John von Neumann Institute for Computing

TollML: A Database of Toll-Like Receptor Structural Motifs

F. Jamitzky, J. Gong, T. Wei, W. M. Heckl, S. C. Rössle

published in

From Computational Biophysics to Systems Biology (CBSB08),
Proceedings of the NIC Workshop 2008,
Ulrich H. E. Hansmann, Jan H. Meinke, Sandipan Mohanty,
Walter Nadler, Olav Zimmermann (Editors),
John von Neumann Institute for Computing, Jülich,
NIC Series, Vol. **40**, ISBN 978-3-9810843-6-8, pp. 241-244, 2008.

http://www.fz-juelich.de/nic-series/volume40

TollML: A Database of Toll-Like Receptor Structural Motifs

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

[1] Leibniz Supercomputing Centre, Garching, Germany
E-mail: jamitzky@lrz.de

[2] Department für Umwelt- und Geowissenschaften, LMU München, Germany

[3] Deutsches Museum, München, Germany

During recent years Toll-like receptors (TLRs) have spearheaded a tremendous research interest and the amount of sequenced relevant proteins grows exponentially. A critical step towards the successful TLR structure modeling is to generate the leucine-rich repeats (LRR) motif aided sequence alignment between the target sequence and the templates. However, because of the irregularity of LRR motifs in TLRs, most TLRs have no LRR annotaions in current databases, and in those TLRs with LRR partitions, the indicated repeat number and the boundaries of LRRs are quite different among databases. In order to provide a useful platform for structure prediction and analysis of these sequences, we developed TollML, an XML based database specialized for TLR structural motifs. Its original TLR sequences were extracted from NCBI's protein database. The LRR motifs as well as transmembrane and TIR motifs for all known TLR sequences are identified and annotated manually and can then be used for the prediction of protein structures via alignment and threading. The resulting database has been used for the structure prediction of TLR7, 8 and 9.

1 Introduction

Toll-like receptors (TLRs) play a crucial role in innate immunity[1]. To date, 13 TLRs have been identified in mammalian, and equivalent forms of many of these have been found in other vertebrate species. Under a structural view, the TLRs consists of 3 parts: the TIR domain inside the membrane, the transmembrane region and the ectodomain (ECD) formed by 18 to 25 leucine-rich repeats (LRRs). It is just the ECD that is directly involved in recognition of a variety of pathogens (ligands). Although for most of these TLRs ligand recognition specificity, downstream adapter molecules and signaling pathway have now been established, we still do not know much about their structural interaction with ligands. In 2005 the crystal structure of TLR3 ectodomain (ECD) was resolved and recently the crystal structures of TLR1/2 and 4 in respective complexes with agonist and antagonist ligands were shown. All these explained how the LRR based platform is adapted to the recognition of ligands. However, with high throughput genome sequencing projects the amount of sequenced TLR proteins continues to grow exponentially. It is clear that the discrepancy between the rate at which novel protein sequences are discovered and the rate at which detailed structural information will be obtained from X-ray diffraction or nuclear magnetic resonance spectroscopy (NMR) will continue for the foreseeable future. For this reason, there is a pressing need for theoretical methods to predict protein structures from their sequence. The understanding of their structural interactions can help us design vaccines, understand autoimmune diseases, and define the correlates of immune

protection. The TLR7, 8 and 9 constitute the TLR7 family which is one of the six major vertebrate TLR families. They are all located in the endosome and recognize nucleic acids. Our objective is to construct models for TLR7, 8 and 9 ECD based on the structure known TLRs and other LRR containing proteins.

2 Database Construction

TollML entries were originally extracted from NCBI proteins database and PDB[2] using the search keys: toll* and tlr*, where the star (*) stands for any suffix. The data were then filtered semi-automatically to exclude TLR related molecules such as adaptors, protein kinases and transcription factors. The metabolic pathways information was then extracted from the KEGG-database for each TLR entry[3]. Aside from the extracted information, TollML contains additional annotations, which must be manually accomplished. These annotations include LRR partitions, ligand information and structural information achieved through other projects or extracted from published articles. The indicated number of LRRs and their boundaries in individual TLRs are quite different among databases or researchers. This difference reflects the irregularity of LRR motifs in TLRs. We standardized the LRR definition and partitioned each TLR ECD into LRRs manually. The generated XML-datafile was then stored in an XML-database[4].

3 Contents of Database Entries

The current version (1.2) contains 2232 entries[5]. 1529 entries are of mammalian and the rest are of non-mammalian. All entries are divided into 24 groups, from TLR1 to TLR23 and others. A special tag named TollML label records which group an entry belongs to as a quick search index. The entry distribution over different TLR families is illustrated in Tab. 1. Each entry in TollML provides information of one TLR protein.

TLR	1	2	3	4	5	6	7	8	9	10-14	other	total
Mammalia	55	92	77	921	47	46	71	55	83	81	2	1529
Non-mam	267	132	22	16	22	1	13	8	12	54	159	703
Total	322	224	99	937	69	47	84	63	95	135	161	2232

Table 1. The entry distribution of TLR families for mammalian/non-mammalian groups.

4 Construction of a Conformational LRR Database

Leucine-rich repeats are an array of 20 to 30 amino acid long protein segments. Every segment is rich in the hydrophobic amino acid leucine. They play an important role in protein-protein interactions, such as signal transduction, cell adhesion, DNA repair, recombination, transcription, RNA processing, disease resistance, ice nucleation, apoptosis and

innate immune response. The first crystal structure of LRR containing protein, a ribonuclease inhibitor, was determined in 1993. It is a horseshoe- shaped structure containing 15 LRRs with a parallel beta-sheet lining the inner circumference and alpha helices flanking the outer circumference.

LRRs are present in over 6000 proteins from viruses to eukaryotes. For more than 80 of them the structure is known. In order to create a convenient workbench to carry out the homology modeling and to manage the structure known LRRs flexibly and efficiently, we decided to construct a conformational LRR database.

All leucine-rich repeats 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 consensus sequence LxxLxLxxN(Cx)xL, in which L stands for Leu, Ile, Val or Phe, N stands for Asn, Thr, Ser or Cys and x is any amino acid. A short beta-sheet begins always at the third position. 4 L residues at position 1, 4, 6, and 11 participate in the hydrophobic core. The side chains of asparagines (N) at position 9 form hydrogen bonds between neighbor LRRs in the loop structure. The VS of LRRs is quite different in length and consensus sequence. It can contain a variety of secondary structures.

5 Protein Comparative Modeling

At present there are over 80 LRR containing proteins whose crystal structures are available in RCSB Protein Data Bank (PDB), including the ECDs of TLR1/2, 3 and 4. They provide useful resources for homology modeling of TLR7, 8 and 9.

Comparative modeling, also called homology search, exploits the fact that evolutionarily related proteins with similar sequences, have similar structures. The process of building a comparative model is conceptually straightforward. First, Blast searches for the sequence to be modeled (the target) against a database of known protein structures is performed to find a most similar sequence (the parent). The similarity is usually greater than 35%. Second, an alignment is generated between the target and the parent. This sequence alignment is used to construct an initial model (sometimes referred to as a framework or template) by copying over some main chain and side chain coordinates from the parent structure based on the equivalent residue in the sequence alignment. At last, the model is improved by energy minimization and molecular dynamics.

Acknowledgments

This work was supported by Graduiertenkolleg 1202 of Deutsche Forschungsgemeinschaft (DFG).

References

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

2. Helen M. Berman, et al. *The Protein Data Bank*,Nucleic Acids Res. **28**, 235-242, 2000.
3. Gong J, Wei T, Jamitzky F, Heckl W M, Roessle S C, *TollML a User-Editable Database for Toll-like Receptors and Ligands*, Proc. Suppl. 2nd IAPR Itermational Workshop on Pattern Recognition in Bioinformatics (PRIB 2007) , Singapore, 2007.
4. `http://exist-db.org/`
5. `http://zeus.krist.geo.uni-muenchen.de/~tollml`.