
**International Workshop on Computational & Integrative Biology**

A satellite meeting of the International Conference of Integrative Biology

**September 18th to 20th, 2009**

**Hangzhou, China**

**http://www.cls.zju.edu.cn/binfo/ib/2009**


# CIB'2009

## Proceedings

**Hanzhou, China**

**Sep 9th, 2009**

# Contents

# Preface

The International Workshop on Computational and Integrative Biology, 2009 (CIB'09) is a satellite meeting of the series Integrative Bioinformatics International Symposium (IB). This meeting is of interest to bioinformaticians, computer scientists, biologists and others working in, or interested in finding out more about, the developing area of integrative biology.

This volume of CIB'09 proceedings contains abstracts or full papers of the talks and demos/posters presented at the event. Full papers will probably appear in the Journal of Integrative Bioinformatics with further recommendation and reviews.

IB series meetings are chaired by Prof. Dr. Ralf Hofestädt, University of Bielefeld, Germany. This CIB is co-organized by the Prof. Dr. Ming Chen, Zhejiang University, China. CIB'09 is sponsored by the Federal Ministry of Education and Research, Germany, The Ministry of Science and Technology, China, and IMBio – Informationsmanagement in der Biotechnologie e.V.

Hangzhou, September 2009

Ming Chen

Ralf Hofestädt

# Models of SIGIRR inhibiting the Toll-like receptor TLR4 signaling pathway

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

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

[2] Leibniz Supercomputing Centre. 85748 Garching, Germany

[3] Deutsches Museum. 80538 Munich, Germany

*Corresponding author. Email: tiandi.wei@informatik.uni-muenchen.de

## Abstract

Toll-like receptors (TLRs) belong to the Toll-like receptor (TLR)/interleukin-1 receptor (IL-1R) superfamily, which is defined by a common cytoplasmic Toll/interleukin-1 receptor (TIR) domain. These receptors recognize pathogen-associated molecular patterns and initiate an intracellular kinase cascade to cause 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 the MyD88-dependent TLR signaling. It attenuates the recruitment of MyD88 adaptors to the receptors with its intracellular TIR domain. Thus, SIGIRR reveals potential significance in the therapy of autoimmune diseases. However, the mechanism how SIGIRR structurally interacts with TLRs and adaptor molecules remains unclear. Here, we developed three-dimensional structures for the TIR domains of TLR4, MyD88 and SIGIRR based on computational modeling. Through protein-protein docking analysis, we suggest models of essential complexes involved in the TLR4 signaling and the SIGIRR inhibiting processes. SIGIRR may exert its inhibitory effect through blocking the molecular interface of TLR4-TLR4 and MyD88-MyD88 dimers mainly via its BB-loop region.

## 1. Introduction

The Toll-like receptor (TLR)/interleukin-1 receptor (IL-1R) superfamily plays an important role in differentially recognizing pathogen products and mediating immune responses. All members of this superfamily possess a conserved cytoplasmic Toll/interleukin-1 receptor (TIR) domain [1], which is connected to an ectodomain through a single transmembrane stretch. The TLR/IL-1R superfamily can be divided into two main groups based on ectodomains: immunoglobulin (Ig) domain-containing receptors and Toll-like receptors (TLRs) [2].

To date, thirteen TLRs have been identified in mammals. Their ectodomains consist of 16 to 28

leucine-rich repeats (LRRs). These LRRs provide a variety of structural frameworks for binding of protein and non-protein ligands including lipopolysaccharide (LPS), lipopeptide, cytosine-phosphate-guanine (CpG) DNA, flagellin, imidazoquinoline and double/single stranded RNA [3]. TLRs are capable of recognizing ligands in a dimer form [4-6]. Upon receptor activation, an intracellular TIR signaling complex is formed between the receptor and downstream adaptor TIR domains [7]. MyD88 (Myeloid differentiation primary response protein 88) is the first characterized intracellular adaptor molecule among all known adaptors in the TLR signaling. It consists of an N-terminal death domain (DD) separated from its C-terminal TIR domain by a linker sequence. MyD88 also forms dimer through DD-DD and TIR-TIR domain interactions when recruited to the receptor complex [8]. Further, 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) that leads 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 signals (MyD88-independent) while the TLR4 can use both pathways to signal.

SIGIRR (single immunoglobulin interleukin-1 receptor-related molecule), also known as TIR8 (Toll/IL-1R 8), was initially identified as an Ig domain-containing receptor of the TLR/IL-1R superfamily in 1998 by Thomassen *et al* [9]. Both the extracellular and intracellular domains of SIGIRR differ from those of other Ig domain-containing receptors. Its single extracellular Ig domain does not support ligand binding. Its intracellular TIR domain cannot activate NF-κB because it lacks two critical 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) [9]. SIGIRR rather acts as an endogenous inhibitor for MyD88-dependent TLR and IL-1R signaling because over expression of SIGIRR in Jurkat or HepG2 cells substantially reduced LPS, CpG DNA or IL-1-induced activation of NF-κB [10-12]. In this regard, SIGIRR may prevent some autoimmune diseases such as systemic lupus erythematosus caused by TLR-mediated induction of type I interferons [13]. Previous mutagenesis investigated three deletion mutants of SIGIRR [12]: 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 [12]. However, detailed structural interaction behaviors of SIGIRR are unknown.

The structures of TIR domains from human TLR1, 2, 10 and IL-1RAPL have been solved so far by X-ray crystallography [14-16]. Thereof, the TLR1 and 2 modules behave as monomers in solution and the packing of the molecules in the crystal lattice did 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. In light of these, we have built three-dimensional structures for TIR domains of TLR4, MyD88 and SIGIRR by homology modeling and protein threading to elucidate the mechanism of SIGIRR inhibiting the MyD88-dependent TLR4 signals. Models of essential molecular complexes involved in the TLR4 signaling and the SIGIRR inhibiting processes are proposed based on results of protein-protein docking studies.

## 2. Methods

### 2.1 Templates identification and sequence alignments

Amino acid sequences of the target proteins, human TLR4 (GenBank Accession No. O00206), MyD88 (AAC50954) and SIGIRR (CAG33619) were extracted from the NCBI protein database [17]. TIR domain structures of TLR4, MyD88, and SIGIRR (Tyr165-Pro308, without the C-tail) were constructued by homology modeling. Due to the high homology of the target proteins, four common templates were obtained via BLAST search against the Protein Data Bank (PDB) [18]. They were TLR1 (PDB code: 1FYV), TLR2 (1FYW), TLR10 (2J67), and IL-1RAPL (1T3G). Multiple sequence alignment of each target with the templates were generated using MUSCLE [19] and analyzed using Jalview [20]. Since the secondary structure of the TIR domain is composed of

| | TLR1 | TLR2 | TLR10 | IL-1RAPL | Avg |
|---|---|---|---|---|---|
| TLR4 | 53.4 | 57.8 | 51.4 | 44.2 | 51.7 |
| MyD88 | 44.5 | 45.3 | 40.6 | 47.4 | 44.5 |
| SIGIRR | 41.8 | 42.3 | 37.7 | 49.0 | 42.7 |

**Table 1: Sequence similarities (%) between targets and templates.**

well organized alternative β-sheet and α-helix, we adjusted the alignments manually according to the secondary structure information to improve the alignment quality. The secondary structures of each target were predicted by PSIPRED [21]. In addition, the C-terminal tail of the TIR domain, which is unique to SIGIRR, has no structure-known homologue to serve as template. In this case we employed the protein threading method THREADER 3.5 [22] to determine an acceptable template structure. The selected template was N-terminal domain of N-ethylmaleimide sensitive factor (NSF-N) (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 [23]. 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 the modeling, gap regions in the alignment constituted loop structures in the model, which impeded the model accuracy. ModLoop [24] was used to modify such loop regions. Finally, we used the model quality assessment programs, ProQ [25], ModFOLD [26] and MetaMQAP [27] to evaluate the output candidate models and select the most reliable one.

## 2.3 Model docking

Pairwise model docking included five complexes of TIR domains: TLR4-TLR4, MyD88-MyD88, TLR4 dimer-MyD88 dimer (tetramer), TLR4-SIGIRR and MyD88-SIGIRR. Protein-protein docking programs GRAMM-X [28] and ZDOCK [29] were used to predict the interactions between these complexes. Both programs can return 10 most probable predictions which are selected from thousands of candidates based on geometry, hydrophobicity and electrostatic complementarity of the molecule surface. We subsequently chose the most reasonable solution from these outputs by considering further qualifications. These qualifications include residue conservation of the interaction sites, and knowledge from published articles [6, 15, 30, 31].

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

**Table 2: Model evaluation. All these displayed scores indicate that the models are reliable in terms of overall packing. 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 favoured 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.**

# 3. Results

## 3.1 Molecular modeling of TLR4, MyD88 and SIGIRR TIR domains

In the secondary structure aided alignments for the modeling, the average target-template sequence similarity of TLR4, MyD88 and SIGIRR is 51.7%, 44.5% and 42.7%, respectively (detailed in Table 1). The resulting structures exhibit a typical TIR domain conformation where five central parallel β-sheets (βA-βE) are surrounded by a total of five α-helixes (αA-αE) on both sides (Figure 1A). Besides, the structure of NSF-N was identified as a template for SIGIRR's C-tail through protein threading. The C-tail contains four parallel β-sheets with an α-helix and

some loop structures on the one side, whereas the other side points to SIGIRR's TIR (Figure 1A).

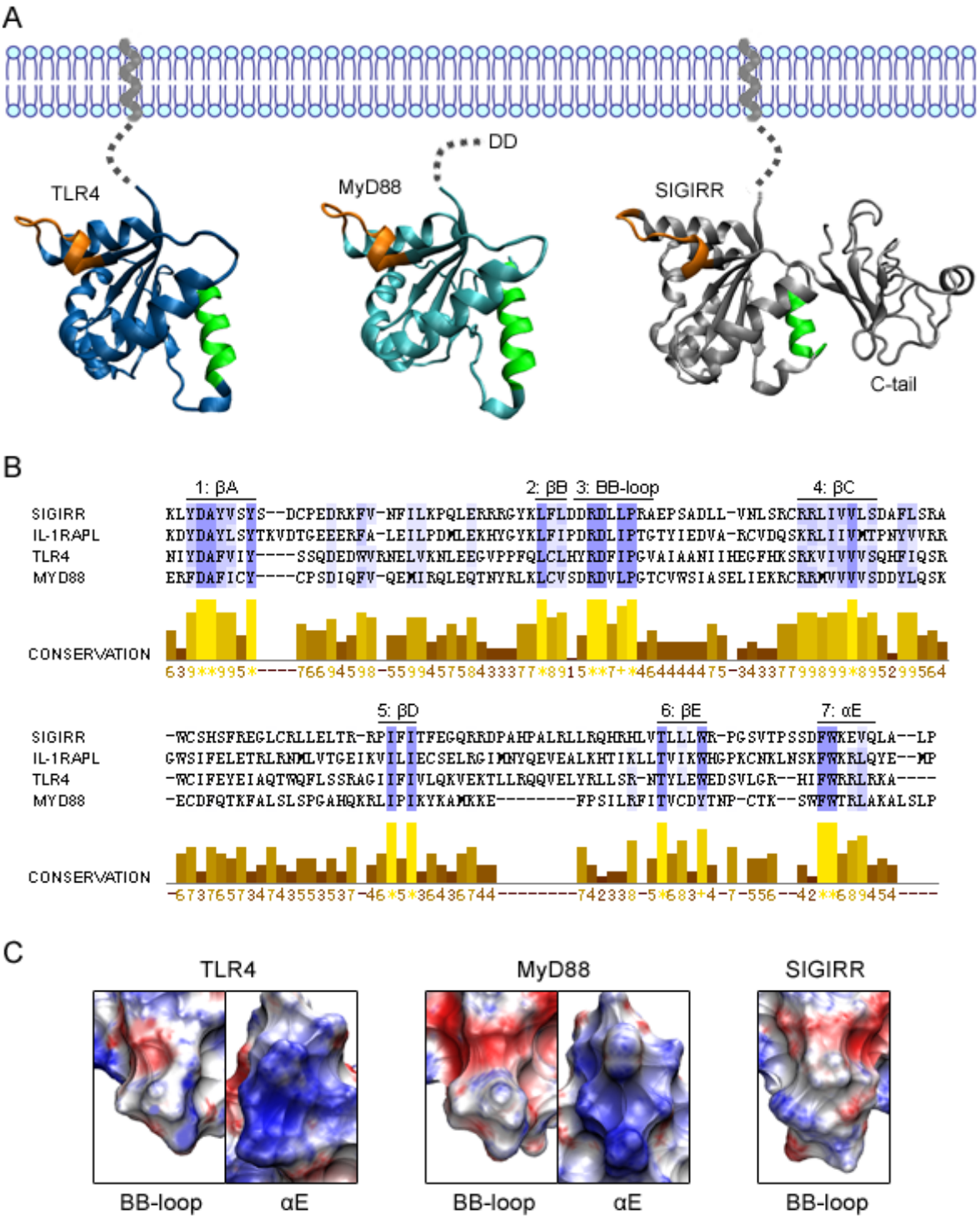



**Figure 1: Three-dimensional structures and conserved regions of TIR domains of TLR4, MyD88 and SIGIRR. (A) The BB-loop and αE regions are highlighted in orange and green respectively. (B) Multiple sequence alignment of different TIRs indicates seven conserved boxes. (C) Surface charge distribution (APBS electrostatics) of BB-loop and αE with red indicating areas of negative charge and blue indicating positive charge.**

These results suggest that the TIR domain and the C-tail of SIGIRR are not an integrative structure, but two interconnected individual modules. Evaluation of the models involved analysis of geometry, stereochemistry and energy distributions of the molecules. The evaluation results (Table 2) are indicative of a good quality of all models.

Multiple sequence alignment of TIR domains from different molecules detected seven conserved boxes in the TIR domain (Figure 1B). Our models show that they correspond to β-sheet A (βA), β-sheet B (βB), BB-loop, β-sheet C (βC), β-sheet D (βD), β-sheet E (βE) and α-helix E (αE), respectively. Functional significance can be usually observed in conserved regions. Nevertheless, the five β-sheets (box 1, 2, 4-6) are embedded structures that form a hydrophobic core of the TIR domain and hence unable 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 and MyD88, along with the BB-loop of SIGIRR may be important to ensure binding specificity achieved by different combinations of TIRs during signaling (Figure 1A). Figure 1C illustrates the electrostatic surface potential of these BB-loops and αEs. Accordingly, all BB-loops can be divided into two parts. The N-terminal (upper region of BB-loops in Figure 1C) is negatively charged, whereas the C-terminal (lower region of BB-loops in Figure 1C) is positively charged. The αEs, by contrast, are predominantly positive.

## 3.2 Pairwise docking of TLR4, MyD88 and SIGIRR TIR domains

As noted above, TLR4, MyD88 and SIGIRR are able to interact heterotypically with each other. To elucidate how SIGIRR disturbs the MyD88-dependent TLR4 signals, it is indispensable to understand the interaction mode of the signaling complex of TLR4 and MyD88 without the presence of SIGIRR. As a result, we performed protein docking analysis for the five TIR complexes as follows. An optimal docking solution was chosen for each complex from large numbers of candidates (detailed in Methods). Molecular surface charge analysis indicates that all the selected models exhibit good electrostatic complementarity (data not shown).

### 3.2.1 TLR4-TLR4

The signaling mechanism of all TLRs is likely to involve receptor dimerization. This can be achieved in various ways by different receptors [32]. TLR4's TIR domain reveals an axially symmetric dimer with the BB-loop (involved redsidues: Pro714-Ala717) of one monomer protruding into the groove formed by the αC (Cys747-Ile748) and DD-loop (Gln782) of the other (Figure 2A). Simultaneously, the αB (Ala719) of each monomer interacts tightly with each other in the middle of both BB-loop connections.

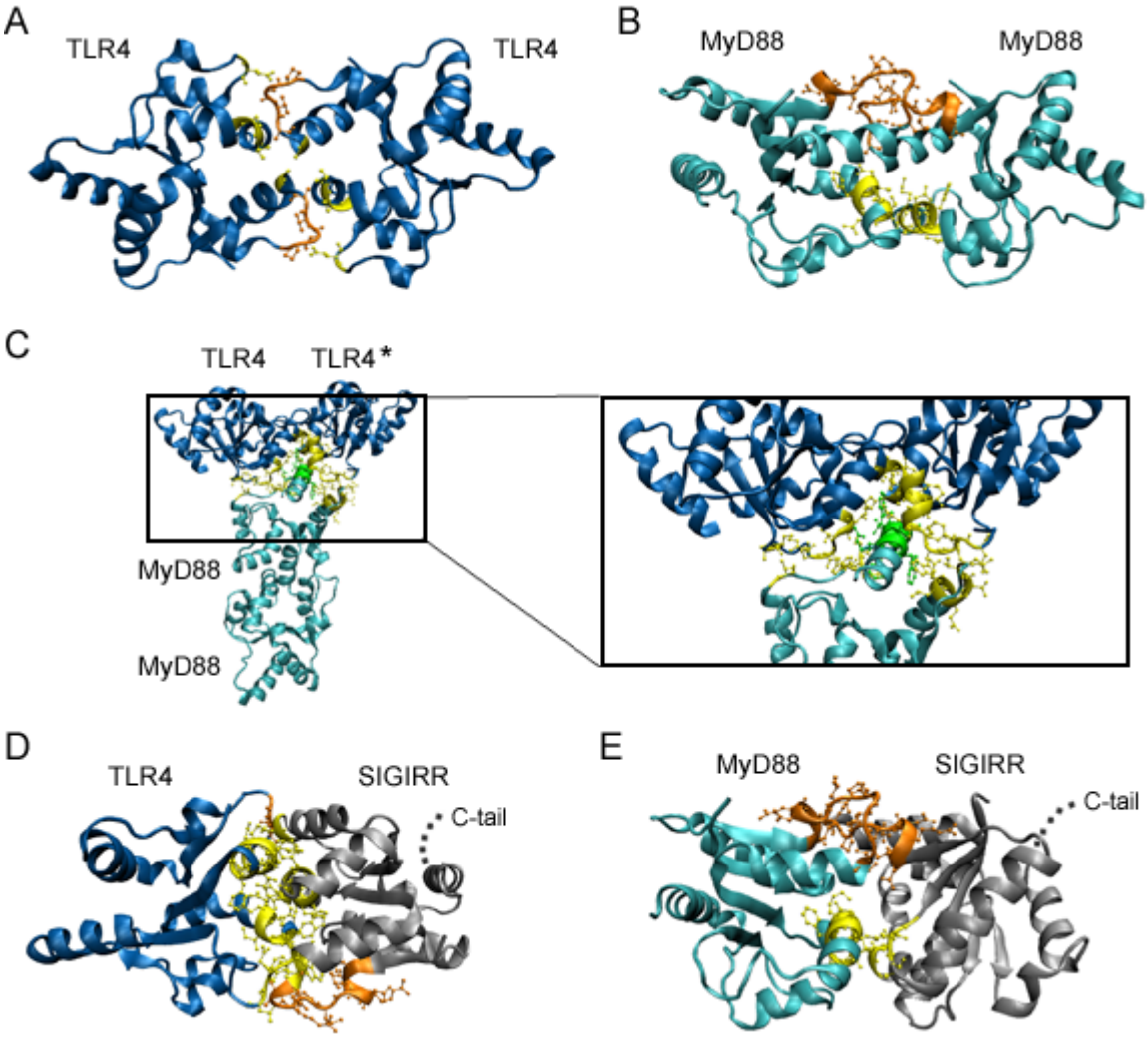



**Figure 2: Essential complexes involved in the TLR4 signaling and the SIGIRR inhibiting processes by protein docking. Interacting regions of BB-loop and αE are labeled in orange and green respectively. Other interacting regions are labeled in yellow. All interacting residues (orange/green/yellow) are extra represented using CPK (Corey, Pauling & Kultun) convention. (A) TLR4-TLR4 dimer. (B) MyD88-MyD88 dimer. (C) TLR4 dimer-MyD88 dimer tetramer. (D) TLR4-SIGIRR dimer. (E) MyD88-SIGIRR dimer.**

### 3.2.2 MyD88-MyD88

MyD88 forms dimer when incorporated into a receptor complex [8]. The BB-loops (Asp195-Cys203) from both monomers were docked together by an antiparallel packing (Figure 2B). Under the BB-loop connection both αCs (Cys233-Lys238) are brought into contact. This model is also axially symmetirc. Our finding is consistent with Loiarro *et al*'s conclusion that a heptapeptide, which mimics the BB-loop of MyD88's TIR domain, strongly interferes with dimerization of MyD88 [30].

### 3.2.3 TLR4 dimer-MyD88 dimer

Both dimers described above were assembled into a tetramer (Figure 2C). The TLR4 dimer provides a binding pocket adjacent to its interface. This pocket is constituted by the αC (Gln755) of a TLR4 monomer as well as the αB (Ala719-His724) and αC (Tyr751-Thr756) of the other monomer (TLR4*). The αE (Cys280-Arg288) of a MyD88 monomer just fills the pocket and makes interactions. This connection is further stabilized by three surrounding joints: 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).

### 3.2.4 TLR4-SIGIRR

As an inhibitor of the TLR signaling, SIGIRR heterodimerizes with TLR4 [12]. Our docked model exhibits an extensive interface composed of three patches, which indicates a strong molecular affinity (Figure 2D). 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). Last, SIGIRR's αD (Pro268-Ala269) interacts with TLR4's BB-loop (Val716-Ala717). 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.

### 3.2.5 MyD88-SIGIRR

SIGIRR interferes with the functional dimer conformation of MyD88 by heterodimerization with MyD88 [12]. Our docked model shows that the molecular interface between MyD88 and SIGIRR is quite large (Figure 2E). SIGIRR's BB-loop (Asp201-Ala208) complements MyD88's BB-loop (Asp195-Val204), by substituting the other BB-loop in the customary MyD88 homodimer (Section 3.2.2). Furthermore, SIGIRR's AA-loop (Ser172-Cys174) and αC (Arg235-Ala236) interacts with MyD88's αC (Gln229-Thr237) under the BB-loops. Likewise, SIGIRR's C-tail does not seem to play any role in this dimer.

## 4. Discussion

So far, the only crystallized dimer structure of TLR's TIR domain is the TLR10 dimer [15], in which the BB-loop and αC of each monomer constitute the major part of the symmetric dimer interface. Miguel *et al* conceived of TLR4's dimerization manner as identical to TLR10's (no evidence given) [31]. However, we do not consider them to be necessarily identical, because the TIR domain inherently has diverse dimer conformations [15, 16, 33] and TLR4 has different ligand-binding and signaling mechanism compared with TLR10. Poltorak *et al* reported that a

single-point mutation (Pro712His) of the TIR domain of murine TLR4 abolished the TLR4 response to LPS [34]. Our TLR4 dimer model supports Poltorak *et al*'s results. The corresponding residue Pro714 is located at the very tip of the BB-loop and interacts tightly with Gln782 of the other monomer.

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. Previously, Dunne *et al* modeled the TLR4-MyD88 heterodimer using TLR4 and MyD88 monomers [35]. This may, however, lead to disregard of important contributions made by other participators during the molecular interactions. The docking of TLR4-dimer and MyD88-dimer allows significant advances made in the structural interpretation over the previous work. The tetramer revealed in our study demonstrates that the stimulus induced dimerization of TIR domains creates a new molecular surface for the recruitment of signaling adaptor proteins.

All results from pairwise docking studies here can be assembled to derive a working hypothesis for the TLR4 signaling transduction and SIGIRR inhibition mode (Figure 3). Receptor activation would trigger the formation of TLR4 TIR dimers recruiting MyD88 TIR dimers and forming a signaling tetramer (middle complex in Figure 3). Model predictions including SIGIRR reveal that SIGIRR binds to TLR4 by occupying TLR4's interacting sites, which should interrupt TLR4

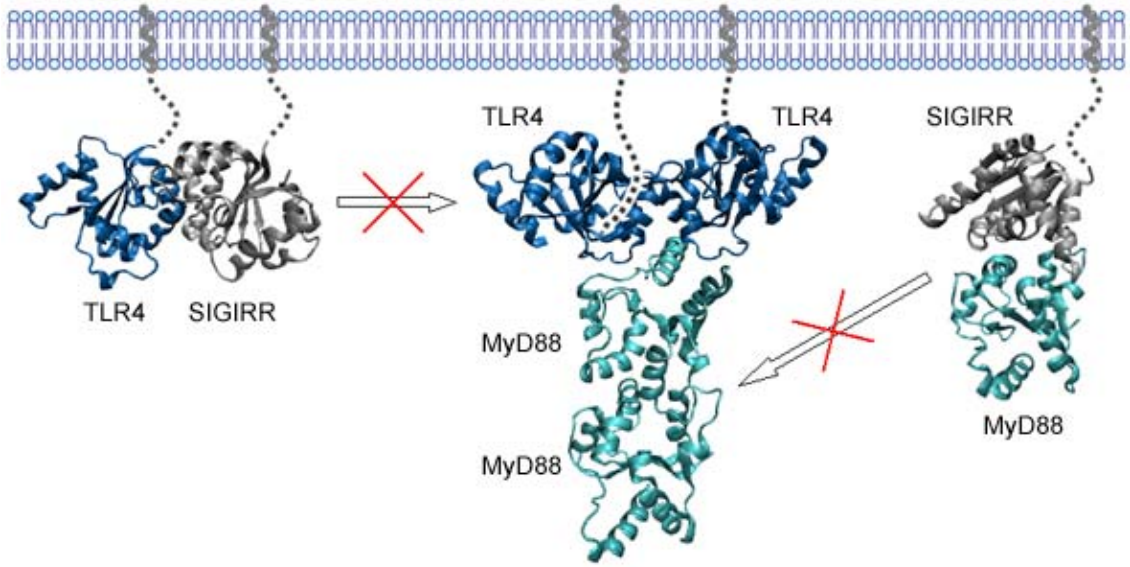



**Figure 3: Model of SIGIRR inhibiting the TLR4 signaling.**

homodimer formation (left complex in Figure 3). On the other hand, the MyD88-SIGIRR dimer shows a resemblance to the MyD88 homodimer. That is, SIGIRR replaces a MyD88 monomer, interrupting MyD88 homodimer formation (right complex Figure 3). In both cases the BB-loop of SIGIRR plays a key role in binding. Remarkably, TLR4 and MyD88 possess a more extensive

molecular interface with SIGIRR (heterodimer) than with themselves (homodimer). This observation highlights the strong molecular affinity of SIGIRR as an inhibitor. In addition, SIGIRR's unique C-tail is located distantly from the active BB-loop according to our model, consistent with the observation that this tail is not required for SIGIRR's inhibitory effect on TLR signaling [12].

During recent years SIGIRR has received a tremendous research interest due to its therapeutic potential in autoimmune diseases. Although the significance of SIGIRR has been widely acknowledged, its inhibition mechanism remains unclear owing to the lack of structural information. This work depicts a residue-detailed structural framework of SIGIRR inhibiting the TLR4 signaling pathway using computational approaches. These results would facilitate efforts to design further site-directed mutagenesis to learn more details about the regulatory role of SIGIRR in inflammatory and innate immune responses.

## 5. Acknowledgements

This work was supported by Graduiertenkolleg 1202 of the Deutsche Forschungsgemein-schaft (DFG) and the DFG excellence cluster Nanosystems Initiative Munich (NIM). We thank Dr. Hans J. Anders and Dr. Lech Maciej for their scientific advice.

## 6. References

[1] A. Bowie, and L. A. O'Neill. The interleukin-1 receptor/Toll-like receptor superfamily: signal generators for pro-inflammatory interleukins and microbial products. *Journal of leukocyte biology,* 67(4):508-14, 2000.

[2] M. U. Martin, and H. Wesche. Summary and comparison of the signaling mechanisms of the Toll/interleukin-1 receptor family. *Biochimica et biophysica acta,* 1592(3):265-80, 2002.

[3] N. J. Gay, and M. Gangloff. Structure and function of Toll receptors and their ligands. *Annual review of biochemistry,* 76:141-65, 2007.

[4] E. Latz, A. Verma, A. Visintin, M. Gong, C. M. Sirois, D. C. Klein, B. G. Monks, C. J. McKnight, M. S. Lamphier, W. P. Duprex, T. Espevik, and D. T. Golenbock. Ligand-induced conformational changes allosterically activate Toll-like receptor 9. *Nature immunology,* 8(7):772-9, 2007.

[5] L. Liu, I. Botos, Y. Wang, J. N. Leonard, J. Shiloach, D. M. Segal, and D. R. Davies. Structural basis of toll-like receptor 3 signaling with double-stranded RNA. *Science,* 320(5874):379-81, 2008.

[6] B. S. Park, D. H. Song, H. M. Kim, B. S. Choi, H. Lee, and J. O. Lee. The structural

basis of lipopolysaccharide recognition by the TLR4-MD-2 complex. *Nature,* 458(7242):1191-5, 2009.

[7] L. A. O'Neill, and A. G. Bowie. The family of five: TIR-domain-containing adaptors in Toll-like receptor signalling. *Nature reviews. Immunology,* 7(5):353-64, 2007.

[8] K. Burns, F. Martinon, C. Esslinger, H. Pahl, P. Schneider, J. L. Bodmer, F. Di Marco, L. French, and J. Tschopp. MyD88, an adapter protein involved in interleukin-1 signaling. *The Journal of biological chemistry,* 273(20):12203-9, 1998.

[9] E. Thomassen, B. R. Renshaw, and J. E. Sims. Identification and characterization of SIGIRR, a molecule representing a novel subtype of the IL-1R superfamily. *Cytokine,* 11(6):389-99, 1999.

[10] N. Polentarutti, G. P. Rol, M. Muzio, D. Bosisio, M. Camnasio, F. Riva, C. Zoja, A. Benigni, S. Tomasoni, A. Vecchi, C. Garlanda, and A. Mantovani. Unique pattern of expression and inhibition of IL-1 signaling by the IL-1 receptor family member TIR8/SIGIRR. *European cytokine network,* 14(4):211-8, 2003.

[11] D. Wald, J. Qin, Z. Zhao, Y. Qian, M. Naramura, L. Tian, J. Towne, J. E. Sims, G. R. Stark, and X. Li. SIGIRR, a negative regulator of Toll-like receptor-interleukin 1 receptor signaling. *Nature immunology,* 4(9):920-7, 2003.

[12] J. Qin, Y. Qian, J. Yao, C. Grace, and X. Li. SIGIRR inhibits interleukin-1 receptor- and toll-like receptor 4-mediated signaling through different mechanisms. *The Journal of biological chemistry,* 280(26):25233-41, 2005.

[13] M. Lech, O. P. Kulkarni, S. Pfeiffer, E. Savarese, A. Krug, C. Garlanda, A. Mantovani, and H. J. Anders. Tir8/Sigirr prevents murine lupus by suppressing the immunostimulatory effects of lupus autoantigens. *The Journal of experimental medicine,* 205(8):1879-88, 2008.

[14] Y. Xu, X. Tao, B. Shen, T. Horng, R. Medzhitov, J. L. Manley, and L. Tong. Structural basis for signal transduction by the Toll/interleukin-1 receptor domains. *Nature,* 408(6808):111-5, 2000.

[15] T. Nyman, P. Stenmark, S. Flodin, I. Johansson, M. Hammarstrom, and P. Nordlund. The crystal structure of the human toll-like receptor 10 cytoplasmic domain reveals a putative signaling dimer. *The Journal of biological chemistry,* 283(18):11861-5, 2008.

[16] J. A. Khan, E. K. Brint, L. A. O'Neill, and L. Tong. Crystal structure of the Toll/interleukin-1 receptor domain of human IL-1RAPL. *The Journal of biological chemistry,* 279(30):31664-70, 2004.

[17] D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. Dicuccio, R. Edgar, S. Federhen, M. Feolo, L. Y. Geer, W. Helmberg, Y.

Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, V. Miller, J. Ostell, K. D. Pruitt, G. D. Schuler, M. Shumway, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner, and E. Yaschenko. Database resources of the National Center for Biotechnology Information. *Nucleic acids research,* 36(Database issue):D13-21, 2008.

[18] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic acids research,* 28(1):235-42, 2000.

[19] R. C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research,* 32(5):1792-7, 2004.

[20] M. Clamp, J. Cuff, S. M. Searle, and G. J. Barton. The Jalview Java alignment editor. *Bioinformatics,* 20(3):426-7, 2004.

[21] K. Bryson, L. J. McGuffin, R. L. Marsden, J. J. Ward, J. S. Sodhi, and D. T. Jones. Protein structure prediction servers at University College London. *Nucleic acids research,* 33(Web Server issue):W36-8, 2005.

[22] D. T. Jones, R. T. Miller, and J. M. Thornton. Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. *Proteins,* 23(3):387-97, 1995.

[23] A. Fiser, R. K. Do, and A. Sali. Modeling of loops in protein structures. *Protein science,* 9(9):1753-73, 2000.

[24] A. Fiser, and A. Sali. ModLoop: automated modeling of loops in protein structures. *Bioinformatics,* 19(18):2500-1, 2003.

[25] B. Wallner, and A. Elofsson. Can correct protein models be identified? *Protein science,* 12(5):1073-86, 2003.

[26] L. J. McGuffin. The ModFOLD server for the quality assessment of protein structural models. *Bioinformatics,* 24(4):586-7, 2008.

[27] M. Pawlowski, M. J. Gajda, R. Matlak, and J. M. Bujnicki. MetaMQAP: a meta-server for the quality assessment of protein models. *BMC Bioinformatics,* 9:403, 2008.

[28] A. Tovchigrechko, and I. A. Vakser. GRAMM-X public web server for protein-protein docking. *Nucleic acids research,* 34(Web Server issue):W310-4, 2006.

[29] R. Chen, L. Li, and Z. Weng. ZDOCK: an initial-stage protein-docking algorithm. *Proteins,* 52(1):80-7, 2003.

[30] M. Loiarro, F. Capolunghi, N. Fanto, G. Gallo, S. Campo, B. Arseni, R. Carsetti, P. Carminati, R. De Santis, V. Ruggiero, and C. Sette. Pivotal Advance: Inhibition of MyD88 dimerization and recruitment of IRAK1 and IRAK4 by a novel peptidomimetic

compound. *Journal of leukocyte biology,* 82(4):801-10, 2007.

[31] R. Nunez Miguel, J. Wong, J. F. Westoll, H. J. Brooks, L. A. O'Neill, N. J. Gay, C. E. Bryant, and T. P. Monie. A dimer of the Toll-like receptor 4 cytoplasmic domain provides a specific scaffold for the recruitment of signalling adaptor proteins. *PLoS ONE,* 2(8):e788, 2007.

[32] R. J. Gibbard, P. J. Morley, and N. J. Gay. Conserved features in the extracellular domain of human toll-like receptor 8 are essential for pH-dependent signaling. *The Journal of biological chemistry,* 281(37):27503-11, 2006.

[33] X. Tao, Y. Xu, Y. Zheng, A. A. Beg, and L. Tong. An extensively associated dimer in the structure of the C713S mutant of the TIR domain of human TLR2. *Biochemical and biophysical research communications,* 299(2):216-21, 2002.

[34] A. Poltorak, X. He, I. Smirnova, M. Y. Liu, C. Van Huffel, X. Du, D. Birdwell, E. Alejos, M. Silva, C. Galanos, M. Freudenberg, P. Ricciardi-Castagnoli, B. Layton, and B. Beutler. Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in Tlr4 gene. *Science,* 282(5396):2085-8, 1998.

[35] A. Dunne, M. Ejdeback, P. L. Ludidi, L. A. O'Neill, and N. J. Gay. Structural complementarity of Toll/interleukin-1 receptor domains in Toll-like receptors and the adaptors Mal and MyD88. *The Journal of biological chemistry,* 278(42):41443-51, 2003.

# Integration of -omics data and networks for biomedical research

**Christian Klukas und Falk Schreiber** [1]

[1] Leibniz Institute of Plant Genetics and Crop Plant Research,

Corrensstr. 3, 06466 Gatersleben, Germany

klukas@ipk-gatersleben.de, schreibe@ipk-gatersleben.de

## Abstract

More and more often research focus in the fields of biology and medicine moves from the investigation of single phenomena to the analysis of complex cause and effect relations. The clarification of complicated relations requires the consideration of different domains, for instance, gene expression, protein, and metabolite data. Furthermore, it is often sensible not to analyze measured data in isolation, but to consider the context of relevant biological networks. In this paper newly developed functionalities of the VANTED system are presented. They allow users from medicine and biology to interactively structure extensive experiment data, to filter, to evaluate, and to visualize the data and the analysis results in context of biological networks and classification hierarchies.

## 1. Introduction

The methodology of biochemical research has strongly changed during the last years. Nowadays large amounts of experimental data is produced by massive-parallel analysis technologies, for instance by automated enzyme-assays, metabolite and transcriptprofiling. Using the right supporting software, the resulting data base provides a comprehensive view on the biochemistry of an organism. The clarification of complicated connections in organisms generally requires the consideration of different domains. To handle this problem, instead of analysing the data in isolation, it is worth to consider the context of relevant biological networks. Available software systems for this task (see [SH07]) are tuned besides a few exceptions to single data domains and/or are firmly coupled to certain databases. In this paper newly developed functionalities of the VANTED system [JKS06] are presented. They allow users from medicine and biology to interactively work with extensive experiment data, to filter, to statistically evaluate, and to visualize the analysis results directly in context of relevant biological networks and classification hierarchies.

## 2. Methods

For the analysis of experimental data integrated views of the measured values and relevant

background information should be generated. This approach corresponds with the idea of system biology – instead of considering single parts, the analysis covers the overall system with all interactions to better understand biological phenomena. In order to fulfil the goal of creating a software system which supports users in the analysis, three aspects are of importance for the design of the VANTED software: 1) data models for experiment data and biological networks, 2) the process of data mapping, by means of connecting experiment data and networks, 3) the analysis and visualisation of the network-integrated data sets. These three points are described in the following.

1) **Data models for experiment data and biological networks** By investigation of common experiment designs the following crucial experimental factors have been identified: information about time series, replicates, environmental conditions, treatments and genetic lines. A data model which is able to handle experiment data, partitioned by the mentioned experiment factors, has been developed and is shown in Figure 1. To simplify the design and implementation, the model does not store information about the experiment procedures, but instead focuses on information required for experiment data mapping, visualization and analysis.

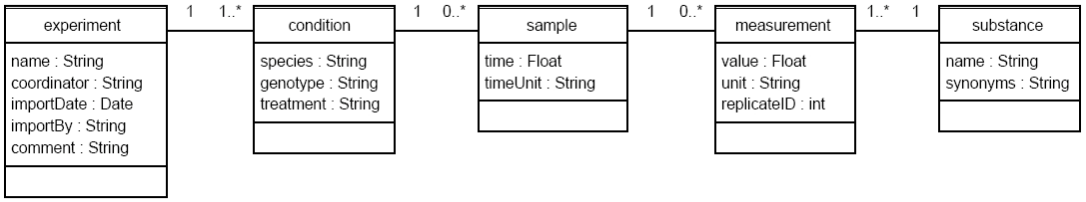



**Figure 1: Data model (UML class diagram).**

In contrast to some other systems VANTED supports dynamic networks. Networks can be loaded into the system from databases (e. g. KEGG) or from files (e. g. GML, SBML, Pajek .net format). In addition, it is possible to construct or edit networks manually with integrated editor functions, thus networks can be easily extended if more substances were measured.

2) **Data mapping** For the integration of measurement data into relevant networks a data mapping is carried out. If measurement data and networks share common identifiers, the data mapping procedure is carried out automatically. In addition, synonyms are considered for network elements as well as for experiment data. Information about synonyms and alternative identifiers is taken automatically from integrated databases (Expasy Enzyme [Bai00], KEGG Compound and KEGG BRITE [KAG+08]) or can be provided by the user. Optionally, data sets which could not be mapped on the basis of substance names and synonyms are mapped to newly generated network

nodes. In this manner new substances can be easily integrated into an existing network.

3) **Histogram functions for classification hierarchies and network-integrated data** The basis of the histogram function are classification hierarchies modelled as graphs (e.g. Gene Ontology or KEGG BRITE) consisting of classification nodes CN and leafnodes representing genes LN. By means of a data analysis function LN of the hierarchy, containing the experiment data are partitioned into several groups depending on the assigned data (e. g. up- or down-regulated gene nodes). In order to get an overview about the classification-specific group assignment within CN, the frequencies of LN groupassignments are determined and a corresponding data set is constructed for every non-leaf CN hierarchy node. This data set is visualized by node-embedded bar- or pie-charts. The most interesting CN nodes are nodes which show an uncommon pattern in the frequency of assigned groups. The significance of a observed frequency distribution in comparison to the overall proportions can be analyzed using Fisher's exact test. The result of this statistical test is a probability value p. If p lies under a user-defined threshold (e. g. p _ 0:05), the observed frequency distribution is regarded as non-random and therefore as significant. The visualization may then be simplified by removing all nodes from the hierarchy from which there is no significant node reachable, see Figure 2 (top).

VANTED also supports the visualization of several values connected to a single network element. While other systems often support only a simple colour code for the representation of a single measured value or the ratio of two values, the integration of diagrams into the network representation enables the visualization of more complex structured data sets. Another advantage in using line- or bar-charts is that such kinds of diagrams are widely used in other areas and thus 3 are easy to understand.

## 3. Application example

Certain human cell lines are used to investigate the development of cancer. For this application example gene expression data of a human cell line, affected by a specific type of carcinoma (human choriocarcinoma BeWo), is compared to a control line (human placenta). The data sets were downloaded from the KEGG EXPRESSION database [KAG+08]. In order to get a general overview about the differences of the lines, the gene expression data can be assigned with the VANTED system to the KEGG pathway hierarchy (using information from the KEGG BRITE system). In the present data set no additional annotation files need to be considered in order to generate a corresponding pathway hierarchy, because the data sets from the KEGG EXPRESSION database already use gene IDs, used also inside the KEGG pathway diagrams. For datasets from a different source additional annotation files may be needed.

At first a data mapping is carried out which generates for each gene of the data set a new node. After that, the automated workflow is started by using the menu command “Hierarchy/Analysis Pipeline”:

1. Depending on the gene expression values, the available network nodes are categorized as down-, up- or not-regulated. A user-specified threshold is used during this procedure.

2. Gene-nodes are related to the KEGG Pathway hierarchy which is constructed as classification tree. Each new node of the classification tree represents a pathway, a BRITE gene function or a (pathway) category. In the present data set 695 out of 836 gene nodes could be connected to at least one node of the classification hierarchy.

3. Histograms are calculated. For each classification node the number of reachable nodes, belonging to a user-selected group as well as the number of remaining reachable gene-nodes is determined. In this example, user selected nodes are unregulated nodes.

4. In step two of the pipeline 695 genes were assigned to 190 different pathways. Most of the pathways show a similar relative proportion of not-regulated to up- or downregulated genes. With the help of Fisher’s exact test those pathways can be identified, which show a non-random divergence to this pattern. By using p _ 0:05, 18 pathway nodes and KEGG BRITE category nodes remain in the visualization and can be easily investigated in more detail.

5. The layout of the result network takes place.

The result of the pipeline is shown in Figure 2 (top). The remaining pathway nodes contain in comparison to the complete data set either a comparatively high or low number of regulated genes. From the classification hierarchy KEGG pathways can be loaded and be investigated in detail. Figure 2 (bottom) shows the ECM-Receptor Interaction pathway which contains a comparatively large number of down-regulated genes and two upregulated genes. The corresponding distribution of the genes within the pathways can be easily recognized.

## 4. Summary

Because VANTED is implemented as an open source Java Web Start application it can be used on most computer platforms such as Linux, Windows and Mac OS X. The combination of functions for the network-integrated visualization and analysis of experiment data of different -omics areas, covering the access to KEGG pathways, Gene Ontology and the flexible visualization of time series data, including different conditions and replicates, make VANTED a valuable tool for research projects in biology, medicine and bioinformatics.

## References


[Bai00] A. Bairoch. The ENZYME database in 2000. Nucleic Acids Research, 28(1):304–305, 2000.

[JKS06] B. H. Junker, C. Klukas und F. Schreiber. VANTED: A system for advanced data analysis and visualization in the context of biological networks. BMC Bioinf., 7:109.1–13, 2006.

[KAG+08] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu et al. KEGG for linking genomes to life and the environment. Nucleic Acids Research, 36(Database issue):D480, 2008.

[SH07] M. Suderman und M. Hallett. Tools for visually exploring biological networks. Bioinfo., 23(20):2651–2659, 2007.

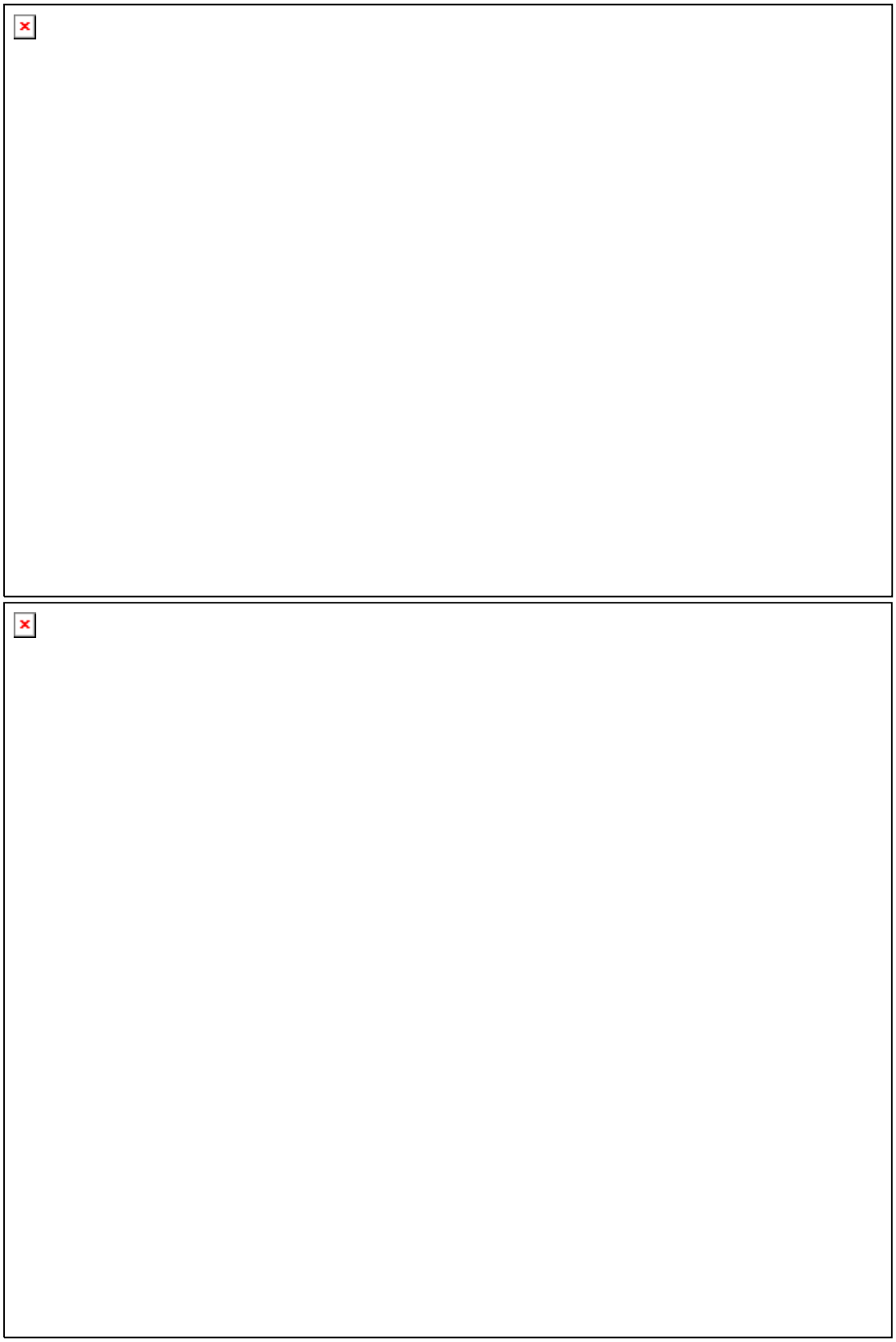


**Figure 2: Top: KEGG BRITE pathway hierarchy (by means of Fisher's exact test as significant recognized pathways). Note that only classification nodes CN are shown. Bottom: ECM-Receptor Interaction pathway with detailed representation of up- (red) and down-regulated (blue) genes.**

# Predicting the Oncogenic Pathway Activities of Individual Samples from Microarray Gene Expression Profiles

**Gaofeng HUANG [1,2,*] and Wee Joo CHNG [2,3]**

[1] Balliol College, University of Oxford, United Kingdom

[2] Dept. of Hematology-Oncology, National University Hospital, Singapore

[3] School of Medicine, National University of Singapore, Singapore

*Corresponding email: g.huang@balliol.oxon.org

## 1. Summary

Cancer is a genetic disease related to DNA mutations in cells. The efforts to identify the key oncogenic mutations help us to understand the causes and progression of various cancers, and therefore open up more treatment options. Nowadays, the development of whole genome microarray expression profiling (GEP) allow us to monitor the expression value of every gene throughout the whole genome. However, there are still two major issues: first of all, most cancers are not due to one single gene mutation, but multiple ones typically involving in several cell signaling pathways related to the control of cell grown and cell fate; secondly, GEPs are measuring mRNA levels, which might not related to the levels of active form proteins, and hence might not related to downstream biological consequences.

Gene Set Enrichment Analysis (GSEA [1]) and BinReg [2] are two recent developments in this area. GSEA predicts pathway activities between two conditions, for example, cancer tissues versus normal controls. But it is unable to predict pathway activities for each individual sample. BinReg can predict pathway activities for each individual one but it need a training GEP matrix for each pathway of interest as input, which makes it less applicable in real clinical settings.

Here, we proposed two novel methods, namely iGSEA and iPASA, to predict the pathway activities of individual samples from microarray gene expression profiles, and overcome the above mentioned shortcomings of GSEA and BinReg. The results of our methods not only are compatible with BinReg but also correlate nicely with clinical experiments.

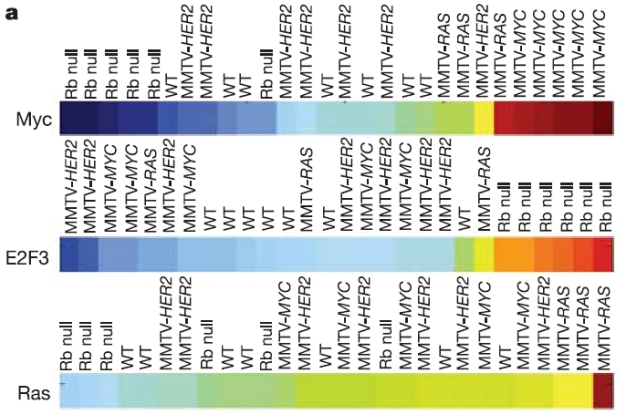
## 2. Results

### Mouse Tumor Dataset

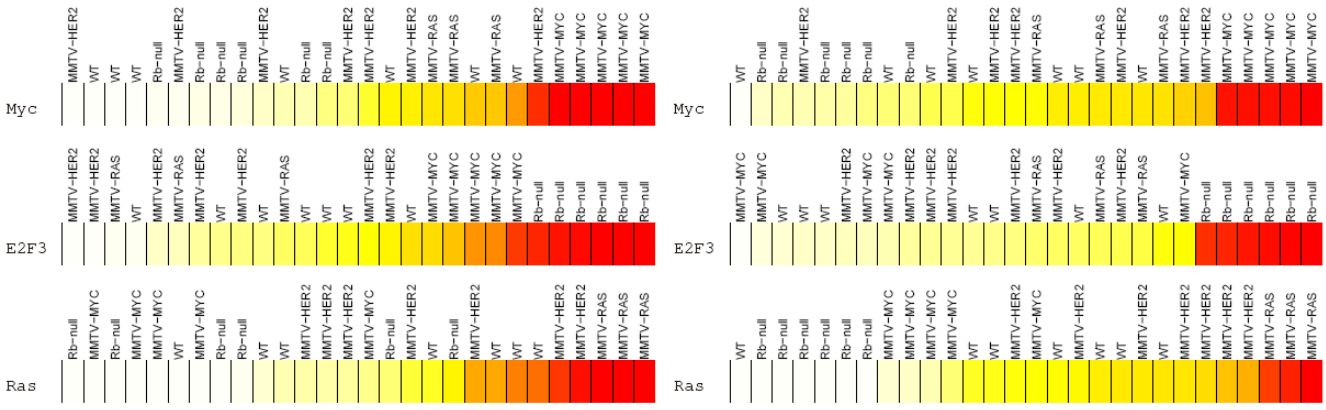
Following BinReg [2], we test our methods on the same mouse tumor dataset (GEO number: GSE3158). The mouse tumor dataset contains 28 samples, including 7 normal wild-type reference samples (WT), 5 Mammary Tumor Virus (MMTV) samples affecting MYC gene (MMTVMYC),

3 MMTV samples affecting HRAS gene (MMTV-HRAS), 7 MMTV-HER2, and 6 samples with gene Rb deleted (Rb null).

Figure 1(a) is the predicted results for 3 pathways (MYC, E2F3 and RAS) from BinReg [2]. For each pathway, the 28 samples are sorted from left to right according to the predicted pathway activities. It is clearly shown in the figure that the 5 MMTV-MYC samples are predicted to have high MYC pathway activities; the 6 Rb-null samples are predicted to have high E2F3 pathway activities; the 3 MMTV-HRAS samples are predicted to have mid to high RAS pathway activities. All of these fit the underlying biology very well. Figure 1(b) and Figure 1(c) show the results of our two methods, iGSEA and iPASA. As you can see, both our two methods produce compatible results with BinReg. The MMTV-MYC, MMTV-RAS and Rb-null samples are exclusively predicted to have high MYC, RAS and E2F3 activities, and iPASA achieves slightly better specificity than iGSEA.




(a) BinReg results [2]




(b) iGSEA results (c) iPASA results

**Figure 1: Predicted pathway activities in mutated mouse model**

**Multiple Myeloma Clinical Patient Sample Dataset**

We further study the MYC pathway activities in a cohort of 57 Multiple Myeloma (MM) patients. The MYC pathway is supposed to play a very important role in MM. However, unlike the previous mouse tumor datset, we don't actually know the "true" MYC pathway activities of each individual MM patient sample. Hence, we conduct the MYC protein staining experiments for these 57 patients samples to get the "true" values that our computational predictions can compare with. Figure 2 shows the prediction results of our iGESA and iPASA algorithms. The 57 samples are sorted from left to right by their predicted MYC pathway activities, and the labels are the results of the MYC protein staining experiments. As reference, we also put "MYC expression" in the figure, which is the expression value of the MYC gene itself; and "core MYC genes", which is the average expression of 6 core MYC related genes manually chosen by our clinical experts.

As mentioned earlier gene expression mRNA levels do not necessarily fit protein expression levels. In the right end of the "MYC expression" track of the figure, there are 4 samples with high MYC gene expression level but low MYC protein staining values (0 or < 5%). Using iGSEA and iPASA, we are able to overcome this to some extend, and as you can see in the figure, the predicted high MYC activities actually fit with the high MYC protein staining values. The computational prediction results of iGSEA and iPASA are compatible with using "core MYC genes", which is the current laborious manual procedure in our clinical lab.

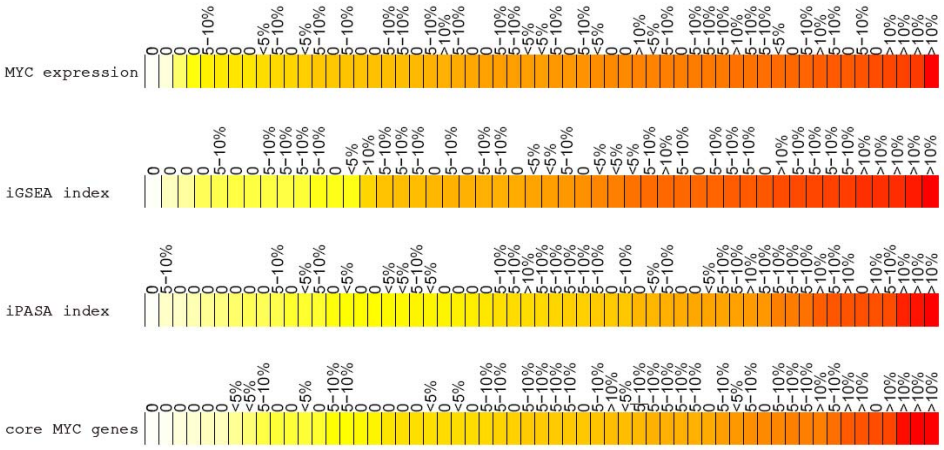
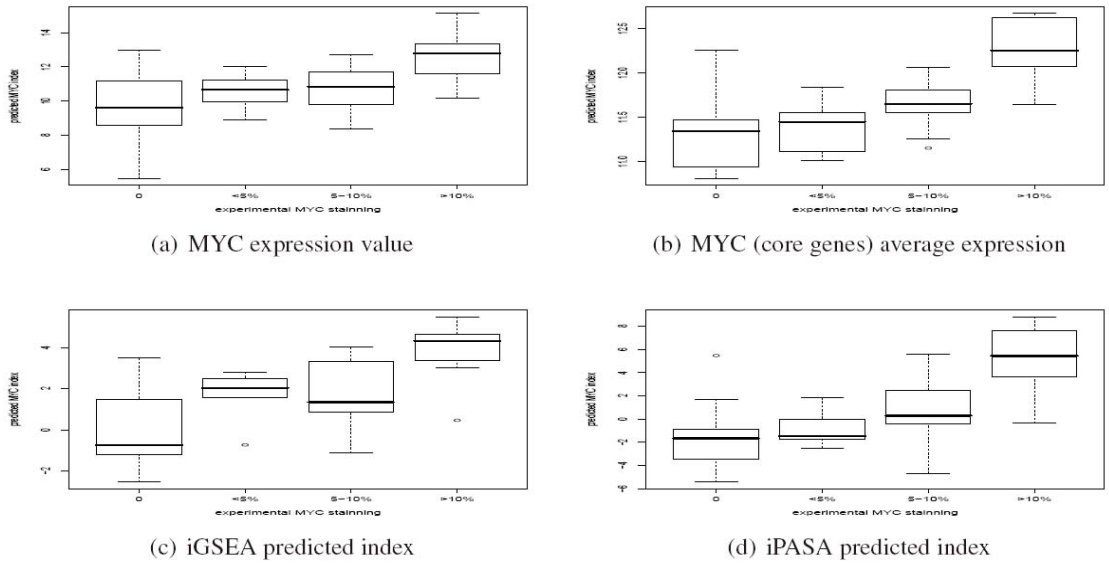



**Figure 2: Predicted MYC pathway activities**

Figure 3 further shows the boxplot over 4 MYC protein staining categories, 0, < 5%, 5%−10% and > 10%. The figure clearly shows that our prediction values correlate with the experimental staining results. Although we cannot directly calculate the correlation coefficients due to the non-linearity, iGSEA and iPASA predictions show more difference between low (0 and < 5%) and high (5% − 10% and > 10%) MYC staining than using only MYC gene expression value. We also conduct leave-one-out cross validation using a simple threshold classifier between low and

high MYC staining; iGSEA and iPASA achieve the prediction accuracy 68.8% and 80.7% respectively, while using MYC gene only achieves 60.7%, and using the 6 core MYC genes also achieves 82.7%.




(a) MYC expression value (b) MYC (core genes) average expression

(c) iGSEA predicted index (d) iPASA predicted index

**Figure 3: Predicted MYC pathway activities vs. Experimental MYC staining results**

## 3. Methods

### iGESA

The analysis of individual sample GSEA (iGSEA) is basically applying the classical GSEA [1] in 1-vs-*n* matter, i.e. to compare each case sample to all *n* controls. Given the gene expression profile matrix $Mp,(n+m)$ of $p$ genes, $n$ normal controls and $m$ case samples, we first use the $n$normal controls to estimate the mean and standard deviation of each gene's "normal" expression state. Then we perform z-score transformation for the $m$ cases using the estimated mean and standard deviation of each gene, and get a matrix $zp,m$. Further given a pathway signature of $k$ gene names, we calculate the Enrichment Score (as described in GSEA [1]) for each case (column of matrix $zp,m$), which indicates how enrich these $k$ genes are in a particular MM sample. This Enrichment Score is our predicted pathway activity index and permutation test can be further conducted to determine the significance of the score. Hence, unlike BinReg [2], our algorithm doesn't need a training gene expression matrix of the pathway of interest, and can predict the pathway activities from simply a predefined pathway signature ($k$ gene names).

### iPASA

The idea of our iPASA (individual Pathway Activity Score Analysis) originates from the fact that using the average expression value of a handful of "core" genes for a pathway often gives very good estimation of the pathway activity. However, to find the "core" genes that can represent a

pathway is a laborious and difficult task which needs the help of clinic experts with in-depth understanding of the pathway of interest. iPASA is designed to tackle this problem.

To start with, iPASA takes a predefined pathway signature of *k* gene names as input. These *k* genes (usually hundreds) are not necessarily to be "core" or verified genes related to the pathway, and are usually curated gene lists from literature or other high throughput gene mutation experiments. The key idea is that if a subset of "core" genes within these *k* genes actually indicates the pathway activities, there should be a reasonable degree of correlation among the expression values of these genes. Furthermore, genes which have higher correlations are likely to be the "core" genes, and therefore deserve a higher weight in the algorithm. Principle Component Analysis (PCA) [3] is a suitable mathematical tool for this task. Hence, iPASA uses the first principle component (PC1) as the predicted pathway activity score.

**Connection between iGESA and iPASA**

There exists intrinsic mathematical connection between iGSEA and iPASA. First of all, the z-score transformation in iGSEA is an analog to the centering and scaling procedure in PCA analysis. The Enrichment Score in iGSEA is $ES \approx \alpha \sum_{i=1}^{k} z_i - \beta \mathrm{rank}(z_k)$, where and $\alpha$ and $\beta$ are constant coefficients, rank(zk) is the sorted rank order of *zk*. The predicted activities score in iPASA is $AS = \sum_{i=1}^{k} \omega_i z_i$, where $\omega i$ are PCA coefficients. So both two scoring methods have similar form, and differ in weighting coefficients.

## 4. Conclusions

In this paper, we proposed two effective computational methods (iGSEA and iPASA) to predict the pathway activities of individual samples from microarray gene expression profiles. Comparing to GSEA [1], our methods can predict the pathway activities for each individual sample. Comparing to BinReg [2], our methods need only a list of gene names as input rather than a training GEP profile matrix for a pathway of interest. The results of our methods are compatible with BinReg and correlate nicely with real clinical experiments.

## References


[1] A. Subramanian, P. Tamayo, et al. *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles* PNAS 102, 15545-15550, (2005)

[2] A. Bild, G. Yao, J. T. Chang, Q. Wang, A. Potti, D. Harpole, et al. *Oncogenic pathway signatures in human cancers as a guide to targeted therapies*, Nature 439, 353-357 (2006)

[3] I.T. Jolliffe *Principal Component Analysis*, Springer, NY, XXIX, 487 p. 28 illus. (2002)

# Using Nearest Neighbor Search (NNS) to Discover Transcription Factor Binding Sites (TFBSs)

**Gaofeng Huang**[1], **Peter Jeavons**[1]

[1] Computing Laboratory, University of Oxford, London, United Kingdom

## Abstract

In this paper, we study the prediction of Transcription Factor Binding Sites, more precisely, the Motif Discovery Problem (MDP). We proposed a novel approach of using algorithmically well-studied Nearest Neighbor Search (NNS) algorithms to tackle this problem. We also integrate several state-of-the-art algorithmic elements from the literature, including Local Sensitive Hashing, Hidden Markov Background Modelling, and Gibbs Sampling, into a uni03ed approach under a seed-and-grow framework. Experimental results shows that our algorithm outperforms other algorithms in term of both speed and solution quality.

# Reconstruction of biological networks based on life science data integration

**B. Kormeier [1], S. Janowski [1], T. Töpel [1], K. Hippe [2], P. Arrigo [3] and R. Hofestädt [1]**

[1] Bielefeld University, Bioinformatics Department PO Box 100131, D-33501 Bielefeld, Germany

[2] Leibniz-Institute of Plant Genetics and Crop Plant Research (IPK), Corrensstr.3, 06466 Gatersleben, Germany

[3] CNR ISMAC Via De Marini 6, Genoa, Italy

## Extended Abstract

Large amounts of high dimensional biological data are generated from different high-throughput experiments and from literature. The rapidly growing number of databases and data types poses the challenge of integrating the heterogeneous data, especially in biology. Currently there are about 1170 important molecular biology databases [1].

Thus, the challenge is to capture, model, integrate and analyze the data in a consistent way to provide a new and deeper insight into complex biological systems.

High throughput sequence investigation tools, array technologies for gene/protein analysis and the expanding electrical infrastructure for the study of molecular data represent the initiation of a virtual cell. The vision of implementation of a virtual cell combines bioinformatics and systems biology today. However, we are still a long way from implementing even a simple virtual cell. The first step in reaching this goal is to understand the metabolism, which is based on gene-controlled biochemical reactions. Therefore, modeling and simulation of metabolic networks is important. Regarding the literature, different methods of modeling biological networks have been introduced. One other problem is the quantitative simulation of these processes. Therefore, it is still an open question to find the most useful method for the simulation of biological networks, which will represent the backbone of a virtual cell. In our paper we will present a new tool which creates a large scale biological network using data integration and data warehousing methods.

BioDWH[2] is implemented in Java and uses a relational database management system in its backend, e.g., Oracle or MySQL. It provides an easy-to-use Java application for parsing and loading the source data into the data warehouse. Several ready-to-use parsers for popular life science information systems are already available, such as: UniProt, KEGG, OMIM, GO, Enzyme, BRENDA, PDB, MINT, SCOP, EMBL-Bank, and PubChem. Furthermore, an XML-configurable

monitor for data source updates is part of the system. For status requests to the data warehouse, we have developed a graphical user interface that works with every web browser. A well-engineered, object-relational mapping tool called Hibernate was used as a persistence layer, which performs well and is independent from manufacturers like MySQL or Oracle. Additionally, the Hibernate framework fits perfectly into the Java-based infrastructure of the data warehouse. A Java interface and the object-relational mapping using Hibernate persistence or Java Persistence Architecture (JPA) constitute an easy plug-in architecture for integration of new parser.

This object-relational mapping (ORM) is an automated and transparent persistence method of Java application for tables in a relational database system, whereas a mapping between objects and metadata of the database is described. In principle, ORM works with reversible transformation of data from one representation into another. An ORM solution consists of four parts: first, an application programming interface (API) that executes simple CRUD (create, retrieve, update, delete) operations using objects of persistent classes; second, a programming language or API to formulate queries that depend on Java entity classes or properties of classes; third, a facility for mapping metadata; finally, techniques of an ORM implementation to handle interactions of dirty checking, lazy association fetching and other optimization functions of transactional objects.

The different features of BioDWH are usable by a graphical user interface. It enables the configuration of the monitor and parser for the different public life science data sources as well as the local database management system. BioDWH is available at http://biodwh.sourceforge.net/.

Based on the CardioWorkBench EU project (http://www.cardioworkbench.eu/) we implemented a platform-independent data warehouse system that integrates multiple heterogeneous data sources into a local database enriched with protein microarrays from human smooth muscle cells that are related to cardiovascular diseases. Based on our VINEdb[3] information system we extended CardioVINEdb (http://agbi.techfak.uni-bielefeld.de/CardioVINEdb/) with more data sources, better data warehouse infrastructure including monitoring and microarray data. In addition, we upgraded the visualization components and web pages for better navigation and exploration. To ensure maximum up-to-dateness of the integrated data, we developed a data warehouse infrastructure including a monitor component. Furthermore, the common web-based user interface provides a visualization component that allows interactive exploration of the integrated data.

Based on the data content of the BioDWH data warehouse, we would like to introduce VANESA (Visualization and Analysis of Networks in System Biology Applications). VANESA, a JAVA based software solution, is an application for modeling and visualization of biological networks.

With the use of VANESA, we were able to model and visualize the most important pathways based on the proteins in the different microarray samples. VANESA is available at http://vanesa.sourceforge.net/.

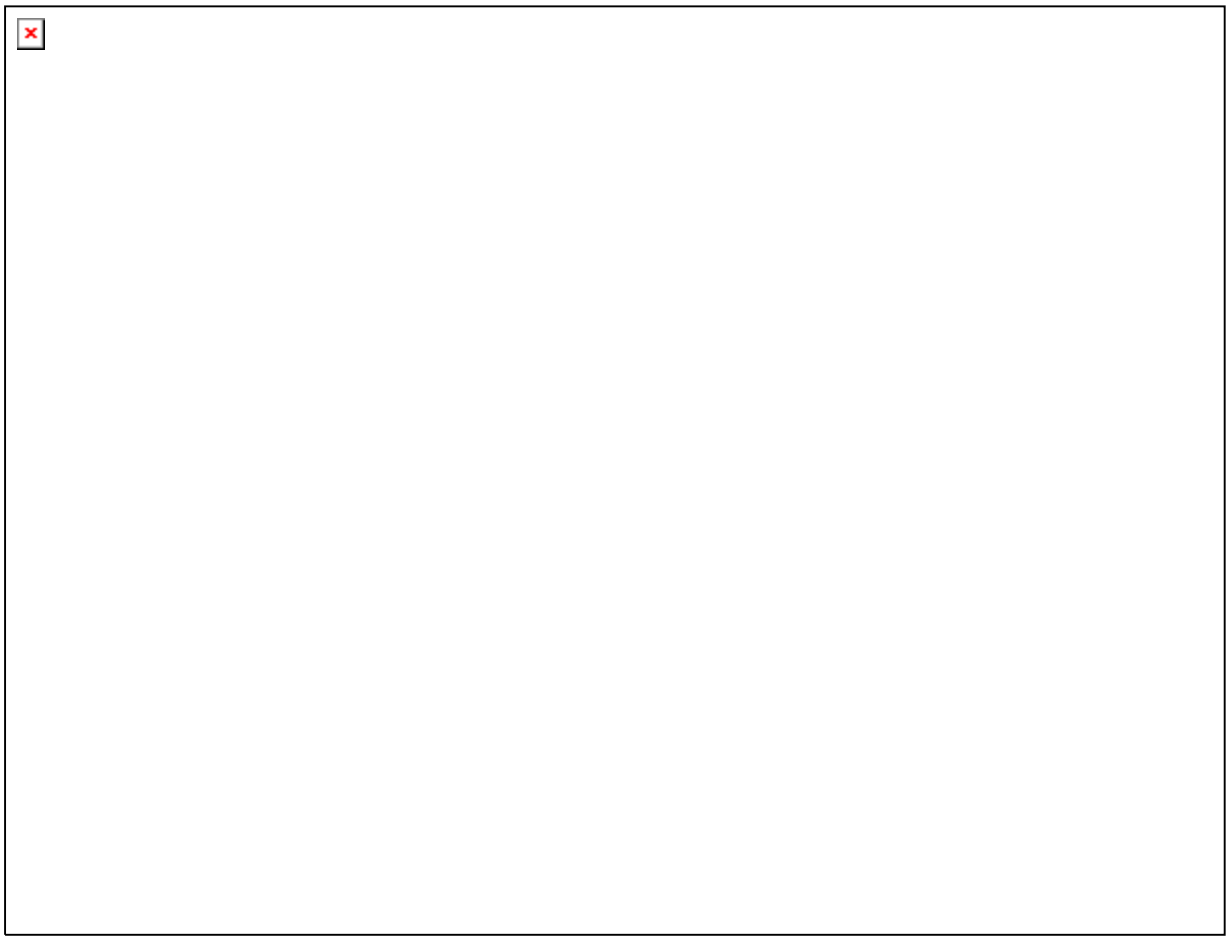


**Figure 1: Visualization of the Tight junction signaling pathway (hsa04530) by VANESA. The red marked place is the relevant protein on the microarry sample.**


[1] M.Y. Galperin. The Molecular Biology Database Collection: 2008 update. Nucleic Acids Research, 36(Database issue):D2-D4, 2008.

[2] T. Töpel, B. Kormeier, A. Klassen and R. Hofestädt. BioDWH: A Data Warehouse Kit for Life Science Data Integration. Journal of Integrative Bioinformatics, 5(2):93, 2008.

[3] S. Hariharaputran, T. Töpel, B. Brockschmidt and R. Hofestädt. VINEdb: a data warehouse for integration and interactive exploration of life science data. Journal of Integrative Bioinformatics, 4(3):63, 2007. Online Journal: http://journal.imbio.de/index.php?paper_id=63

# Simulation-based Model Checking Approach for Cell Fate Specification - Using Cell Illustrator Online: A Computational Platform for Systems Biology

**Chen Li, Masao Nagasaki, Ayumu Saito, Satoru Miyano**

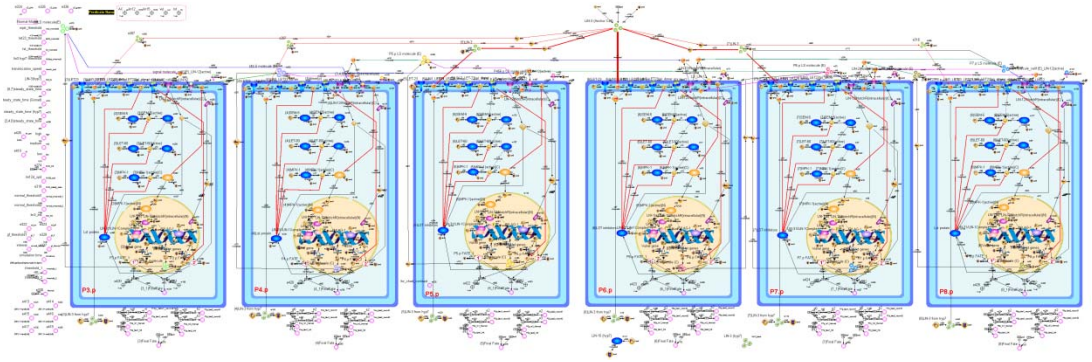
Human Genome Center, Institute of Medical Science

The University of Tokyo

4-6-1 Shirokanedai, Minatoku, Tokyo 108-8639, Japan

## 1. Simulation-based Model Checking

Model checking is a powerful technique for automatically verifying the requirements of finite state concurrent systems. Since the size of the state space grows exponentially with the number of processes when dealing with continuous values, this problem is generally intractable and serves as a major obstacle hampering further advancement. Till now, researchers generally deal with such problem by means of discretizing discrete model or continuous model of biological networks under specific abstraction criteria. We establish a quantitative methodology by handling quantitative values without such discretization, to model and analyze an *in silico* model incorporating the use of model checking based on a biosimulation tool Cell Illustrator Online (CIO).



We construct above large-scale quantitative vulval precursor cell fate specification model of *C. elegans* with CIO. This probabilistic model involves totally 1761 components (place: 426, transition: 442, arc: 780). We performed 480,000 simulations and examined the consistency and correctness of the model under 48 sets of genotypes that are the combinations of four genes and one anchor cell. This method is proved to be a useful means to give researchers valuable biological insights and better understandings of biological systems and observation data that are hard to capture with the qualitative approach.

## 2. Cell Illustrator Online

We developed a Cell System Ontology (CSO) (http://www.csml.org/) and Cell System Markup Language (CSML) for visualizing, modeling and simulating biological pathways. Based on CSO and CSML, we developed a modeling and simulation tool Cell Illustrator Online (CIO, https://cionline.hgc.jp/) that enables wet lab biologists to draw, model, elucidate and simulate complex biological processes and systems such as metabolic pathways, signal transduction cascades, gene regulatory pathways and dynamic interactions of various biological entities (e.g. genomic DNA, mRNA and proteins). The architecture employs Hybrid Functional Petri Net with extension (HFPNe) which is an enhanced Petri net enabling intuitive modeling for biologists. CIO comes preloaded with TRANSPATH® pathways and chains, providing immediate access to signal transduction and metabolic pathway representations derived from the scientific literature. The integration of TRANSPATH reactions provides direct access to thousands of experimentally demonstrated binding and regulatory relationships – providing a unique set of building blocks for drawing custom networks and pathways. Furthermore, we have developed a method for automatic parameter estimation for HFPNe models by using a technolgy called data assimilation which "blends" simulation models and observational data "rationally". This data assimilation method is more suited for high performance computing systems and we plan to enroll this function into CIO in the near future for high performance computing environment.

# Towards identification of human disease phenotype-genotype association via a network-module based method

**Jeffrey Q. Jiang** [1,2]

[1] CAS-MPG Partner Institute for Computational Biology, Shanghai, China

[2] Department of Bioinformatics, College of Life Sciences, Zhejiang University, Hangzhou 310058, China

## Abstract

Inspired by recent discovery that human dis- ease phenome shows a modular organization on the genetic landscape, we introduce a network-module based method towards phenotype-genotype association inference and disease gene identification. This approach integrates protein-protein interaction network, phenotype similarity network and known phenotype-genotype associations, and then decomposes the resulted assembled network into modules (or communities) wherein we identified and prioritized the disease genes from the candidates within the loci associated with the query disease using a linear regression model and concordance score. For the known phenotype-gene associations in the OMIM database, we used the leave-one-out validation to evaluate the feasibility of our method, and successfully ranked known disease genes at top 1 in 887 out of 1807 cases. Moreover, applying this approach on 850 OMIM loci characterized by an unknown molecular basis, we propose high-probability candidates for 81 genetic diseases.

# An Algorithm to Produce Conditional Equation for Smooth Signal Flows in the Petri Net Model of a Signaling Pathway

**Yoshimasa Miwa** [1]

[1] Biopathway Analysis Center, Faculty of Science, Yamaguchi University, Yamaguchi, Japan

## Abstract

Parameter determination is a critical problem in modeling and simulating biological pathways such as signaling pathways. Signaling pathways are information cascades of enzyme reactions from transmembrane receptors to the nucleus DNA, which ultimately regulate intracellular responses such as programmed cellular proliferation, gene expression, differentiation, secretion and apoptosis.

Basic facts for deciding parameter can be obtained from biological experiments and scientific common principles. However, in the majority of cases, reliable data of detailed reactions have not been reported in biological literature.This leads us to develop a method that determines parameter of model without experimental data based on biological literature.

In this study, we propose an algorithm to produce conditional equations that estimated parameters realize smooth signal flows in the model of a signaling pathway. We have used gPetri neth for modeling signaling pathways.Petri net is a powerful tool in modeling and simulating various concurrent systems, and recently have been widely accepted as a description method for biological pathways. We have used an example of IL-1 signaling pathway to demonstrate our proposed method.

Firstly, we have modeled a discrete Petri net model of IL-1 signaling pathway.Then, we have determined the firing frequency of each transition by applying the proposed method for a part of the Petri net model of IL-1 signaling pathway. Finally, we have simulated Petri net model of IL-1 signaling pathway to confirm the appropriateness and validity of proposed method by using Cell Illustrator 3.0 with the decided delay times.

# Gene identification and regulation of phenylpropanoid pathways in plants

**Jie Luo [1,*], Eugenio Butelli [2], Lionel Hill [2], Adrian Parr [3], and Cathie Martin [2]**

[1] National Key Laboratory of Crop Genetic Improvement, National Center of Plant Gene Research (Wuhan), Huazhong Agricultural University, Wuhan 430070, China

[2] Department of Metabolic Biology, John Innes Centre, Colney, Norwich, NR4 7UH, United Kingdom

[3] Technologies for System Biology, Institute of Food Research, Colney, Norwich, NR4 7UA, United Kingdom

*Correspond author: jie.luo@mail.hzau.edu.cn

## Abstract

Phenylpropanoid pathway is one of the most important secondary metabolic pathways in plants. Many of the products from different branches of this pathway, such as anthocyanins and flavonols, exhibit a broad spectrum of health-promoting effects when consumed as part of the diet. There is considerable interest in elucidating the regulation of this pathway and in enhancing the levels of these bioactive molecules in plants used as foods. Using an integration of targeted metabolomics and functional genomics approach, we have functionally identified a number of genes encoding BAHD acyltransferases involved in the phenylpropanoid-related pathways. By incorporating co-expression profiling with metabolite accumulation, genes encoding three anthocyanin acyl transferases were identified from Arabidopsis thaliana (1). By employing metabolomics-oriented reverese genetic approach, three genes encoding spermdine hydroxycinnamoyl transferases were identified with distinctive specificity (2)!

For the metabolic engineering to enchance the accumulation of health-promoting compounds such as polyphenols derived from the phenylpropanoid pathway, fruit-specific expressing AtMYB12 (a MYB transcritption factor from A. thaliana) in tomato resulted in organe fruit with extremely high levels of multiple polyphenolic antioxidants (3). Simulaneously expressing two transcription factors (AmDel and AmRos1 from Antirrhinum majus) in a fruit-specific manner in tomato resulted in very high levels of anthocyanins in fruit with intense purple coloration in both peel and flesh. Cancer-susceptible Trp53C/C mice fed a diet supplemented with the purple tomatoes showed a significant extension of life span (4).

(1) Luo, J et al., Plant J, 50, 678 (2007); (2) Luo, J et al., Plant Cell, 21, 318 (2009);
(3) Luo, J et al., Plant J, 56, 316 (2008); (4) Butelli, E et al., Nat Biotechnol, 26, 1301 (2008)

# Using surveys of Affymetrix GeneChips to study antisense expression

**Olivia Sanchez-Graillet, Maria A. Stalteri, Joanna Rowsell, Graham J.G. Upton and Andrew P. Harrison***

***- corresponding author: harry@essex.ac.uk**

Departments of Mathematical Sciences and Biological Sciences

University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ

## Abstract

We have used large surveys of Affymetrix GeneChip data in the public domain to conduct a study of antisense expression across diverse conditions.

We derive correlations between groups of probes which map uniquely to the same exon in the antisense direction. When there are no probes assigned to an exon in the sense direction we find that many of the antisense groups fail to detect a coherent block of transcription. We find that only a minority of these groups contain coherent blocks of antisense expression suggesting transcription.

We also derive correlations between groups of probes which map uniquely to the same exon in both sense and antisense direction. In some of these cases the locations of sense probes overlap with the antisense probes, and the sense and antisense probe intensities are correlated with each other. This configuration suggests the existence of a Natural Antisense Transcript (NAT) pair. We find the majority of such NAT pairs detected by GeneChips are formed by a transcript of an established gene and either an EST or an mRNA.

In order to determine the exact antisense regulatory mechanism indicated by the correlation of sense probes with antisense probes, a further investigation is necessary for every particular case of interest. However, the analysis of microarray data has proved to be a good method to reconfirm known NATs, discover new ones, as well as to notice possible problems in the annotation of antisense transcripts.

## 1. Introduction

Our knowledge of the transcriptome is rapidly evolving and it is becoming increasingly clear that RNA plays a range of diverse roles in regulating gene expression [1]. Natural antisense transcripts (NATs) are endogenous RNAs whose sequences are complementary to other transcripts. Antisense

transcripts are implicated in transcription, processing, stability, transport and translation of their complementary RNAs [2]. NATs have now been found in many organisms, but we have little knowledge of the functions of many of these transcripts [3]. Bioinformatic approaches show a large number of potential NATs in genomic sequences, but provide no information about the expression of NATs in specific cell or tissue types [4]. It is therefore important to experimentally verify the expression of NATs in order to unravel their biology.

Affymetrix GeneChip technology [5] is a widely used resource in the life sciences. GeneChips provide multiple measures of the expression level for each gene. Each probe is a 25-nt oligomer (25mer) and each probeset, designed to represent a different gene transcript, typically consists of eleven perfect match (PM) probes as well as corresponding mismatch (MM) probes. The widespread popularity of GeneChips, with large data-sets stored in public repositories such as the Gene Expression Omnibus [6], makes them particularly suited for unravelling aspects of the transcriptome across many conditions, for diverse conditions, developmental stages, phenotypes and diseases. However, such studies will be limited to a sample of the transcriptome, that for which there are probes with the appropriate sequence. A huge number of expressed sequence tags (ESTs) were used in the design of the Affymetrix arrays [5] and due to the extensive use of such sequences Stalteri and Harrison [7] predicted that some probes may be mapping to exotic RNA sequences. We are not the only group to consider the use of Affymetrix data for exploring the biology of the transcriptome and there have already been searches for antisense expression using mouse arrays [8, 9].

Experimental artefacts in the preparation of targets, such as spurious synthesis of complementary strands, may act to confuse the interpretation of genome-wide experiments [2]. In some cases it is likely that a significant number of postulated NATs may be artefacts produced by genomic priming with contaminant genomic DNA during cDNA library construction [3]. Given the potential for confusion resulting from artefacts, it is imperative that care is taken in analysing Affymetrix data when searching for NATs. It is widely assumed that on a GeneChip multiple probes from within the same probeset measure the same thing. However, we find there are a number of probesets that contain probes behaving inconsistently with the rest of the probeset [10]. Some of these discrepancies may result from interesting biological processes such as alternative splicing and alternative polyadenylation [7]. However, other problems result from spatial flaws in hybridization [11]. Moreover, some probes may not measure expression reliably, due to particular sub-sequences, or motifs, within their 25 bases. Wu et al. [12] reported that probes containing runs of guanine were typically outliers in probesets and also showed abnormal binding affinities. We recently confirmed that such probes are outliers [13], but further discovered that the probes

containing runs of guanine are unusually well correlated with each other across many thousands of experiments. We associate this effect with the formation of G-quadruplexes occurring on the surface of a GeneChip [13]. Studying correlations in expression across many experiments is informative because coordinated biases affecting many probes simultaneously can be identified.

The regulation of antisense transcription might be tailored to its type of action [14], and the expression patterns of NATs and their targets might indicate the regulatory mechanism that is occurring. For example, when both sense and antisense probes are correlated with each other they should measure the same thing. This might indicate bidirectional transcription, particularly as [15] discovered that antisense transcripts are mainly located in the promoter and terminator regions of genes.

## 2. Materials and methods

We use a pipeline to analyse tens of thousands of Affymetrix GeneChips [10] downloaded from the Gene Expression Omnibus [6]. Our pipeline brings together unique mappings of probes, quality control analysis on each GeneChip and data-mining signal intensities across many experiments.

We first identify spatial flaws in individual GeneChips [11,16,17] that leads us to blank out signals from a fraction of each chip. We group all probes aligning to the same exon together, and we calculate the correlations between each of the probe pairs. The intensities are transformed onto a log scale and the signals are correlated across all experiments for one chip type. All the pair-wise probe correlations for each exon are collated into a matrix that is colour-coded according to the correlation value. The original correlation values are multiplied by ten, and then rounded, so that we express the correlations as integers. Heatmaps are symmetrical matrices in which the diagonal represents the perfect correlation of each probe with itself (correlation with value 10).

We wish to only study unique probes, those that only target one place on one exon [10]. This means that we only utilise a fraction of the probes available on the GeneChip, but the fraction we use will have been chosen to provide reliable measurements. We proceed by calculating the alignment “value” for each probe through multiplying the alignment length and the percentage sequence identity, e.g. a probe that aligns to a sequence with 25 bases and percentage sequence identity of 80% has an alignment value of 20 (25x0.8). A probe is considered to be mapping uniquely to an exon if it: aligns exactly (25 bases, 100% identity) to only one exon and to any of its synonyms (i.e. the exons in the same genomic region but with different Ensembl identifiers); maps to only one place on the exon; does not map to any exon-exon junctions; does not map

substantially to any other exon (i.e. does not have an alignment value between 20 and 25 for any other exon). We also identify probes that map uniquely to exons in an antisense direction. Such probes map uniquely to the reverse complement of the exon sequence. An example of an antisense probe is illustrated in Figure 1. We can observe that the reverse complement of the sequence of probe 201427_s_at:294:1093 aligns to exon ENSE00001435187. Thus if the NAT to ENSE00001435187 is expressed, it will be detected by the probe 201427_s_at:294:1093.

For the present study, we analyse the CEL files obtained from GEO for experiments that used the Human GeneChip HG-U133_Plus_2. We obtain our genomic coordinates and exon definitions from Ensembl (release 48). Probes containing the motif CCTCC or runs of four or more contiguous guanines were taken out of the exon heatmaps since they produce misleading information [18].


**Figure 1. Probe 201427_s_at maps in an antisense direction to exon ENSE00001435187 (only a fragment is shown).**

## 3. Results

In this section we present our analysis of the heatmaps generated for exons containing only antisense probes and for exons containing both sense and antisese probes.

### 3.1 Exons with antisense probes only

Almost all groups of antisense probes that map uniquely to an exon, but for which there are no sense probes mapping, showed low or negative correlation with each other. As an example, the antisense probes in Figure 2 are not correlated although they are from the same probeset (222076_at) and map to the same exon ENSE00000764836. The probes are not detecting a coherent signal. Figure 3 shows that there are RefSeq and Ensembl transcripts only on the positive strand. There are no transcripts on the negative strand, to which the probes in the probeset

222076_at map.


**Figure 2. The antisense probes from probeset 222076_at are not correlated. The columns indicate the probe order in the heatmap, probe identifier (in which pm means perfect match, followed by the order of the probe in its probeset), x-coordinate of probe location on the array, y-coordinate of probe location on the array, interrogation position of probe on Affymetrix consensus sequence, probe sequence, geometric mean of the intensities across GEO, and standard deviation (of the logs of intensities), respectively. The numbers in each of the cells represent the rounded correlation x 10.**

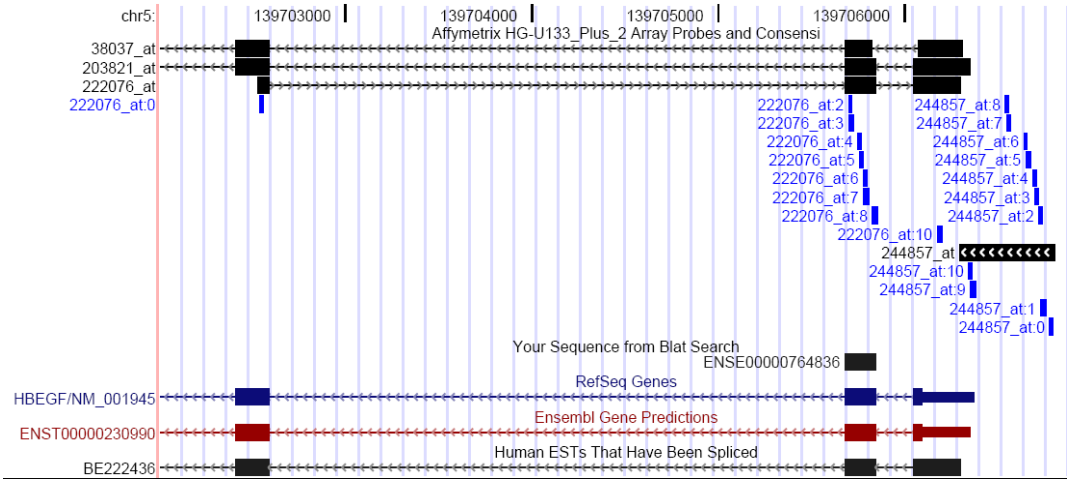



**Figure 3. Screen-shot of the UCSC browser [19] shows that the probes in the probeset 222076_at are on the positive strand and exon ENSE00000764836 (in ENST00000230990) is on the negative strand. There are no transcripts aligning to the positive strand in this region.**

There are only a few cases in which antisense probes mapping uniquely to the same exon (or from the same probeset) are highly correlated. As an example, the heatmap in Figure 4 refers to a group of highly correlated (average correlation 0.84) antisense probes in the probeset 201427_s_at. The probes illustrated are the only probes in the probeset that uniquely map to the exon ENSE0001435187. The high correlations among these antisense probes suggest that there may be a real biological signal. Figure 5 shows that the probes in probeset 201427_s_at map to a region where there is overlap between the 3' ends of the CCDC152 and SEPP1 genes, which are

transcribed from opposite strands. Probeset 201427_s_at aligns to the negative strand, and thus it aligns sense to the SEPP1 transcripts, and antisense to the RefSeq transcript CCDC152 and to the Ensembl transcript ENST00000361970 (through exon ENSE00001435187) that are on the positive strand. Affymetrix assigns probeset 201427_at to SEPP1, with an annotation grade of "A", i.e., at least 9 of the 11 probes perfectly match the associated transcripts [20], which in this case include the 3 RefSeq transcripts for SEPP1. The Affymetrix annotation for probeset 201427_s_at also includes several cross-hybridising transcripts assigned as having 11/11 Negative Strand Matching Probes. One of these is NM_001134848, the RefSeq transcript for CCDC152. The antisense (with respect to exon ENSE0001435187) transcription being detected is expected to be from the RefSeq transcripts corresponding to the SEPP1 gene. The NATsDB database [21] describes SEPP1 as belonging to a SA (sense-antisense) pair with the mRNA BC039102 that is on the positive strand.

| | | | | | | | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 6 | 201427_s_at.pm11 | 57,837 | 1986 | GGATACAGTACGGATTTGTCCAAAT | 1769 | 3.97 | 10 | 10 | 9 | 9 | 10 | |
| 5 | 201427_s_at.pm10 | 72,441 | 1935 | CCTGACCTCCTTTATGGTTAATACT | 1410 | 2.90 | 10 | 10 | 9 | 10 | | 10 |
| 4 | 201427_s_at.pm9 | 291,449 | 1927 | CCTATAAACCTGACCTCCTTTATGG | 1493 | 2.58 | 9 | 9 | 9 | | 10 | 9 |
| 3 | 201427_s_at.pm6 | 627,217 | 1816 | AAACTTGAGTGGCTGTCTTAAAAGA | 705 | 2.81 | 10 | 10 | | 9 | 9 | 9 |
| 2 | 201427_s_at.pm3 | 545,253 | 1733 | AAGACTCATCCTGATTTTTACTATC | 945 | 3.20 | 10 | | 10 | 9 | 10 | 10 |
| 1 | 201427_s_at.pm2 | 294,1093 | 1722 | TTTGTATATCTAAGACTCATCCTGA | 795 | 3.21 | | 10 | 10 | 9 | 10 | 10 |

**Figure 4. Highly correlated antisense probes from probeset 201427_s_at. The probes map in antisense direction to exon ENSE00001435187.**

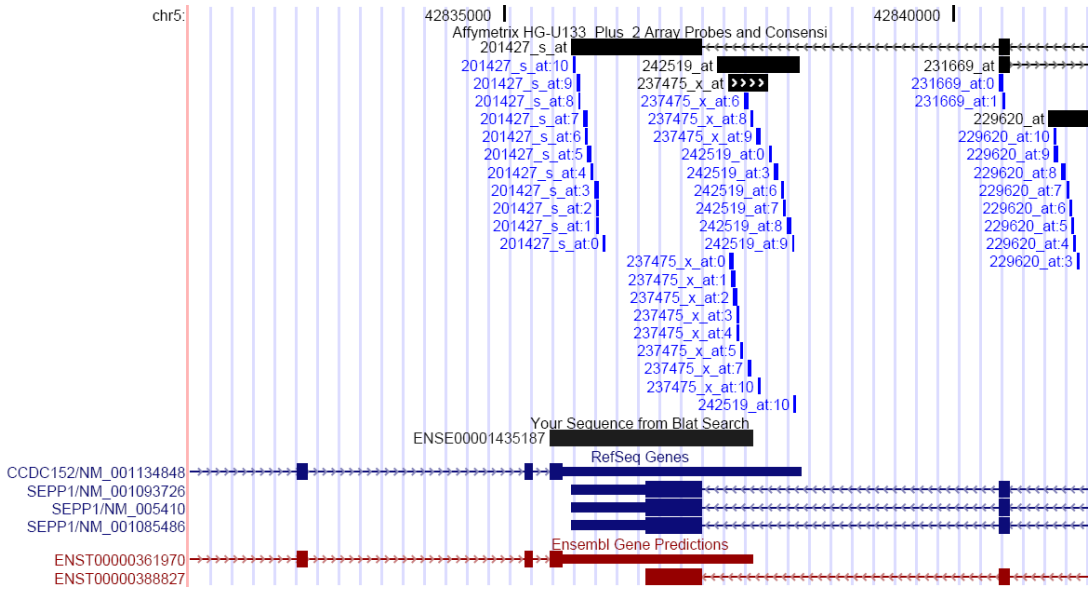


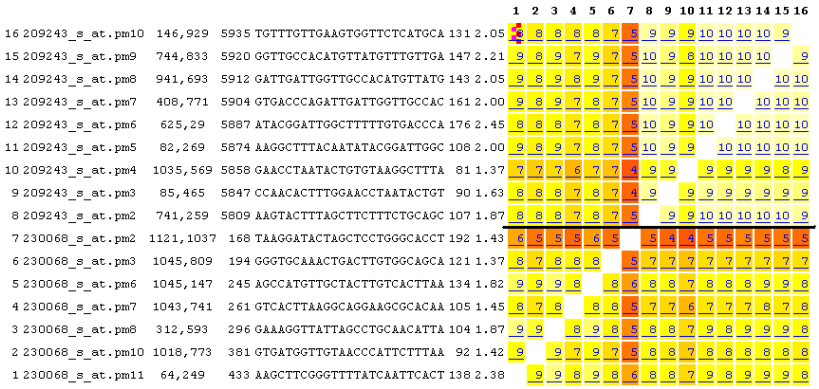

**Figure 5. Screen-shot of the UCSC browser [19] showing that probes in probeset 201427_s_at map to the SEPP1 transcripts (negative strand).**

## 3.2 Exons with antisense and sense probes

### 3.2.1 Classification by correlation heatmap

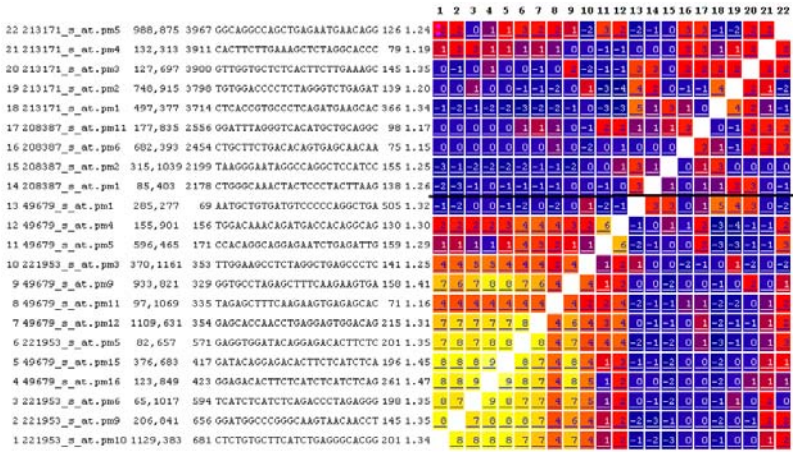
Out of 1,048 exons that have sense and antisense probes mapping uniquely to them, we selected 433 exons that have at least 4 sense probes and at least 4 antisense probes. We analysed a total of 100 exons randomly selected out of the 433 exons with probes in both senses. The exons can be classified into the general patterns shown in Table 1. Detailed descriptions of the biology of example transcripts matching these patterns are presented in the Supplementary material.

**Table 1. Examples of general patterns of exons containing probes in sense and antisense directions. In each heatmap, the probes below the line align in the antisense direction and the probes above the line align in the sense direction. The antisense and sense probes do not necessarily overlap.**

| Pattern Description | Number of cases | Example |
|---|---|---|
| **1:** The sense and antisense probes are correlated. | 14 |  |

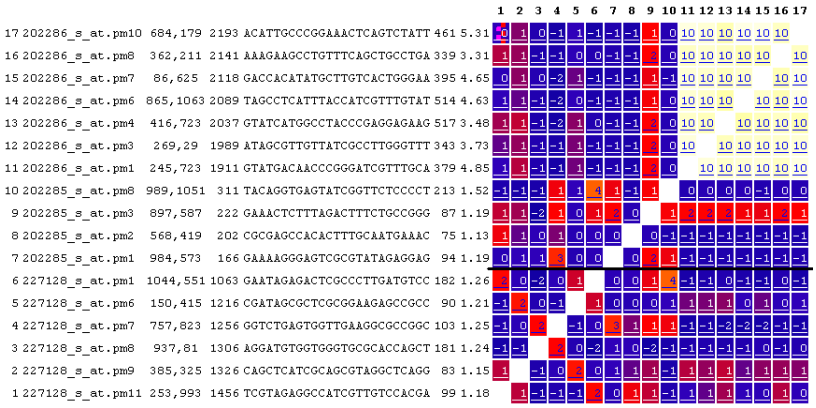
**ENSE00000876661**

**2:** Only the 4 antisense probes are correlated.



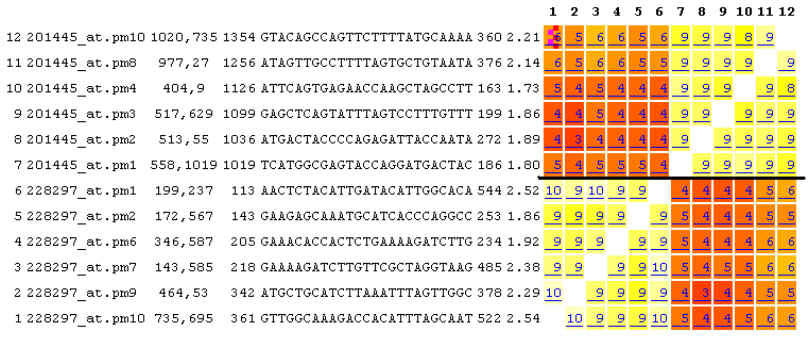
**ENSE00000860190**

**3:** Only the 40 sense probes are correlated. A sub-group of sense probes may not be correlated with another sub-group of sense probes.



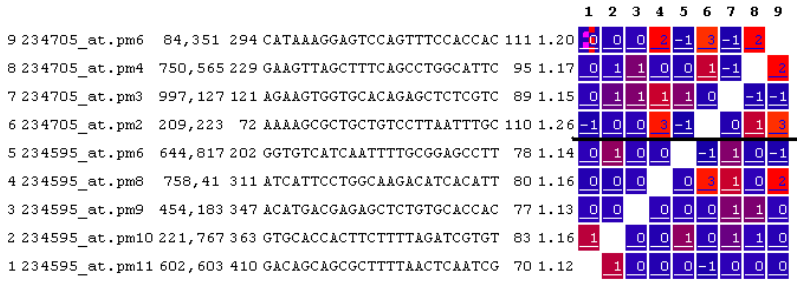
**ENSE00001454677**

**4:** The 13 antisense probes are correlated, the sense probes are correlated. The antisense probes are not correlated or



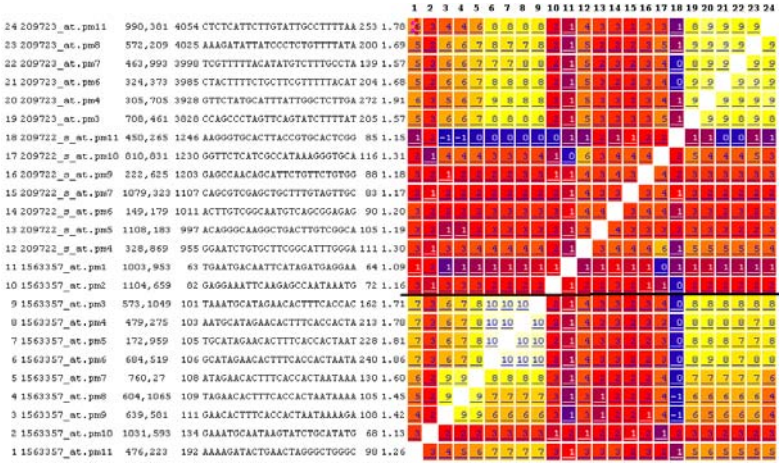
**ENSE00001452067**

**5:** The probes 26 are not correlated or are negatively correlated.



**ENSE00001222306**

**6:** The 3 antisense probes are correlated. Some of the sense probes are correlated. The antisense probes are correlated with a sub-group within the



**ENSE00001485956**

### 3.2.2 Antisense and sense probes heatmaps across different GEO experiments

In order to check whether the pattern presented by exon ENSE00001452067 (in pattern 4) was constant across different GEO experiments, we generated the heatmaps for this exon in 40 GSE experiment series that had at least 10 GSM cel files each. Figure 6 depicts these heatmaps and Table 2 contains the descriptions of the types of GSE experiments corresponding to each heatmap. The first 13 heatmaps from left to right starting at the top of Figure 6 correspond to experiments related to cancer. The remaining 27 heatmaps are not related to cancer.

In general, we observe that the antisense probes are highly correlated across the different experiments either with cancer or not. The antisense probe correlations are represented by the first 6 probes on the bottom left of the heatmaps. It seems that the presence of cancer does not determine the pattern of the correlation heatmaps.

### 3.2.3 General remarks

When an exon has sense and antisense probes mapping uniquely to it, there is usually a SA (sense antisense) pair in the NATsDB database [21] associated to the gene which contains that exon.

When sense and antisense probes overlap, the existing SA-pair is mainly formed by an established gene and by an mRNA or EST. This is confirmed by the work of Yelin et al. [22] who detected that the overlap between genes is frequently established by complementary ESTs, even when mRNAs are present in the clusters. They also observed that around 70% of the SA genes overlapped in their 5'-most and 3'-most exons (here called external exons) which supports the idea that SA overlap could be involved in gene regulation since the external exons contain UTRs of mRNAs. We find that ~70% of the exons containing sense and antisense probes are external exons and that ~23% of the exons containing only antisense probes are external exons.

The most frequent heatmap patterns found in our study are patterns 3 and 5. The exons which follow pattern 5 (nothing is correlated) tend to be NBD (non-bidirectional) NATs (i.e., the complementary sequences are found on the same strand going in the same direction) [23] or are not part of a SA-pair (in the NATsDB database). Pattern 3 involves exons in which only the sense probes are totally or partially correlated with each other. The fact that only sense probes are correlated suggests that there is no transcription in the antisense gene/RNA.

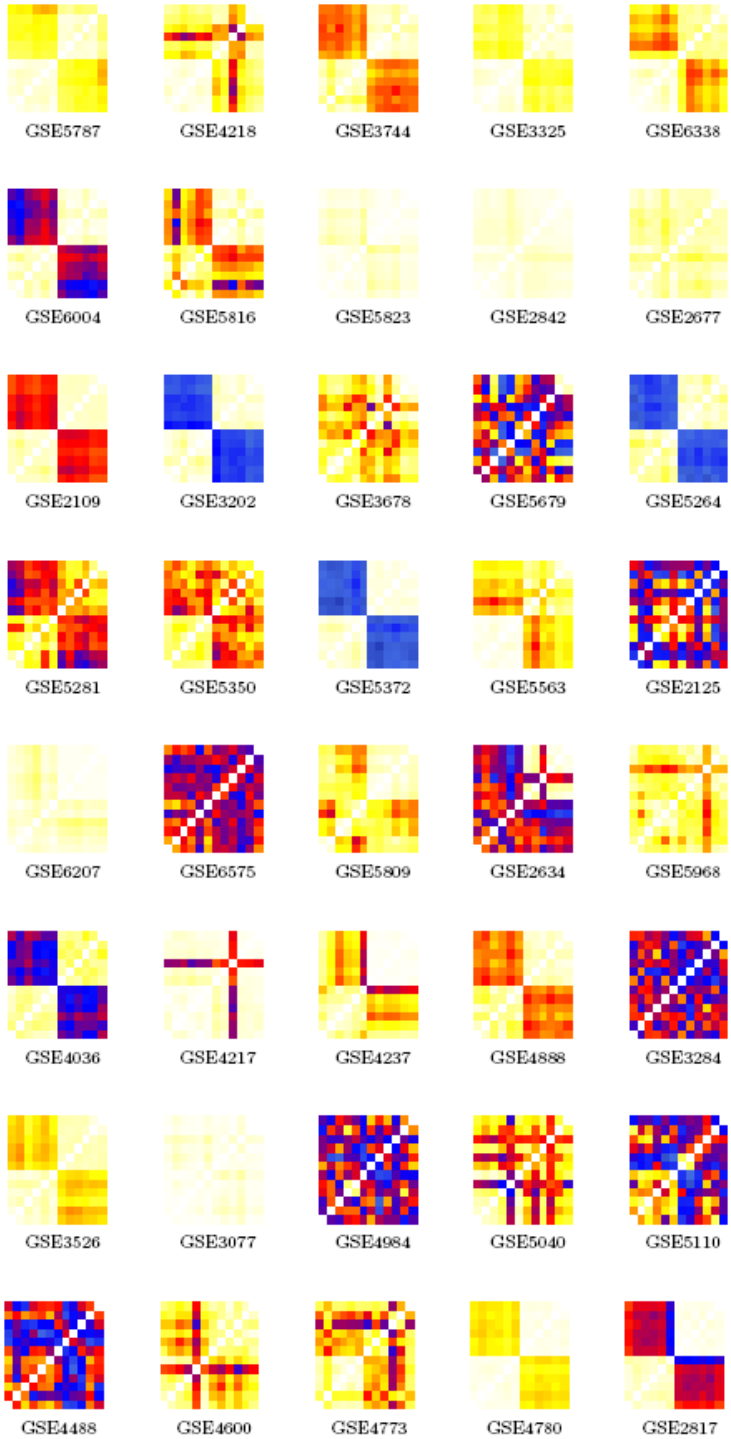
## 4. Conclusions

We find that for exons with only antisense probes mapping uniquely to them, the antisense probes are typically not correlated with each other. This suggests that the expression seen in each of these probes is not coherently detecting an antisense transcript. Just because strong signal is seen in a small fraction of the probes within an antisense probeset does not mean that the transcript is detected. Careful analysis of each probeset is required on a case by case basis.

For most of the cases where exons contain both sense and antisense probes, there is a SA-pair in which one transcript belongs to an established gene and the other is an EST or mRNA. In some cases there might be a novel antisense transcript since there are high correlations in the locus where the annotations indicate that there is not a transcript/RNA. In other cases there is a transcript in the antisense direction that does not cover the locus where the highly correlated probes align. This suggests that the transcript might be longer than the current annotation indicates. Care has to be taken when analysing the data considering that some probes might behave in different ways according to the type of experiment.

Our method for studying the correlations of sense and antisense Affymetrix probes has been shown to be useful for finding possible NATs, confirming existing ones, suggesting the existence of novel transcripts, or suggesting the reannotation of transcripts.

**Figure 6. Heatmaps with data from different experiments for exon ENSE00001452067. The heatmap representing the overall experiments is in Table 1 pattern 4 (the antisense probes are correlated, the sense probes are correlated, and the sense probes are not correlated or are weakly correlated with the antisense probes).**



| Experiment | Type | Cancer |
|---|---|---|
| GSE5787 | cervical cancer | y |
| GSE4218 | glioblastoma cancer cells in culture (different states) | y |
| GSE3744 | breast cancer | y |
| GSE3325 | prostate cancer and control | y |
| GSE6338 | peripheral T-cell lymphoma and control | y |
| GSE6004 | thyroid cancer and control | y |
| GSE5816 | cancer cell lines, treatment and control | y |
| GSE5823 | cancer cell lines, c-myc knockdown and control | y |
| GSE2842 | childhood ALL, treated and untreated and controls | y |
| GSE2677 | childhood ALL, treated and untreated and controls | y |
| GSE2109 | Cancer in different tissues | y |
| GSE3202 | non-small cell lung cancer cell lines, treatment and control | y |
| GSE3678 | PTC (papillary thyroid carcinoma) and paired controls | y |
| GSE5679 | dendritic cells (monocytes) ligand treatment and control | n |
| GSE5264 | bronchial epithelial cells, differentiation time course | n |
| GSE5281 | LCM-capture cells in Alzheimer's brain and normal controls | n |
| GSE5350 | microarray quality control project, reference human RNA, reference human brain RNA and mixtures of the two | n |
| GSE5372 | airway epithelial cells before and after injury | n |
| GSE5563 | vulvar intraepithelial neoplasia and control | n |
| GSE2125 | alveolar macrophages | n |
| GSE6207 | human liver cell line (HepG2), transfected with miR-124, and controls | n |
| GSE6575 | whole blood from children with autism and controls | n |
| GSE5809 | endometrial stromal cells, treatment and control | n |
| GSE2634 | human and non-human primate blood | n |
| GSE5968 | HepG2 cell line transfected with PGC-1 transcription factor mutants, and controls | n |
| GSE4036 | cerebellar tissues of schizophrenic patients and controls | n |
| GSE4217 | spheroid formation and recovery of human foreskin fibroblasts at ambient temperature | n |
| GSE4237 | pituitary adenomas (benign brain tumor) | n |
| GSE4888 | endometrium sampled across the cycle in 28 normo-ovulatory women | n |
| GSE3284 | blood leukocyte receiving inflammatory stimulus and controls | n |
| GSE3526 | normal post-mortem tissue samples | n |
| GSE3077 | dilution series of blood and placenta, comparison of Illumina and Affymetrix platforms | n |
| GSE4984 | monocyte derived dendritic cells, treatment and control | n |

| | | |
|---|---|---|
| GSE5040 | lymphoblast cell lines from patients with Freidriech's ataxia and normal controls, treated and untreated | n |
| GSE5110 | skeletal muscle biopsies from men before and after 48 h knee immobilization | n |
| GSE4488 | blood from affected and obligatory carriers of pituitary adenoma predisposition (PAP) and controls | n |
| GSE4600 | SH-SY5Y neuroblastoma cell line, undifferentiated, differentiated, transfected with MeCP2 decoy oligonucleotide and controls | n |
| GSE4773 | SK-N-MC neuroblastoma cell line, treatment with rotenone and controls | n |
| GSE4780 | benign (grade 1) and aggressive (grades 2 and 3) meningiomas | - |
| GSE2817 | gliomas (brain tumors) | - |

**Table 2. Description of the GSE experiments**

## Acknowledgements

OSG and MS are supported by a grant from the BBSRC (BB/E001742/1). JR is supported by a Strategic Studentship from the BBSRC (BBS/S/H/2005A/11996A). We are grateful to Dr. William Langdon for the development of a number of software tools used in this research.

## References

[1] J. Mattick, RNA regulation: a new genetics?, Nature Reviews Genetics, 5:316, 2004

[2] F. Perocchi, Z. Xu, S. Clauder-Münster and L. Steinmetz, Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D., Nucleic Acids Research, 35:e128, 2007

[3] P. Galante, D. Vidal, J. de Souza, A. Camargo and S. Souza, Sense-antisense pairs in mammals: functional and evolutionary considerations, Genome Biology, 8:R40, 2007

[4] Ø. Røsok and M. Sioud, Systematic search for natural antisense transcripts in eukaryotes., International Journal of Molecular Medicine, 15:197-203, 2005

[5] Affymetrix Inc. Design and performance of the GeneChip Human Genome U133 Plus 2.0 and Human Genome U133A 2.0 Arrays., Technical Note Part No. 701483 Rev.2., 2003

[6] T. Barrett, D. Troup, S. Wilhite, P. Ledoux, D. Rudnev, C. Evangelista, I. Kim, A. Soboleva, M. Tomashevsky and R. Edgar, NCBI GEO: mining tens of millions of expression profiles – database and tools update., Nucleic Acids Research, 35: D760-D765, 2007

[7] M. Stalteri and A. Harrison, Interpretation of multiple probe sets mapping to the same gene in Affymetrix GeneChips, BMC Bioinformatics, 8, 13, 2007

[8] S. Oeder, J. Mages, P. Flicek and R. Lang R. Uncovering information on expression of natural antisense transcripts in Affymetrix MOE430 datasets., BMC Genomics, 8:200, 2007

[9] A. Werner, G. Schmutzler, M. Carlile, C. Miles and H. Peters, Expression profiling of antisense transcripts on DNA arrays., Physiol. Genomics, 28:294, 2007

[10] O. Sanchez-Graillet, J. Rowsell, W.B. Langdon, M. Stalteri, J. Arteaga-Salas, G. Upton and A.

Harrison, Widespread existence of uncorrelated probe intensities from within the same probeset on Affymetrix GeneChips., Journal of Integrative Bioinformatics, 5(2):98, 2008

[11] J. Arteaga-Salas, H. Zuzan, W. Langdon, G. Upton and A. Harrison, An overview of image processing methods for Affymetrix GeneChips., Briefings in Bioinformatics, 9(1):25, 2008

[12] C. Wu, H. Zhao, K. Baggerly, R. Carta and L. Zhang, Short oligonucleotide probes containing G-stacks display abnormal binding affinity on Affymetrix microarrays, Bioinformatics, 23:2566-2572, 2007

[13] G. Upton, W. Langdon and A. Harrison, G-spots cause incorrect expression measurement in Affymetrix microarrays., BMC Genomics, 9:613, 2008

[14] M. Lapidot and Y. Pilpel, Genome-wide natural antisense transcription: coupling its regulation to its different regulatory mechanisms., *EMBO Rep*., 7(12):1216–1222, 2006

[15] Y. He, B. Vogelstein, V. Velculescu, N. Papadopoulos and K. Kinzler, The Antisense Transcriptomes of Human Cells. Report., Science,322(5909):1855–1857, 2008

[16] G. Upton and C. Lloyd, Oligonucleotide arrays: information from replication and spatial structure., Bioinformatics, 21:4162, 2005

[17] W. Langdon, G. Upton, R. Camargo and A. Harrison, A survey of spatial defects in Homo Sapiens Affymetrix GeneChips, Transactions on Computational Biology and Bioinformatics, TCBB.2008.108, 2008

[18] G. Upton, O. Sanchez-Graillet, J. Rowsell, J. Arteaga-Salas, N. Graham, M. Stalteri, F. Memon, S. May and A. Harrison, On the causes of outliers in Affymetrix GeneChip data., (Submitted)

[19] W. Kent, C. Sugnet, T. Furey, K. Roskin, T. Pringle, A. Zahler and D. Haussler, The Human Genome Browser at UCSC., Genome Res., 12:996-1006, 2002

[20] Affymetrix Inc. Affymetrix GeneChip IVT Array Whitepaper Collection. Transcript Assignment for NetAffx Annotations. Revision Date: 2006-3-24. Revision Version: 2.3. Transcript_Assignment_whitepaper.pdf, 2006

[21] Y. Zhang, J. Li, L. Kong, G. Gao, Q.R. Liu and L. Wei., NATsDB: Natural Antisense Transcripts DataBase., Nucleic Acids Res., 35(Database issue):D156-61, 2007

[22] R. Yelin, D. Dahary, R. Sorek, E. Levanon, O. Goldstein, A. Shoshan, A. Diber, S. Biton, Y. Tamir, R. Khosravi, S. Nemzer, E. Pinner, S. Walach, J. Bernstein, K. Savitsky, G. Rotman, Widespread occurrence of antisense transcription in the human genome. Nat Biotechnol., 21(4):371-2, 2003

[23] J. Chen, M. Sun, W. Kent, X. Huang, H. Xie, W. Wang, G. Zhou, R. Shi, J. Rowley. Over 20% of human transcripts might form sense-antisense pairs. Nucleic Acids Research,32(16):4812–20, 2004

# Genome-wide survey of rice microRNAs and microRNA–target pairs in the root of a novel auxin-resistant mutant

**Yijun Meng · Fangliang Huang · Qingyun Shi · Junjie Cao · Dijun Chen · Jinwei Zhang · Jun Ni ·Ping Wu · Ming Chen**

Y. Meng · F. Huang · D. Chen · J. Zhang · J. Ni ·P. Wu · M. Chen

State Key Laboratory of Plant Physiology and Biochemistry, College of Life Sciences, Zhejiang University, 310058 Hangzhou, People's Republic of China.

Q. Shi · J. Cao · D. Chen · M. Chen: Department of Bioinformatics, College of Life Sciences, Zhejiang University, 310058 Hangzhou, People's Republic of China.

## Abstract

Auxin is one of the central hormones in plants, and auxin response factor (ARF) is a key regulator in the early auxin response. MicroRNAs (miRNAs) play an essential role in auxin signal transduction, but knowledge remains limited about the regulatory network between miRNAs and protein-coding genes (e.g. ARFs) involved in auxin signalling. In this study, we used a novel auxin-resistant rice mutant with plethoric root defects to investigate the miRNA expression patterns using microarray analysis. A number of miRNAs showed reduced auxin sensitivity in the mutant compared with the wild type, consistent with the auxinresistant phenotype of the mutant. Four miRNAs with signiWcantly altered expression patterns in the mutant were further conWrmed by Northern blot, which supported our microarray data. Clustering analysis revealed some novel auxin-sensitive miRNAs in roots. Analysis of miRNA duplication and expression patterns suggested the evolutionary conservation between miRNAs and protein-coding genes. MiRNA promoter analysis suggested the possibility that most plant miRNAs might share the similar transcriptional mechanisms with other non-plant eukaryotic genes transcribed by RNA polymerase II. Auxin response elements were proved to be more frequently present in auxin-related miRNA promoters. Comparative analysis of miRNA and protein-coding gene expression datasets uncovered many reciprocally expressed miRNA–target pairs, which could provide some hints for miRNA downstream analysis. Based on these Wndings, we also proposed a feedback circuit between miRNA(s) and ARF(s). The results presented here could serve as the basis for further in-depth studies of plant miRNAs involved in auxin signalling.

**Keywords:** Auxin response element · Auxin response factor · Auxin signaling · Microarray · MicroRNA–target pairs · Rice root

# Gene expression profile displaying up-regulation and down-regulation of amino acid nutritional metabolic modules

**Jing Li** [1]

[1] Tianjin University of Science and Technology, Tianjin, China

## Abstract

The raw gene expression data of yeast were excavated by SMD (Stanford Microarray Database) and amino acid metabolic pathway from KEGG (Kyoto Encyclopedia of Genes and Genomes). The methionine and cysteine nutritional metabolic modules were selected and analyzed, expecting to establish the relationship between the modules, and find out the important genes, which made predominant contribution for the up-regulation and down-regulation of modules, trying to identify the gene expression differences and synergies.

The results show that four common genes exist between these two modules, comprising YAL012W (4.4.1.1), YFR055W (4.4.1.8), YJR130C (2.5.1.48) and YLR303W (2.5.1.49). For methionine modules, the five important genes were YAL012W (4.4.1.1), YDR502C (2.5.1.6), YER043C (3.3.1.1), YGR155W (4.2.1.22) and YLR303W (2.5.1.49), playing the decisive role in the gross of gene expression, while for cysteine modules, YAL012W (4.4.1.1), YLR303W (2.5.1.49), YNL247W (6.1.1.16) and YCL064C (4.3.1.17) were important. By common genes and important genes, maybe the cooperative up-regulation relationship exists between these two modules. Our investigation also found that certain crucial metabolite can be accumulated by controlling the important genes in the metabolic process.

# Identifying the impact of G-Quadruplexes on Affymetrix exon arrays using cloud computing

**Farhat N. Memon, Olivia Sanchez-Graillet, Graham J. G. Upton, Anne M. Owen and Andrew P. Harrison***

Departments of Mathematical Sciences and Biological Sciences, University of Essex, Wivenhoe Park, Colchester, Essex, CO4 3SQ, United Kingdom

*Corresponding author:harry@essex.ac.uk

http://bioinformatics.essex.ac.uk/

## Summary

A tetramer quadruplex structure is formed by four parallel strands of DNA/ RNA containing runs of guanine. These quadruplexes are able to form because guanine can Hoogsteen hydrogen bond to other guanines, and a tetrad of guanines can form a stable arrangement. Recently we have discovered that probes on Affymetrix GeneChips that contain runs of guanine do not measure gene expression reliably. We associate this finding with the likelihood that quadruplexes are forming on the surface of GeneChips.

Our original discovery was made using 3′ arrays. We have now extended our analysis to look at Affymetrix Exon arrays. In order to cope with the rapidly expanding size of Exon array datasets in the public domain, we are exploring the use of cloud computing. This is a recently introduced high-performance solution that takes advantage of the computational infrastructure of large organisations such as Amazon and Google.

We expect that cloud computing will become widely adopted because it enables bioinformaticians to avoid capital expenditure on expensive computing resources and to only pay a cloud computing provider for what is used. Moreover, as well as financial efficiency, cloud computing is an ecologically-friendly technology, it enables efficient data-sharing and we expect it to be faster for development purposes. Here we propose the advantageous use of cloud computing to perform a large data-mining analysis of public domain Exon arrays.

## 1. Introduction

### 1.1 G-Quadruplex

The binding of guanine to cytosine and adenine to thymine usually occurs through the famous Watson-Crick interactions in double stranded DNA. However, in single-stranded DNA sequences, a guanine can bind to another guanine through a Hoogsteen hydrogen bond. A tetrad of guanines

can then form a loop, in which each guanine can bind to two other guanines at 90 degrees (similar to the edges of a square). Indeed, this occurs throughout a genome because single-stranded DNA sequences that have frequent occurrences of guanine runs are capable of forming four-stranded structures, known as G-Quadruplexes, G-tetrads, or G4 DNA [1].

In a single strand of DNA, a G-quadruplex consists of four runs of guanines (called the stems of G-quadruplex) with three loops in between the four stems. GGGAGCGGGTTGACGGGAAGGG, a segment of single stranded DNA sequence for instance, can form a G-quadruplex in which the four sets of underlined Gs represent four stems of guanine and the nucleotides in between these stems create loops. Both the stem size and loop size have biological significance. As the stem size increases, a G-quadruplex becomes more stable; whereas an increase in loop size weakens the stability of quadruplexes [1].

[2] demonstrated that G-rich nucleic acid sequences can adopt quadruplex structures that are stabilised by the presence of G-quartets (Figure 1). A G-quadruplex may not necessarily form through a single nucleic acid sequence; sometimes two or four parallel nucleic acid sequences may form a G-quadruplex collectively.

Figure 1(B) illustrates a number of different topologies for G-quadruplexes. For example, the Monomer Chair and Monomer Basket show G-quadruplexes that are formed in a single nucleic acid sequence, whilst the Dimer Chair and Dimer Basket illustrate that two G-rich nucleic acid strands are capable of forming a G-quadruplex. Indeed a tetramer can result from four parallel strands forming a G-quadruplex. A quadruplex that forms through more than one sequence falls into the category of Intermolecular Quadruplex structures. Thus, the Dimer and Tetramer are both examples of Intermolecular Quadruplex structures [3]. Keeping tetramer quadruplex structures in mind, we are investigating the implications for microarrays that are used to analyse genomic data.

## 1.2 Affymetrix GeneChips enable whole-transcriptome studies of the Genome

The production of messenger RNA reflects the activity level of a gene, and many biologists are interested in the conditions in which a specific gene is turned on or turned off. Microarray technology allows the simultaneous study of many genes in parallel, providing a snapshot of how a genome is operating. A microarray usually consists of a glass slide, containing a 2D array of an orderly arrangement of fragments of single-stranded DNA, referred to as probes, that represent the genes of an organism. Each DNA fragment representing a gene is assigned a specific location on the array. A fluorescently labeled DNA or RNA (target sequence) will stick through hybridisation to its complementary probe. The genes that are active are detected through measuring the light

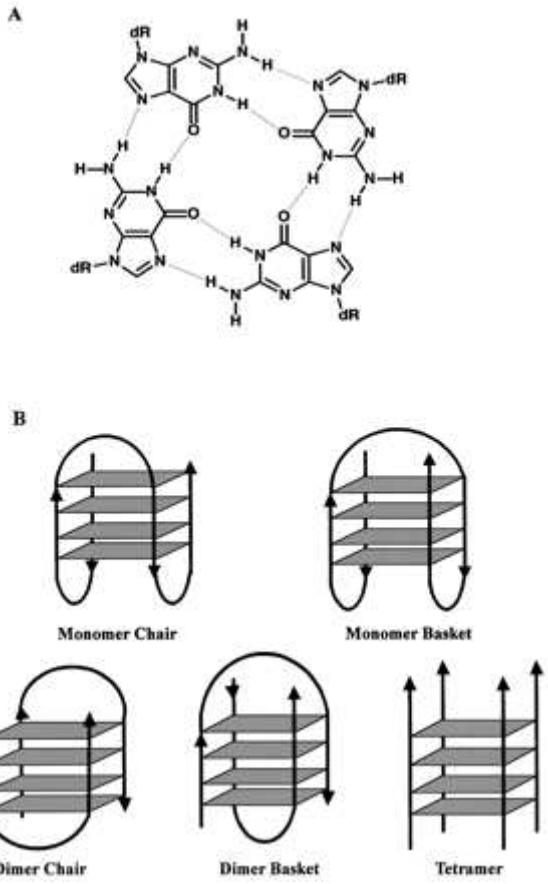



**Figure 1: Schematic presentation of G-quartet structures. (A) G-quartet. (B) Different layouts/topologies and loop orientation of quadruplexes (Source: http://nar.oxfordjournals.org/cgi/content/full/31/8/2097)**

from the excited fluorescence of the labelled DNA or RNA.

There are many types of microarray that are commercially available. However, in this study we focus on the Affymetrix GeneChip, a high density oligonucleotide array. An Affymetrix GeneChip consists of 25-mer oligonucleotide probes which have been synthesised in-situ through photolithographical methods. Each gene is represented by several probes, collectively called a probe set. The size of a GeneChip covered by an array of probes is 1.28cm×1.28cm. Due to improvements in array manufacturing technology, the number of distinct probe sequences within this area has increased over time, with some of the latest designs having over 5 million different cells, each containing many thousands of copies of a distinctive probe sequence. Figure 2 shows the basic construction of an Affymetrix GeneChip.

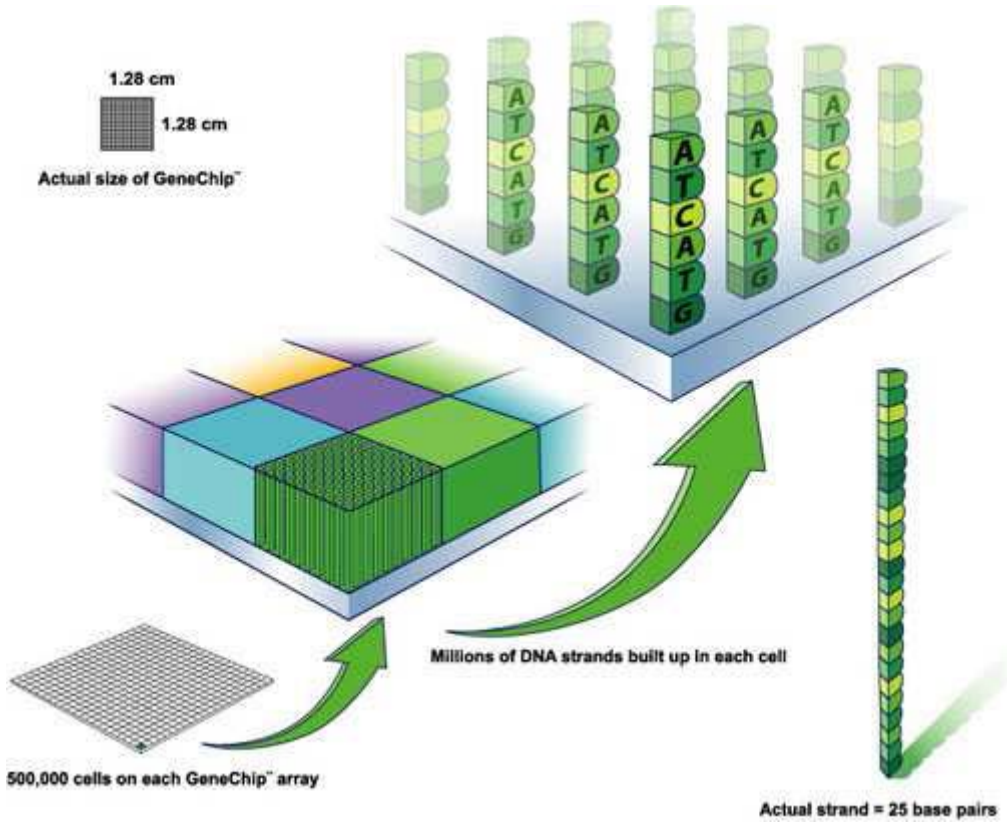



**Figure 2: Basic Structure of Affymetrix GeneChip**

**(Source: http://electronicdesign.com/Files/29/10603/Figure 06.jpg)**

Affymetrix has released GeneChips for most major model organisms. One of their most widely used designs is known as a 3′ **Array**, because most probes are selected towards the 3′ region of a gene. Some cross-hybridization to other transcripts can occur even though the probes are selected to ideally avoid such cross-hybridisation. This led to the Affymetrix 3′ design including, for each gene-specific probe, a probe that is identical in sequence except for a complementary base at its centre (13th base). These mismatch (MM) probes are placed immediately adjacent to their perfect match (PM) probes. In this way, each gene is represented by 22 different probes (11 Perfect match probes and 11 Mismatch probes). The design philosophy is that 11 signal intensities measure a particular gene fragment plus a sequence-specific background; while 11 mismatch probes report a close approximation to the sequence-specific background. The intention is that subtraction of the MM signal from the PM signal will result in a measure of a genes expression, though strategies are required to deal with the not infrequent cases where it is the MM signal that is the greater. The multiple measurements of gene expression are collated into one composite expression measure.

Affymetrix has introduced another chip design, the **Exon array**, which is designed to investigate exon-level expression. They have smaller probesets and these probesets detect exons across the gene, not just towards the 3′ end. Mismatch probes do not exist in Exon arrays. There are

approximately four probes per exon and roughly 40 probes per gene. Exon arrays enable ”exon-level” analysis, which allows us to distinguish between different isoforms of a gene, and to detect specific alterations in exon usage, some of which may play a central role in disease mechanism and etiology. Exon arrays also allow ”gene-level” expression analysis, that summarises multiple probes on different exons into an expression value of all transcripts from the same gene.

## 1.3 Identifying problems in GeneChip data

Affymetrix report that over 10,000 published papers have used or described their technology. As each typical study comprises multiple GeneChips, there are now many tens of thousands of GeneChips in the public domain that are now available for meta-analysis. Although the power of GeneChip technology is widely recognised, many open questions remain about the appropriate analysis of GeneChip data. This is particularly true now that we have the opportunity to mine large GeneChip datasets in order to discover novel signatures associated with diseases.

It is expected that if a particular gene is highly expressed then all the probes in a probe set representing that gene will be consistent in showing the presence of that particular gene. However, [4] found that probes containing runs of guanine show abnormal affinities; they tend to have increased cross-hybridisation signals and reduced target-specific hybridisation signals, presumably due to multiplex binding forming G-quartet structures. We recently confirmed that probes having a sequence of four or more guanines, which we termed G-spots, typically have poor correlation with other probes in their probeset [5]. However, we went further in discovering that the intensities reported from these G-spot probes are correlated with each other. We suggested that the intensities reported by these probes should not be used in the calculation of gene expression values and these G-spot probes should not be included within future array designs.

We have proposed that structures closely resemblingG-quadruplexes are forming on GeneChips, and this is why probes containing runs of guanine are not fit for purpose [5]. Neighbouring probes with the same sequence can come into physical contact on a GeneChip. For most sequences which lack complementary sections they will not be expected to hybridise to each other. But for probes containing runs of guanine, it is possible that a stack of Hoogsteen hydrogen bonds can occur [5]. A grouping of four probes can then form a stable tetrad at each guanine, and the resulting stack of tetrads forms a G-quadruplex. In such a G-quadruplex the guanines face inwards and are not available to hybridise to target sequences. But in the interpretation of [5], the formation of a G-quadruplex frees up space in the immediate surroundings of the four probes. This reduction in probe density increases the rate, and strength, of hybridization between target RNA sequences

containing runs of cytosines and the neighbouring probes, all of which contain runs of guanine. This results in cross-hybridisation dominating for these probes, and the G-spot probes not detecting the target RNA for which they were chosen. This accounts for why the G-spot sequences are poorly correlated with other probes that are able to measure target RNA reliably.

We aim to focus on Exon arrays in order to examine whether the problems found in 3′ arrays, specially the misbehaviour of G-spot probes, also affect Exon arrays. Although we have only used Human Exon 1.0 ST V2 arrays in our study, our results should apply to any Affymetrix Exon array.

## 2. Method

Section 2.1 explains our approach to analyse Affymetrix Exon arrays and section 2.2 describes cloud computing, a high-performance technology we have adopted for this study.

### 2.1 Our approach

We have designed a pipeline to analyse Affymetrix exon arrays, downloaded from NCBI's Gene Expression Omnibus (GEO). Our pipeline processes CEL files, the data files that contain average fluorescence intensity of each probe in the array. The pipeline includes uniquemapping of probes to exons, calibration processes for quality control analysis, and creation of heatmaps for all Ensembl-defined exons. We have used a similar methodology to that of our previous work on 3′ arrays [6], but necessarily ignore the contribution from MM probes (which are missing on Exon arrays).

#### 2.1.1 Unique probe mappings

Rather than using information from all the probes on an array, we are selective and only use probes which are uniquely mapping to an exon, in order to reduce the effects of crosshybridisation. We have described previously [6] that we consider a probe to be uniquely mapping to an exon if it completely aligns with 25 bases to only one exon and to any of its synonymous exons (i.e. exons located on the same genomic region, although they have different Ensembl identifiers). Moreover, we insist that the alignment of completely 25 bases should only be at one place on the exon. Furthermore, the probes should not map partially or totally (20 or more bases) to any other exon. [6] provides more details about our way of establishing unique mappings.

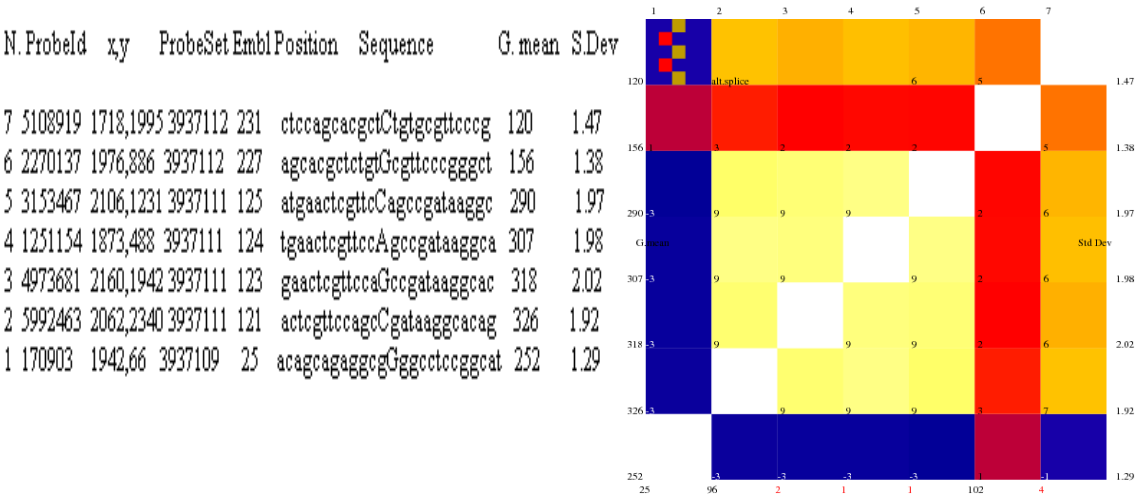
#### 2.1.2 Calibration process

[7] reported that many gene expression measures more than doubled when they introduced typical levels of spatial noise seen in raw GeneChip data. Thus an important issue in the analysis of

microarray is quality control, and we apply a calibration process that includes normalization and detection of spatial flaws in all CEL files [8].

### 2.1.3 Generation of exon Heatmaps

The last phase is to generate heatmaps of all Ensembl exons. We are currently using Ensembl Release 48. We assume that if an exon is within transcripts in the sample then all the probes detecting that exon should, ideally, respond in the same way, i.e. the fluorescent signals from these probes should be correlated. We examine the correlation coefficient value between the pairs of probes using the processed CEL files and then generate the heatmap for quick visualization and easy analysis. We use a heatmap as a graphical representation of the correlations between the levels of expression of different probes across a number of samples. Each cell in a heatmap is colour coded according to the cell's correlation coefficient value. A bright white to yellow cell indicates highly correlated pairs of measurements while a dark blue cell represents a poor/low correlation. The cells on the diagonal always correspond to correlation coefficients of one, because we are comparing a probe with itself.



| N. | ProbeId | x,y | ProbeSet | Embl | Position | Sequence | G. mean | S.Dev |
|---|---|---|---|---|---|---|---|---|
| 7 | 5108919 | 1718,1995 | 3937112 | 231 | | ctccagcacgctCtgtgcgttcccg | 120 | 1.47 |
| 6 | 2270137 | 1976,886 | 3937112 | 227 | | agcacgctctgtGcgttcccgggct | 156 | 1.38 |
| 5 | 3153467 | 2106,1231 | 3937111 | 125 | | atgaactcgttcCagccgataaggc | 290 | 1.97 |
| 4 | 1251154 | 1873,488 | 3937111 | 124 | | tgaactcgttccAgccgataaggca | 307 | 1.98 |
| 3 | 4973681 | 2160,1942 | 3937111 | 123 | | gaactcgttccaGccgataaggcac | 318 | 2.02 |
| 2 | 5992463 | 2062,2340 | 3937111 | 121 | | actcgttccagcCgataaggcacag | 326 | 1.92 |
| 1 | 170903 | 1942,66 | 3937109 | 25 | | acagcagaggcgGggcctccggcat | 252 | 1.29 |

**Figure 3: Collection of all probes that represent Ensembl exon ENSE00000330846. First column contains sequence number in which these probes appear in heatmap. Next column have probe IDs followed by X andY positions, probe set ID, their position in Ensembl exon, sequence of nucleotide, geometric mean, and standard deviation. The values inside each cell of heatmap represent the rounded value of (correlation x 10).**

In our previous work on 3′ arrays, G-spot probes are poorly correlated with other members of their probesets but are highly correlated with each other [5]. We are expecting to see the same behaviour of G-spot probes in Exon arrays. This is suggested by Figure 3, which illustrates a heatmap representing the correlations between pairs of seven probes uniquely mapping to the Ensembl exon ENSE00000330846. Probes 2, 3, 4, and 5 are highly correlated relatively bright

cells. However, probes 1 and 6 are behaving as outliers and are poorly correlated with other probes, but partially correlated with each other. Probe 1 contains a sequence of four guanines whilst probe 6 contains three consecutive guanines.

## 2.2 Cloud Computing

There was a time when companies used their own generators to produce electricity for running their factories or plants. This usually required a large capital expenditure when purchasing dynamos, and also required maintenance costs. This business model was quickly dropped and companies started to buy electricity from a utility supplier of electricity, because it proved to be cheaper and easier to buy electricity as a commodity without worrying about maintenance and updating equipment.

It is becoming increasingly apparent that computing is performing a similar transition at present, with computing, and other information technologies, being sold as a commodity which can be purchased from utility suppliers. The availability of significant computational opportunities is being provided by several companies, such as Amazon, Google, and Microsoft. They provide high-performance solutions that enable users to utilise their computational infrastructure, and to only pay for the resources used. The web services platform of these organisations are suitable for user groups of any size, including individuals. The "Cloud computing" concept is very simple: the computing resources are located somewhere (not in your office/ computer room) and you will connect to them and use them according to your requirement.

Cloud computing enables bioinformaticians to avoid capital expenditure on computers which rapidly decrease in value. It also minimises the time and effort required to maintain large clusters and removes the requirement for space and cooling systems needed to house the computers. We expect that cloud computing will be widely adopted by bioinformaticians in the near future. Furthermore, cloud computing is a green technology, as the carbon footprint of one large datacentre is much less than that of many groups housing their own inefficient computational infrastructure. Moreover, many users can easily gain access to shared data on the cloud, and don't have to worry about the inconvenience of managing, and paying for, lots of data transfer.

We have begun to explore the use of cloud computing through Amazon's platform, though Amazon does not require any long term commitment of its users. They provide us with the flexibility to choose any development platform or programming model that is most appropriate for the problems to be solved. Amazon Web Services (AWS) provide different services which includes Amazon Elastic Compute Cloud (Amazon EC2), Amazon Simple DB, Amazon Simple

Storage Service (Amazon S3), Amazon CloudFront, Amazon Simple Queue Service (SQS), Amazon Elastic MapReduce, AWS Premium Support.

AWS is already hosting some public data sets, including Ensembl and some of the NCBI databases [9]. We expect that Ensembl and NCBI will continue their practice of uploading all their data, as it grows beyond the petabyte scale[10]. This is beneficial to our work, as we already use several of these databases, and we do not need to cover the costs of uploading this data.

To get high computing power, we use Amazon Elastic Compute Cloud (EC2) that provides an environment to run virtual servers on demand, Amazon Simple Storage Service (S3) is used to store our own data; whilst Amazon's public data sets enable us to use some of the Ensembl and NCBI data freely. To use Amazon EC2 service, an Amazon Machine Image (AMI) is required. An AMI is an encrypted machine image (a file) that contains all the information required to boot an instance of our software and it stores in Amazon Simple Storage Service (S3). One can either create its own AMI or use public AMIs (public AMI can be used as it is or with some modification). The next step, bundling an AMI, performs certain tasks related to confidentiality and authentication which include the compression of AMI in order to minimise bandwidth usage and storage requirements, encryption of the AMI, breaking down the encrypted AMI into smaller chunks to upload, and creation of a file that contains the details about the image's small chunks with their checksum values. Then one or more instances can be launched for that AMI and finally we administer these instances as we do on our server. The block diagram to show the flow of EC2 is depicted in Figure 4.

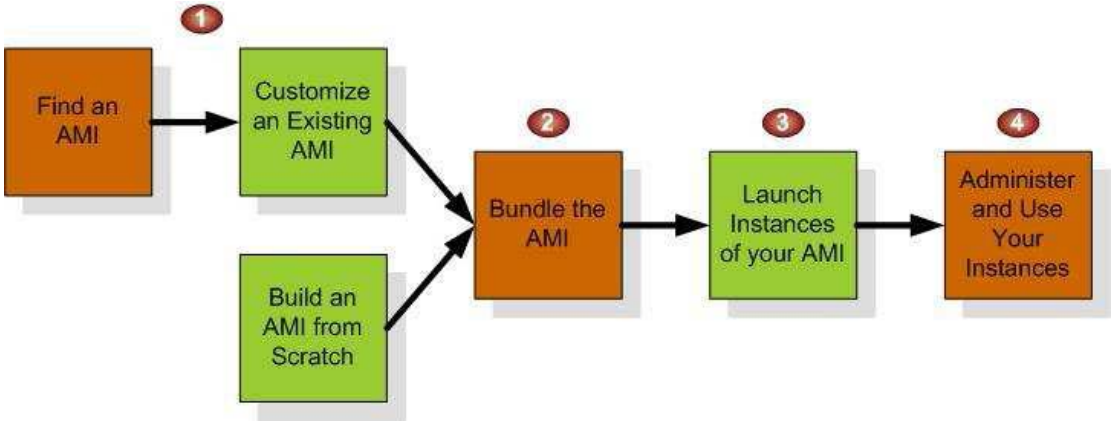



**Figure 4: Amazon Elastic Cloud Compute (EC2) Flow (Source: http://aws.amazon.com/)**

## 3. Conclusion and Future Work

Our previous study concluded that G-spot probes in 3′ arrays usually have poor correlations with other members of their probe set but are typically highly correlated with each other. For this

purpose, we have already generated the heatmaps for 3' arrays that include eleven human arrays, twelve mouse arrays, as well as some plant arrays, all of which can be obtained via http://bioinformatics.essex.ac.uk/.

We have now generated heatmaps for all the Ensembl exons to analyse Human Exon array ST V2 in order to find misbehaving probes and their causes. The Exon array ST V2 occupies more space and resources, principally because 3′ arrays contain far fewer examples of exons than do Exon arrays. The heatmaps for one human Exon array design occupys 8.6 GB whilst the heatmaps for eleven human 3′ array designs require 5.5 GB.

We have found examples of exons on the Exon array that contain more than 100 uniquely mapping probes but for which the majority of the probes are not correlated with each other. We are currently investigating the causes of this unexpected behaviour. We have also collated probes contain runs of guanines, and are in the process of using the cloud to derive correlations among these probes, as well as the correlations between these probes and the other members of their respective probesets. This will enable us to directly compare the behaviour of G-spot probes in Exon arrays with those in 3′ arrays. We expect to find similar behaviour within the different designs.

We require an ever-increasing amount of computational resources in order to carry mining of large biological datasets. It is our opinion that we, and other groups with similar interests, will increasingly turn towards cloud computing solutions. To our knowledge, this paper is the first study of using cloud computing for performing microarray analysis.

## References

[1] J.L. Huppert and S. Balasubramanian. Prevalence of quadruplexes in the human genome. *Nucleic Acids Research*, 33(9):2908-2916, 2005
*Online Journal*: http://nar.oxfordjournals.org/cgi/content/abstract/33/9/2908

[2] V. Dapic, V. Abdomerovic, R. Marrington, J. Peberdy, A. Rodger, J.O. Trent and P.J. Bates. Biophysical and biological properties of quadruplex oligodeoxyribonucleotides. *Nucleic Acids Research*, 31(8):2097-2107, 2003
*Online Journal*: http://nar.oxfordjournals.org/cgi/content/abstract/31/8/2097

[3] V. K. Yadav, J. K. Abraham, P. Mani, R. Kulshrestha and S. chowdhury. QuadBase: genome-wide database of G4 DNA - occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes. *Nucleic Acids Research*, 36(Database issue):D381-D385, 2008

*Online Journal*: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2238983

[4] C. Wu, H. Zhao, K. Baggerly, R. Carta and L. Zhang. Short Oligonucleotide probes containing G-stacks display abnormal binding affinity on Affymetrix microarrays. *Bioinformatics*, 23(19):2566-2572, 2007

*Online Journal*:

http://bioinformatics.oxfordjournals.org/cgi/content/abstract/23/19/2566?ck=nck

[5] G. J. G. Upton, W. B. Langdon and A. P. Harrison. G-spots cause incorrect expression measurement in Affymetrix microarrays. *BMC Genomics*, 9:613, 2008

*Online Journal*: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2628396

[6] O. Sanchez-Graillet, J. Rowsell, W. B. Langdon, M. Stalteri, J. M. Arteaga-Salas, G. J. G. Upton and A. P. Harrison.Widespread existence of uncorrelated probe intensities from within the same probeset on Affymetrix GeneChips. *Journal of Integrative Bioinformatics*, 5(2):98, 2008

*Online Journal*: http://journal.imbio.de/index.php?paper_id=98

[7] M. Reimers and J. N. Weinstein. Quality assessment of microarrays: Visualization of spatial artifacts and quantitation of regional biases. *BMC Bioinformatics*, 6:166, 2005

*Online Journal*: http://www.ncbi.nlm.nih.gov/pubmed/15992406

[8] W. B. Langdon, G. J. G. Upton, R. S. Camargo and A. P. Harrison. A Survey of Spatial Defects in Homo Sapiens Affymetrix GeneChips. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2008

*Online Journal*: http://www2.computer.org/portal/web/csdl/doi/10.1109/TCBB.2008.108

[9] Amazon Web Services. http://aws.amazon.com/publicdatasets/

[10] A. Bateman and M. Wood. Cloud Computing. *Bioinformatics*, 25(12):1475, 2009

*Online Journal*: http://bioinformatics.oxfordjournals.org/cgi/reprint/25/12/1475

# Centroid extensions of de novo motif detection algorithms

**Hendrik Mehlhorn [1], Ivo Grosse [2]**

Leibniz Institute of Plant Genetics and Crop Plant Research Gatersleben, Germany

Martin–Luther–Universitaet Halle–Wittenberg, Germany

## Abstract

The detection of *cis-regulatory modules* (*CRM*s) is of central importance for many branches of molecular biology . One of the most powerful algorithms solving this problem is the *Centroid Gibbs Sampler* which is based on a realistic promoter model, information from *phylogenetic footprinting*, and the concept of predicting the centroid solution rather than the *maximum a posteriori* solution. To evaluate the idea of *centroids* we apply it to other stochastic algorithms. This and the variation of further computational properties leads to nine novel algorithms for the detection of *CRM*s. We compare the prediction quality of these algorithms to that of four common algorithms by examining different data sets. We find that the *centroid* approach improves the prediction of the studied algorithms. This is evidence to suggest that the *centroid* approach might be useful in other areas of modern genomics and epigenomics.

# Simulation of the phase decision processes of clock gene expressions by promoter transcriptional regulations

**Kenji Miyamoto** [1]

[1] Biopathway Analysis Center, Faculty of Science, Yamaguchi University, Yamaguchi, Japan

## Abstract

So far, we have analyzed mammalian circadian clock systems by simulating and modeling these underlying clock gene regulation with hybrid fnctional Petri net (HFPN). The established HFPN model consists of six clock genes Per, Cry, Rev-Erb, Ror, Bmal, and Clock. In this model, BMAL/CLOCK protein activates gene transcription of Per, Cry, Rev-Erb, and Ror, but PER/CRY protein inhibit these genes. Therefore, the coupling of the positive regulation by BMAL/CLOCK and the negative regulation by PER/CRY on these genes produce self-sustained oscillations of mammalian circadian rhythms. However, the peaks of these gene expressions obtained from the constructed HFPN model are different to the results of the biological experiment. This means that phase relation of gene expressions cannot be determined only by the coupling of negative and positive regulations.

The gene being controlled by three kinds of promoters, E-box, D-box, and RRE has been expressed at different time in the morning, daytime, and nighttime, respectively. So we thought that the gene expression phases are decided by the combination of promoters that switches gene transcriptional activity on and off.

Here, we investigated the transcriptional regulation by promoters that influence to the phases of clock gene expressions. First, we incorporated these three type of promoters into the established HFPN model and simulate this HFPN model with Cell Illustrator. As a result, we confirmed that gene transcriptions regulated by the same promoter peaks at the same time. Furthermore, we examined the phases of clock gene oscillations by changing the stabilities between transcription factors and promoters. Then, we revealed that the increase of the stabilities between MAL/CLOCK and E-box produces the correct phase relations among gene oscillations in biological experiment.

# The LAILAPS search engine: relevance ranking in life science database

**Matthias Lange** [1]

[1] Leibniz Institute of Plant Genetics and Crop Plant Research, Germany

## Abstract

With the growing data available in life science databases, search engines and retrieval systems are common tools at the life science desktop. Hereby, not the number of query results for a data query matters, but the relevance does. Consequently, the extraction of information and ranking of thousands database entries should be addressed by algorithms and tools for bioinformatics.

In this paper, we present the LAILAPS search engine for life science databases. The concept is a combination of an intuitive and slim Web user interface on top of a machine learning (ML) ranking system. Using an inverse text index, query terms are searched in life science databases. With a set of features, extracted from each database hit, ML algorithms compute user specific relevance scores. Using expert knowledge as training data for the ML ranking system, a reliable relevance ranking of database hits have been developed. The manual expert training of the ML-engine is complemented by a user feedback and interaction profiling system for logged-in users, that automatically re-trains the ML ranking engine based on the user behavior and application background.

LAILAPS shows that a combination of a easy to use Web fronted, text indexing, feature extraction, artificial intelligence, user profiling, phrase searching, and synonym based query expansion is a useful approach for information retrieval in life science. LAILAPS is public available for SWISSPROT data at \\url{http://lailaps.ipk-gatersleben.de}.

# LIMS lite: a system for management and search of primary lab data

**Matthias Klappers** [1]

[1] Leibniz Institute of Plant Genetics and Crop Plant Research, Germany

## Abstract

Nowadays modern scientific instituts produce a high amount of primary data in consequence of using high throughput technology (htpt). This primary data is processed and analyzed by bioinformatic software tools. The permanent and centralized storage of the primary data becomes a very important task. Software methods and tools are always in constant development and can change during the period of a project. Also the primary data are the first step of analysis and with that create a part of the whole process. These data must be kept to complete the chain of scientific evidence.

Currently all primary data is stored on distributed and heterogenous systems without further description.

We developed a tool called mph {LIMS lite}, whose primary goal is the storing, managing and searching of primary lab data. It is an easy to use web based user interface which allows storage of the data with additional freetext and controlled vocabulary tagging in a relational database in a hierarchical structure. This structure is based on the structural basis of the workflows in a scientific institute.

# RiboNucleic acid tertiary structure comparison using graph theory

**Zain-ul-Abdin Khuhro*, Farhat N. Memon and Andrew P. Harrison**

Department of Mathematical Sciences, University of Essex, Wivenhoe Park, Colchester, Essex, United Kingdom, CO4 3SQ

*Corresponding author : zkhuhr@essex.ac.uk

http://bioinformatics.essex.ac.uk/

## Summary

The study of Ribose Nucleic Acid (RNA) has implications for many diseases, as well as deepening our understanding of evolution. Groups of RNA structures take similar forms, some of which indicate shared function. With the aim of addressing questions relating to RNA structural diversity, we have developed CORONATION (Comparison Of RibOse Nucleic Acid Tertiary's Involving Overlapping Networks), an algorithm which rapidly finds similarities between RNA tertiary structures. CORONATION works by creating graphtheoretic descriptions of geometrical patterns within each RNA structure. It identifies 3D arrangements of bases shared between different structures through searching for cliques in a graph of the overlap between the two structures' graphs. CORONATION is efficient, fast and its performance has been successfully tested on many RNA structures. This is beneficial for structure comparison and for gaining insight into structure-function relationships.

## 1. Introduction

Comparing motifs within structures that have common properties is a frequent problem in RNA research. Such RNA structures include the primary structure, the linear sequence of nucleotide bases. The secondary structure of RNA is usually defined in terms of hydrogen bonds between bases and there has been a significant amount of bioinformatics research directed at the RNA secondary structure prediction and comparison problem [1, 2]. However, the functional form of RNA molecules frequently require a specific tertiary structure. This paper describes a new approach to comparing tertiary similarities between known RNA structures.

A variety of computational procedures have already been applied to the problem of tertiary structural comparisons. In PRIMOS [3] each nucleotide is represented by two pseudo-torsion angles ($\eta$ and $\theta$) and a whole structure as a sequence of $\eta$-$\theta$ values, which was called an RNA worm. PRIMOS detects structural differences between molecules with the same number of nucleotides by comparing their worms. It is mainly suitable for examining different conformations

of the same molecule and searching structures for a specified continuous motif. Another method NASSAM [4] is a graph theoretic method that searches nucleic acid structures for a given 3D pattern. It represents each base by two vectors and a whole nucleic acid structure as a labelled graph, in which the nodes are the vector representations of the bases and the edges are the distances between them. After representation of the given 3D pattern and the structure to be searched as graph, the exponential-time Ullman algorithm for subgraph isomorphism is used to x approach to comparing tertiary similarities between known RNA structures.

Both PRIMOS and NASSAM are useful motif searching methods, but they are unsuitable for detecting new motifs that are not specified in advance. To partially overcome this limitation, the COMPADRES method [5] employs PRIMOS's worm representation and searches structures for new motifs consisting of at least five continuous nucleotides. Specially, for a given dataset of RNA structures, COMPADRES generates an RNA worm representation for the entire dataset by concatenating the worms of all chains. The resulting worm is then plotted against itself so that the $\eta$ and $\theta$ values of each nucleotide are compared to those of the other nucleotide. Finally, the plot is scanned and all diagonals with at least five continuous matches of nucleotides are considered as new motif candidates. The limitation for their method is that discovered motifs are restricted to be sequential.

Databases are now being developed in order to aid the structural classification of RNA. Within SCOR [6], RNAs are dissected into structural elements and all examples of structures containing specific structural elements may be easily traced. A new classification database of RNA tertiary structures, DARTS [7], clusters RNA structures mainly on the basis of their global resemblances. The classification of structures is hierarchical, and helps to reveal the current structural repertoire of RNA, exposing common global folds and local tertiary motifs. DARTS also enables comparison between newly determined RNA structures and structures previously classified.

This present paper is focused on comparing tertiary similarities between known RNA structures. A particularly important application is the ability to compare the largest common geometric arrangement of nucleotides in different RNA structures. We describe an algorithm, CORONATION, aimed at searching for 3D patterns of bases, taking advantage of existing databases of RNA structures. In particular, we compare structures using a graph-theoretic description of their 3D geometry. CORONATION is based on a similar methodology to that of GRATH [8], which built on earlier algorithms to classify structures [4]. The advantage of this methodology is that it works without initial assumption about what structural motifs might be in common, and it does not require the similarities to be sequential.

## 2. Method

### 2.1 Transforming a RNA structure into a graph

A graph is a set of objects called points, nodes or vertices, connected by links called lines or edges. A graph describes both the layout of the network and how parts of the network interact with each other. [8, 9] described how biomolecular structures can be transformed into graphs. In these methods [8, 9, 10], a vector passing along the axis of a protein secondary structure is assigned to a node, and the geometric relationships between pairs of vectors, that is the distance and angles, define the edges of the graph.

CORONATION transforms a RNA tertiary structure into a graph through assigning nodes to be a vector through each nucleotide. A nucleotide in RNA (Figure 1) is built from three basic components: ribose, phosphate, one of the four nitrogenous bases, with the bases attached to the C1′ atom of each ribose via a glycosidic linkage. These nitrogenous bases are either purine derivatives (guanine and adenine) or pyrimidine derivatives (cytosine and uracil), and so differ in their atomic components. However, the sugar-phosphate backbone of RNA always contains main atoms (P,O5′,C5′,C4′,C3′,O3′), and these can be used to define a frame of reference for each nucleotide. In CORONATION, we assign a vector to run from the O5′ atom to the C5′ atom in each nucleotide. The *x*, *y* and *z* coordinates for each atom are obtained from the Protein Data Bank (PDB) file [11].

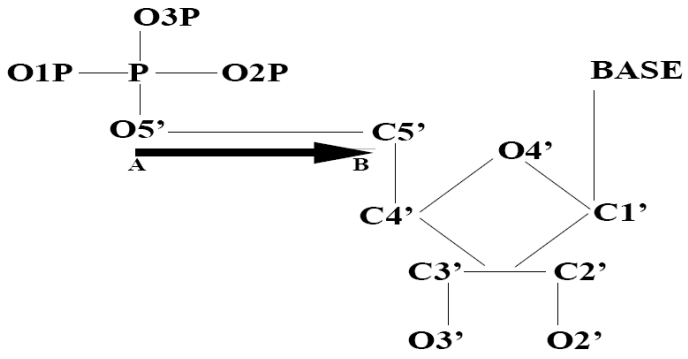



**Figure 1: RNA Nucleotide Structure labelling the sugar phosphate backbone with the structure of the base (BASE), varying; where O3P exists at 5' end.**

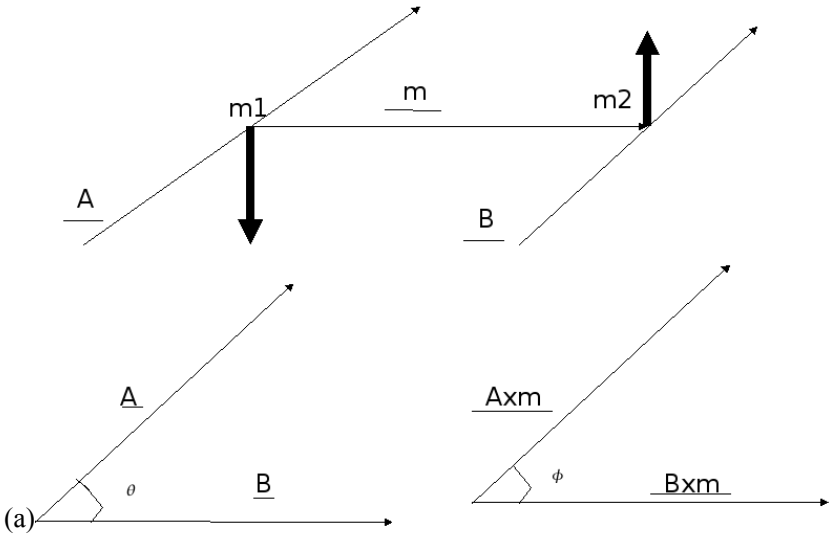
CORONATION labels the edges of the graph, each of which runs between two nodes which correspond to base vectors, by geometric measures. In the CORONATION representation, similar to that of GRATH, the distance is chosen to be that between the two mid-points of each of the vectors. The angle theta $\theta$ between two vectors, A and B, is formed from taking the dot product angle, [8], and is defined to be between 0° and 180°. But in order to describe the shape in three

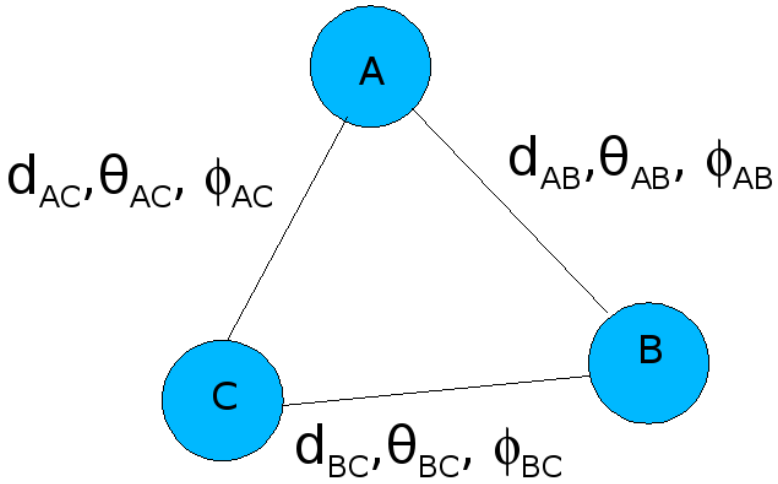
dimensions another angle is required. Each vector (A or B) and their midpoint vector m creates a plane. The vectors A×m and B×m are perpendicular to their respective planes. We can create an angle, the dihedral angle, from the dot product angle between A×m and B×m, (Figure 2a).

By considering all the pairwise relationships between each of the nucleotides in the structure, we can transform a tertiary structure into a fully connected graph. Such a graph gives us the geometric relationship between pairs of vectors of atoms and their relevant bases. This graph representation can also be considered as a matrix.

## 2.2 Comparison of RNA structures

CORONATION works by generating a sequence of matrices to compare a pair of structures, each with their molecular sequences *n1n2n3...nN* and *m1m2m3...mM*, where N, M are the number of bases in the two structures. The first matrix in Table 1, called the SEQ(sequences) matrix of two RNA molecules, contains information about the matches in the types of their bases between two RNA molecules. In this example, the first structure has six bases and the second structure has five bases, so the 2D-matrix contains six columns and five rows. The matrix contains eleven matches, e.g. position (1, 2) has guanine (G) in structure one and a guanine (G) in structure two, as seen in their corresponding row and column. There are 19 mismatches, e.g position (3, 3) has cytosine (C) in structure one and adenine (A) in structure two. We have used consecutive numbers in the matrix starting from 1 and onwards to order these matches, e.g. position (1, 2) is assigned match two and position (3, 6) is assigned match seven.






(b)

**Figure 2: (a): The dot product angle theta and torsion angle are calculated where the midpoint vectors and their perpendicular vectors which give us the idea of dihedral angle Phi; (b): A graph of three RNA structures where nodes are vectorial representation of the axis through each structure. Node represents bases and edges between vectorsA and B which are labelled by: the distance,** *dAB***; the dot product angle,** *θAB***; the dihedral angle,***φAB*.

The second matrix in Table 2, called as the correspondence matrix, gives us information about the allowable relationships between pairs of bases, which may correspond to common geometric pairings in both graphs. The matrix is completed by studying pairs of matches in the SEQ matrix, table 1. As an example, position (1, 7) in the correspondence matrix results from comparing bases one and six in structure one and comparing bases one and three in structure two.

A RNA sequence runs from 5′ to 3′ end, and so this means that many pairs of points in the correspondence matrix can be ignored. In SEQ matrix for example the match pairing (3, 5) can be ignored because it corresponds to moving backward in structure one, from base five to base two, but forwards in structure two, from base one to base two. Because of these topological constraints the correspondence matrix is anti-symmetric, meaning only the upper right part of the matrix is available for matches. Further conditions can be imposed on the positions allowable in the correspondence matrix because only different bases are compared in the two RNAs. So the match pairing (1, 2) can be ignored in the correspondence matrix because this corresponds to comparing two bases in graph one but only one base in graph two. Similarly, we can ignore the match pairing (4, 9) in the correspondence matrix.

| | G | G | C | C | G | A |
|---|---|---|---|---|---|---|
| *G* | 1 | 2 | | | 3 | |
| *G* | 4 | 5 | | | 6 | |
| *A* | | | | | | 7 |
| *A* | | | | | | 8 |
| *G* | 9 | 10 | | | 11 | |

**Table 1: The Sequence matrix (SEQ)**

Two graphs, which have same structure are said to have ”isomorphic” structure, i.e. if there is a correspondence or mapping between the nodes of graph G and graph H such that adjacent pairs of nodes in G are mapped to adjacent pairs of nodes in H [9, 12]. In the case of RNA, the edges of the graph represented the relative geometries between pairs of bases. Any RNA graphs which share isomorphic regions results from a geometric arrangement of bases which are common between two structures. In practise, the geometrical description of two RNA molecules will not be identical, but related RNA structures can have similar 3D arrangements, with similar distances and angles between corresponding vectors. An important criteria of similarity is the allowed tolerances in the differences between distances and angles for two structures [8]. We choose a distance tolerance of 2Å (angstrom), an angular tolerance for theta $\theta$, the dot product angle, of 30°, and the tolerance for dihedral angle phi$\varphi$to be 30°.

The method proceeds to study positions in the correspondence matrix. Each position corresponds to two nodes in both the first and second RNA graphs. The edges between these two nodes have a distance, angle, and dihedral angle associated with them. As an example we consider position (1, 6) in the correspondence matrix. CORONATION checks whether the distance, angle, and dihedral angle between bases one (guanine) and five (guanine) in the first RNA structure are the same, within the tolerances, as those between bases one (guanine) and two (guanine) in the second RNA structure. The answer is stored in a new correspondence matrix, Table 3, with one indicating agreement and zero disagreement.

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 2 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Table 2: The correspondence matrix before checking tolerances**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

**Table 3: The correspondence matrix after checking tolerances are met between edges of the two structures**

Each node in the correspondence matrix, Table 3, results from a shared geometric relationship, within some tolerances, between pairs of bases in each of the two tertiary structures being compared. In order to find common motifs shared between the two RNAs it is necessary to find cliques, that is matching subgraphs, in the correspondence matrix. A clique is a sub graph of a graph in which every node is connected to every other node, Figure 3. Any clique discovered corresponds to a set of bases in structure one, between which the geometry in terms of distance and angles, correspond to a set of bases in structure two. The nodes in the correspondence matrix within the clique map to pairs of bases in each of the structures, as stored in Table 1.

As an example, Table-3 indicates there is a clique of size three between the nodes 1, 5 and 11, i.e. we have found one (agreement) in each of the elements (1,5), (1,11) and (5,11). An examination of Table 1 reveals that bases one, two and five in structure one, have the same geometry, within x constraints in common.

The number of edges, Ne, in a clique of size C is $Ne = C(C-1)/2$ . If we need to satisfy P constraints per edge then the total number of constraints will be $Nc = P \times C \times (C-1)/2$ . As we have three constraints, a distance and two angles, so an overlap clique of three means there are nine constraints in common.

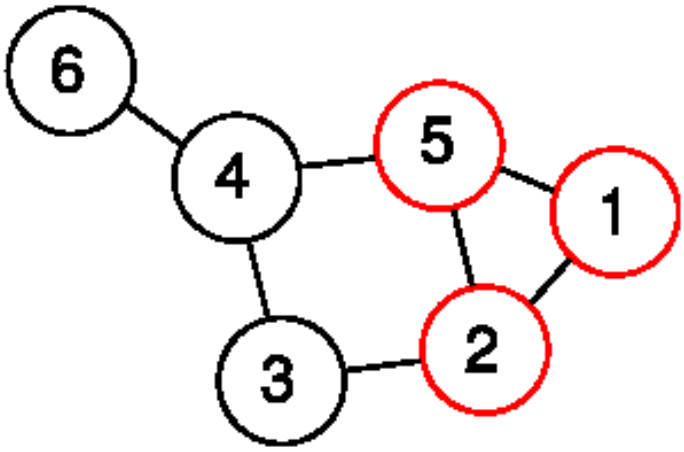



**Figure 3: This graph gives us a maximal clique of size three out of six nodes (1,2,5).**

For finding a clique in a graph means to examine each subgraph with at least k vertices and check to see if it forms a clique. The GRATH algorithm used the clique detection technique of [13]. However, this algorithm is not optimised to search for cliques in very sparse matrices, as is the case here. So instead, CORONATION uses Cliquer [14]. This can search for the maximum cliques or cliques whose size is within the range, optionally limiting the search to maximal cliques.

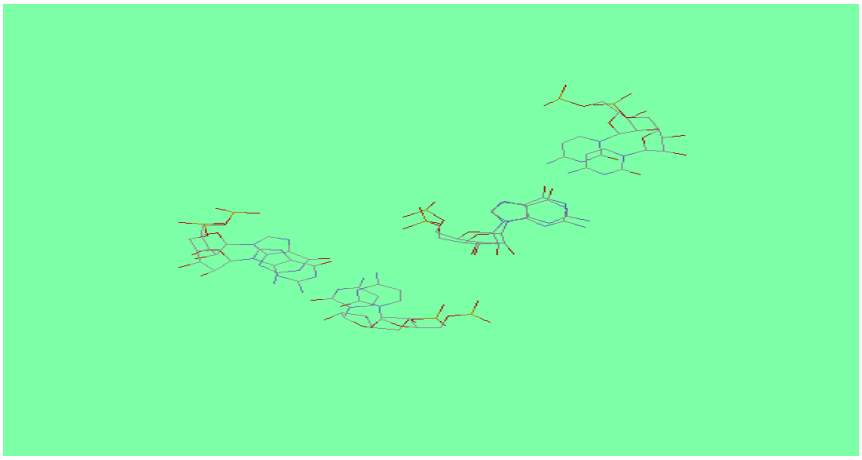


**Figure 4: Two RNA structures comparison of nucleotide bases which are embossed on each other. There is slight difference which gives us the idea that there is an error which has been occurred because of threshold range values in the length, theta and phi.**

## 2.3 Visual Structure comparison tool

Our procedure for analysing RNA structures begins with a detailed structural comparison of all nucleotides in their structures to identify characteristic and variable structural features. The results can be viewed by using Rasmol [15] which is a powerful tool to view 3D appearances of the PDB files. By comparing the two structures we can see how the structures resemble each other (Figure

4). There will be differences in the coordinates of two structures. But these will be sufficiently small to result in the geometric distance and angles between pairs of vectors being less than the thresholds set in CORONATION when generating the correspondence matrix.

## 3. Results

### 3.1 PDB data and Structural analysis

The Protein Data Bank (PDB) [11] contains information about the 3D structures of large biological molecules, primarily proteins and nucleic acids. The PDB contained over 50,000 such structures of macromolecules in 2008, with RNA featuring in 4% of these structures. The RNA structures vary in size, from those with just a few nucleotides to those with many hundreds. We have begun to use CORONATION to identify the structural similarities between each of these structures. Typically we find that each comparison produces several maximal cliques, but that these cliques have many common bases of nucleotides. We identify the largest number of common bases of nucleotides, and call this the common clique. As an example, we compare the structure of PDB:1ehz [16], a phenylalanine tRNA, against 963 RNA structures, Figure 5. Figure 5, shows the majority of structures have a common clique of ten bases or less, but there are some structures show much greater similarity. We are presently exploring how best to analyse the statistics of these structural similarities, and to identify whether any unusually significant matches between structures relate to common functions.

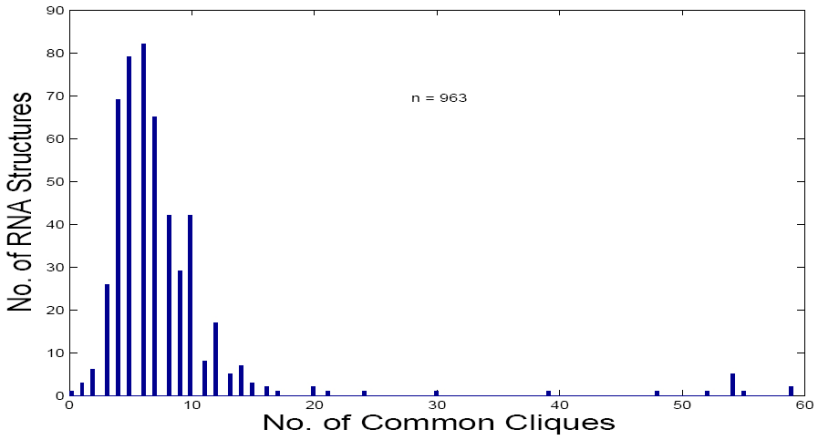



**Figure 5: A comparison of the structure 1ehz with 963 other RNA structures. When there are more than one example of a maximal clique, the common clique is that which contains the largest number of bases shared between the cliques.**

## Conclusion

We have described CORONATION, a novelmethod for detecting common substructures shared between RNA tertiary structures. For each structure we transform its PDB-formatted file into a graph, in which the nodes corresponds to a vector running from O5′ to C5′ for each nucleotide,

and the edges are labelled by the midpoint distance, dot-product angle and dihedral angle between the two vectors. CORONATION generates an overlap graph in order to find cliques, which correspond to shared geometric arrangements between groups of bases in the two structures.

CORONATION is highly efficient and fully automated, and a typical comparison of two RNA molecules takes only a few seconds on a standard PC. The method can be readily scaled to perform an all-against-all comparison of all the RNA 3D structures currently available in the PDB. CORONATION has been developed in C and PERL.

## References


[1] J. Allali and M. F. Sagot. A New Distance for High Level RNA Secondary Structure Comparison. *IEEE/ACMTransaction on computational biology and bioinformatics* , 2(1), 2005.

[2] P. D. Rijk and R. D. Wachter. RnaViz, a program for the visualisation of RNA secondary structure. *Nucleic Acids Research*, 25(22):4679-4684, 1997.

[3] C. M. Duarte, L. M. Wadley, A. M. Pyle. RNA structure comparison, motif search and discovery using a reduced representation of RNA conformational space. *Nucleic Acids Research*, 31(16):4755-4761, 2003.

[4] A. M. Harrison, D. R. South, P. Willett, P. J. Artymiuk. Representation, searching and discovery of patterns of bases in complex RNA structures. *Journal of Computer-Aided Molecular Design*, 17:537-549, 2003.

[5] L. M. Wadley, A. M. Pyle. The identification of novel RNA structural motifs using COMPADRES: an automated approach to structural discovery. *Nucleic Acids Research*, 32(22):6650-6659, 2004.

[6] M. Tamura, D. K. Hendrix, P. S. Klosterman, N. R. B. Schimmelman, S. E. Brenner, S. R. Holbrook. SCOR: a Structural Classification of RNA Database. *Nucleic Acids Research*, 30(1):392-394, 2004.

[7] M. Abraham, O. Dror, R. Nussinov, et al. Analysis and classification of RNA tertiary structures. *RNA*, 14(11), 2008. *Online Journal*: http:www.rnajournal.org

[8] A. Harrison, F. Pearl, I. Sillitoe, T. Slidel, R. Mott, J. Thornton and C. Orengo. Recognising the fold of a protein structure. *Bioinformatics*, 19(14):1748-1759, 2003.

[9] H. M. Grindley, P. J. Artymiuk, D. W. Rice, P. Willett. Identification of tertiary structure resemblance in proteins using a maximal common sub-graph isomorphism algorithm. *Journal of Molecular Biology*, 229:707-721, 1993.

[10] L. Holm, C. Sander. Protein Structure Comparison by Alignment of Distance Matrices. *Journal of Molecular Biology*, 233:123-138, 1993.

[11] H. Berman, K. Henrick, H. Nakamura, J. L. Markley. The worldwide Protein Data Bank(wwPDB):ensuring a single, uniform archive of PDB data. *Nucleic Acids Research*, 35:D301-D303, 2007.

[12] D. Kozen. A clique problem equivalent to graph isomorphism. *ACM,SIGACT News*, 10(2):50-52, 1978.

[13] C. Bron, J. Kerbosch. Algorithm 457: Finding all cliques of an undirected graph[H]. *Communication ACM*, 16(9):575-577, 1973.

[14] S. Niskanen, P. R. J. stergrd. Cliquer User's Guide, Version 1.2. *Communications Laboratory, Helsinki University of Technology, Espoo, Finland,Tech. Rep. T48*, 2008.

[15] R. Sayle and A. Bissell. RasMol: A program for Fast, Realistic Rendering of Molecular Structures with Shadows. *Proceedings of the 10th Eurographics UK*, 1993.

[16] H. Shi and P. B. Moore. The crystal structure of yeast phenylalanine tRNA at 1.93 °A resolution: A classic structure revisited. *RNA*, 6:1091-1105, 2000.

# Spectral markers for knotted core of proteins

**Lissy Anto P. [1,2] and Achuthsankar S. Nair [1]**

[1] Centre for Bioinformatics, University of Kerala, Thiruvananthapuram, Kerala, India, Pin 695581.

[2] St.Joseph's College,Irinjalakuda, Thrissur, Kerala, India, Pin 680121.

Email: lissyantop@gmail.com

## Summary

Only a few percentage of known protein structures have shown knotted configurations in their native fold. Still the functions of these configurations are not understood. Protein knot localization has become possible in single molecule experiments. Here, we investigate peptide knot characteristics in detail with the amino acid indices hydrophobicity and isoelectric point which have crucial role in retaining the stability of proteins.Water capture and release is found to be controllable by the tightening force in knots. In this work we analysed protein sequences having knotted core with the help of Fourier analysis. The set of knot proteins from proteinKNOT web server(pKNOT ) has been used for the experimentation.

## 1. Introduction

In mathematics, knots are closed curves and usually categorized according to the minimal number of crossings in a projection onto a plane. There are several algorithms, like the Alexander polynomial or the Homfly polynomial, which are able to distinguish between knots. Knots are rare though the reason is not well understood. It has been hypothesized that knotted structures are difficult to fold[1] and would essentially preserve their unknotted state after the initial collapse. The presence of a knot may alter the enzymatic activity of the protein.

## 1.1 Knots in protein:

Though we now know the structures of a large number of proteins, only a few have knotted structures. The knotted regions have been shown to be important in both ligand binding and enzyme activity. Knotted proteins have become more common in recent years due to the enormously growing number of structures deposited in the Protein Data Bank (PDB). Real proteins do not have their termini joined. This presents a technical problem, as knots are only properly (mathematically) defined in circular strings. However, one common definition of a knot is "a loop in a string that tightens when pulled". This can be applied to a protein by repeatedly smoothing while keeping the two termini fixed in place and seeing if a straight line is obtained. Using this method, the ends can also be progressively trimmed to find the exact location of the

knotted core. Protein knots can be quantified by the number of residues on either side of them[2].

### 1.2 Identifying knots in protein:

The first web server to detect the knots in proteins as well as provide information on knotted proteins in PDB is the proteinKNOT(pKNOT) web server[3]. The pKNOT web server detects the knot in a protein by smoothing the protein chain using the Taylor's algorithm[1]. The algorithm first fixes both N and C termini in space, then repeatedly smoothes and straightens the protein chain. The chain is reduced in such a way that, with details of the chains eliminated, the knot can be easily detected. If the protein does not contain a knot, the chain will simply shrink into a straight line.

## 2. Materials and Methods

### 2.1 Biosequence Processing using Digital Signal Processing

For the structural and functional study of biosequences, spectral analysis has been used to detect latent periodicities of biological sequences. The spectral analysis are largely based on Fourier transform which can reveal long range periodicities. When the window length is chosen appropriately for the STFT, the information related to a protein's biological properties for the chosen protein example could be explored[4].

In Genomic Signal Processing, DNA sequences are mapped into digital signals in a variety of ways thereby enabling the use of digital signal processing, a popular tool in engineering field, to study the biosequences. For a DNA string x[n] of N characters (with alphabets A, G, C & T), the four binary indicator sequences $u_A[n]$, $u_G[n]$, $u_C[n]$, & $u_T[n]$ can be defined as follows:

Let the DNA sequence be x[n] = [ A G T C G A T G C A T C ] . The indicator sequences are,

$u_A[n]$ = [ 1 0 0 0 0 1 0 0 0 1 0 0],

$u_G[n]$ = [ 0 1 0 0 1 0 0 1 0 0 0 0],

$u_C[n]$ = [ 0 0 0 1 0 0 0 0 1 0 0 1] and

$u_T[n]$ = [ 0 0 1 0 0 0 1 0 0 0 1 0].

(i.e. each indicator sequence has a 1 if the corresponding base exists at the position 'n', otherwise a 0) The sum of all binary indicators at any position n is 1 for all n.

i.e. $u_A[n] + u_G[n] + u_C[n] + u_T[n] = 1$ for n=0, 1, 2,....N-1.

Let $U_A[k]$, $U_G[k]$, $U_C[k]$ and $U_T[k]$ be the Discrete Fourier Transforms (DFT) of the binary sequences $u_A[n]$, $u_G[n]$, $u_C[n]$ & $u_T[n]$ respectively. The power spectrum of x[n] is given by

$R[k] = \Sigma |U_X[k]|^2$ for X=A, G, C or T & k = 0, 1, 2, … (N-1).

This power spectrum approach is used to locate exons by the special feature of showing period three peak in the power spectrum of the exon regions [4]. This well established method has been improved with the help of electron ion interaction pseudo potential (EIIP) [5]. Instead of using the four indicator sequences to find the power spectrum, a numeric sequence formed by the substitution of the EIIP values for A, G, C & T in a DNA string is used. This numerical sequence represents the distribution of the free electrons' energies along the DNA sequence. This sequence is named as the 'EIIP indicator sequence'. It has been reported that EIIP as a physico chemical parameter is meaningful in revealing coding regions of genomic signals in a better way than the four indicator sequences[5]. The experimentation with EIIP indicator sequence has given high discrimination between coding and non-coding regions. Also this method is used to find false exons. i.e. introns showing period three peak behavior like exon regions. The power spectrum method is applicable to amino acid sequences too. Resonent Recognition Model(RRM) analysis involves converting the amino acids that constitute a protein into a "discrete time series." The position of an amino acid in the sequence can be thought of as the time. The datum associated with each time in our study is hydrophobicity which is a measure of an amino acid's tendency to avoid water. After the conversion of the amino acid sequence into a numeric sequence based on hydrophobicity,it is made into the protein time (space) series signal (which we call a "AA signal"), which is analyzed to locate the dominant frequencies. It has been shown that a particular function in a protein is represented by one RRM characteristic frequency that can be determined by Fourier analysis [6]. There exists a significant correlation between the spectra of numerical representations of amino acids and their biological activity [6]. More specifically, the biological function of a protein is characterized by certain frequencies of its signal representation. In this study, we consider the AA signal which are used to analyse the Knot feature is then.

## 2.2 Analysis of Knots:

For the present work, we have used the published knot proteins from pKNOT webserver. These are of trefoil knot which is the simplest knot of all, and is characterized by three crossings. It is mathematically denoted as $3_1$ knot. The proteins with a trefoil knot are methyltransferase, transcarbamylase, methionine adenosyltransferase, carbonic anhydrase and YMPa superantigen (NMR). In Table 1, ''Species'' refers to the scientific name of the organism from which the protein was taken for structure determination. ''PDB code'' give Protein Data Bank entry for each knotted protein. "Knotted core" of a knot is the minimum configuration which stays knotted after a series of deletions from each terminus(in brackets we indicate how many amino acids can be removed from either side before the structure becomes unknotted). Table 2 gives the amino acid index value of hydrophobicity by which we made the digital signals for experimentation. In some studies ,the statistical analysis of knotted structures shows a higher occurrence for Leu, Phe, Trp, Gly, His,

Gln,Asp, Lys and Pro[7].

**Table 1. Knotted core of Organisms**

| Species | PDB Code | Knotted Core | Species | PDB Code | Knotted Core |
|---|---|---|---|---|---|
| E.coli | 1ns5 | 69-121(32) | Bos Taurus | 1hcb | 29-256(2) |
| T.maritime | 1o6d | 68-117(30) | Dunaliella salina | 1lug | 30-256(3) |
| S.aureus | 1vh0 | 73-126(31) | Rattus norv. | 1v9e | 37-270(4) |
| B.subtilis | 1to0 | 64-116(32) | H.sapiens | 1y7w | 32-270(4) |
| H.influenza | 1uaj | 93-138(92) | H.sapiens | 1flj | 30-256(3) |
| E.coli | 1p9p | 90-130(89) | Mus musculus | 1z93 | 28-254(9) |
| T.thermophilus | 1v2x | 96-140(51) | Mus musculus | 1znc | 32-261(1) |
| E.coli | 1j85 | 77-114(42) | H.sapiens | 2znc | 32-246(3) |
| A.aeolicus | 1ipa | 185-229(29) | Mus musculus | 1keq | 7-234(4) |
| S.viridochromog | 1gz0 | 172-214(28) | H.sapiens | 1jd0 | 28-257(3) |
| H.influenza | 1zjr | 95-139(58) | Mus musculus | 1rj6 | 29-257(2) |
| B.subtilis | 1x7p | 192-234(31) | E.coli | 1fug | 33-260(32) |
| T.thermophilus | 1nxz | 165-216(30) | Rattus norv. | 1qm4 | 30-253(29) |
| A.M Thermoautotr | 1vhk | 158-208(27) | Spinacia oleracea | 1yve | 239-451(62) |
| N.gonorrhoeae | 1v6z | 103-202(25) | E.coli | 1yrl | 220-435(52) |
| H.sapiens | 1k3r | 48-234(28) | B.fregilis | 1js1 | 169-267(57) |
| H.sapiens | 1kop | 36-223(0) | X.campestris | 1yh1 | 171-272(62) |

**Table 2. Hydrophobicity values**

| Amino Acid | Hydrophobicity | Amino Acid | Hydrophobicity |
|---|---|---|---|
| Ala(A) | 0.13134 | Leu (L) | 0.05671 |
| Arg(R) | 1 | Lys (K) | 0.78059 |
| Asn(N) | 0.53134 | Met (M) | 0.01940 |
| Asp(D) | 0.80597 | Phe (F) | 0 |
| Cys(C) | 0.10597 | Pro (P) | 0.24328 |
| Gln(Q) | 0.48805 | Ser (S) | 0.19402 |
| Glu(E) | 0.74328 | Thr (T) | 0.15671 |
| Gly(G) | 0.16865 | Trp (W) | 0.11343 |
| His(H) | 0.41940 | Tyr (Y) | 0.27462 |
| Ile (I) | 0.03731 | Val (V) | 0.06865 |

## 5. Results & Discussion

In our investigation, the frequency percentage of amino acids in knotted structures occupies a maximum for Leu, Gly, Asp and Val. The frequency distribution of amino acids in the knotted core and the whole sequence are shown in Fig.1 and Fig. 2. The higher percentage for the above four amino acids is evident in Fig. 1. Amino acids are assigned positive integer values from 1 onwards when they are taken in alphabetic order.

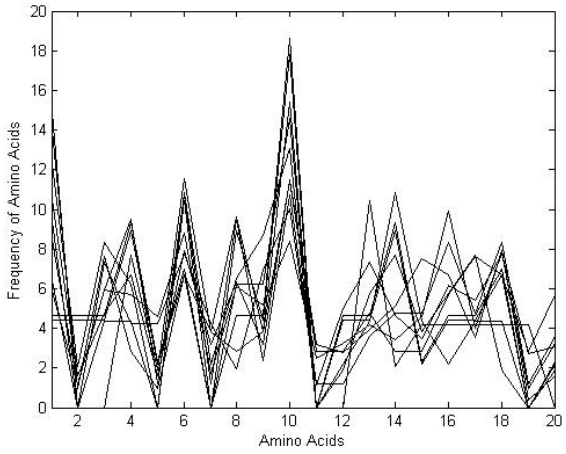



**Fig 1. Freq. Distribution in Knotted core**

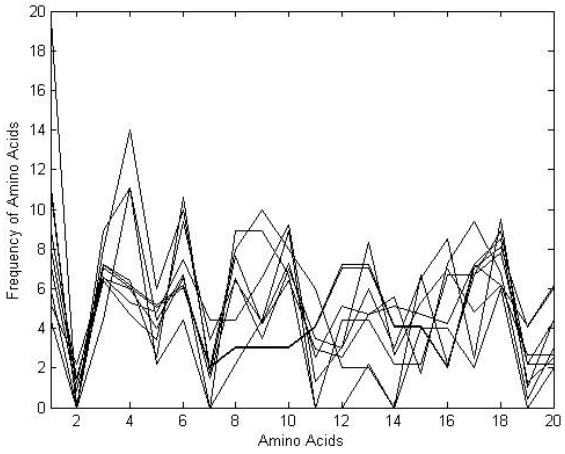



**Fig 2. Freq. Distribution in the whole sequences**

Proteins can be either positively or negatively charged based on pH conditions. When the positive and negative charges on protein are equal, the net charge is zero. The characteristic pH of a solution at which the net charge on protein is zero (positive and negative charges are equal) is defined as the isoelectric point. The isoelectric point of a protein is an important property because it is at this point that the protein is least soluble, and therefore unstable. When the isooelectric point of knotted core and that of the whole sequence of the data set is calculated, it was interesting to see that they differ only a little. So the stability of a knotted core measures the stability of the whole protein. Fig. 3 shows the variation of isooelectric point of the knotted core and the sequences from the dataset.

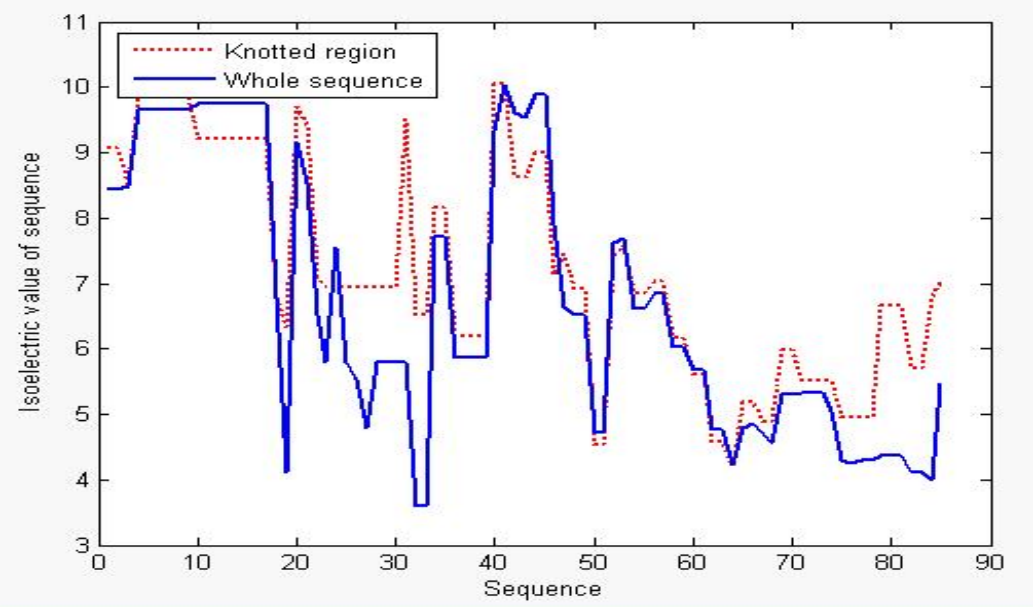



**Fig 3. Plot of Isoelectric value of knotted core and full sequence**

Hydrophobic interactions are the most important non-covalent forces that are responsible for different phenomena such as structure stabilization of proteins. The role of hydrophobicity is a determinant of protein-protein interactions. Results suggest that surface hydrophobicity can be used to identify regions of a protein's surface most likely to interact with a binding ligand. It is generally accepted that the hydrophobic effect is the main factor in stabilizing the folded structure of globular proteins[9,10]. The identification of knotted area of proteins is thus relevant in deriving these hydrophobic core of a protein. Knots contribute to thermal stability as long as they maintains the conformation of the folded state.

The structural features of protein sequences is hidden in periodicity of symbolic sequences[11]. The packing of the residues in a knotted core typifies the rest of the protein interior. What most distinguishes these regions from others is the synergism between good packing, strong hydrogen bonding, and the optimal interaction among peptide dipoles.Knot residues can pack very

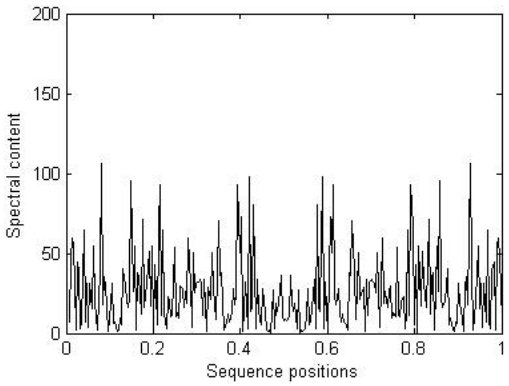



**Fig. 4 Spectrum of a full sequence(pdb code:1yrl)**

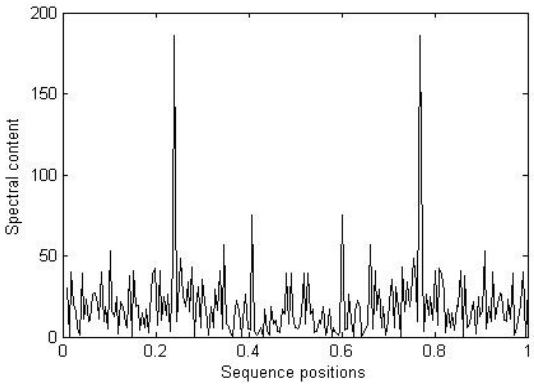



**Fig. 5 Spectrum of a knotted core(pdb code:1yrl)**

efficiently without sacrificing good hydrogen bonding[12].The protein sequences(Table1) are converted into numeric sequences by replacing amino acids by their respective hydrophobicity values(Table2). The analysis of FFT spectrum creates a special appearance for knotted region. A periodic pattern is observed at the knotted core as there is higher concentration of hydrophobic residues(Fig. 4 & Fig. 5).

With the huge number of proteins, there is a large enough collection of folds to get a better idea of how frequent and important all forms of knots are to protein structure and function. The analysis of these should provide useful insight into how proteins fold.

In evolutionary context, homologous structures of protein tend to retain topological features. The trefoil knot in carbonic anhydrase can be found in isozymes ranging from bacteria and algae to humans. Class II ketol-acid reductoisomerase comprises a figure-eight knot present in Escherichia coli and spinach, and S-adenosylmethione synthetase contains a deep trefoil knot in E. coli and rat. It reveals that particular knots have indeed been preserved throughout evolution, which suggests a crucial role for knots in protein enzymatic activity and binding.

## References


1. W. R. Taylor, A deeply knotted protein and how it may fold, Nature 406: 916-919,2000.
2. Peter Virnau, Leonid A. Mirny and Mehran Kardar, Intricate Knots in Proteins: Function and Evolution, PLoSComputational Biology, 2(9): e122, 2006.
3. Yan-Long Lai, Shih-Chung Yen, Sung-Huan Yu and Jenn-Kang Hwang, pKNOT: the protein KNOT web server, Nucleic Acids Research, 2007 Vol. 35, Web Server issue.

**4.** Silverman, B. D. and Linsker R., A measure of DNA DNA periodicity, J Theor Biol., 3,118, 1986.

**5.** Achuthsankar S. Nair, Sreenadhan S., Coding measure scheme employing electron-ion interaction pseudopotential(EIIP), Bioinformation, 1(6): 197-202, 2007.

**6.** Cosic, I., Can short time Fourier transform detect the localized latent periodicity of a protein sequence?, IEEE EMBS Asian-Pacific Conference 66 – 67, 2003.

**7.** Wang Xiang-Hong et al, Structural statistical properties of knotted proteins, Chinese Phys. B 18**:** 1684-1690, 2009.

**8.** Dartmouth College Computer Science Technical Report, Analysis of Protein Sequences Using Time Frequency and Kolmogorov-Smirnov Methods, Kobby Essien.

**9.** Bing-Yan Zhu et al, Packing and hydrophobicity effects on protein folding and stability, Protein Science 2: 383-394. 1993.

**10.** Kurt Wagschal et al, The role of position a in determining the stability and oligomerization state of a-helical coiled coils: 20 amino acid stability coefficients in the hydrophobic core of proteins, Protein Science 8:2312–2329,1999.

**11.** Eugene Korotkov and Nikolay Kudryaschov,Latent Periodicity of Many Genes, Genome Informatics 12: 437–439, 2001.

**12.** Roger B. Gregory, Protein-solvent interactions, American Chemical Society, 1995.

# Prediction of RNA-binding sites in a protein sequence using concurrently conserved pattern mining

**Chen-Ming Hsu** [1]

[1] Computer Science and Information Engineering, Ching Yun University, Taiwan

## Abstract

The identification of RNA-binding residues (RBRs) in proteins is important in molecular recognition. In the absence of structures for RNA-protein complexes, it is desirable to predict RBRs by protein sequences alone. In this study, we present a concurrently conserved pattern mining approach named WildSpan to tackle this problem. It is observed that a functional site in protein structure usually consists of several local blocks linked with long wildcard regions and flexible among homologous sequences. The WildSpan is invoked to discover concurrently conserved patterns spanning large wildcard regions (W-patterns) in homologous sequences and the discovered W-patterns are used to identify RBRs in a protein sequence. We compare with the multiple-sequence-alignment based method (ConSurf) on a dataset of 132 RNA-binding proteins, WildSpan and ConSurf achieves maximum Matthews Correlation Coefficients (MCC) of 0.309 and 0.224, respectively. The performance of WildSpan further improved the MCC from 0.309 to 0.323, when the information of interface propensities has integrated into the WildSpan. Besides, the predicting power of WildSpan using all of discovered W-patterns achieves a MCC of 0.214, which is also better than an optimal MCC of 0.206 predicted by the structure-based Naive Bayes classifier (RNABindR). Further analysis for discovered W-patterns, there are about 70% of concurrently conserved blocks in a W-pattern are observed to be clustered in space and about 80% are found to be near the RNA-binding interfaces within a 5 anstron distance. Conclusively, the efficiency of sequence-based WildSpan is not only favorable in predicting complex-structure-unknow protein but also largely desired in large-scale proteomics.

# Protein structure quality analyzer

**V. Amardev Rajesh [1,*], Lubna Sulthana [1], T. Venumadhav [2], M. Bhaskar [3]**

[1] Department of Biotechnology and Bioinformatics, Dravidian University, Kuppam, Andhra Pradesh, India.

[2] Onan biotech, Hyderabad, Andhra Pradesh, India

[3] Department of Zoology, S. V. University, Tirupati, Andhra Pradesh, India

*Corresponding / presenting author: e-mail: rajesh.v.amardev@gmail.com

## Abstract

Proteins are large, complex molecules that play many critical roles in the body. If the proteins have similar sequences, as they share common ancestor then the proteins will have similar three-dimensional structures, so they infer the functional relationship. It often becomes necessary to compare the similarity of a model of a protein with that of the predicted model to optimize the functional performance. The most critical method to find similarity is to measure the Root Mean Square Distance (RMSD). However a single measure cannot evaluate the similarity between two structures, it also involves many other measures such as obtaining C-Alpha matches from the tool C-Alpha Matcher, performing superimposition from Swiss PDB viewer, getting an optimal superposition from Superpose tool and obtaining LG score from MaxSub and tools to build and visualize the protein structural models. In view of the importance, protein quality analyzer tool is constructed using Html and java script linked to many online tools. This enables quickly access to find the similarity between two structures and get results of the desired protein's structure quality. The tool is effective and is accurate because of its integration of many tools, which are necessary to find the quality of protein structure.

# Establishment of promoter sequences and annotations database

**Yong, Li** [1]

[1] College of Life Sciences Northeast Agricultural University, Harbin, China

## Abstract

Gene expression regulation is an important research field, especially to understand gene tissue-specific expression, inducible expression and so on. And these are the keys of organism grow and develop orderly. To understand gene expression regulation, many cis-acting elements were found and studied, but how they work, how they co-operate to make promoter tissue-specific or inducible, are still unknown, mainly because of lack of annotated promoter sequences.

To collect promoters from GenBank and annotate them, a perl script was written based on BioPerl. It could retrieve DNA sequences containing promoter from GenBank, and find promoter, transcription start site (TSS), translation initiation site (TIS) according to sequence features, and find expression specification according to sequence annotations. Using this script, plant promoters were retrieved from GenBank and annotated. Then every promoter was outputted into a GenBank format file. Finally all found promoters were put into a database constructed with MySQL based on BioSQL schemas.

This database contains 5744 promoter sequences, 2821 TSS and 2623 TIS. Of these promoters, 338 were tissue-specific or inducible, including cold-, salt-, wound, auxin-, ethylene-inducible, and root-, fruit-, anther-, seed-, pod-specific. The web interface of this database is coming soon.

# Identification of SNPs by 454 sequencing and convertion of CAPS markers in soybean

**Yong, Li** [1]

[1] College of Life Sciences Northeast Agricultural University, Harbin, China

## Abstract

To discover new SNPs and develop easy assay method in soybean, we have compared the high-through sequences of variety Asgrow A3237 with the whole genome sequences of Williams 82. 3899 SNPs were identified between two genotypes, the most mutations were transitions, such as A→G and C→T, which would influence the genes expression by methylation. The SNPs were widely distributed in the soybean genome, targeting numerous genes involved in various physiological and biochemical processes influencing important agronomic traits of soybean. A set of 36 SNPs displayed as potential CAPS candidate were resequenced, and 16 SNPs were validated in the nine soybean varieties, and seven SNPs were converted into CAPS. The novel SNPs discovery and CAPS markers convertion system developed in this study was fast and cost effective for identification and application of SNPs, and holds great promise for molecular assisted breeding of soybean.

# A multi-level model accounting for the effects of JAK2-STAT5 signal modulation in Erythropoiesis

**Xin, Lai** [1]

[1] Systems Biology and Bioinformatics Group, Department of Computer Science, University of Rostock, Ulmen Str.69, Haus 3, 3.OG, Raum 401, 18057 Rostock, Germany

## Abstract

We develop a multi-level model, using ordinary differential equations, based on quantitative experimental data, accounting for murine erythropoiesis. At the sub-cellular level, the model includes a description of the regulation of red blood cell differentiation through Epo-stimulated JAK2-STAT5 signalling activation, while at the cell population level the model describes the dynamics of (STAT5-mediated) red blood cell differentiation from their progenitors. Furthermore, the model includes equations depicting the hypoxia-mediated regulation of hormone erythropoietin blood levels. Take all together, the model constitutes a multi-level, feedback loop-regulated biological system, involving processes in different organs and at different organisational levels.

We use our model to investigate the effect of deregulation in the proteins involved in the JAK2-STAT5 signalling pathway in red blood cells. Our analysis results suggest that down-regulation in any of the three signalling system components affects the hematocrit level in an individual considerably. In addition, our analysis predicts that exogenous Epo injection (an already existing treatment for several blood diseases) may compensate the effects of single down-regulation of Epo hormone level, STAT5 or EpoR/JAK2 expression level, and that it may be insufficient to counterpart a combined down-regulation of all the elements in the JAK2-STAT5 signalling cascade.

# Simulation analysis of the enzyme expression patterns in E.coli towards the understanding of biological systems

**Tomoaki, Yamamotoya [1]**

[1] Biopathway Analysis Center, Faculty of Science, Yamaguchi University, Yamaguchi, Japan

## Abstract

Metabolism is the system constituted by many enzymic reactions, maintaining the activity of a living organism. Until now, individual functions of genes and these products related to the metabolism have been analyzed. However, the whole mechanism of metabolism can not be uncovered with only the studies of these individual functions.

Computer simulation allows us to integrate these individual functions, leading us to the complete understanding of the metabolism. Therefore, we have tried to construct a computational model of E.coli metabolic pathways with the data obtained from biological experiments. At the beginning, we constructed a framework model of the central metabolism with hybrid functional Petri net based on the structural information of metabolic pathways presented in KEGG and Ecocyc databases. The next task should be to incorporate kinetic parameters of enzyme expressions into that framework model.Then, we made biological experiments using Flow Cite Meter with GFP fusion strain to obtain time course data of enzymic expressions.

Further, in order to acquire more various information of metabolic pathways, we conducted a biological experiment in which a carbon source was changed from glucose to mannose. Based on the provided experimental results, by using Cell Illustrator, we constructed a computational model and analyzed the enzyme expression patterns which are different between the glucose and mannose cultures in terms of rates of enzymic productions. This analysis implies that this difference is caused by a transcription factor that controls the productions of some specific enzymes.

# Experimentation and evaluation of SOM-based classification of cancer cells with the information of proteins

**Maya, Tachibana [1]**

[1] Biopathway Analysis Center, Faculty of Science, Yamaguchi University, Yamaguchi, Japan

## Abstract

Cancer cells are different in progress speed and in the effect of the therapeutic drug even if we extracted cancer cells from the same part. Therefore it tend not to be correct that a classification method of cancer cells depend on a morphologic characteristic. A study to characterize cancer cells in molecular levels has been performed with a laser scanning cytometer (LSC). Although some studies that tried to classify cancer cells with protein amount and cohesion level in cancer cells are performed, effective method is still unavailable. Hence, we have been developing a classification method of cancer cells that depend on a correlation between protein amount and cohesion level in cancer cells by using self-organizing map (SOM).

In this context, we checked relation of this correlation to the mapped-location of a protein by SOM and found that cancer cells of the similar correlations locate at the near positions in SOM. Based on this fact, we conducted this SOM-based classification of cancer cells on 20 proteins, confirming that SOM has a possibility to extract the features of a cancer cell with respect to the protein amount and the cohesion level.

# Prediction Method of Essential Points in a Biological Pathway for Cell System Stability by using Recurrent Neural Network

**Hironori Kitakaze** [1]

[1] Biopathway Analysis Center, Faculty of Science, Yamaguchi University, Yamaguchi, Japan

## Abstract

Hybrid functional Petri net (HFPN) was proposed as a method for the modeling of biological pathways. This method can adopt appropriate functions to continuous and discrete events in a biological pathway, and can model any biological pathway without loosing the relation of the connection of substances and reactions, which is usually depicted by a figure in the literature. Sequential deletion of HFPN elements, which corresponds to the operation in a biological experiment such as gene knockout, enables the essential point prediction. That is, by repeating these deleting operations in the HFPN model, we can identify essential points in the cell system. However, the time for simulation and the time for the effect confirmation of the deleted element would increase as the size of a biological pathway become larger.

In this poster presentation, we propose a computational method using recurrent neural network (RNN) that predicts essential points in a cell system. RNN is a neural network that reflects the development of time in the network. The procedures that convert a given HFPN to the RNN and predict essential points in an HFPN from the converted RNN are presented. Although the learning process of BPTT (back propagation through time) method in the RNN takes a long time, it was confirmed that the proposed method can produce the prediction result of essential points totally faster than the conventional method using the HFPN.