Toll-like Receptor Structural Determinants: Variability Analysis by Digital Signal Processing Methods

Norbert Maggi, *Member, IEEE,* Patrizio Arrigo, and Carmelina Ruggiero, *Member, IEEE*

***Abstract*— The design of new drugs delivery systems is strongly dependent on the capability to maximize biocompatibility and reduce immunotoxicity. The minimization of foreign body reactions is one of the critical step, and toll-like receptors play a pivotal role in sensing and activation of response against exogenous elements. The complexity of these molecules brings about the need to identify those local structural elements that preferentially interact with different classes of compounds. We have applied Digital Signal Processing (DSP) methods to identify the regions containing these structural elements. DSP analysis has been carried out on 'wild-type' and several allelic forms of Toll-like receptor 1 cDNA. DSP has enabled the screening of allele specific nucleotide domains that could have an effect on allele-specific response against exogenous compounds.**

INTRODUCTION

THE innate immunity system is an evolutionary conserved defence system mainly devoted to respond against attacks by pathogens (virus and bacteria). The metabolic pathway of natural immunity is very complex and embeds several reaction cascades. An alteration of this surveillance mechanism has been demonstrated in many pathological conditions, such as cancer and cardiovascular diseases. As relates to cancer conventional treatment, such as radiotherapy [1] and chemotherapy [2-4], it can be noted that immunotherapy approaches use the specificity of the immune system in order to provide a more efficacious and a better tolerated therapy [5]. Moreover, relevant evidence indicates that the role of innate immunity in cardiovascular diseases may be very important [6]. A class of proteins called Toll-like receptors plays a critical role, in the activation of an innate response. The human family of Toll-like receptors encompasses 10 members. Some of them (TLR1, TLR2, TLR4, TLR5, TLR6 and TLR10) are located on the external part of the cellular membrane. The others (TLR3, TLR7, TLR8 and TLR9) are settled in the endosome or in other internal cellular compartments. TLRs are present in several cell types, such as macrophages, NK cells, T cells [6, 7]. The efficacy of individual natural responses against the detrimental effects of pathogen's attacks or exogenous compounds is tuned by intrinsic genetic variability. Table I shows the very high genetic variability in all TLR. Considering the high number of single nucleotide polymorphisms (SNPs) in each of the members of the toll like family, the evaluation of TLRs function, in human disease, is very complex [8].

TABLE I
NUMBER OF VARIATION FOR TOLL-LIKE RECEPTORS

Toll-like receptor	Number of SNPs registered in SNPs database
TLR1	335
TLR2	493
TLR3	389
TLR4	616
TLR5	763
TLR6	855
TLR7	550
TLR8	276
TLR9	391
TLR10	366

The structural information available about TLRs mostly relates to the TIR domain, which is inside the cell. This limits the capability to deeply investigate the interaction between immunogenic compounds with an elective domain on the receptor. The extracellular region of TLR contains leucine-rich repeat (LRR) motifs, which are known to have an important role in the activation of the innate immunity. This activation could be regarded as the result of a combinatory set of complex ligand-protein interaction that induces the response activation. This response depends on the type of ligand that TLRs recognize, which allows for a fine-tuned immune response. The investigation of fine-tuning of ligand-TLRs binding has not so far been completed. Taking into account genetic variability, it is important to elucidate which is the protein region influenced by the related mutation. Considering the complexity of these molecules it seem worthwhile to single out potential hot spots of mutagenesis in order to focus on specific smaller domains for the analysis of ligand receptor interactions. These regions can be 3D modelled and virtually screened, similarly to what is often done in the process of drug discovery [9-12].

DSP methods have been successfully used to analyse DNA or protein sequences [13, 14] and they seem to be a quite simple and robust approach to pre-process biosequences. We applied DSP methods to identified the presence of conserve regions and, moreover, those domains that are likely to be relevant in a mutational event. Our analysis has been carried out on cDNA of Toll-like receptors. In the past [15] the analysis of protein structural

Manuscript received July, 25 2013. Patrizio Arrigo is with the CNR Institute of Macromolecular Studies (ISMAC), Section of Genoa, Via De Marini 6, 16149 Genoa, Italy (E-mail: arrigo@ge.ismac.cnr.it). All other authors are with the Department of Informatics, Bioengineering, Robotics and Systems engineering, Nanobiotechnology and Medical Informatics Laboratory University of Genoa, Via all'Opera Pia, 13, 16145 Genoa, Italy (*corresponding author, phone: +390103532991; fax: +390103532154; E-mail: carmel@dibris.unige.it).

978-1-4799-3163-7/13/$31.00 ©2013 IEEE

motif through cDNA screening has be performed by using ANN. The DSP seems to ensure a greater reproducibility of results. Here we apply DSP to face the same problem of mapping protein characteristics on the corresponding cDNA. The proposed system has been tested on Toll-like receptor1. In the past we have performed a docking analysis of some chemicals against the whole TLR3 [16]. This previous screening was focused on the endocellular effect. Now we would like to deeply investigate the interaction of chemical with specific domains on a surface TLR. because these molecule is on the cell surface and it preferentially interact with bacterial components except nucleic acids. TLR1 is a critical element in the inflammatory process associated with bacterial infection.

I. METHODS

In order to use DSP tools to identify possible regions of interest in a DNA or RNA sequence, it is necessary to map the sequence from a symbolic form in term of a character string to a numerical form. In order to achieve this a numeral is assigned to each nucleotide in the sequence. Different techniques have been proposed, each of which aims to enhance information, which is hidden in the sequence for further analysis.

In many cases nucleotides have been represented by lexicographic order (A=1, C=2, G=3, T/U=4) or by different binary codes that can represent not only the nucleotide but also its properties such as number of H-bonds or weak (A, T/U) or strong (G, C). A very promising method to code nucleotide is based on electro-ion interaction potential (EIIP) [17], which is defined as the average energy of the localized electrons of the nucleotide. Resonant Recognition Model (RRM) [18, 19] is a physical and mathematical model which interprets biological sequence linear information using signal analysis methods in order to treat the primary sequence as a discrete signal. In this case, we have applied RRM to a RNA sequence. From this model, it is possible to generate the EIIP values that are assigned to the nucleotides, and that originates a numerical sequence, which represents the distributions of the free electrons' energies within in the nucleotide sequence. This method have been successfully used for several problems, such as the identification of proteins' hot spots, the identification of coding regions and peptide design.

EIIP it is biologically more meaningful than a lexicographic method as it represents a physical property. In that methods the corresponding values only represent the presence or absence of a nucleotide.

TABLE II
ELECTRO ION INTERACTION PSEUDO POTENTIALS OF NUCLEOTIDES

Nucleotide	EIIP values
A	0.1260
G	0.0806
C	0.1340
T	0.1335

EIIP is an estimate of average energy states of all valence electrons in a particular nucleotide. Table 1 shows the values of EIIP for each nucleotide. The nucleotide sequence can be converted into a numerical sequence by replacing each nucleotide with the corresponding EIIP value. For example, if $x[n]$= TGTGCTTGCA, then using the values from Table 1, the corresponding EIIP numerical sequence will be $x[n]$=[0.1335 0.0806 0.1335 0.0806 0.1340 0.1335 0.1335 0.0806 0.1340 0.1260]. The next steps use digital signal analysis methods applied to the obtained numerical series in order to investigate the potential.

In our study we used a frequency domain method to point out possible regions in sequence. The STFT introduce time-localization by dividing a signal into a number of short overlapping sections using a sliding window. Fourier transform is not apply to the whole signal, but only to the section selected by the sliding window.

Thus, we have a series of frequency spectra with each spectrum corresponding to a short interval of time. The STFT of the corresponding EIIP numerical sequence and is given by is defined by

$$U[k] = \sum_{m=0}^{L-1} x[m + rR]w[m]e^{\frac{-j\pi km}{N}} \quad (1)$$

where L is the window length, N is the DFT length, R is the shift interval, and r and k are integers such that

$$-\infty < r < \infty \quad \text{and} \quad 0 \le k \le N-1$$

Then the spectral content at k^{th} instant is:

$$S\ [k]\ =\ \left|U\ [k]\right|^2 \quad (2)$$

The time and frequency resolution depends by the length of the window. An increase in the window length enhances the frequency resolution at the expense of the time resolution while a decrease in the window length improves the time resolution at the expense of the frequency resolution.

Generally DSP methods have been applied to DNA sequence for the identification of regions that code the protein. It has been observed that there is a prominent power spectrum peak at frequency $f = 1/3$ in the coding regions of a DNA sequence. This characteristic is related to the nucleotide triplet arrangement (or codon) that encodes protein amino acids.

The regions can be identified by evaluating S[L/3] over a window of L samples, then sliding the window by one samples and recalculating S[L/3]. This process is carried out over the entire cDNA sequence. The peaks in the spectra obtained by the sliding window could correspond to the active regions of the sequence. The window has been splitted into segmentation element dividing the sampling window by 3.

TABLE III
REGION OF VARIATION OF ALLELES RESPECT TO TLR1 WILD TYPE

Alleles	Accession number	Variation regions								
1	DQ012254		499-544	637-682		781-826				1905-1950
2	DQ012255		499-544	637-682		781-826	1648-1693			1905-1950
3	DQ012256		499-544	637-682		781-826	1579-1624			1905-1950
4	DQ012257		499-544	637-682		781-826	1538-1583			1905-1950
5	DQ012258		499-544	637-682		781-826	1281-1326			1905-1950
6	DQ012259		499-544	637-682	697-742	781-826		1472-1517		1905-1950
7	DQ012260		499-544	637-682	697-742	781-826		1472-1517	1759-1804	1905-1950
8	DQ012261	193-238	499-544	637-682	697-742	781-826		1472-1517	1759-1804	1905-1950
9	DQ012262	306-351	499-544	637-682		781-826				1905-1950
10	DQ012263	68-113	499-544	637-682		781-826	868-913			1905-1950
11	DQ012264	125-170	499-544	637-682		781-826				1905-1950

Many types of windowing have been set up in the past. It has been shown that for coding region identification the window length affects the results and its optimal value depends on the length and the number of exons in sequence [20-22].
At the moment we have no information about the optimal type and the length of the window. Hence, in order to tune the parameters for the DSP analysis we have selected the window amplitude on the basis of medium length of a functional LRR. Their length is about 24 bp. To analyse a wide region of TLR1, we have decided to double this region, using a Blackman window of length L=48 that is shown in figure 1.

Specifically, we have applied a Short Time Fourier Transform (STFT) based method to the cDNA sequence of human TLR1 (wild type) and its registered 11 alleles to evidence possible different function related to the variation in the sequences. Sequences have been retrieved from NCBI Nucleotide Database (http://www.ncbi.nlm.nih.gov/nuccore). For each sequence an energy profile has been obtained. In figure 2 the resulting wild type profile is shown.

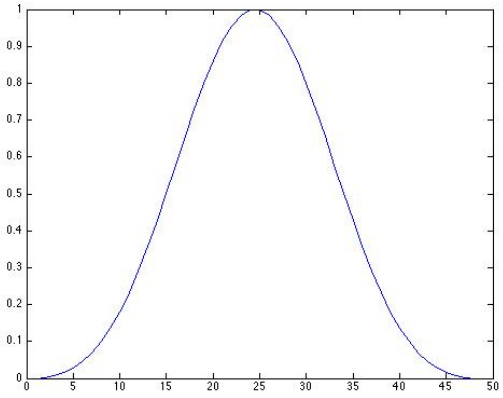

Fig. 1 Blackman window

Each allele energy profile has been subtracted with wild type one. This procedure has highlighted the functional conserved regions and the varied regions. In table III are reported the regions where the variations have been detect.

II. DISCUSS AND CONCLUSION

In a previous paper we have evaluated the binding efficiency of some polymers, used in drug delivery, against the whole ectodomain of TLR3 receptor [16]. Those results have been affected by the lack of reliable complete structural information about TLR3. Taking those results into account, we have decided to perform refined analysis only on those protein domains filtered out by DSP. The TLRs have shown very interesting characteristics that could be valuable to design new synthetic biology system for advanced diagnostic [23]. Currently, the majority of tools are based on antigen-antibody reaction. This approach has several limiting factors such as cross-reaction with other proteins or lack of binding efficiency dues to genetic variation. Some authors have demonstrated the effect of mutation in some domains of a Toll-like receptor [24, 25], but a systematic analysis is currently not yet performed.
In order to restrain more than it is possible, the influence of these unfavourable factors, we have tested our DSP system on the screening of TLR1 domains that can preserving the

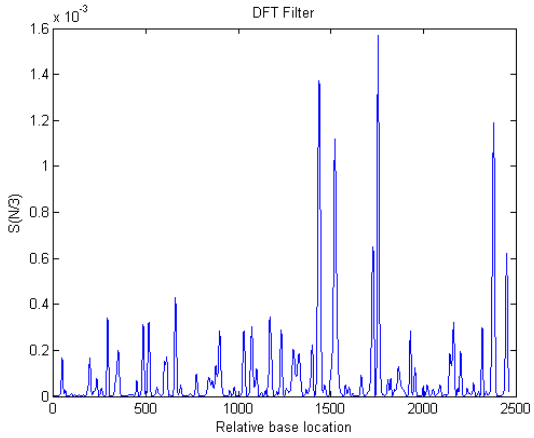

Fig. 2 Energy profile of human TLR1 c DNA

maximal recognition capability (MRC) of the molecules. The MRC can depend by mutation on specific domains such as LRR but also by the modification in the hinge regions that are essential to preserve topological integrity of the ectodomain. The detection of these critical domains allows to estimate the MRC by virtual screening and therefore to plan the experimental validation. This approach allows also to estimate the effects of single nucleotide polymorphisms on the MRC. It is well known that a single mutation can dramatically reduce the capability to bind exogenous (chemical and biological) immunological triggers. We guess that both antibody or biosensor design optimization must consider the effects of mutational events into account. The application of DSP to map the protein conserved domains onto the cDNA is a simplest way to bridge together the genomic and proteomic aspect of protein engineering.

REFERENCES

[1] G. Lee, N. Harnett, L. Zychla, and R. E. Dinniwell, "Radiotherapy Treatment Review: A Prospective Evaluation of Concordance between Clinical Specialist Radiation Therapist and Radiation Oncologist in Patient Assessments," *Journal of Medical Imaging and Radiation Sciences*, vol. 43, pp. 6-10, 2012.

[2] B. A. Chabner and T. G. Roberts, Jr., "Timeline: Chemotherapy and the war on cancer," *Nat Rev Cancer*, vol. 5, pp. 65-72, Jan 2005.

[3] B. Al-Lazikani, U. Banerji, and P. Workman, "Combinatorial drug therapy for cancer in the post-genomic era," *Nat Biotechnol*, vol. 30, pp. 679-92, Jul 2012.

[4] S. Gaglio, C. Ruggiero, G. Spinelli, G. Bonadonna, P. Valagussa, and C. Nicolini, "BREASTCAN: an expert system for postoperative breast cancer therapy," *Computers and Biomedical Research*, vol. 19, pp. 445-461, 1986.

[5] J. M. Silva, M. Videira, R. Gaspar, V. Preat, and H. F. Florindo, "Immune system targeting by biodegradable nanoparticles for cancer vaccines," *J Control Release*, vol. 168, pp. 179-99, Jun 10 2013.

[6] R. Spirig, J. Tsui, and S. Shaw, "The Emerging Role of TLR and Innate Immunity in Cardiovascular Disease," *Cardiology Research and Practice*, vol. 2012, pp. 1-12, 2012.

[7] M. L. Alegre, J. Leemans, A. Le Moine, S. Florquin, V. De Wilde, A. Chong, *et al.*, "The multiple facets of toll-like receptors in transplantation biology," *Transplantation*, vol. 86, pp. 1-9, Jul 15 2008.

[8] D. N. Cook, D. S. Pisetsky, and D. A. Schwartz, "Toll-like receptors in the pathogenesis of human disease," *Nat Immunol*, vol. 5, pp. 975-9, Oct 2004.

[9] N. Maggi, P. Arrigo, and C. Ruggiero, "Comparative analysis of Rac1 binding efficiency with different classes of ligands: morpholines, flavonoids and imidazoles," *IEEE Trans Nanobioscience*, vol. 11, pp. 181-7, Jun 2012.

[10] A. Lavecchia and C. Di Giovanni, "Virtual Screening Strategies in Drug Discovery: A Critical Review," *Curr Med Chem*, May 2 2013.

[11] M. Lill, "Virtual screening in drug design," *Methods Mol Biol*, vol. 993, pp. 1-12, 2013.

[12] Y. Tanrikulu, B. Kruger, and E. Proschak, "The holistic integration of virtual screening in drug discovery," *Drug Discov Today*, vol. 18, pp. 358-64, Apr 2013.

[13] S. Barman, S. Saha, A. Mandal, and M. Roy, "Prediction of protein coding regions of a DNA sequence through spectral analysis," in *Informatics, Electronics & Vision (ICIEV), 2012 International Conference on*, 2012, pp. 12-16.

[14] R. Kakumani, O. Ahmad, and V. Devabhaktuni, "Identification of CpG islands in DNA sequences using statistically optimal null filters," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2012, pp. 1-14, 2012/08/29 2012.

[15] P. Arrigo, P. Fariselli, and R. Casadio, "Can functional regions of proteins be predicted from their coding sequences? The case study of G-protein coupled receptors," *Gene*, vol. 221, pp. GC65-110, Oct 9 1998.

[16] N. Maggi, P. Arrigo, and C. Ruggiero, "In silicon evaluation of nanoparticle cell interaction via human TLR3," in *Nanotechnology (IEEE-NANO), 2010 10th IEEE Conference on*, 2010, pp. 995-998.

[17] A. S. Nair and S. P. Sreenadhan, "A coding measure scheme employing electron-ion interaction pseudopotential (EIIP)," *Bioinformation*, vol. 1, pp. 197-202, 2006.

[18] E. Pirogova, G. P. Simon, and I. Cosic, "Investigation of the applicability of dielectric relaxation properties of amino acid solutions within the resonant recognition model," *IEEE Trans Nanobioscience*, vol. 2, pp. 63-9, Jun 2003.

[19] I. Cosic, "Macromolecular bioactivity: is it resonant interaction between macromolecules?--Theory and applications," *IEEE Trans Biomed Eng*, vol. 41, pp. 1101-14, Dec 1994.

[20] M. Akhtar, J. Epps, and E. Ambikairajah, "Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 2, pp. 310-321, 2008.

[21] M. Akhtar, E. Ambikairajah, and J. Epps, "Digital Signal Processing Techniques for Gene Finding in Eukaryotes," in *Image and Signal Processing*. vol. 5099, A. Elmoataz, O. Lezoray, F. Nouboud, and D. Mammass, Eds., ed: Springer Berlin Heidelberg, 2008, pp. 144-152.

[22] M. Akhtar, E. Ambikairajah, and J. Epps, "Optimizing period-3 methods for eukaryotic gene prediction," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 621-624.

[23] I. H. Cho, J. W. Jeon, S. H. Paek, D. H. Kim, H. S. Shin, U. H. Ha, *et al.*, "Toll-like receptor-based immuno-analysis of pathogenic microorganisms," *Anal Chem*, vol. 84, pp. 9713-20, Nov 20 2012.

[24] B. Beutler, Z. Jiang, P. Georgel, K. Crozat, B. Croker, S. Rutschmann, *et al.*, "Genetic analysis of host resistance: Toll-like receptor signaling and immunity at large," *Annu Rev Immunol*, vol. 24, pp. 353-89, 2006.

[25] A. E. Medvedev, "Toll-Like Receptor Polymorphisms, Inflammatory and Infectious Diseases, Allergies, and Cancer," *J Interferon Cytokine Res*, May 15 2013.