Rewiring the T-cell: signaling defects and novel prospects for the treatment of SLE. *Trends Immunol* 2003;**24:**259–263.
5. Cohen PL, Eisenberg RA. Lpr and gld: single gene models of systemic autoimmunity and lymphoproliferative disease. *Annu Rev Immunol* 1991;**9:**243–269.
6. Botto M, Dell'Agnola C, Bygrave AE, Thompson EM, Cook HT, Petry F, *et al*. Homozygous C1q deficiency causes glomerulonephritis associated with multiple apoptotic bodies. *Nat Genet* 1998;**19:**56–59.
7. Hibbs ML, Tarlinton DM, Armes J, Grail D, Hodgson G, Maglitto R, *et al*. Multiple defects in the immune system of Lyn-deficient mice, culminating in autoimmune disease. *Cell* 1995;**83:**301–311.
8. Napirei M, Karsunky H, Zevnik B, Stephan H, Mannherz HGMoroy T. Features of systemic lupus erythematosus in Dnase1-deficient mice. *Nat Genet* 2000;**25:**177–181.
9. Shull MM, Ormsby I, Kier AB, Pawlowski S, Diebold RJ, Yin M, *et al*. Targeted disruption of the mouse transforming growth factor-β1 gene results in multifocal inflammatory disease. *Nature* 1992;**359:**693–699.
10. Lartigue A, Courville P, Auquit I, Francois A, Arnoult C, Tron F, *et al*. Role of TLR9 in anti-nucleosome and anti-DNA antibody production in lpr mutation-induced murine lupus. *J Immunol* 2006;**177:**1349–1354.

11. Mohan C, Alas E, Morel L, Yang P, Wakeland EK. Genetic dissection of SLE pathogenesis. Sle1 on murine chromosome 1 leads to a selective loss of tolerance to H2A/H2B/DNA subnucleosomes. *J Clin Invest* 1998;**101:**1362–1372.
12. Shimizu S, Sugiyama N, Masutani K, Sadanaga A, Miyazaki Y, Inoue Y, *et al*. Membranous glomerulonephritis development with Th2-type immune deviations in MRL/lpr mice deficient for IL-27 receptor (WSX-1). *J Immunol* 2005;**175:**7185–7192.
13. Yin Z, Bahtiyar G, Zhang N, Liu L, Zhu P, Robert ME, *et al*. IL-10 regulates murine lupus. *J Immunol* 2002;**169:**2148–2155.
14. Polentarutti N, Rol GP, Muzio M, Bosisio D, Camnasio M, Riva F, *et al*. Unique pattern of expression and inhibition of IL-1 signalling by the IL-1 receptor family member TIR8/SIGIRR. *Eur Cytokine Netw* 2003;**14:**211–218.
15. Thomassen E, Renshaw BR, Sims JE. Identification and characterization of SIGIRR, a molecule representing a novel subtype of the IL-1R superfamily. *Cytokine* 1999;**11:**389–399.
16. Wald D, Qin J, Zhao Z, Qian Y, Naramura M, Tian L, *et al*. SIGIRR, a negative regulator of Toll-like receptor-interleukin 1 receptor signaling. *Nat Immunol* 2003;**4:**920–927.
17. Qin J, Qian Y, Yao J, Grace C, Li X. SIGIRR inhibits interleukin-1 receptor- and toll-like receptor 4-mediated signaling through different mechanisms. *J Biol Chem* 2005;**280:**25233–25241.
18. Garlanda C, Di Liberto D, Vecchi A, La Manna MP, Buracchi C, Caccamo N, *et al*. Damping excessive inflammation and tissue damage in *Mycobacterium tuberculosis* infection by Toll IL-1 receptor 8/single Ig IL-1-related receptor, a negative regulator of IL-1/TLR signaling. *J Immunol* 2007;**179:**3119–3125.
19. Garlanda C, Riva F, Polentarutti N, Buracchi C, Sironi M, De Bortoli M, *et al*. Intestinal inflammation in mice deficient in Tir8, an inhibitory member of the IL-1 receptor family. *Proc Natl Acad Sci USA* 2004;**101:**3522–3526.
20. Huang X, Hazlett LD, Du W, Barrett RP. SIGIRR promotes resistance against *Pseudomonas aeruginosa* keratitis by down-regulating type-1 immunity and IL-1R1 and TLR4 signaling. *J Immunol* 2006;**177:**548–556.
21. Xiao H, Gulen MF, Qin J, Yao J, Bulek K, Kish D, *et al*. The Toll-interleukin-1 receptor member SIGIRR regulates colonic epithelial homeostasis, inflammation, and tumorigenesis. *Immunity* 2007;**26:**461–475.
22. Lech M, Kulkarni OP, Pfeiffer S, Savarese E, Krug A, Garlanda C, *et al*. Tir8/Sigirr prevents murine lupus by suppressing the immunostimulatory effects of lupus autoantigens. *J Exp Med* 2008;**205:**1879–1888.
23. Reap EA, Leslie D, Abrahams M, Eisenberg RA, Cohen PL. Apoptosis abnormalities of splenic lymphocytes in autoimmune lpr and gld mice. *J Immunol* 1995;**154:**936–943.
24. Lau CM, Broughton C, Tabor AS, Akira S, Flavell RA, Mamula MJ, *et al*. RNA-associated autoantigens activate B cells by combined B cell antigen receptor/Toll-like receptor 7 engagement. *J Exp Med* 2005;**202:**1171–1177.
25. Savarese E, Chae OW, Trowitzsch S, Weber G, Kastner B, Akira S, *et al*. U1 small nuclear ribonucleoprotein immune complexes induce type I interferon in plasmacytoid dendritic cells through TLR7. *Blood* 2006;**107:**3229–3234.
26. Pawar RD, Ramanjaneyulu A, Kulkarni OP, Lech M, Segerer S, Anders HJ. Inhibition of Toll-like receptor-7 (TLR-7) or TLR-7 plus TLR-9 attenuates glomerulonephritis and lung injury in experimental lupus. *J Am Soc Nephrol* 2007;**18:**1721–1731.
27. Christensen SR, Shupe J, Nickerson K, Kashgarian M, Flavell RA, Shlomchik MJ. Toll-like receptor 7 and TLR9 dictate autoantibody specificity and have opposing inflammatory and regulatory roles in a murine model of lupus. *Immunity* 2006;**25:**417–428.
28. Kumagai Y, Kumar H, Koyama S, Kawai T, Takeuchi O, Akira S. Cutting edge: TLR-dependent viral recognition along with type I IFN positive feedback signaling masks the requirement of viral replication for IFN-α production in plasmacytoid dendritic cells. *J Immunol* 2009;**182:**3960–3964.
29. Calvani N, Caricchio R, Tucci M, Sobel ES, Silvestris F, Tartaglia P, *et al*. Induction of apoptosis by the hydrocarbon

oil pristane: implications for pristane-induced lupus. *J Immunol* 2005;**175:**4777–4782.

30. Lee PY, Kumagai Y, Li Y, Takeuchi O, Yoshida H, Weinstein J, *et al*. TLR7-dependent and FcγR-independent production of type I interferon in experimental mouse lupus. *J Exp Med* 2008;**205:**2995–3006.

31. Savarese E, Steinberg C, Pawar RD, Reindl W, Akira S, Anders HJ, *et al*. Requirement of Toll-like receptor 7 for pristane-induced production of autoantibodies and development of murine lupus nephritis. *Arthritis Rheum* 2008;**58:**1107–1115.

32. Patole PS, Schubert S, Hildinger K, Khandoga S, Khandoga A, Segerer S, *et al*. Toll-like receptor-4: renal cells and bone marrow cells signal for neutrophil recruitment during pyelonephritis. *Kidney Int* 2005;**68:**2582–2587.

33. Allam R, Pawar RD, Kulkarni OP, Hornung V, Hartmann G, Segerer S, *et al*. Viral 5′-triphosphate RNA and non-CpG DNA aggravate autoimmunity and lupus nephritis via distinct TLR-independent immune responses. *Eur J Immunol* 2008;**38:**3487–3498.

34. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, *et al*. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2008;**36:**D13–21.

35. Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, Weissig H, *et al*. The Protein Data Bank and the challenge of structural genomics. *Nat Struct Biol* 2000;**7:**(suppl): 957–959.

36. Fiser A, Do RK, Sali A. Modeling of loops in protein structures. *Protein Sci* 2000;**9:**1753–1773.

37. Jones DT, Taylor WR, Thornton JM. A new approach to protein fold recognition. *Nature* 1992;**358:**86–89.

38. Wallner B, Elofsson A. Identification of correct regions in protein models using structural, alignment, and consensus information. *Protein Sci* 2006;**15:**900–913.

39. Pawlowski M, Gajda MJ, Matlak R, Bujnicki JM. MetaMQAP: a meta-server for the quality assessment of protein models. *BMC Bioinformat* 2008;**9:**403.

40. Tovchigrechko A, Vakser IA. GRAMM-X public web server for protein–protein docking. *Nucleic Acids Res* 2006;**34:**W310–314.

41. Chen R, Li L, Weng Z. ZDOCK: an initial-stage protein-docking algorithm. *Proteins* 2003;**52:**80–87.

42. Mackay F, Silveira PA, Brink R. B cells and the BAFF/APRIL axis: fast-forward on autoimmunity and signaling. *Curr Opin Immunol* 2007;**19:**327–336.

43. Marshak-Rothstein A, Rifkin IR. Immunologically active autoantigens: the role of toll-like receptors in the development of chronic inflammatory disease. *Annu Rev Immunol* 2007;**25:**419–441.

44. Pomeranz Krummel DA, Oubridge C, Leung AK, Li J, Nagai K. Crystal structure of human spliceosomal U1 snRNP at 5.5 Å resolution. *Nature* 2009;**458:**475–480.

45. Barrat FJ, Meeker T, Gregorio J, Chan JH, Uematsu S, Akira S, *et al*. Nucleic acids of mammalian origin can act as endogenous ligands for Toll-like receptors and may promote systemic lupus erythematosus. *J Exp Med* 2005;**202:**1131–1139.

46. Lech M, Garlanda C, Mantovani A, Kirschning CJ, Schlondorff D, Anders HJ. Different roles of TiR8/Sigirr on toll-like receptor signaling in intrarenal antigen-presenting cells and tubular epithelial cells. *Kidney Int* 2007;**72:**182–192.

Paper 5

Homology modeling of human Toll-like receptors TLR7, 8 and 9 ligand-binding
domains

Tiandi Wei, Jing Gong, Ferdinand Jamitzky, Wolfgang M. Heckl, Robert W. Stark and
Shaila C. Rössle

Homology modeling of human Toll-like receptors TLR7, 8, and 9 ligand-binding domains

Tiandi Wei,[1] Jing Gong,[1]* Ferdinand Jamitzky,[1,2] Wolfgang M. Heckl,[1,3] Robert W. Stark,[1] and Shaila C. Rössle[1]

[1]Department of Earth and Environmental Sciences, Center for Nanoscience, Ludwig-Maximilians-Universität München, 80333 Munich, Germany

[2]Leibniz Supercomputing Centre, 85748 Garching, Germany

[3]Deutsches Museum, 80538 Munich, Germany

Abstract: Toll-like receptors (TLRs) play a key role in the innate immune system. The TLR7, 8, and 9 compose a family of intracellularly localized TLRs that signal in response to pathogen-derived nucleic acids. So far, there are no crystallographic structures for TLR7, 8, and 9. For this reason, their ligand-binding mechanisms are poorly understood. To enable first predictions of the receptor–ligand interaction sites, we developed three-dimensional structures for the leucine-rich repeat ectodomains of human TLR7, 8, and 9 based on homology modeling. To achieve a high sequence similarity between targets and templates, structural segments from all known TLR ectodomain structures (human TLR1/2/3/4 and mouse TLR3/4) were used as candidate templates for the modeling. The resulting models support previously reported essential ligand-binding residues. They also provide a basis to identify three potential receptor dimerization mechanisms. Additionally, potential ligand-binding residues are identified using combined procedures. We suggest further investigations of these residues through mutation experiments. Our modeling approach can be extended to other members of the TLR family or other repetitive proteins.

Keywords: Toll-like receptor; leucine-rich repeats; protein-nucleic acid interaction; homology modeling

Introduction

Toll-like receptors (TLRs) play an essential role in the innate immunity, recognizing invasion of microbial pathogens and initiating intracellular signal transduction pathways to trigger expression of genes, the products of which can control innate immune responses.[1] To understand how these receptors work, it is crucial to investigate them from a structural perspective. To date, only the crystal structures of the ectodomains of human TLR1/2/3/4 and mouse TLR3/4 have been determined.[2–6] The progress of genome projects, however, already led to the identification of 13 TLRs in mammalian and more than 20 TLRs in nonmammalian. A total of more than 2000 TLR proteins has been sequenced.[7] Thus, the structures of most TLRs are still unknown because structure determination by X-ray diffraction or nuclear magnetic resonance spectroscopy experiments remains time-consuming. Here, computational methods can help to bridge the gap between sequence determination and structure determination. To this end, homology modeling is a powerful tool to predict the three-dimensional structure of proteins.

Homology modeling is based on the assumption that similar sequences among evolutionarily related proteins share an overall structural similarity. The modeling procedure can be divided into a number of

*Correspondence to: Jing Gong, Theresienstr. 41, 80333 Munich, Germany. E-mail: gongj@informatik.uni-muenchen.de

A

2Z80(A)	NT	01												
3CIG(A)		09	10	11	12	13								
3CIG(A)			15	16	17	18	19							
2Z64(A)				01	02	03	04	05	06					
2A0Z(A)							21	22	23	CT				
TLR7	NT	15	16	17	18	19	20	21	22	23	24	25	CT

B

2Z66(A)	NT	01										
2A0Z(A)		15	16									
2Z7X(A)		01	02	03	04							
2A0Z(A)				19	20	21						
2A0Z(A)				03	04	05	06	07				
2A0Z(A)							22	23	CT			
TLR8	NT	15	16	17	18	19	20	21	22	23	24	25	CT

C

| 2A0Z(A) | NT | 01 | | | | | | | | | | |
2A0Z(A)		15	16	17	18	19	20	21				
2Z63(A)				18	19	20						
2A0Z(A)						21	22	23	CT			
TLR9	NT	15	16	17	18	19	20	21	22	23	24	25	CT
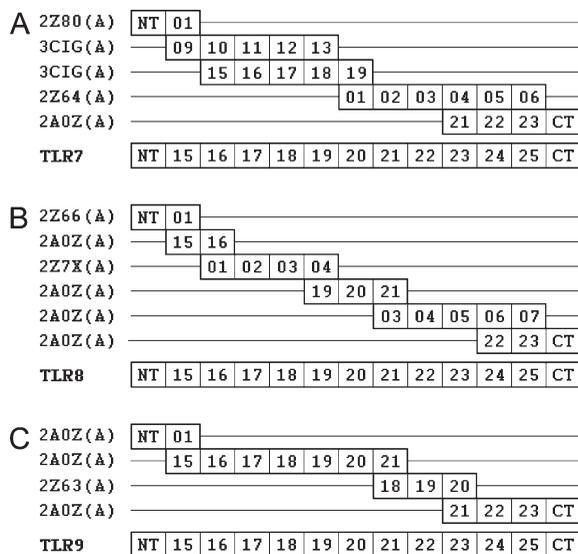
Figure 1. Models of multiple alignments between targets and templates. The numbers 01–25 denote the canonical LRRs; NT and CT denote N-/C-terminal LRRs. (A) Five segments selected from four structures (2Z80 chain A: human TLR2; 3CIG chain A: mouse TLR3; 2Z64 chain A: mouse TLR4; 2A0Z chain A: human TLR3) were used as templates for the human TLR7 ectodomain. (B) Six segments selected from three structures (2Z66 chain A: human TLR4; 2A0Z chain A: human TLR3; 2Z7X chain A: human TLR2) were used as templates for the human TLR8 ectodomain. (C) Four segments selected from two structures (2A0Z chain A: human TLR3; 2Z63 chain A: human TLR4) were used as templates for the human TLR9 ectodomain.

These ectodomains are highly variable. They are directly involved in the recognition of a variety of pathogen-associated motifs including lipopolysaccharide, lipopeptide, cytosine–phosphate–guanine (CpG) DNA, flagellin, imidazoquinoline, and ds/ssRNA.[13] Upon receptor activation, a TIR signaling complex is formed between the receptor and adaptor TIR domains.[14]

The receptors TLR7, 8, and 9 compose a family[15] with a longer amino acid sequence than other TLRs. They are localized intracellularly and signal in response to nonself nucleic acids. They also contain an irregular segment between their LRR14 and 15. A recent study showed that the ectodomains of TLR9 and 7 are cleaved in the endolysosome to recognize ligands.[16] Only the cleaved forms can recruit MyD88 on activation. In the absence of the crystallographic structures, we developed structural models of cleaved ligand-binding domains of TLR7/8/9 by homology modeling. From the structural model we predict potential ligand-binding sites and infer possible configurations of the receptor–ligand complex.

Results

Template identification

Our target structures are the cleaved functional ectodomains of the human TLR7/8/9 comprising LRR15−25 and N/C-terminal LRRs. All the six structure-known TLR homologues were employed as template sources: human TLR1/2/3/4 and mouse TLR3/4. The TLR ectodomain is composed of strictly organized LRRs. Nevertheless, the LRR number of cleaved ligand-binding domain of human TLR7/8/9 is 13 (LRR15−25 and N/C-terminal LRR),[16] whereas the LRR number of the structure-known TLRs varies from 20 to 25. Therefore, none of the structure-known TLRs is suitable to serve as a full length template. To overcome this limitation, LRR segments with higher sequence similarity to the individual LRRs in the target were selected from the six complete homologous structures. The segments were then combined into the multiple templates. Figure 1 shows the multiple alignment models for the three proteins TLR 7/8/9, presenting the relationship between target and template segments. The sequence similarity between each LRR pair (target/template LRR) is listed in Table I. The average target–template similarities of TLR7/8/9 are 47.70, 47.20, and 46.78%, respectively.

Remarkably, the group of TLR7/8/9 has a unique structural character that is absent in other TLRs. A specific segment (26−32 residue long) is located

steps.[8,9] First, selection of suitable template(s) related to the target sequence. A template segment assembly can usually improve the model quality.[10] Second, alignment of the target sequence to the template(s). Third, building coordinates of the three-dimensional model based on the alignment. Fourth, evaluation of the model and its refinement. The resulting model can then be used to infer biological functionalities or to generate hypotheses for new experiments. A recent study on TLR4[11] highlighted the reliability and the significance of homology modeling applied to TLRs.

The structure of a TLR consists of a leucine-rich repeat (LRR) ectodomain, a helical transmembrane domain, and an intracellular Toll/IL-1 receptor homology (TIR) signaling domain.[12] The ectodomain contains varying numbers of LRRs and resembles a solenoid bent into a horseshoe shape. At both ends there is a terminal LRR that shields the hydrophobic core of the horseshoe.

Table I. *Sequence Similarities (%) of Target–Template LRR Pairs*

	NT	15	16	17	18	19	20	21	22	23	24	25	CT	Avg
TRL7	28.60	58.30	60.00	41.70	47.10	58.30	46.20	52.00	50.00	33.30	46.20	48.30	50.00	47.70
TRL8	29.40	41.70	52.00	50.00	57.70	60.00	40.60	44.00	52.00	36.00	60.00	42.30	53.30	47.20
TRL9	32.30	58.30	48.00	50.00	44.40	41.70	46.90	51.90	36.00	54.20	50.00	52.00	42.40	46.78

In the header line, 15−25 denote canonical LRRs. NT and CT denote N-/C-terminal LRRs. Avg denotes the average values.

```
                              LxxLxLxxNxL
               ...TMGEADG GEKVWLQPGDLAPAPVDTPSSEDFRPNCSTL NFTLDLS...
PredictProtein ...LLLLLLL LLEEELLLLLLLLLLLLLLLLLLLLLHHHHHH LLLLLLL...
    NNPREDICT ...H------ --HEEE------------------------ --EEH--...
        SSPro ...HCCCCCC CCEEEECCCCCCCCCCCCCCCCCCCCCCCCCC EEEEECC...
       GOR IV ...HHHCCCC CCEEEECCCCCCCCCCCCCCCCCCCCCCCCCC CEECCCC...
               LRR14          Irregular Region          LRR15
```

Figure 2. Irregular region analysis of TLR9. Four methods (PredictProtein,[20] NNPREDICT,[21] SSPro,[22] and GOR IV[23]) were used to predict the secondary structures of the irregular region of TLR9. The results (italic letters) indicate a short β-sheet at position 3–5 of this region. Besides, this region matches the LRR pattern at three important positions (bold letters). These features support the presumption that this irregular region is a beginning N-terminal LRR after the ectodomain cleavage.

before LRR15, which was described as an undefined region.[17–19] The sequence similarity search against Protein Data Bank (PDB) provided no significant results. Thus, we carried out secondary structure predictions for this region with four different methods. As example, the results for TLR9 are shown in Figure 2. All methods indicated a short β-sheet at position 3–5 of the segment, which is a prominent characteristic of LRRs. In addition, we compared its amino acid sequence with the consensus sequence of LRRs. The most significant positions of the LRR consensus sequence, LxxLxLxxNxL, are the four L residues which form the hydrophobic core of a LRR structure. Here, the letter L not only stands for leucine but also for other highly hydrophobic residues. As illustrated in Figure 2, the specific segment of TLR9 contains three of the four highly hydrophobic residues. Also, the corresponding segment of TLR7/8 has the same features. Thus we regard this segment as an irregular LRR. Because the N-terminal LRR together with LRR1–14 of the receptor ectodomain are deleted upon arriving in endolysosome, this irregular LRR may become a new N-terminal LRR of the truncated structure. Moreover, multiple alignments of all known mammalian sequences showed that this region is very variable within each of the TLR7/8/9 groups. The structure of this LRR may be relatively relaxed, because it lacks the first L residue that participates in forming the hydrophobic core of a LRR structure and the N residue that forms hydrogen bonds between neighboring LRRs. These features also support the hypothesis that this irregular LRR is an N-terminal LRR. For this reason, a N-terminal LRR with known structure was selected as corresponding template (Fig. 1).

Structure modeling and evaluation

The three-dimensional coordinates of the models were created by MODELLER[24] and modified by ModLoop.[25] The final structures of the ectodomains of TLR7/8/9 reveal a large, arc-shaped assembly consisting of 11 canonical LRRs and two terminal LRRs, which adopted a right-handed solenoid structure (Fig. 3). The TLRs are distinct from other LRR proteins in that their LRR consensus motifs are often interrupted by extended insertions.[26] Two 4–7-residue-long insertions protuberate from the structure surface at LRR18 and LRR20, respectively. These insertions are well conserved in length and position on the sequence level in the three TLRs. The models show that the insertions are all located on one face of the arc, whereas the other face is insertion-free (Fig. 3). The convex site β-sheets are directed toward the insertion face. This feature is consistent with the known structures of TLR1/2/3/4. Because all the known ligand-binding sites of TLR1/2/3/4 are on the insertion face of the structure, the insertions suggest some functional significance. In addition, the human TLR7/8/9 are glycosylated as it is the case for other TLRs. The glycans were shown to be nonfunctional for ligand binding.[2–6] The NCBI protein database provides seven predicted N-linked glycosylation sites for TLR7/8 cleaved form and six for TLR9. All sites are located on the insertion-free faces. The PDB format files of the three final models are provided as Supporting Information Files 1–3. Evaluation of the models involved analysis of geometry, stereochemistry, and energy distributions in the models. The evaluation results (Table II) are indicative of a good quality of all three models.

Table II. *Model Evaluation*

	TLR7	TLR8	TLR9	TLR3
ProQ_LG/MS	5.340/0.461	4.613/0.402	4.355/0.339	7.923/0.526
PROCHECK	97.4%	96.2%	97.5%	99.6%
ModFOLD_Q/P	0.7588/0.01	0.7100/0.0126	0.7166/0.0121	0.7116/0.0124
MetaMQAP_GDT/RMSD	57.534/3.049 Å	53.908/3.121 Å	54.645/3.244 Å	79.322/1.566 Å

All these displayed scores indicate the models to be reliable in terms of overall packing. For comparison purpose, the values of TLR3 crystal structure (PDB code: 2A0Z) were also listed. ProQ_LG: >1.5 fairly good; >2.5 very good; >4 extremely good. ProQ_MS: >0.l, fairly good; >0.5, very good; >0.8, extremely good. PROCHECK: percentage of residues in most favored regions and additional allowed regions. ModFOLD_Q: >0.5, medium confidence; >0.75, high confidence. ModFOLD_P: <0.05, medium confidence; <0.01, high confidence. MetaMQAP_GDT/RMSD: an ideal model has a GDT score over 59 and a RMSD around 2.0 Å.

Homology Modeling of hTLR Ligand-Binding Domains

Figure 3. Structural models and ligand-binding regions of TLR7/8/9. Insertions are located on one face of the horseshoe, whereas the other face is insertion-free. The reported essential residues are located on the insertion face (labeled in blue). The orange regions are potential ligand-binding regions on the insertion face. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]
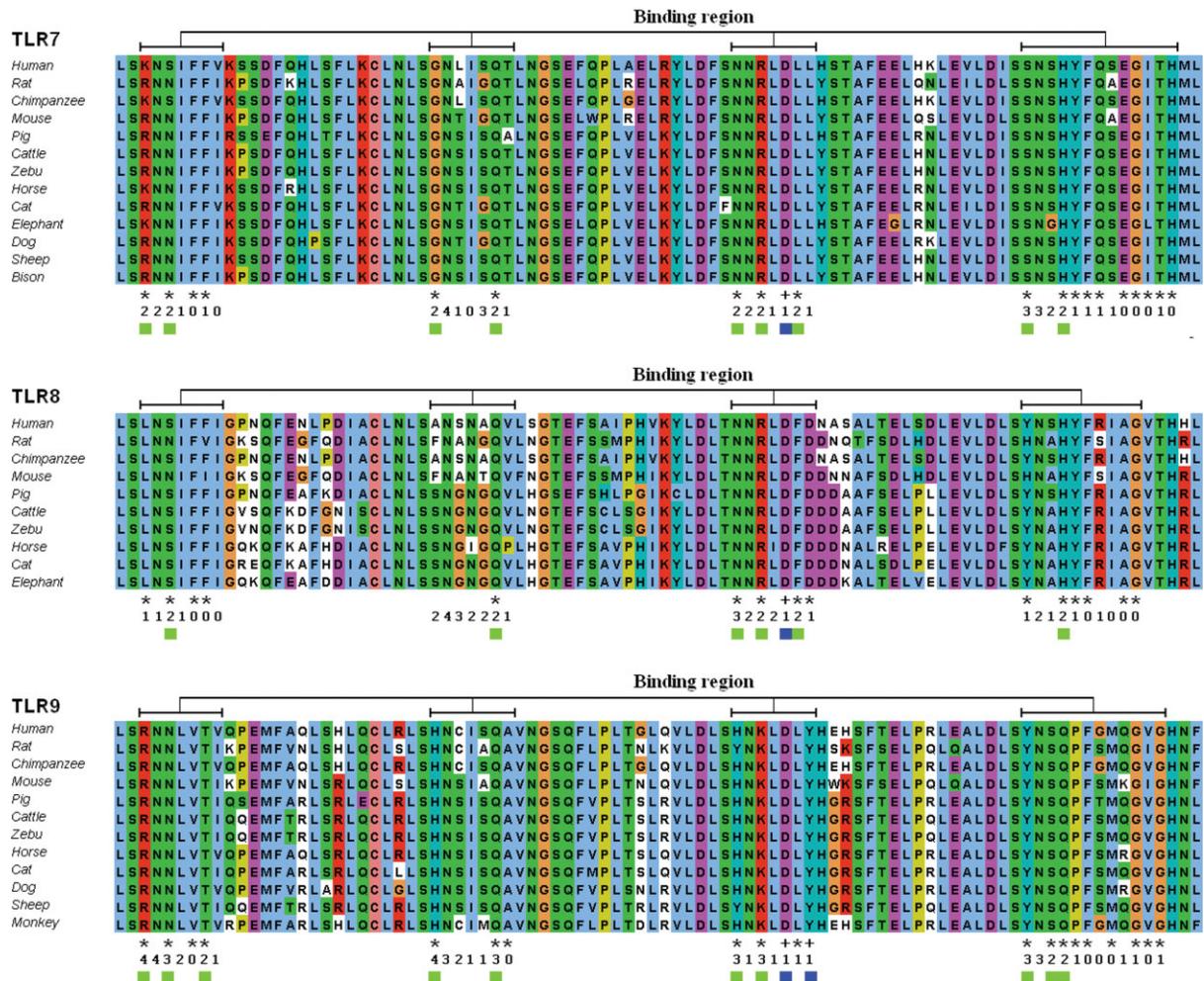
Figure 4. Partial multiple sequence alignments of different mammalian TLR7/8/9. The multiple sequence alignments represent the conservation of each residue in the potential ligand-binding regions (corresponding to the orange regions in Fig. 3). In the first line below the alignments, plus signs designate important residues as reported in the literature and the asterisks designate highly conserved positions. In the second line, the number of positive docking predictions of each position is indicated. In the third line, blue squares designate important residues as reported in the literature and green squares indicate the suggested ligand-binding residues. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]
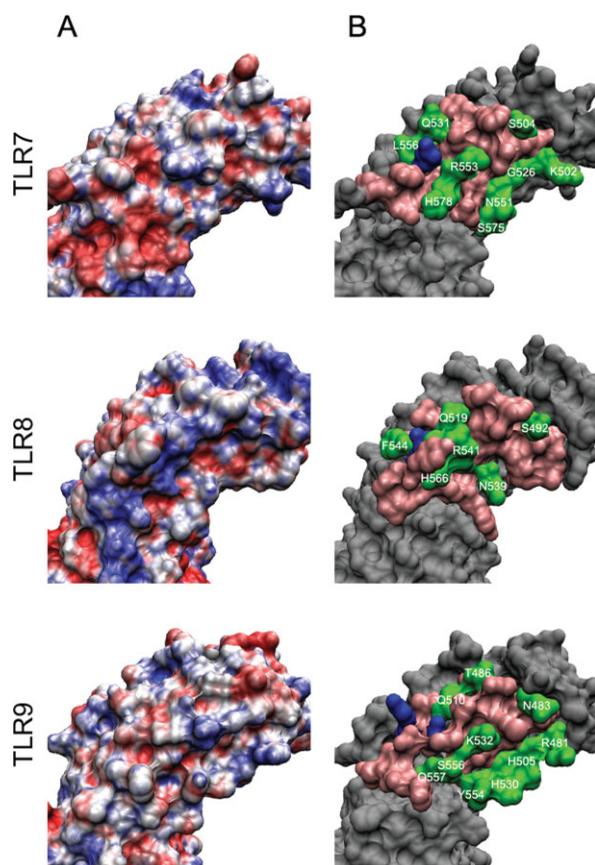
Figure 5. Surface analysis of ligand-binding regions of TLR7/8/9. (A) Surface charge distribution (APBS electrostatics) of ligand-binding regions of TLR7/8/9. Blue: positive charge; white: neutral; red: negative charge. (B) Important residues in ligand-binding regions of TLR7/8/9. Blue: important residues as reported in the literature; pink: residues close the blue ones but excluded from the potential ligand-binding residues through investigating processes; green: suggested potential ligand-binding residues (residue name and number are labeled). [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Potential ligand-binding residues

Several residues are essential for the ligand recognition: Asp543 in TLR8; Asp535 and Tyr537 in TLR9.[18,27] Our models can help to understand the biological function of these residues (Fig. 3). According to these reported residues and the sequence comparison of TLR7/8/9, we inferred a ligand-binding region for TLR7/8/9, respectively (detailed in the Discussion section). It is located at the insertion face of the ectodomain around LRR17 (Fig. 3). Because of the considerable size of the nucleic acids, the ligand-binding region should contain more interacting residues. We identified potential ligand-binding residues in the ligand-binding region aside from the experimentally determined ones. To accomplish this goal we integrated results from manual analyses and automatic docking programs.

TLR3 is closely related to the TLR7/8/9 family because of its intracellular localization and nucleic acid ligand. Therefore, we used the recently published crystal structure of the mTLR3-dsRNA 2:1 complex[6] as a guide to predict the essential interacting residues in TLR7/8/9. From all interacting residues of mTLR3, we identified three principles for the essential residues:

1. The essential residues are located on the protein surface and spatially close to each other.
2. They are highly conserved among species.
3. They create a nonnegatively charged environment.

On basis of these principles, we searched for additional residues that might be essential for ligand recognition. At first, surface residues that were spatially close (within two LRRs) to the experimentally determined essential residues were marked on the predicted models (orange regions in Fig. 3). These residues can be far from each other on the sequence level. Then, multiple alignments of all known mammalian TLR7/8/9 sequences were generated to select the highly conserved residues (columns with an asterisk in Fig. 4) from the marked ones. Notably, the L (or I, V) and N residues of the LRR consensus sequence LxxLxLxxNxL are conserved, but they cannot interact with ligands, because they are buried to form the hydrophobic core of an LRR. These residues are not labeled with asterisks in Figure 4.

Four protein-RNA docking programs and five protein-DNA docking programs (listed in the Materials and Methods section) were used to predict ligand-binding residues in TLR7/8 and TLR9. A residue from the prefiltered regions was marked as a ligand-binding residue, if it was positively predicted by at least two programs. In Figure 4, the number of positive predictions is listed for each target residue. The surface charge distributions of the regions of interest were calculated to verify the charge pattern in the predicted ligand-binding regions [Fig. 5(A)]. The resulting residues correspond to positively charged or neutral environments.

Figure 5(B) illustrates the protein surface residues from the different steps of our investigation for TLR7/8/9, respectively. All final predicted ligand-binding residues are summarized in Table III. These residues are indicated in green in both Figures 4 and 5(B).

Discussion

All three resulting models revealed similar conformations. This supports the assumption that TLR7/8/9 share

Table III. *Potential Ligand-Binding Residues of TLR7/8/9*

|	
TLR7	K502 S504 G526 Q531 N551 R553 L556 S575 H578
TLR8	S492 Q519 N539 R541 F544 H566
TLR9	R481 N483 T486 H505 Q510 H530 K532 Y554 S556 Q557
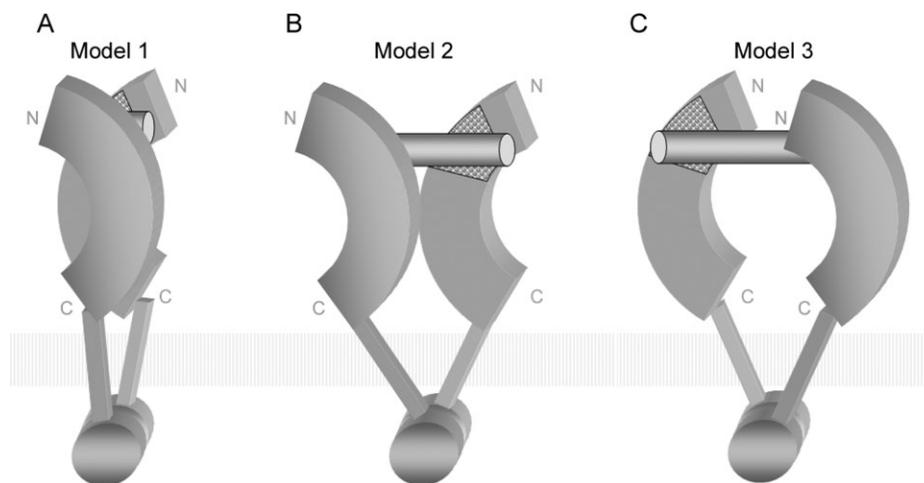
Figure 6. Proposed models of receptor–ligand 2:1 complex.

a common ligand-binding and signaling mechanism.[18] We compared and analyzed the predicted structures to suggest the receptor–ligand 2:1 complex models.

Ligand-binding region

The mouse TLR9 contains a short fragment in its LRR17 that is homologous to the methyl CpG DNA binding domain protein.[27] The mutant of Asp535 and Tyr537 in this fragment abolished the TLR9 function.[27] In the human TLR8, the Asp543 that corresponds to TLR9's Asp535 was determined to be required for the TLR8 function.[18] Through sequence comparison, the Asp residue was found to be highly conserved in the TLR7/8/9 family but not in other TLRs. We considered this Asp to be significant for TLR7, because the TLR7/8/9 are highly homologous and their ligands are all pathogen-derived nucleic acids. In particular, the TLR7 and 8 are present as tandem duplication in many studied genomes discussed by Roach *et al.*[15] In this regard, TLR7/8/9 have a ligand-binding region located spatially around the Asp residue.

We can further exclude the necessity of other ligand-binding regions on the ectodomains, because the minimum size of stimulatory oligonucleotides is six bases.[28] These oligonucleotides are not large enough to reach another ligand-binding region on the receptor.

Receptor dimerization

The signaling mechanism of all TLRs is likely to involve dimerization of the ectodomains.[18] However, this can be achieved in various ways by using different receptors and stimuli. TLR9 is a preformed dimer. The distance between both monomers is reduced upon contact with CpG DNA.[29] TLR1/2 are activated and connected into a heterodimer by triacylated lipopeptide.[4] TLR4 recognizes lipopolysaccharide indirectly

through the coreceptor protein MD-2 and is induced to form a TLR4-MD-2 homodimer.[5] In the TLR3 homodimer the dsRNA interacts with two regions of each receptor ectodomain. Direct protein–protein interactions between both receptors occur at their C-terminal LRRs, whereas the other regions are separated by the dsRNA.[6]

The structures obtained by the homology modeling together with the identification of possible ligand-binding sites can be used to derive a working hypothesis for the structure of the receptor–ligand complex. We propose three possible receptor–ligand 2:1 complex models for the TLR7/8/9 family (Fig. 6). In all three models, the ssRNA or CpG DNA ligand interacts with the binding region on the insertion surface of both receptor ectodomains. The ectodomains are on opposite sides of the ligand. Simultaneously, the intracellular TIR domains are also in a dimer configuration. Thus the C-terminal LRRs of each monomer, which are connected to the TIR through a 20-amino-acid-long transmembrane stretch, are spatially close to each other. The main difference between the three models is the relative position of the ectodomains. In the first model [Fig. 6(A)], both C-terminal LRRs are brought into proximity, forming a protein–protein contact. Both binding regions sandwich the ligand. In the second model [Fig. 6(B)], both receptors are shifted apart along the ligand extending directions back to back. In the third model [Fig. 6(C)], both receptors are shifted in opposite directions face to face. Obviously, the minimum ligand size required by the first model is the smallest. Therefore, a CpG DNA of six bases is already long enough to stimulate TLR9.[28] The minimum size required by the second and third models is larger. These two models, however, cannot be excluded, because there is so far no evidence that TLRs have only one dimer form. Without the crystal structure of their ligands, it is difficult to determine a more precise

model for the receptor dimerization. Hence, it remains interesting to study the atomic structure of the stimulatory ssRNA/CpG DNA and to further determine the detailed interactions between ligands and receptors.

Materials and Methods

Template identification and sequence alignments

Amino acid sequences with LRR motif partitions of human TLR7/8/9 ectodomain were extracted from TollML.[7] TollML is a specialized database of TLR sequence motifs, derived from the NCBI protein database.[30] Multiple sequence alignments of all individual LRRs of TLR7/8/9 to the LRR consensus sequence are provided as Supporting Information File 4. Because the TLR ectodomain is a repetitive protein (LRRs), we selected and combined segments from all the six known TLR ectodomain structures into multiple templates to optimize the sequence similarity between targets and templates. The six candidate templates were human TLR1/2/3/4 and mouse TLR3/4 and were obtained from the PDB.[31] The PDB codes are 2Z7X, 2Z80, 2A0Z, 2Z63, 2Z66, 3CIG and 2Z64, respectively. Three steps led to the identification of structural templates. First, we partitioned the known structures into a total of 136 individual LRRs. Because of the irregularity of the LRR sequences, the partition according to the LRR consensus sequences was performed manually. Second, the LRRs were collected into the LRRML database,[32] which can return the most similar LRR for an input LRR sequence through similartiy search. Third, optimal template pieces for each target were found and combined to generate multiple alignments. Because the TLR LRRs follow common characteristic consensus sequences, target–template alignments were generated more accurately by hand than through software.

Structure construction and analysis

The initial three-dimensional coordinates of the models were generated by the fully automated program MODELLER 9v3.[24] The input files were the multiple alignment file and the coordinate files of the templates. The ligand-binding domains of TLR7/8/9 contain two 4–7-residue-long insertion regions, which correspond to gaps in the multiple alignment. During the modeling these regions became loop structures, which limited the model accuracy. ModLoop[25] was used to modify these loop regions. The resulting models were evaluated by PROCHECK,[33] ProQ,[34] ModFOLD,[35] and MetaMQAP.[36]

The detection of potential ligand-binding sites was achieved through residue conservation analysis, surface charge analysis, and several automatic docking programs. BindN,[37] DP-Bind,[38] DBS-PRED,[39] DBS-PSSM,[40] and PreDs[41] were used for protein-DNA

docking of TLR9. BindN, Pprint,[42] RNAbindR,[43] and RISP[44] were used for protein-RNA docking of TLR7/8.

Conclusions

We predicted three-dimensional structures of the closely related TLR7/8/9 ligand-binding domains by homology modeling. LRR segments were selected from known TLR structures, which are locally optimal for the target sequences. These segments were then combined into multiple templates.

To predict essential residues in the ligand-binding region, sequence conservation and charge distributions were examined. Only highly conserved nonnegative residues that are positively predicted by at least two docking programs can be considered as potential ligand-binding residues. Based on these models we also suggest three possible receptor dimerization schemes which require different minimum ligand sizes.

In summary, our models provide a structural framework that can act as a guide to develop a functional hypothesis to interpret experimental data of TLR7/8/9. They may also facilitate efforts to design further site-directed mutagenesis to learn the ligand recognition and the downstream signaling mechanisms. The presented modeling approach can be extended to other repetitive protein domains.

References

1. Takeda K, Akira S (2005) Toll-like receptors in innate immunity. Int Immunol 17:1–14.
2. Bell JK, Botos I, Hall PR, Askins J, Shiloach J, Segal DM, Davies DR (2005) The molecular structure of the Toll-like receptor 3 ligand-binding domain. Proc Natl Acad Sci USA 102:10976–10980.
3. Choe J, Kelker MS, Wilson IA (2005) Crystal structure of human toll-like receptor 3 (TLR3) ectodomain. Science 309:581–585.
4. Jin MS, Kim SE, Heo JY, Lee ME, Kim HM, Paik SG, Lee H, Lee JO (2007) Crystal structure of the TLR1-TLR2 heterodimer induced by binding of a tri-acylated lipopeptide. Cell 130:1071–1082.
5. Kim HM, Park BS, Kim JI, Kim SE, Lee J, Oh SC, Enkhbayar P, Matsushima N, Lee H, Yoo OJ, Lee JO. (2007) Crystal structure of the TLR4-MD-2 complex with bound endotoxin antagonist Eritoran. Cell 130:906–917.
6. Liu L, Botos I, Wang Y, Leonard JN, Shiloach J, Segal DM, Davies DR (2008) Structural basis of toll-like receptor 3 signaling with double-stranded RNA. Science 320: 379–381.
7. Gong J, Wei T, Jamitzky F, Heckl WM, Rössle SC (2007) TollML-a user editable database for Toll-like receptors and ligands. In: Proceedings of the Second IAPR International Workshop on Pattern Recognition in Bioinformatics, International Association of Pattern Recognition, Singapore, Paper ID 231.
8. Ginalski K (2006) Comparative modeling for protein structure prediction. Curr Opin Struct Biol 16:172–177.
9. Kopp J, Schwede T (2004) Automated protein structure homology modeling: a progress report. Pharmacogenomics 5:405–416.

Homology Modeling of hTLR Ligand-Binding Domains

10. Bujnicki JM (2006) Protein-structure prediction by recombination of fragments. Chembiochem 7:19–27.
11. Kubarenko A, Frank M, Weber AN (2007) Structure-function relationships of Toll-like receptor domains through homology modelling and molecular dynamics. Biochem Soc Trans 35:1515–1518.
12. Brodsky I, Medzhitov R (2007) Two modes of ligand recognition by TLRs. Cell 130:979–981.
13. Gay NJ, Gangloff M (2007) Structure and function of Toll receptors and their ligands. Annu Rev Biochem 76: 141–165.
14. O'Neill LA, Bowie AG (2007) The family of five: TIR-domain-containing adaptors in Toll-like receptor signalling. Nat Rev Immunol 7:353–364.
15. Roach JC, Glusman G, Rowen L, Kaur A, Purcell MK, Smith KD, Hood LE, Aderem A (2005) The evolution of vertebrate Toll-like receptors. Proc Natl Acad Sci USA 102:9577–9582.
16. Ewald SE, Lee BL, Lau L, Wickliffe KE, Shi GP, Chapman HA, Barton GM (2008) The ectodomain of Toll-like receptor 9 is cleaved to generate a functional receptor. Nature 456:658–662.
17. Bell JK, Mullen GE, Leifer CA, Mazzoni A, Davies DR, Segal DM (2003) Leucine-rich repeats and pathogen recognition in Toll-like receptors. Trends Immunol 24: 528–533.
18. Gibbard RJ, Morley PJ, Gay NJ (2006) Conserved features in the extracellular domain of human toll-like receptor 8 are essential for pH-dependent signaling. J Biol Chem 281:27503–27511.
19. Matsushima N, Tanaka T, Enkhbayar P, Mikami T, Taga M, Yamada K, Kuroki Y (2007) Comparative sequence analysis of leucine-rich repeats (LRRs) within vertebrate toll-like receptors. BMC Genomics 8:124.
20. Rost B, Yachdav G, Liu J (2004) The PredictProtein server. Nucleic Acids Res 32:W321–W326.
21. Kneller DG, Cohen FE, Langridge R (1990) Improvements in protein secondary structure prediction by an enhanced neural network. J Mol Biol 214:171–182.
22. Cheng J, Sweredoski M, Baldi P (2005) Accurate prediction of protein disordered regions by mining protein structure data. Data Min Knowl Discov 11:213–222.
23. Garnier J, Gibrat JF, Robson B (1996) GOR method for predicting protein secondary structure from amino acid sequence. Methods Enzymol 266:540–553.
24. Fiser A, Do RK, Sali A (2000) Modeling of loops in protein structures. Protein Sci 9:1753–1773.
25. Fiser A, Sali A (2003) ModLoop: automated modeling of loops in protein structures. Bioinformatics 19:2500–2501.
26. Bell JK, Askins J, Hall PR, Davies DR, Segal DM (2006) The dsRNA binding site of human Toll-like receptor 3. Proc Natl Acad Sci USA 103:8792–8797.
27. Rutz M, Metzger J, Gellert T, Luppa P, Lipford GB, Wagner H, Bauer S (2004) Toll-like receptor 9 binds single-stranded CpG-DNA in a sequence- and pH-dependent manner. Eur J Immunol 34:2541–2550.
28. He G, Patra A, Siegmund K, Peter M, Heeg K, Dalpke A, Richert C (2007) Immunostimulatory CpG oligonucleotides form defined three-dimensional structures: results from an NMR study. ChemMedChem 2:549–560.
29. Latz E, Verma A, Visintin A, Gong M, Sirois CM, Klein DC, Monks BG, McKnight CJ, Lamphier MS, Duprex WP, Espevik T, Golenbock DT. (2007) Ligand-induced conformational changes allosterically activate Toll-like receptor 9. Nat Immunol 8:772–779.
30. Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Khovayko O, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Ostell J, Pruitt KD, Schuler GD, Shumway M, Sequeira E, Sherry ST, Sirotkin K, Souvorov A, Starchenko G, Tatusov RL, Tatusova TA, Wagner L, Yaschenko E (2008) Database resources of the national center for biotechnology information. Nucleic Acids Res 36:D13–D21.
31. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242.
32. Wei T, Gong J, Jamitzky F, Heckl WM, Stark RW, Roessle SC (2008) LRRML: a conformational database and an XML description of leucine-rich repeats (LRRs). BMC Struct Biol 8:47.
33. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 26: 283–291.
34. Wallner B, Elofsson A (2006) Identification of correct regions in protein models using structural, alignment, and consensus information. Protein Sci 15:900–913.
35. McGuffin LJ (2008) The ModFOLD server for the quality assessment of protein structural models. Bioinformatics 24:586–587.
36. Pawlowski M, Gajda MJ, Matlak R, Bujnicki JM (2008) MetaMQAP: a meta-server for the quality assessment of protein models. BMC Bioinformat 9:403.
37. Wang L, Brown SJ (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res 34:W243–W248.
38. Hwang S, Gou Z, Kuznetsov IB (2007) DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins. Bioinformatics 23: 634–636.
39. Ahmad S, Gromiha MM, Sarai A (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics 20:477–486.
40. Ahmad S, Sarai A (2005) PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformat 6:33.
41. Tsuchiya Y, Kinoshita K, Nakamura H (2005) PreDs: a server for predicting dsDNA-binding site on protein molecular surfaces. Bioinformatics 21:1721–1723.
42. Kumar M, Gromiha MM, Raghava GP (2008) Prediction of RNA binding sites in a protein using SVM and PSSM profile. Proteins 71:189–194.
43. Terribilini M, Sander JD, Lee JH, Zaback P, Jernigan RL, Honavar V, Dobbs D (2007) RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. Nucleic Acids Res 35:W578–W584.
44. Tong J, Jiang P, Lu ZH (2008) RISP: a web-based server for prediction of RNA-binding sites in proteins. Comput Methods Programs Biomed 90:148–153.

Paper 6

A leucine-rich repeat assembly approach for homology modeling of human TLR5-10 and mouse TLR11-13 ectodomains

Tiandi Wei, Jing Gong, Ferdinand Jamitzky, Wolfgang M. Heckl, Shaila C. Rössle and Robert W. Stark

A leucine-rich repeat assembly approach for homology modeling of human TLR5-10 and mouse TLR11-13 ectodomains

Tiandi Wei [a,b], Jing Gong [a,b,]*, Shaila C. Rössle [b], Ferdinand Jamitzky [a,c], Wolfgang M. Heckl [a,d,e] and Robert W. Stark [a,b]

[a]*Center for Nanoscience, Ludwig-Maximilians-Universität München. 80799 Munich, Germany.*

[b]*Department of Earth and Environmental Sciences, Ludwig-Maximilians-Universität München. 80333 Munich, Germany.*

[c]*Leibniz Supercomputing Centre. 85748 Garching, Germany.*

[d]*Deutsches Museum. 80538 Munich, Germany.*

[e]*TUM School of Education, Technische Universität München. 80799 Munich, Germany.*

*Corresponding author:

Add: Theresienstr. 41, 80333 Munich, Germany

Email: gongj@informatik.uni-muenchen.de

Tel: +49-89-2180-4359

Fax: +49-89-2180-4334

Abstract

So far, 13 groups of mammalian Toll-like receptors (TLRs) have been identified. Most TLRs have been shown to recognize pathogen-associated molecular patterns from a wide range of invading agents and initiate both innate and adaptive immune responses. The TLR ectodomains are composed of varying numbers and types of leucine-rich repeats (LRRs). As the crystal structures are currently missing for most TLR ligand-binding ectodomains, homology modeling enables first predictions of their three-dimensional structures on the basis of the determined crystal structures of TLR ectodomains. However, the quality of the predicted models that are generated from full-length templates can be limited due to low sequence identity between the target and templates. To obtain better templates for modeling, we have developed an LRR template assembly approach. Individual LRR templates that are locally optimal for the target sequence are assembled into multiple templates. This method was validated through the comparison of a predicted model with the crystal structure of mouse TLR3. With this method we also constructed ectodomain models of human TLR5, TLR6, TLR7, TLR8, TLR9, and TLR10 and mouse TLR11, TLR12, and TLR13 that can be used as first passes for a computational simulation of ligand docking or to design mutation experiments. This template assembly approach can be extended to other repetitive proteins.

Key words

1. Introduction

Cells of the innate immune system, such as macrophages and dendritic cells, express a limited number of germline-encoded pattern-recognition receptors (PRR) that specifically recognize pathogen-associated molecular patterns (PAMPs) within microbes. These molecular patterns are unique to these microbes and are absent in the host [1]. Toll-like receptors (TLRs) are currently the best-characterized members of the PRRs [2]. The progress of genome sequencing projects has led so far to the identification of 13 groups of TLRs in mammalian genomes, ten in humans and 13 in mice [3], and more than 20 in non-mammalian genomes [4]. All TLRs have a common domain organization, with an extracellular ectodomain, a helical transmembrane domain, and an intracellular Toll/IL-1 receptor homology (TIR) domain [5]. The extracellular domain (ectodomain) is responsible for the recognition of common structural patterns in various microbial molecules. For example, lipoproteins or lipopeptides are recognized by TLR2 complexed with TLR1 or TLR6, viral double-stranded RNAs by TLR3, lipopolysaccharides by TLR4, bacterial flagellins by TLR5, single-stranded RNAs by TLR7 or TLR8, and microbial CpG DNAs by TLR9 [6, 7]. The TIR domains of TLRs are associated with the intracellular signaling cascade leading to the nuclear translocation of the transcription factor NF-κB [8].

A TLR ectodomain contains 19 to 27 consecutive leucine-rich repeat (LRR) motifs sandwiched between two terminal LRR modules (LRRNT and LRRCT) [4]. LRRs exist in more than 6000 proteins and more than 100 crystal structures of these proteins have been deposited in the Protein Data Bank (PDB) [9, 10]. In every case, the protein adopts an arc or horseshoe shape. An individual LRR motif is defined as an array of 20 to 30 amino acids that is rich in the hydrophobic amino acid leucine. All LRR sequences can be divided into a conserved segment and a variable segment. The conserved segments, LxxLxLxxNxL, generate the concave surface of the LRR arc or horseshoe by forming parallel β-strands, while the variable parts form its convex surface consisting of helices or loops. The terminal LRRNT and LRRCT modules stabilize the protein structure by shielding its hydrophobic core from exposure to solvent.

To date, only the crystal structures of the ectodomains of human TLR1 through 4 and mouse TLR2 through 4 have been determined [11-15]. High-throughput genome sequencing projects, however, have led to the identification of more than 2000 TLR sequences. Thus, the structures of most TLRs are still unknown because structure determination by X-ray diffraction or nuclear magnetic resonance spectroscopy experiments remains time-consuming. Protein structure prediction methods are powerful tools to bridge the gap between sequence determination and structure determination.

Homology modeling, also referred to as comparative modeling, is currently the most accurate computational method for protein structure prediction. This approach constructs a three-dimensional model for a target protein sequence from a three-dimensional template structure of a homologous protein. Thus, the quality of the homology model strongly depends on the sequence identity between the target and template. Below 30% identity, serious errors may occur [16]. Due to different repeat numbers and distinct arrangements of LRRs in the TLR ectodomains, a proper full-length template with a sufficiently high sequence identity to the target is often missing. This limitation can be overcome by assembling multiple LRR templates. In this approach the most similar (on the sequence level) LRR with a known structure is

searched as a local template for each LRR in the target sequence. Such an LRR template may be derived from TLRs or from other proteins. Thereby, a suitable template may be found even for an insertion-containing irregular LRR. All local template sequences are then combined to generate a multiple sequence alignment for the complete target sequence. Thus, a high-quality model can be created, even if no adequate single template is available. To facilitate a multiple template assembly of LRR proteins, we have developed the LRRML database [9], which archives individual LRR structures manually identified from all known LRR protein structures. In addition, we have developed TollML [4], a database of sequence motifs of TLRs. In TollML, all known sequences of TLR ectodomains were semi-automatically partitioned into LRR segments and are made available for query. For newly sequenced TLRs that are not yet archived in TollML, we have implemented an LRR prediction program named LRRFinder on the TollML webpage. It requires as input an LRR-containing amino acid sequence and returns the number and positions of LRRs in the input sequence. LRRFinder recognizes LRR motifs based on a position-specific weight matrix scan, with the sensitivity and specificity both higher than 93%. With the help of these two databases, LRR partitions of a TLR ectodomain can be directly obtained, and an optimal structure template for each LRR segment can be quickly found. A schematic flowchart of the modeling procedure is shown in Figure 1.

In this study, we apply the multiple template assembly approach to TLRs. To demonstrate the potential of the method we constructed two models of the mouse TLR3 ectodomain as a test case using our LRR template assembly method and a standard profile-profile alignment-aided full-length template recognition method. Both models were then compared with the crystal structure of mouse TLR3. The overall and ligand-binding site conformation of the template assembly-based model is closer to that of the crystal structure than that based on the standard method. We also modeled the human TLR5 through 10 and mouse TLR11 through 13 ectodomains, which represent mammalian TLR ectodomains with unknown structures. A comparison of the model for human TLR6 with the very recently reported crystal structure of mouse TLR6 shows a very good structural agreement.

2. Methods

2.1 Template selection and sequence alignment

Amino acid sequences of mouse TLR3, human TLR5 through 10, and mouse TLR11 through 13 ectodomains were extracted from TollML release 3.0 (IDs 627, 531, 571, 992, 575, 1022, 851, 703, 705, and 704). Their LRR partitions were annotated by TollML. For each LRR sequence contained in each target TLR, the three-dimensional LRR structure with the highest sequence identity was selected as a template from LRRML through a sequence similarity search. Then, a multiple sequence alignment of a target with all its local LRR templates was generated with each template comprising one alignment line. For instance, the mouse TLR3 has a total of 25 LRRs and accordingly required 25 templates. The associated multiple sequence alignment then has 26 lines (Figure 1). Because of the characteristic consensus sequences of LRRs, these alignments were made more accurately manually than automatically. To generate an alternate model with standard methods, the widely acknowledged template recognition program pGenTHREADER [17] was executed to find templates for mouse TLR3. This method calculates sequence profiles from an input sequence and combines profile-

profile alignments with secondary structure specific gap-penalties, pair potentials, and solvation potentials using a linear combination. The output is the complete PDB structures that serve as candidate templates ranked by P-values. Each candidate sequence is aligned with the target sequence.

2.2 Model construction and validation

The initial three-dimensional coordinates of all models were calculated by MODELLER 9v7 [18]. The above-described alignment file and the corresponding template structures of a target model were inputted into the default 'model' routine of MODELLER. A given number of three-dimensional models were calculated. The ectodomains of TLRs contain a number of insertion regions. Some of them corresponded to four to 15 amino acid-long gaps in the alignments because their templates do not contain a corresponding insertion. During modeling, these gaps produced loop structures in the model, thus deteriorating the model accuracy. ModLoop [19] was used to rebuild the coordinates of these loop regions. Finally, we used the model quality assessment programs ProQ [20] and MetaMQAP [21] to evaluate the output candidate models and select the one with the best scores as the final model. The structure superimpositions and molecular electrostatics involved in the structural analysis were carried out using SuperPose v1.0 [22] and VMD [23], respectively. The docking studies of TLR11 and its ligand profilin were performed with GRAMM-X [24].

3. Results

3.1 LRR templates

The number of LRRs in the full-length ectodomains of mouse TLR3, human TLR5 through 10, and mouse TLR11 through 13 is 25, 23, 21, 28, 28, 28, 21, 26, 25, and 27, respectively. Consequently, a total of 252 individual LRR templates sourced from 41 different PDB structures were selected from LRRML. Their sequence identities with the targets vary from 26.0% to 95.7% (43.8% on average), and similarities from 39.0% to 100% (58.2% on average). Remarkably, all cases of relatively low sequence identity (< 35%) were caused by highly irregular target LRRs. These highly irregular sequences include LRRNT/CTs, the highly mutated LRR15 of TLR7/8/9, and the insertion-containing LRRs whose templates do not include a similar insertion. The sources (LRRML IDs) and sequence identities of all LRR templates are listed in Table 1.

As the modeling of mouse TLR3 was carried out to verify our approach, we assumed that the crystal structure of the mouse TLR3 ectodomain was unknown and excluded the corresponding LRR entries of mouse and human TLR3 ectodomains from LRRML before the template search. The selected individual LRR templates for mouse TLR3 were associated with 18 PDB structures, 14 of which were from non-TLR proteins. The target-template sequence identities range from 33.3% to 50.0% (44.1% on average). By contrast, pGenTHREADER provided only complete PDB structures of LRR proteins as candidate templates, with each candidate possessing a pairwise sequence alignment with the target. Because no single template covered the entire sequence of mouse TLR3, we selected the first seven candidates by rank (except mouse and human TLR3) and combined them into a multiple alignment to avoid template gaps. These templates included PDB structures 2Z64, 2Z81, 1O6V, 3FXI, 2Z7X, 1JL5, and 3BZ5, which covered the closest homologues (mouse TLR2/4 and human TLR1/4) to mouse TLR3

among all proteins with known structures. Nevertheless, the sequence identities of the seven templates to mouse TLR3 range from 16.1% to 21.2% (18.7% on average), which fell much below the cut-off value 30% for homology modeling [16].

3.2 Structural models

3.2.1 Model of mouse TLR3

Recently, we constructed a model of mouse TLR3 with the LRR template assembly method as a test case of the LRRML database [9]. It revealed a horseshoe-shaped assembly adopting a regular solenoid structure without disordered regions. The model was superimposed with the crystal structure of mouse TLR3 ectodomain (PDB code: 3CIG) [14] at its both ligand-binding regions, LRRNT through LRR3 and LRR19 through LRR21. The backbone root mean square deviation (RMSD) is 1.96 Å and 1.90 Å, respectively [9]. To verify the improvements of the database and the modelling process, we reconstructed the mouse TLR3 model (Figure 2b) with up-to-date LRR templates. Compared with the old model, four of the 25 LRRs of TLR3 in the new model were assigned new templates with higher sequence identities (Table 1). Because these four LRRs are not involved in the TLR3 ligand-binding sites, the corresponding RMSD values of the new model were the same as for the previous model. These values indicate that the model predicted with our method well matches the crystal structure and can be used to predict potential ligand-binding sites [25].

For comparison purposes, the mouse TLR3 ectodomain was also modeled with a standard profile-profile alignment-aided full-length template recognition method. All of the ten output models obtained from MODELLER for the full-length templates-based standard alignment showed a serious structural disorder spanning from LRR6 through LRR10 (Figure 2a). The LRR6 through LRR10 on the crystal structure form a regular solenoid structure with an α-helix in the variable segment of LRR8 (Figure 2c). By contrast, the corresponding LRRs on the model completely lost the proper LRR shape and interwove with one another. The disorder was caused by mismatches or target gaps in the alignment, where only two to four template LRRs were assigned to five target LRRs (Figure 2d). The standard alignment could not create a one-to-one correspondence between the target and template LRR units due to the irregularity of the LRRs. ProQ and MetaMQAP were used to evaluate the quality of the different models of mouse TLR3 (Table 2). Both programs make an integrative assessment of the structure quality considering geometry, stereochemistry, and energy distribution of the structures. Both template assembly-based models received better scores than the model based on standard method.

3.2.2 Models of human TLR5 through 10 and mouse TLR11 through 13

With the LRR template assembly method we modeled the human TLR5 through 10 and mouse TLR11 through 13 ectodomains. All of the resulting models are provided in Supplementary File 1. Ramachandran plots of these models were created with PROCHECK [26] and are provided in Supplementary File 2. Model evaluation data by ProQ and MetaMQAP are listed in Table 2. The models reveal a horseshoe shape (Figure 3), where a longer or shorter sequence (more or less LRR units) implies a smaller or larger horseshoe opening, e.g., TLR7/8 (smaller opening) and TLR6/10 (larger opening). Their overall structural similarity reflects the phylogenetic

relationships among these TLRs. For example, TLR6 is similar to TLR10 while TLR7 is similar to TLR8, consistent with the molecular tree proposed by Roach and co-workers [27]. Mammalian TLRs are distinct from other LRR proteins in that they contain two to seven insertion-containing irregular LRRs, which may be necessary for ligand-binding and receptor dimerization. Our models show that all insertions are located on that face of the horseshoe, to which the convex site β-strands point, whereas the other face is completely insertion-free.

To highlight the availability of these models for an analysis of receptor-ligand interaction mechanisms, we performed molecular electrostatics calculations of the mouse TLR11 model with a ligand. TLR11 can recognize profilin of some apicomplexan protozoa parasites. This protein is involved in parasite motility and invasion [28]. Expression of TLR11 is suppressed in humans [29]. The electrostatic analysis (Figure 4a) shows that the entire surface of the profilin of *P. falciparum* (PDB code: 2JKF) is predominantly negatively charged, whereas TLR11 exhibits several positively charged patches (Figure 4b). Protein docking studies using GRAMM-X showed that profilin and the positive patches on TLR11 possess compatible sizes and electrostatic complementarity (Supplementary File 3).

Very recently, the crystal structure of the TLR6 ectodomain complexed with TLR2 and a Pam2CSK4 ligand was released in PDB (PDB code: 3A79). The TLR6 structure is a hybrid structure of mouse and inshore hagfish, where 18 mouse LRRs (LRRNT through LRR17) were hybridized with two hagfish LRRs (LRR18 and LRRCT) [30]. This crystal structure served as an additional benchmark for our template assembly approach. The superimposition of the mouse part of the crystal structure with our human TLR6 model yielded a backbone RMSD of 1.94 Å, which indicates that the model is very similar to the crystal structure (Figure 5). A second model of human TLR6 was generated with pGenTHREADER in a similar procedure as described for the mouse TLR3 (Supplementary File 1). The backbone RMSD between the crystal structure and this model is 1.89 Å. The only full-length template used for this model was the structure of human TLR1 (PDB code: 2Z7X). Because TLR6 and TLR1 possess the same number of LRRs and have a very high sequence identity (63.3%), the structure of TLR1 serves as an excellent full-length template. Under these very good conditions, both, the standard and the template assembly approaches, provided high-quality models.

4. Discussion

In template-based protein modeling, the overall sequence identity between the target and template is an important criterion for the selection of suitable templates [31]. For repetitive LRR proteins, however, there is often no appropriate full-length template available due to different repeat numbers and distinct arrangements. This problem can be solved by combining individual repeating units that are locally optimal for the target sequence. In the method validation with mouse TLR3, the average target-template sequence identity achieved by our method was 44.8%, which was significantly higher than that (18.7%) achieved by a standard profile-profile alignment-aided template recognition method. However, both the standard and the template assembly methods produced models of human TLR6 very well matched the crystal structure of mouse TLR6. The comparison between the models obtained with both methods highlights the potential of the template assembly approach. It can produce models with a similar quality as the standard profile-profile alignment method. The template assembly method, however, reveals its particular strength in situations where no adequate full-length

templates are available. In the case of TLR3, the standard profile-profile alignment method failed to predict a reliable model due to significant gaps in the template. Here, the template assembly method overcomes the difficulties and generates a realistic model.

In a previous work, we constructed models of human TLR7/8/9 ligand-binding domains by combining LRR segments that were extracted from all known crystal structures of TLRs [32]. The average target-sequence similarities for TLR7/8/9 were 47.7%, 47.2%, and 46.8%, respectively. The resulting models supported experimentally determined ligand-binding residues [33, 34] and provided a reliable basis to identify potential ligand-binding residues and potential receptor dimerization mechanisms. Here, we went a step further and extended the scope of the approach by searching LRR segments from all LRR-containing proteins with known structures with the LRRML database because the same type of LRRs can exist in different proteins [9]. Consequently, 33 of the 41 source PDB structures are non-TLR proteins (numbers derived from Table 1). The average target-sequence similarities for TLR7/8/9 increased to 55.9%, 58.2%, and 59.2%.

Another key issue in LRR protein modeling is the sequence-level LRR partition of the target TLR sequence. The indicated number and beginning/end positions of LRRs in TLRs vary largely across different databases or research reports due to the irregularity and periodicity of LRRs. TollML reports the most complete and accurate LRR motifs for TLRs as compared with a number of databases [4]. In addition, TollML provides a statistics-based LRR prediction program LRRFinder for new TLR entries that are not yet collected in TollML. It can recognize LRRs from an input amino acid sequence with high confidence.

In conclusion, this work depicts an LRR template assembly approach for protein homology modeling. The comparison of a mouse TLR3 model with its crystal structure underlined feasibility and reliability of the method. With this method, a series of mammalian TLR ectodomains were modeled. These models can be used to perform ligand docking studies or to design mutagenesis experiments and hence to investigate TLR ligand-binding mechanisms. Our modeling approach can be extended to other repetitive proteins.

Acknowledgements

References

[1]   Kawai T, Akira S (2008) Toll-like receptor and RIG-I-like receptor signaling. Ann N Y Acad Sci 1143:1-20.

[2]   Medzhitov R (2007) Recognition of microorganisms and activation of the immune response. Nature 449:819-826.

[3]   Kumar H, Kawai T, Akira S (2009) Pathogen recognition in the innate immune response. Biochem J 420:1-16.

[4]   Gong J, Wei T, Zhang N, Jamitzky F, Heckl WM, Rössle SC, Stark RW (2009) TollML: a database of Toll-like receptor structural motifs. J Mol Model in press.

[5]   Brodsky I, Medzhitov R (2007) Two modes of ligand recognition by TLRs. Cell 130:979-981.

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

[7]   Akira S, Hemmi H (2003) Recognition of pathogen-associated molecular patterns by TLR family. Immunol Lett 85:85-95.

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

[9]   Wei T, Gong J, Jamitzky F, Heckl WM, Stark RW, Rössle SC (2008) LRRML: a conformational database and an XML description of leucine-rich repeats (LRRs). BMC Struct Biol 8:47.

[10]   Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. Nucleic Acids Res 28:235-242.

[11]   Jin MS, Kim SE, Heo JY, Lee ME, Kim HM, Paik SG, Lee H, Lee JO (2007) Crystal structure of the TLR1-TLR2 heterodimer induced by binding of a tri-acylated lipopeptide. Cell 130:1071-1082.

[12]   Choe J, Kelker MS, Wilson IA (2005) Crystal structure of human toll-like receptor 3 (TLR3) ectodomain. Science 309:581-585.

[13]   Bell JK, Botos I, Hall PR, Askins J, Shiloach J, Segal DM, Davies DR (2005) The molecular structure of the Toll-like receptor 3 ligand-binding domain. Proc Natl Acad Sci U S A 102:10976-10980.

[14]   Liu L, Botos I, Wang Y, Leonard JN, Shiloach J, Segal DM, Davies DR (2008) Structural basis of toll-like receptor 3 signaling with double-stranded RNA. Science 320:379-381.

[15]   Kim HM, Park BS, Kim JI, Kim SE, Lee J, Oh SC, Enkhbayar P, Matsushima N, Lee H, Yoo OJ, Lee JO (2007) Crystal structure of the TLR4-MD-2 complex with bound endotoxin antagonist Eritoran. Cell 130:906-917.

[16]   Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294:93-96.

[17]   Lobley A, Sadowski MI, Jones DT (2009) pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. Bioinformatics 25:1761-1767.

[18]   Fiser A, Do RK, Sali A (2000) Modeling of loops in protein structures. Protein Sci 9:1753-1773.

[19]   Fiser A, Sali A (2003) ModLoop: automated modeling of loops in protein structures. Bioinformatics 19:2500-2501.

[20]   Wallner B, Elofsson A (2003) Can correct protein models be identified? Protein Sci 12:1073-1086.

[21]   Pawlowski M, Gajda MJ, Matlak R, Bujnicki JM (2008) MetaMQAP: a meta-server for the quality assessment of protein models. BMC Bioinformatics 9:403.

[22]   Maiti R, Van Domselaar GH, Zhang H, Wishart DS (2004) SuperPose: a simple server for sophisticated structural superposition. Nucleic Acids Res 32:W590-594.

[23]   Humphrey W, Dalke A, Schulten K (1996) VMD: visual molecular dynamics. J Mol Graph 14:33-38, 27-38.

[24]   Tovchigrechko A, Vakser IA (2006) GRAMM-X public web server for protein-protein docking. Nucleic Acids Res 34:W310-314.

[25]   Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T (2007) Assessment of CASP7 predictions for template-based modeling targets. Proteins 69 Suppl 8:38-56.

[26]   Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst 26:283-291.

[27]   Roach JC, Glusman G, Rowen L, Kaur A, Purcell MK, Smith KD, Hood LE, Aderem A (2005) The evolution of vertebrate Toll-like receptors. Proc Natl Acad Sci U S A 102:9577-9582.

[28]   Yarovinsky F, Zhang D, Andersen JF, Bannenberg GL, Serhan CN, Hayden MS, Hieny S, Sutterwala FS, Flavell RA, Ghosh S, Sher A (2005) TLR11 activation of dendritic cells by a protozoan profilin-like protein. Science 308:1626-1629.

[29]   Lauw FN, Caffrey DR, Golenbock DT (2005) Of mice and man: TLR11 (finally) finds profilin. Trends Immunol 26:509-511.

[30]

[31]   Kopp J, Schwede T (2004) Automated protein structure homology modeling: a progress report. Pharmacogenomics 5:405-416.

[32]   Wei T, Gong J, Jamitzky F, Heckl WM, Stark RW, Rössle SC (2009) Homology modeling of human Toll-like receptors TLR7, 8, and 9 ligand-binding domains. Protein Sci 18:1684-1691.

[33]   Rutz M, Metzger J, Gellert T, Luppa P, Lipford GB, Wagner H, Bauer S (2004) Toll-like receptor 9 binds single-stranded CpG-DNA in a sequence- and pH-dependent manner. Eur J Immunol 34:2541-2550.

[34]   Gibbard RJ, Morley PJ, Gay NJ (2006) Conserved features in the extracellular domain of human toll-like receptor 8 are essential for pH-dependent signaling. J Biol Chem 281:27503-27511.

Table 1: Source and sequence identity (%) of the LRR templates.

	mTLR3		hTLR5		hTLR6		hTLR10		mTLR12	
LRR	LRRML ID	Iden	LRRML ID	Iden	LRRML ID	Iden	LRRML ID	Iden	LRRML ID	Iden
NT	406	47.6	125	34.6	333	72.7	150	40.9	125	28.6
1	65	45.8	284	45.8	334	54.2	334	45.8	90	37.5
2	212	45.8	264	42.3	335	70.8	335	62.5	178	58.3
3	465	41.7	259	54.2	336	81.0	578	42.9	465	40.7
4	151	50.0	464	46.2	337	80.0	337	68.0	578	40.9
5	177	50.0	259	48.0	338	60.9	338	47.8	21	35.5
6	110	46.2	252	34.6	339	36.0	111	33.3	497	56.0
7	275	45.8	494	29.0	340	33.3	340	41.7	202	40.9
8	8	41.4	135	37.0	341	33.3	28	34.6	39	34.6
9	203	42.3	359	26.6	342	48.1	342	44.4	356	45.8
10	484	54.2	357	41.7	343	34.6	343	46.4	179	45.8
11	293	50.0	140	50.0	344	48.3	365	29.0	177	42.3
12	270	33.3	316	50.0	345	45.5	345	40.9	566	54.2
13	357	41.7	151	45.8	346	41.7	346	41.7	66	50.0
14	65	42.9	216	45.0	347	61.5	358	48.0	468	45.8
15	152	50.0	153	38.1	348	40.0	369	41.7	501	44.0
16	259	40.0	253	36.0	349	68.2	349	45.5	581	34.8
17	316	37.5	356	40.0	350	95.7	350	43.5	372	30.8
18	152	38.5	566	41.4	351	52.0	372	45.5	283	37.0
19	239	50.0	72	58.3	128	37.5	513	36.0	396	38.7
20	239	45.8	579	45.5	-	-	-	-	260	37.5
21	92	50.0	484	40.9	-	-	-	-	496	41.4
22	80	41.7	-	-	-	-	-	-	315	45.8
23	628	46.2	-	-	-	-	-	-	110	35.7
CT	575	42.9	149	31.7	149	31.7	514	29.9	149	20.0

	hTLR7		hTLR8		hTLR9		mTLR11		mTLR13	
LRR	LRRML ID	Iden	LRRML ID	Iden	LRRML ID	Iden	LRRML ID	Iden	LRRML ID	Iden
NT	257	40.6	250	30.0	125	29.0	125	36.7	125	26.5
1	293	45.8	509	45.8	314	41.7	76	38.5	491	54.2
2	107	36.0	573	48.0	101	41.7	191	37.5	107	50.0
3	106	52.4	581	47.6	219	55.0	288	44.0	254	54.2
4	261	45.8	509	45.8	310	41.7	559	36.4	152	50.0
5	140	44.0	564	44.0	157	41.7	557	40.0	4	53.8
6	583	47.6	106	52.4	106	47.6	238	30.4	419	52.0
7	105	45.8	509	45.8	501	41.7	463	44.0	285	44.0
8	573	45.8	573	50.0	259	37.5	271	45.8	135	41.7
9	79	45.8	494	41.7	357	50.0	157	45.8	488	52.0
10	500	46.2	573	42.3	367	37.0	506	44.0	498	36.7
11	500	52.0	316	54.2	337	40.0	237	45.8	128	33.3
12	488	40.7	261	37.0	274	31.0	509	46.2	288	48.0
13	129	44.0	92	37.5	92	33.3	500	54.2	505	45.8
14	386	37.9	259	44.0	357	34.6	7	50.0	573	54.2
15	36	33.3	300	29.6	139	15.6	111	45.8	260	50.0
16	316	41.7	494	41.7	170	37.5	38	42.3	148	36.0
17	358	46.2	158	40.0	316	40.0	484	43.5	262	45.8
18	316	41.7	314	41.7	135	45.8	360	29.2	293	50.0
19	494	50.0	296	37.5	316	44.0	73	37.5	105	45.8
20	503	41.7	128	45.8	493	40.0	403	28.6	260	48.0
21	92	41.7	77	45.8	252	41.7	102	30.4	466	46.2
22	78	48.0	573	60.0	288	44.0	121	34.5	565	41.7
23	552	45.8	169	36.0	111	37.5	316	50.0	93	50.0
24	210	50.0	283	53.8	289	41.7	35	36.0	79	50.0
25	133	40.7	271	45.8	259	45.8	-	-	573	41.7
26	274	48.3	144	42.3	357	52.0	-	-	-	-
CT	149	36.7	149	33.9	149	28.8	469	26.2	149	36.5

Table 2: Evaluation of the crystal structure and models of the TLRs. Higher ProQ_LG/MS and MetaMQAP_GDT values indicate higher model qualities; higher MetaMQAP_RMSD values indicate lower model qualities. [a] Model of Wei *et al.* (2008).

Structure/Model	ProQ_ LG/MS	MetaMQAP_ GDT/RMSD
mTLR3 crystal	7.215/0.469	69.840/1.962
mTLR3 pGenTHREADER	4.013/0.357	41.943/4.691
mTLR3 [a]	5.136/0.423	54.068/3.080
mTLR3 new	5.349/0.405	54.542/3.030
hTLR5	4.707/0.358	54.803/3.126
hTLR6	5.807/0.439	73.230/1.907
hTLR7	4.980/0.381	52.816/3.224
hTLR8	5.053/0.408	53.490/3.113
hTLR9	5.025/0.386	53.774/3.181
hTLR10	4.835/0.362	59.918/2.883
mTLR11	4.371/0.351	49.178/3.433
mTLR12	4.546/0.337	46.131/3.493
mTLR13	4.827/0.407	55.709/2.982

Figure 1: Flowchart of the LRR template assembly method.

Figure 2: Homology models and crystal structure of the mouse TLR3 ectodomain. (a) The homology model based on the standard method. The framed region exhibits serious disorder. (b) The homology model based on template assembly method. (c) The crystal structure (PDB code: 3CIG). The dotted region is an insertion on LRR20 that is missing in the crystal structure. (d) The target-template sequence alignment of the disordered region of the standard method based model. The mismatches and target gaps resulted in the disorder in (a), where two to four template LRRs were wrongly assigned to five target LRRs.
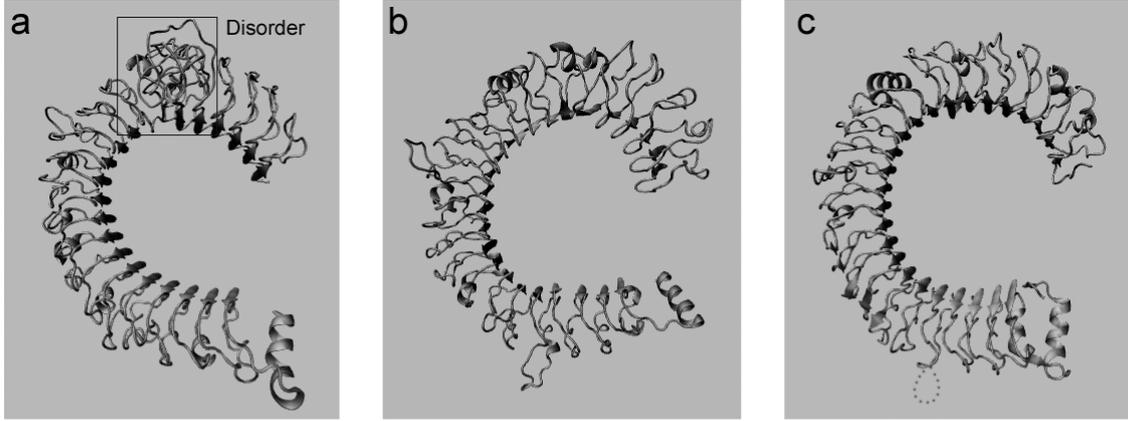

```
          LRR6                        LRR7                        LRR8
mTLR3  |LQELLLAKNKILALRSEELEFLGNSS|LRKLDLSSNPLKEF-SPGCFQTIGK|LFALLLNNAQLNPHLTEKLCWELSNTS|
2Z64   |LVHVDLSYNYIQTITVNDLQF-----|-----LRENPQVNL-----------|--SLDMSLNPIDFIQ---------DQA|
2Z81   |--------------------------|------------------------|---------PIDFIQ---------DQA|
1O6V   |--------------------------|--------------------GSATI|TQDTPINQIFTDTALAEKMKTVLGKTN|
3FXI   |LQKLVAVETNLASLEN--FPIGHLKT|LKELNVAHNLIQSFKLPEYFSNLTN|LEHLDLSSNKIQSIYCTDLRVLHQM--|
2Z7X   |--------------------------|------------------------|---------------------------|
1JL5   |--------------------------|-----KSKTEYYNA-WSEWERNA--|-----PPGNGEQREMAVSRLRDCLDRQ|
3BZ5   |--------------------------|-------------------TLKA|GQTQSFNDWFPDDNFASEVAAAFEMQA|

          LRR9                              LRR10
mTLR3  |IQ-------NLSLANNQLLATSESTFSGLK---------WTN|LTQLD---LSYNNLHDV-GNGSFSYLPS|
2Z64   |FQGIKLHELTLRGNFNSSNIMKTCLQNLAGLHVHRLILGEFK|---------DERNLEIF-EPSIMEGLCD|
2Z81   |TF-------SEIRRI-----------DFAG---------LTS|LNELEIKALSLRNYQS-------QSLKS|
1O6V   |VT-------DTVS--------------QTD---------LDQ|VTTLQ---ADRLGIKSI-D--GVEYLNN|
3FXI   |------------------------------------PLL|NLSLD---LSLNPMNFI-QP---GAFKE|
2Z7X   |---------KISCH--------------P---------TVN|LKHLD---LSFNAFDALPICKEFGNMSQ|
1JL5   |AH-------ELELNNLGLSSLP------EL---------PPH|LESLV---ASCNSLTEL-P----ELPQS|
3BZ5   |TD-------TISEEQLA----------------------T|LTSLD---CHNSSITDM-TG--IEKLTG|
```

Figure 3: Models of human TLR5 through 10 and mouse TLR11 through 13 ectodomains. The N-linked glycan sites of these TLRs were obtained from the NCBI protein database and are labelled with black balls.
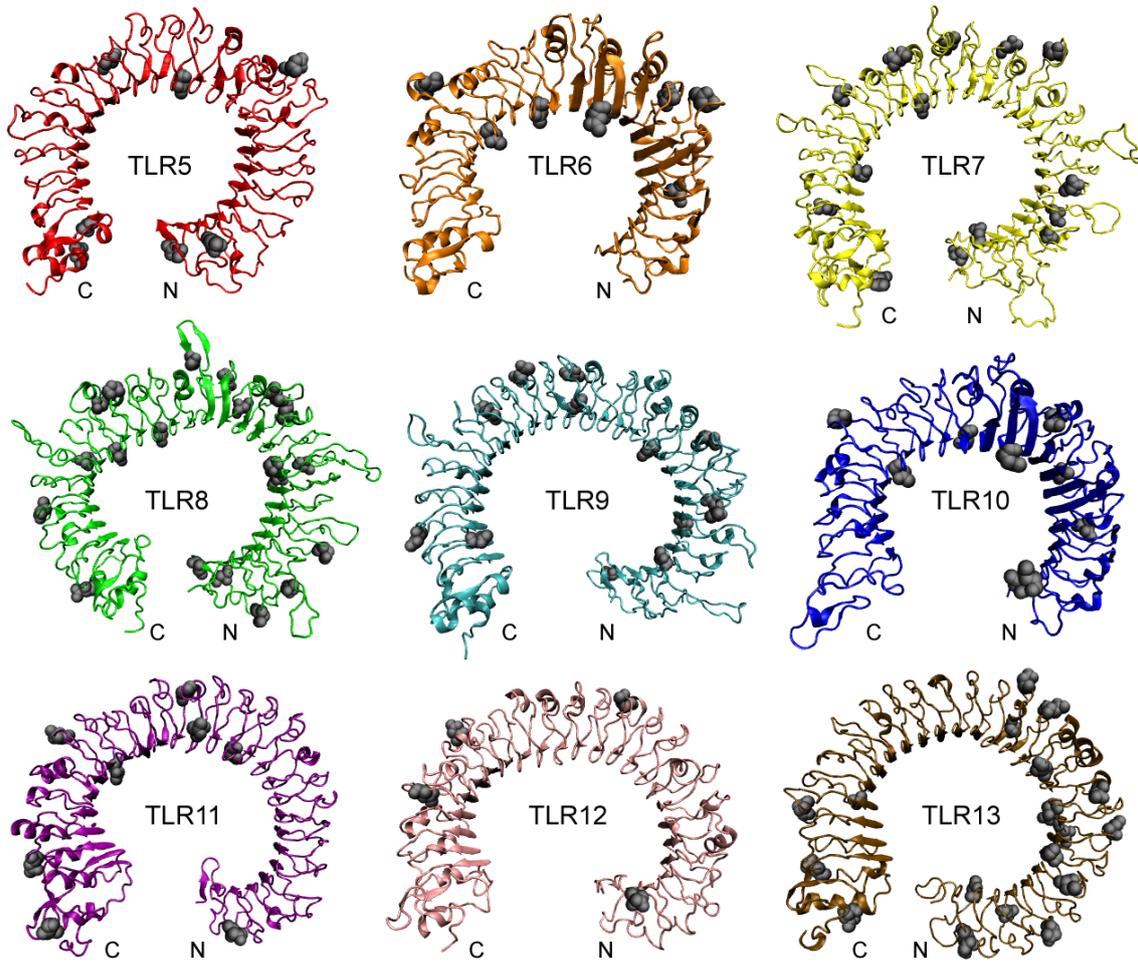
Figure 4: Surface charge analysis (APBS electrostatics) of (a) the crystal structures of profilin (PDB code: 2JKF) and (b) the model of mouse TLR11 ectodomain. Blue: positive charge; red: negative charge.
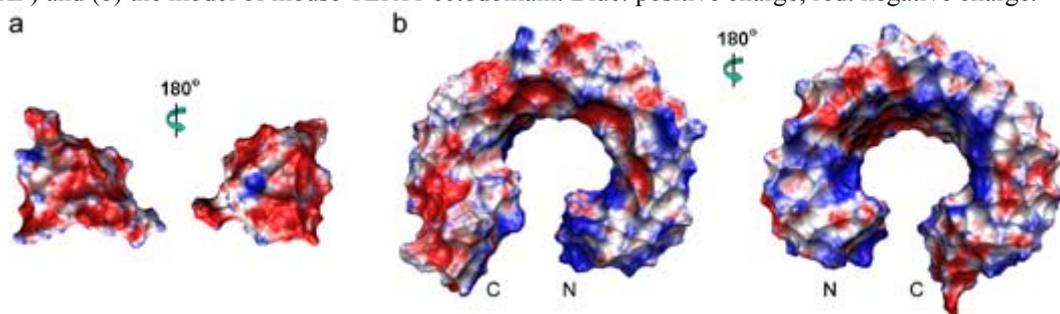
Figure 5: Superimposition of the homology model and the crystal structure of the TLR6 ectodomain. Green: homology model of human TLR6 (LRRNT through LRR17); orange: crystal structure of mouse TLR6 (LRRNT through LRR17). The Pam2CSK4 ligand-bind site is located on the variable parts of LRR10 through LRR12. The overall backbone RMSD is 1.94 Å and the backbone RMSD of the ligand-binding region is 1.18 Å.
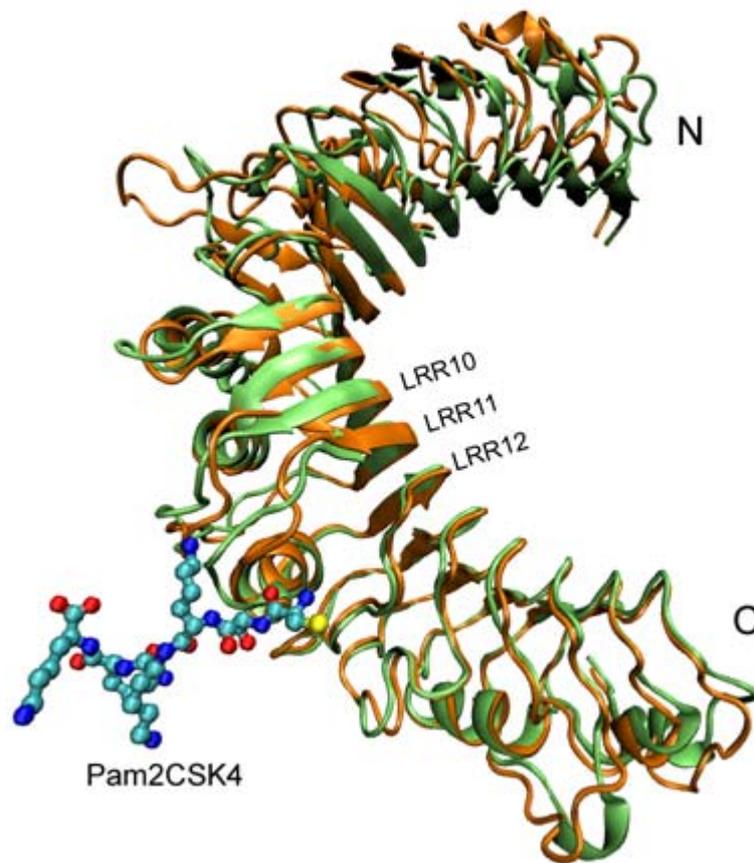
Supplementary file 1: Three-dimensional models of human TLR5-10 and mouse TLR10-13 ectodomains.

Supplementary file 2: Ramachandran plots of the models.

Supplementary file 3: Docking models of TLR11 and profiling ligand.

CV

Personal

Name: Jing Gong
Sex: Female
Born: 1980 in Tianjin, China
Nationality: Chinese

Education

since 02/2007	PhD candidate Ludwig-Maximilians-Universität München Dissertation: Structural databases and molecular interaction models of Toll-like receptors
10/2003-01/2007	Diplom (BS+MS), Bioinformatics Munich Technical University & Ludwig-Maximilians-Universität München Thesis: Prediction methods for subcellular localization of proteins
11/2002-06/2003	DSH German Certificate Ludwig-Maximilians-Universität München
09/1998-07/2002	BS, Marine Biology Ocean University of China Research project: The effects of different N/P Ratios on the growth of *Isochrysis galbana*
09/1986-07/1998	Primary, middle and high schools Final grade: A

Journal publications

1. Tiandi Wei, **Jing Gong**, Ferdinand Jamitzky, Wolfgang M. Heckl, Shaila C. Rössle, Robert W. Stark: A leucine-rich repeat assembly approach for homology modeling of human TLR5-10 and mouse TLR11-13 ectodomains. J. Mol. Model., in press.

2. Johannes Aßfalg, **Jing Gong**, Hans-Peter Kriegel, Alexey Pryakhin, Tiandi Wei, Arthur Zimek: Investigating a correlation between subcellular localization and fold of proteins. J. Univ. Comput. Sci., 2010, 16(5):604-621.

3. **Jing Gong**, Tiandi Wei, Ferdinand Jamitzky, Wolfgang M. Heckl, Robert W. Stark, Shaila C. Rössle: TollML: a Database of Toll-like Receptor Structural Motifs. J. Mol. Model., 2010, 16(7):1283-1289.

4.   Maciej Lech, Veronika Skuginna, Onkar P. Kulkarni, **Jing Gong**, Tiandi Wei, Robert W. Stark, Cecilia Garlanda, Alberto Mantovani, Hans-Joachim Anders: Lack of SIGIRR/TIR8 aggravates hydrocarbon oil-induced systemic lupus. J. Pathol., 2010, 220(5):596-607.

5.   **Jing Gong**, Tiandi Wei, Robert W. Stark, Ferdinand Jamitzky, Wolfgang M. Heckl, Hans J. Anders, Maciej Lech, Shaila C. Rössle: Inhibition of the Toll-like receptors TLR4 and 7 signaling pathways by SIGIRR: a computational approach. J. Struct. Biol., 2010, 169(3):323-330.

6.   Tiandi Wei, **Jing Gong**, Ferdinand Jamitzky, Wolfgang M. Heckl, Robert W. Stark, Shaila C. Rössle: Homology modeling of human Toll-like receptors TLR7, 8 and 9 ligand-binding domains. Protein Sci., 2009, 18(8):1684-1691.

7.   Johannes Aßfalg, **Jing Gong**, Hans-Peter Kriegel, Alexey Pryakhin, Tiandi Wei, Arthur Zimek: Supervised ensembles of prediction methods for subcellular localization. J. Bioinform. Comput. Biol., 2009, 7(2):269-85.

8.   Tiandi Wei, **Jing Gong** (joint first author), Ferdinand Jamitzky, Wolfgang M. Heckl, Robert W. Stark, Shaila C. Rössle: LRRML: a conformational database and an XML description of leucine-rich repeats (LRRs). BMC Struct. Biol., 2008, 8:47.

9.   Dongyan Liu, Jun Sun, **Jing Gong**, Shuben Qian: The effects of different N/P Ratios on the growth of *Isochrysis galbana*. Prog. Fish. Sci., 2002, 1(23): 29-32.

Conference proceedings

1.   **Jing Gong**, Tiandi Wei, Ferdinand Jamitzky, Wolfgang M. Heckl, Robert W. Stark, Shaila C. Rössle: Moldels of SIGIRR inhibiting the Toll-like receptor TLR4 signaling pathway. In Proc. International Workshop on Computational and Integrative Biology, 2009, Hangzhou China, page:1-13.

2.   Johannes Aßfalg, **Jing Gong**, Hans-Peter Kriegel, Alexey Pryakhin, Tiandi Wei, Arthur Zimek: Supervised Ensembles of Prediction Methods for Subcellular Localization. In Proc. 6th Asia Pacific Bioinformatics Conference, 2008, Kyoto, Japan, page: 29-38.

3.   **Jing Gong**, Tiandi Wei, Ferdinand Jamitzky, Wolfgang M. Heckl, Shaila C. Rössle: TollML – a User Editable Database for Toll-like Receptors and Ligands. In Proc. 2nd IAPR International Workshop on Pattern Recognition in Bioinformatics, 2007, Singapore, paper ID: 231.

Presentations and posters

1.   Poster, International Workshop on Computational and Integrative Biology, Hangzhou, China, "Homology modeling of human Toll-like receptors 7, 8 and 9 ligand-binding domains", September, 2009

2.   Invited speaker, 3rd Annual Workshop on Oligonucleotides in Cell Biology and Therapy, Graduated School 1202 of the DFG, Frauenchiemsee, Germay, "Databases of

Toll-like receptors", September, 2008

3. Poster, European Conferences on Computational Biology, Sardinia, Italy, "LRRML – a conformational database of leucine-rich repeats (LRRs)", September, 2008

4. Poster, 2nd IAPR International Workshop on Pattern Recognition in Bioinformatics, Singapore, "TollML – a user editable database of Toll-like receptors and ligands", October, 2007

Awards and honors

2007-2010   Ph.D. Scholarship, German Research Foundation (DFG)

2007-2010   Travel Award, German Research Foundation (DFG)

2001-2002   Bronze Medal Student Scholarship, Ocean University of China

1999-2001   Silver Medal Student Scholarship, Ocean University of China