# Contents

<table>
<thead>
<tr>
<th>Page</th>
<th>Content Area</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>02</td>
<td>Executive Director's Message</td>
<td></td>
</tr>
<tr>
<td>04</td>
<td>For Industry and Clinical Partners</td>
<td></td>
</tr>
<tr>
<td>06</td>
<td>Programme on Human Infectious Diseases</td>
<td></td>
</tr>
<tr>
<td>08</td>
<td>Programme on Cancer Biomarkers</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>Biomolecular Modeling and Design Division</td>
<td>Chandra VERMA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Igor BEREZOVSKY</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Peter J. BOND</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FAN Hao</td>
</tr>
<tr>
<td>34</td>
<td>Imaging Informatics Division</td>
<td>LEE Hwee Kuan</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CHENG Li</td>
</tr>
<tr>
<td></td>
<td></td>
<td>CHIAM Keng Hwee</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LOO Lit Hsin</td>
</tr>
<tr>
<td>42</td>
<td>Translational Research Division</td>
<td>NG Siew Bee</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Yoganathan KANAGASUNDARAM</td>
</tr>
<tr>
<td>46</td>
<td>Bio-Computing Centre and Administration Teams</td>
<td>Samuel K.E. GAN</td>
</tr>
<tr>
<td>47</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Executive Director's Message

Welcome to the Bioinformatics Institute in Singapore!

Slightly more than a decade ago, the first draft of the essentially complete human genome was published. Whereas the presentation of the draft was celebrated with great pomp, the reviews summarizing the achievements thereafter hardly made it into the headlines. Not surprisingly, the outcome with regard to cures for not yet treatable diseases or new biotechnologies has not nearly reached the expectations.

For the insider, this development was not a major surprise (see also JBCB 10(5):1271001, 2012). In 2001, about half of the known protein-coding genes in human was functionally not characterized and, although our biological knowledge is as large as never before in human history, the list of known gene function has not become much longer in the mean time. Most of the human genome (98.5%) is not protein-coding and this “rest” is also actively transcribed; yet, the world of non-coding RNAs’ functions remains enigmatic for the most part. On the positive side though, the knowledge of the human genome sequence allows assessing how much of the human biology at the molecular mechanism level is still unknown and it appears that it is surely more than half of it.

As this situation does not promise immediate success for many pharmaceutical and biotechnological applications at the moment, it provides great opportunities for bioinformatics and computational biology. Although life sciences are not truly a theoretical discipline since the extrapolation depth is small due to the fragmentary knowledge of biomolecular mechanisms, there are a few increasingly important research areas such as studies of sequences, expression profiles, 3D structures and bioimages where the application of quantitative, mathematical concepts has become instrumental for the discovery and for progress in biological theory, for the prediction of function of genes and their interaction in pathways and networks. For example, the concept of sequence homology as common evolutionary ancestry leading to sequence similarity with resembling protein structure and function of proteins was considered obscure when it was first developed; yet, it is at life science’s main stage today.

This development is heralded by the availability of complete genomes and has mainly been fuelled by DNA sequencing but also by other high-throughput experimental techniques. The key task in life sciences now is the interpretation of non-understood genomic sequences in terms of biological function and mechanisms and especially the characterization of functionally not yet annotated genes. We can jokingly say that computational biologists would have lots of biological data for analysis for decades to come even if experimentation in life science had stopped from now on completely.

The Bioinformatics Institute (BII), originally founded by Dr. Gunaretnam Rajagopal in 2001 as an IT services and bioinformatics support unit, has experienced a transformation into a biological research organization since my arrival in August 2007. The scientific mission involves computationally biology driven life science research aimed at the discovery of biomolecular mechanisms. Besides the actual theoretical studies on biological data, this includes also the development of appropriate computer-based theoretical research tools.

Our work would be incomplete without the experimental verification of our own hypotheses and the application of the results. For this purpose, we have two options. On the one hand, we extensively collaborate with experimental and clinical groups from academia in Singapore and abroad as well as with pharmaceutical and biotechnological industry. Alternatively, BII also has its own experimental facilities that will further improve in 2014 in cooperation with IMCB in the Proteos building.

BII has currently almost 20 small and medium-sized independent research teams, most of them led by first-time principal investigators (among them 9 new since 2012-2014), one of them jointly run with SBIC and another one with IMCB, two other A*STAR institutes. All groups are organized in four basic research divisions (including (i) analysis of genome sequences, gene expression and RNA biology, (ii) protein sequence analysis and function prediction of uncharacterized genes, (iii) protein 3D structure modeling and (iv) imaging informatics – computer-supported analysis of microscopic images of cells and tissues with labelled molecules and a fifth, new division of translational research. The latter division also houses a large library of more than 100,000 microbiological, fungal and plant species for future genomics and system biology research. In addition, we also have two cross-divisional programs with clinical focus (“Cancer Biomarkers” headed by Vladimir A. Kuznetsov and “Human Infectious Diseases” headed by Sebastian Maurer-Stroh). To note, BII’s influenza research has generated the first mentioning of BII in the general press for scientific successes in context.

Dr. Frank EISENHA BER
Executive Director
Bioinformatics Institute
with discoveries regarding the H1N1 influenza virus mutation-phenotype relationships (see pages 6-9).

Important scientific discoveries are made with contributions from BII researchers. To name just one story, Vivek Tanavde was involved in the discovery of the role of the primate-specific exonic microRNA-198 in wound healing (Nature, 2013, doi: 10.1038/nature11890). We celebrated Chandra Verma having published more than 100 scientific papers during his first ten years at BII. Further hopeful signs that BII is on the right track towards scientific excellence are provided by the general outcome of scientific publications. That has about doubled since 2007 and has reached a new plateau since 2010 with well above 70 scientific publications per year, quite an achievement for about 100 faculty members, scientists, supporting staff, students, etc. working at BII at present. ~72% of all papers published by BII in 2001-2013 have come out during the years 2008-2013 (392 out of 548). Among the papers with impact factor>5, the fraction is even ~80% (97 out of 121). This is the more remarkable since almost no bioinformatics journal falls into this category; thus, we increasingly outreach into general biological/medical publications. Finally, the quality of BII’s research leads to a handful of patents every year and has attracted biotech and pharma companies for collaborations with research collaborations agreements worth of many millions SGD.

BII is an attractive place for spending the sabbatical year. Prof. Andreas Wagner from the University of Zurich was the first professor taking the opportunity during 2012/2013. We also want to increase the number of PhD students in BII. Therefore, we joined with the School of Computer Engineering (SCE) of the Nanyang Technological University (NTU) in Singapore and launched a PhD program for Computational Biology and Bioinformatics that is now open for applicants (see page 49). Interested good students are encouraged to apply.

The annual life cycle of BII culminates in the annual BII Scientific Conference, the BII Scientific Advisory Board visit and the BII Annual Dinner, all during three consecutive days in February/March. Traditionally, the concluding festive evening is crowned with the Added Dimension Lecture where renowned scientists speak about their very personal experiences and views on science, life and society. In 2010, Sir Tom Blundell talked about his endeavours in communal politics, world travel and crystallography. In 2011, Prof. Wong Limsoon informed us about the difficult beginnings of bioinformatics in Singapore and his personal role in the process two decades ago. In 2012, Nobel Laureate Sydney Brenner contemplated about serendipity and the selection of research tasks. In 2013, Prof. Bertil Andersson, President of Nanyang Technological University, made us laugh with anecdotes from his work in the Nobel Committee.

The members of our institute are united in making BII a success story and I invite you to join us in this endeavour that will open new frontiers in biology and other life sciences as well as their applications for the benefit of society. In this context, enhanced cooperation with clinical research and life science-related industry will go hand in hand with the growing reputation of BII for good science.

Executive Director’s Biography
Frank Eisenhaber’s research interest is focused on the discovery of new biomolecular mechanisms with theoretical and biochemical approaches and the functional characterisation of yet uncharacterised genes and pathways. Frank Eisenhaber is one of the scientists credited with the discovery of the SET domain methyltransferases, ATGL, kleisins, many new protein domain functions and with the development of accurate prediction tools for posttranslational modifications and subcellular localisations. He studied mathematics at the Humboldt-University in Berlin and biophysics and medicine at the Pirogov Medical University in Moscow. He received the PhD from the Engelhardt Institute of Molecular Biology in Moscow. After postdoctoral work at the Institute of Molecular Biology in Berlin-Buch (1989-1991) and at the EMBL in Heidelberg (1991-1999), he worked as teamleader at the Institute of Molecular Pathology (IMP) in Vienna (1999-2007). He joined as Director of the Bioinformatics Institute A*STAR Singapore in August 2007 and he guides the research in the Biomolecular Function Discovery division. He is now the Executive Director since August 2013.
Clinical Collaborations of BII

Cancer Biomarker Discovery

BII has developed methods for molecular diagnostics and prognosis of human cancers by implementing integrative medicine, computational and statistical bioinformatics methods. Adequate multivariate and multifactorial models of high-throughput data combined with appropriate de-noising of diverse sources of complex data are used in the biological and clinical interpretation of such data. Data from TCGA/NCI and other massive genome-wide sequencing, gene expression and clinical oncology databases is used. Novel interactive feature selection methods based on genetic tumour aggressiveness grading and survival models are used to identify cancer patient’s clinical sub-groups with different disease recurrence risks. These are being used to (i) discover novel clinically relevant molecular process, pathways, and cellular programs of ovarian cancer (serous ovarian carcinoma), breast cancer and glioblastoma, (ii) stratify cancer patients according a risk of disease development, (iii) identify relatively small number of specific, sensitive and reliable biomarkers (ncRNAs, mRNA, CNV/ SNP) which could identify clinical subtypes of cancers and lead to the development of reliable kits/assays for personalized diagnostics, prognostic and treatment.

Influenza

BII plays an integral role in the national influenza surveillance by providing a tool for detailed analysis of sequences from routine samples but also occasional severe cases. Hospitals send samples of severe respiratory disease cases, for example from patients in intensive care units, to the National Public Health Laboratory of the Ministry of Health for identification of the pathogen. If it is confirmed as Influenza and sequences can be derived from the sample, these sequences are analyzed through BII’s dedicated pipeline FluSurver to quickly identify known molecular markers of disease severity as well as highlight any new mutations. These tools are used locally and internationally by influenza centres that collaborate with the WHO.

Computational Meibography

Dry eye disease is defined as ocular surface dryness and discomfort associated with an abnormal tear film from either tear deficiency or excessive tear evaporation. In healthy patients, oil glands in the eyelids (Meibomian glands) produce lipids that help to reduce tear evaporation. In patients with Meibomian gland dysfunction (MGD) the oil glands are obstructed or degenerated leading to tear evaporation and dry eye symptoms. MGD is highly prevalent with a prevalence rate of up to 60% in...
Asian populations. Meibography is a way to document the health of these glands by using an infrared transmitting filter and video camera mounted on a slit lamp microscope. Using this method, morphological changes, such as dropouts, shortening, dilation and distortion of the meibomian glands can be detected that are characteristic in MGD. In collaboration with the Singapore Eye Research Institute (SERI), we are developing software that can determine features such as number, length, diameter or twistedness of glands. These parameters can be followed up in a longitudinal fashion to identify subtle changes in gland health, which may occur with increasing degree of disease severity and treatment.

**Mutations in disease**

Dr. Terri Young and her colleagues from Duke University and Duke-NUS Singapore are studying hereditary eye diseases through exome sequencing of patients. BLI is helping with the analysis of effects of identified mutations on the protein structure and cellular function and how they possibly could contribute to the disease mechanism.

Together with I2R (A*STAR) and National University Hospital Singapore (Dr. Yong Wei Peng and Dr. Ross Soo), BLI is developing methods to understand the molecular mechanisms and predict the clinical responses to chemotherapy arising from SNP distributions in patients. In a related project with the Singapore General Hospital and National Cancer Centre (Dr. Daniel Tan), methods are being developed to identify the mechanisms that underpin the effects of clinically used kinase inhibitors on silent-SNP distributions.

In a long-term collaboration with clinicians in Vall d’Hebron hospital (Barcelona), Memorial Sloan Kettering (Prof. Jose Baselga/ Dr. Maurizio Scaltriti), the mechanisms that underpin the effects of combining kinase inhibitors and antibodies in breast cancer are being explored.

**Antimicrobial Development**

In collaboration with Prof. Roger Beuerman at the Singapore Eye Research Institute and colleagues at NUS and NTU, BLI has been participating in the successful development of a portfolio of antimicrobial peptides and small molecules, which represent new classes of antibiotics.

A new mechanism whereby these antibiotics work in strong synergy with existing antibiotics, reducing the dosage, has been found. As seen in medical device coatings, there is potential for preservative systems to extend the life of such devices. 3 patents have been granted and another 10 are filed/being filed on various aspects of these antibiotics, including a platform technology disclosure to design such antibiotics. This work is being funded by ~2.78 million SGD from ETPL (A*STAR), ~4.5 million SGD from a TCR grant from NMRC/A*STAR and ~0.97 million SGD from a CBRG grant from NMRC (contact chandra@bii.a-star.edu.sg).

**Interactions of BLI with Industry Partners**

Next generation sequencing methodologies have brought down the cost of sequencing but have also created new challenges in analyzing the large amounts of generated data. The expertise of the Bioinformatics Institute (BII) in various aspects of sequence analysis complements the commercial sequencing services provided by a local biotechnology company, AIT Biotech. A research collaboration agreement facilitates, with significant financial and in-kind support, joint projects on analyzing next generation sequencing data with the company.

We are also collaborating with a major multinational in the personal care sector to engage bioinformatics expertise of BLI for assessing allergy potential of proteins using their amino acid sequence and tertiary structures.

A working relationship has been established with a subsidiary of another major pharmaceutical company through our long term collaborations with the p53 lab on investigating the therapeutic potentials of stapled peptides in inhibiting the translation initiation cascade.

A collaboration has been set up with a slit lamp camera manufacturer to integrate our software into a slit lamp camera. Clinical validation is underway for routine use of our technology in the eye clinics.

BLI is developing automated computational platforms for the industry to characterize the biophysical properties of cells and tissues. Such characterizations may be useful in, for example, studying how various consumer care products may modify the biophysical properties of skin tissues upon application, such as their size and shape, contractility as well as ability to migrate. In particular, we have developed an automated computational platform to characterize the contractility of cells by calculating the traction stresses that cells exert on their substrates. The computational platform is designed to be autonomous and is suitable for industry partners desiring “one-touch” solutions (contact: chiamkh@bii.a-star.edu.sg).

A collaboration has been established with a consumer care company to use the automated computational platform of calculating traction stresses of skin cells and tissues.

Dr. Igor Kurochkin’s team at BLI has developed Exosome Purification Kit which is suitable for the preparation of exosomes from cell culture media and various biological fluids including blood plasma/serum, urine and saliva. The kit is simpler, faster and more efficient than the classical methods and delivers cleaner exosomes than other commercial kits. The kit has been commercialized under the brand name Exo-spin™ Exosome Purification Kit by Cell Guidance Systems.

Dr. Samuel Gan’s team at BLI has developed a genomic DNA and RNA extraction kits licensed to a local SME Quintech enterprise. This kit brings down the cost of nucleic acid purification while providing better yields than kits from a leading manufacturer. Further his team has also developed an Android app DNAApp V1.0 for analyzing DNA sequencing data on a smartphone. An iOS version of the app is also being developed.

The Natural Product Library (NPL) acquired by BLI in 2014 is a valuable resource for discovery of new metabolites, enzymes drugs etc. This is a collection of bioactive molecules isolated from various plants and microbes. The NPL Chemistry and Biology groups work together to provide a complete suite of purification, structure elucidation and large scale production of bioactive molecules along with assays for screening for specific phenotypes. This library is available for outright licensing or for collaborative development projects with interested companies.
**Programme on Human Infectious Diseases**

**Cross-division programme on human infectious diseases**

BII’s unique combination of research groups from different disciplines (gene expression, protein sequence, protein structure, bioimaging) opens the opportunity to cover a wide range of aspects of collaborative research on infectious diseases that are not easily found under the same roof in other institutes. For example, while aspects of protein sequence mutations and viral evolution are covered by the Biomolecular Function Discovery Division, detailed structural simulations and docking studies are executed by the Biomolecular Modeling and Design Division. The Genome and Gene Expression Data Analysis Division contributes, for example, statistical analysis of patient gene expression or genotype to phenotype correlations. Also the Imaging Informatics Division is part of the infectious disease programme through a video surveillance project and image analysis of infected cells. To coordinate our efforts, we have established a cross-division research programme on human infectious diseases. Interested clinicians, industry partners, researchers and health authorities are welcome to contact the programme director, Sebastian Maurer-Stroh (sebastianms@bii.a-star.edu.sg), for collaborations or more information.

Below follows a short summary of the main ongoing infectious disease directions at BII, with several more projects and collaborations coming up in the near future.

**Viruses**

- Molecular influenza surveillance of drug resistance, vaccine efficacy and emerging mutations in collaboration with NPHL/ MOH Singapore, INMEGEN Mexico City, IAL Sao Paulo and WHO CC Australia
- Modeling and simulation of mutations in viral structures
- Virus-host interaction pathways (with NTU Singapore; other project with Paul Ehrlich Institute, Germany)
- NovProt – novel viral proteins derived from alternative reading frames (in collaboration with Oxford University)
- Dengue: Structure-guided in silico drug design of dengue protease inhibitors (in collaboration with Duke-NUS, ETC, Goethe University Frankfurt and University of Ulm)
- Recognition of respiratory disease symptoms from video surveillance data (Cheng Li at BII)
- HIV (whole genome NGS, new recombinants, phylogeny, co-receptor usage, apoptosis induction and drug resistance in collaboration with TTSH and Texas Tech University)
- Coronaviruses (phylogenetic classification of novel strains, immune response to SARS)
- Flu PCR kits (with ETC and TTSH)
- Pathogen Chip for Respiratory Tract Infections (with GIS and UCD)

**Bacteria**

- Peptoidic antimicrobials with a focus on eye infections (in collaboration with Singapore Eye Research Institute Dr. Roger Beuerman (rwbeuer@mac.com), NTU, NUS, Singapore General Hospital Pathology)
- Beta-lactamase classification and bacterial drug resistance
- Genome analysis of novel bacterial strains (drug targets, resistance and virulence factors, phylogenetics in collaboration with NPHL/MOH Singapore and NUHS)
- Legionella F-box proteins modification by and interaction with host cell machinery (Sharmila Adhikari at BII)
- Pseudomonas motility (Chiam Keng Hwee at BII)

**Some Highlights:**

BII continues to contribute strongly to regional and global influenza surveillance efforts through our tool FluSurver (http://flusurver.bii.a-star.edu.sg/). It facilitates analysis of influenza sequences, summarizing occurrence of new mutations, their geospatial context, 3D structure mapping, vicinity to structural interaction sites as well as prediction of phenotypic effects through our literature-curated database of prior known effects of influenza mutations.

In 2013, a new avian-origin H7N9 influenza strain emerged and caused a high profile outbreak in China. BII quickly contributed analyses shared with health authorities and we made our FluSurver tool available to all H7N9 sequence submitters in collaboration with the GISAID database. We also found and characterized a strain that turned out to be a missing puzzle piece in the understanding of the origins of this new virus.

---

**Figure 1.** Analysis of first 3 cases of the H7N9 outbreak using FluSurfer in GISAID. The tool rapidly identifies important host specificity mutations that distinguish this strain from previous avian influenza viruses, allowing it to infect humans.
H7N9 was not the only avian influenza virus infecting humans recently and a BII team found that the highly pathogenic H7N3 avian influenza strain from July 2012 in Mexico acquired an extended cleavage site through recombination with host 28S rRNA. Additional cleavage sites increase the tissue range the virus can infect and recombination with host RNA has never been seen in naturally occurring sequences before. Also in 2013, collaborating teams from WHO CC Australia, University of Melbourne, and BII identified mutations that can cause antigenic drift of the pandemic 2009 A(H1N1) influenza virus in a ferret model.

Besides work on influenza, teams at BII also identified genetic signatures of HIV-1 envelope mediated bystander apoptosis and contributed a study using powerful sequence similarity search methods and in-depth manual analyses to identify remote homologs in many apparently “orphan” viral proteins. Another BII team conducted an interesting bacterial tethering analysis which reveals a “run-reverse-turn” mechanism for Pseudomonas species motility.

Another important infectious disease topic at BII is the long-standing collaboration of the Biomolecular Modeling and Design Division on Defensin-derived antimicrobials with a focus on eye infections together with the Singapore Eye Research Institute (rwbeuer@mac.com), NTU, NUS and Singapore General Hospital. Besides the already available Defensins Knowledgebase (http://defensins.bii.a-star.edu.sg/), new molecules are now in pre-clinical testing and show activity against a broad spectrum of clinical strains of both gram positive and gram negative bacteria including Pseudomonas, MRSA, fungi (Fusarium, Candida) and Tb. These molecules are fast acting, stable, and do not give rise to resistance and additionally their toxicity is very low, thus characterizing them with a high therapeutic index. Some of these compounds demonstrate a high level of synergy with existing antibiotics, leading to potential reductions in dosage. Clinical trials are planned for eye infections.

SELECTED PUBLICATIONS

Programme on Cancer Biomarkers

Vladimir A. Kuznetsov

Recent progress in biotechnology, nanotechnology, genomics, proteomics and bioinformatics has opened new avenues for understanding the pathophysiology of cancer as well as for the discovery and validation of new cancer-specific molecular targets and clinically-significant biomarkers. To address the above-mentioned issues, we at the Bioinformatics Institute bring together computer scientists tackling real-world biological and biomedical problems, and biologists and clinical researchers using basic analytical methods of computational genomics, systems biology and bioinformatics tools in the post-genomic era.

We explore computational cancer genomics, statistical bioinformatics, microarray and next-generation sequencing (NGS) data analysis, computational imaging, perform wet-lab validation of in silico predictions and subsequently translate the acquired knowledge towards technology development (Figure 1). Data-driven feature selection procedures, mathematical analysis and modeling and disease model based approaches are developed towards discovery of novel biomarkers for accurate molecular classification of the cancer subtypes, early detection of cancers, evaluation of personalized risk of disease recurrence and selection of perspective molecular targets for optimal therapeutic intervention.

Trends, Opportunities, Challenges

Vladimir A. Kuznetsov

Recent progress in biotechnology, nanotechnology, genomics, proteomics and bioinformatics has opened new avenues for understanding the pathophysiology of cancer as well as for the discovery and validation of new cancer-specific molecular targets and clinically-significant biomarkers. To address the above-mentioned issues, we at the Bioinformatics Institute bring together computer scientists tackling real-world biological and biomedical problems, and biologists and clinical researchers using basic analytical methods of computational genomics, systems biology and bioinformatics tools in the post-genomic era.

We explore computational cancer genomics, statistical bioinformatics, microarray and next-generation sequencing (NGS) data analysis, computational imaging, perform wet-lab validation of in silico predictions and subsequently translate the acquired knowledge towards technology development (Figure 1). Data-driven feature selection procedures, mathematical analysis and modeling and disease model based approaches are developed towards discovery of novel biomarkers for accurate molecular classification of the cancer subtypes, early detection of cancers, evaluation of personalized risk of disease recurrence and selection of perspective molecular targets for optimal therapeutic intervention.

Figure 1.

Cancer Biomarker Discovery via Integrative Medicine and Bioinformatics Analysis of Sequencing, Gene Expression and Clinical Oncology Data

Vladimir A. Kuznetsov, Anna V. Ishina

This project investigates molecular diagnosis and prognosis of human cancers via developing and implementing integrative medicine, computational and statistical bioinformatics methods. Adequate multivariate and multifactorial models of high-throughput data de-noising of diverse sources of complex data are used to correct biological and clinical interpretation of the data and hence disease prognosis. We use TCGA/NCI/NIH and other massive genome-wide sequencing, gene expression and clinical oncology data to develop a novel feature selection methodology based on genetic tumour aggressiveness grading and survival models.

The important components of our studies are long-term collaboration projects with A*STAR Biomedical Institutes, clinicians, clinical researchers of Singapore and other countries. In Singapore, we are successfully developing our joint projects with IMCB/A*STAR, IMB/A*STAR, Immunology Network/A*STAR, NTU, NUS, National Cancer Center Singapore, Singapore General Hospital, KK Women's and Children's Hospital, Duke-NUS Graduate Medical School. We are holding the joint projects with SUNY Upstate Medical University, Syracuse, NY, USA (Wenyi Feng), The Methodist Hospital Research Institute in Houston, Texas, USA (Nancy Jenkins, Neal Copeland), and QIMR Berghofer Medicine Res. Institute, Brisbane, Australia (Martin F. Lavin), INSERM U869 - Universite Bordeaux Segalen, Institut Euopeen de Chimie et Biologie (J.-L. Mergny), as well as with a number of our industry partners.

Sense-antisense gene pairs meta-analysis identifies next generation regulatory pathways and biomarkers for classification and prognosis of breast cancer patients

Oleg Grinchuk, Efthimios Motakis, Surya Pavan Yenamandra, Piroon Jenjaroenpun, Qw Ghim Song, Tang Zhiqin, Anna V. Ishina, Vladimir A. Kuznetsov

Bioinformatics Institute, Agency for Science Technology & Research (A*STAR), Singapore

Many recent studies strongly support the idea that gene pairs and their transcripts with sense-antisense overlapping regions utilize their own specific co-regulatory mechanisms which are not applicable to the genes without sense-antisense overlaps. Such sense-antisense gene pairs (SAGPs) may have their own specific structural and regulatory roles in normal and cancerous cells. Original workflow for multivariate correlation analysis and highly intensive survival analysis was applied to massive breast cancer microarray data and validated by QRT-PCR. We identified the set of robustly correlated breast cancer-relevant (BCR) SAGPs. They demonstrated predominantly positive correlation pattern in breast tumors as compared with normal breast tissue. We showed that the observed positive shift of correlations of gene expressions in breast tumors can be independent from DNA copy numbers variation. Unexpectedly, survival analysis using the recently developed 2-D RDDg procedure applied to BCR SAGPs revealed their extreme efficiency in tumors reclassification within diverse breast tumors subpopulations which are currently believed to be relatively homogeneous. The novel phenomenon discovered on the example of BCR SAGPs in fact may reflect the global and yet specific changes in gene expression in the large sub-fraction of genome/transcriptome associated with sense-antisense gene architectures. These changes are strongly associated with breast tumors as well as with distinct breast tumors subpopulations. Therefore, SAGPs as a distinct part of transcriptome could be a novel promising source of breast tumors biomarkers.

Towards personalized cancer chemotherapy: Selecting non-synonymous SNPs (nsSNPs) for possible effects on gemcitabine response using comprehensive integractome, pathway, evolutionary and structure analyses

Limviphuvadh V, Ooi HS, Jenjaroenpun P, Xiang S, Frank Eisenhaber, Sebastian Maurer-Stroh

Collaborators: Konishi Fi(Tokyo Tech), Mah TL, Tong JC (I2R), Yong WP, Soo RA (NUH).

This study aims to select nsSNPs potentially related to patient outcome for gemcitabine treatment against non-small cell lung cancer in the Singaporean population. First, we collected proteins directly involved in gemcitabine transport/metabolism by literature search. Most of them were associated with the KEGG pyrimidine metabolism pathway. Then, we searched for additional proteins that can be linked to this pathway from protein-protein interaction data. Altogether, we found 178 proteins with 2661 nsSNPs. Next, we filtered the nsSNPs for common occurrence among the Singaporean population, and required them to be in a region with a known structure or high similarity to a known structure for structural modeling. Finally, we rank the nsSNPs by their estimated effects on protein stability and vicinity to functional sites/ligands[1]. The new candidate nsSNPs will be screened from a Singaporean patient cohort with known gemcitabine treatment outcome, and statistically tested for individual or epistatic contributions to drug response.

POLA2+1747=GG/GA (POLA2 G583R) a novel determinant of
survival outcome and mortality in Non Small Cell Lung Cancer (NSCLC) patients treated with Gemcitabine (ETPL ref: 12R/Z/06960; TD2011067).

**MECOM/Evi1 as a perspective clinical biomarker of ovarian cancer**

Arsen Batagov, Surya Pavan Yenamandra, Anna V. Ivshina, Vladimir A. Kuznetsov

Epithelial ovarian cancer (EOC) is the deadliest gynecological malignancy. The overall 5-year survival rate of women infected by EOC of only 46%, despite improvements in surgical techniques and therapeutics. The low survival rate is explained by the lack of diagnosis of EOC at an early stage, acquired resistance to chemotherapy, and the lack of effective therapies. By analyzing the data on over 800 patients from four cohorts we discovered that MDS1-EVI1 complex locus (MECOM) is amplified in over 80% of EOC tumors, and expression of transcripts from could be used as strongly specific (Sp) and sensitive (Se) biomarkers for EOC detection (Se=100%, Sp=80%), as well as differential diagnostics, e.g. to discriminate EOC from breast cancer metastases in the ovaries (Sp=100%, Se=94%) [1]. Uniquely, MECOM could be used to classify the patient to one of four prognostic groups (post-operative survival (months): a) 23, b) 7-25, c) 183-167, d) 83, as well as to predict primary chemotherapy outcome (P=0.0024).

1. Patent application BII/P/06614/00/SG. Published in May 2013. http://patentscope.wipo.int/search/en/WO2013074044 (Supported by grant of GAP program of ETPL; A*STAR)

**Computational image analysis of prostate cancer**


**Collaborators:** Kalaw E, Chong K T, Giron D (Tan Tock Seng Hospital, Singapore)

Prostate cancer is the third most common male cancer in Singapore. Accurate diagnosis is essential for early detection. An important strategy is to identify potentially dangerous pre-malignant prostate lesions like atypical small acinar proliferations (ASAP) or high-grade prostatic intra-epithelial neoplasia (HG-PIN) that may become higher-risk cancers when they become malignant.

We aim to build a computer aided diagnosis system to help pathologists in their predictive and diagnostic assessments. We aim to detect local image patterns that are relevant and important for the pathologist, and highlight these regions so that pathologist could easily focus on these regions for prompt diagnosis.

**Automated characterization of biophysical properties of cells and tissues**

Chiam Keng Hwee

We are developing automated computational platforms for the industry to characterize the biophysical properties of cells and tissues. Such characterizations may be useful in, for example, studying how various consumer care products may modify the biophysical properties of normal and cancer cells and tissues upon application, such as their size and shape, contractility as well as ability to migrate. In particular, we have developed an automated computational platform to characterize the contractility of cells by calculating the traction stresses that cells exert on their substrates. The computational platform is designed to be autonomous and is suitable for industry partners desiring “one-touch” solutions (contact: chiamkh@bi.a-star.edu.sg). A collaboration has been established with a consumer care company to use the automated computational platform of calculating traction stresses of skin cells and tissues.

**Understanding and Developing Therapies in Oncology**


**Development of exosome based biomarkers for non-invasive diagnosis and evaluation of progression and treatment efficacy of cancer**

Jenjaroenpun P, Vladimir A. Kuznetsov, Kurochkin I V

Exosomes are nanosized extracellular vesicles secreted by multiple cell types. Tumor cells release large quantities of exosomes into blood and other body fluids. Since exosomes possess characteristic protein and RNA signatures of host tumor cells, analysis of exosomes in various body fluids can be potentially utilized for non-invasive cancer diagnosis. Toward this goal we analyze the complete transcriptomes of exosomes secreted by various cancer cell lines using next generation sequencing techniques (RNA-Seq). Computational approaches for comparison of expression levels of RNA across different samples and experiments using next generation sequencing are still under active development. We are customizing bioinformatics workflows for the analysis of exosomal RNA. Biomarkers discovered with the use of cancer cell lines will be tested on patients’ blood plasma exosome samples. Tumor cell secreted exosomes are very rare in the circulation where exosomes secreted by normal tissues are predominant. Therefore we are developing technological platforms for selective capturing of tumor cell derived exosomes to increase sensitivity of biomarker detection. Successful capturing of cancer exosomes will also facilitate detailed molecular and functional analysis of their biological properties.

**Collaborators:** Basegia J, Scalfritt M, Violeta Sera (Ival of Hebron Memorial Sloan Kettering Cancer Centre), Tan D, Sabapathy K (National Cancer Centre/SGH), Prof. Sir David Lane/ Brown C/ Ghadessy F (p53Lab A*STAR), ETC

We have a cross-disciplinary effort exploring biomolecular mechanisms that are associated with deregulation of cellular processes in cells and investigate the development of novel therapeutics in oncology that includes: use of existing drugs in combination therapy; design of novel small molecules/peptides/stapled peptides to develop new and novel inhibitors of protein-protein interactions; understanding mechanisms that govern the response of different drugs to SNP among patients; predicting the emergence of resistance mutations. Together with the p53Lab of Prof. Sir David Lane, we have developed novel stapled peptides that are currently being pursued as next generation therapeutics targeting p53 and the translational cascade through eIF4E; a major breakthrough has been the finding that the stapled peptides potentially work against mutants that are resistant to small molecules. The work has raised several grants from A*STAR (totaling ~ SGD 6 million and an agreement with a pharmaceutical company to co-develop stapled peptides) to develop inhibitors of other pathways in oncology and to develop predictive platforms together with IP disclosures. Working with colleagues in Barcelona and MSKCC, we have identified that combining lapatinib and herceptin has a synergistic effect in inhibition of Her2 in Her2-overexpressing breast cancers; work that was validated in a large scale worldwide clinical trial. This leads to the development of mechanisms underlying synergy between two commonly used antibodies and this is now the basis for designing new antibodies against Her2 and c-Met.

**SELECTED PUBLICATIONS**

The Bioinformatics Institute (BII) was set up by the Agency for Science, Technology and Research (A*STAR) in July 2001. Located in the Biopolis, it was conceived as the computational biology research and postgraduate training institute as well as a national resource centre in bioinformatics within the Biomedical Research Council (BMRC) of A*STAR.

It was re-launched in late 2007 as a research institute for biomolecular mechanism discovery guided by computational biology methods.

The spectrum of research activities in BII includes theoretical approaches aimed at understanding biomolecular mechanisms that underlie biological phenomena, the development of computational methods to support this discovery process, and experimental verification of predicted molecular and cellular functions of genes and proteins with biochemical methods.

Together with the BMRC/ A*STAR research institutes and multinational R&D organizations in the Biopolis, the BII is situated in a conducive environment for exchange of scientific knowledge and friendly interaction that will prompt greater collaborations, and position the Biopolis as a notable biomedical R&D hub in Asia and in the world.

Scientific Advisory Board

The Executive Director of BII is advised by a Scientific Advisory Board consisting of eminent scientists in the field of bioinformatics/computational biology and experimental life sciences, with respect to the institute’s research directions, recruitment of staff and international research collaborations.

The presiding members are:

**Professor Sir Tom BLUNDELL**
(Chairman)
Director of Research and Professor Emeritus
Department of Biochemistry
University of Cambridge

**Professor Jason SWEDLOW**
Professor of Cell Biology
Howard Hughes Medical Institute Investigator
Department of Cell Biology
Harvard Medical School

**Professor Tom RAPOPORT**
Professor of Cell Biology
Howard Hughes Medical Institute Investigator
Department of Cell Biology
Harvard Medical School

**Professor Koichiro TAMURA**
(from April 2014)
Director, Research Center for Genomics and Bioinformatics
Chairman, Department of Biological Sciences
Tokyo Metropolitan University
Overview of Research Divisions

**Basic Research Divisions**

- **Genome & Gene Expression Data Analysis**
  (genome/transcriptome sequence and expression, RNA biology)
  Vladimir A. KUZNETSOV
  Igor KUROCHKIN
  Brian J. PARKER
  Vivek TANAVDE

- **Biomolecular Function Discovery**
  (from protein sequence to function)
  Birgit & Frank EISENHABER
  Dmitry IVANOV
  Sebastian MAURER-STROH

- **Biomolecular Modeling and Design**
  (3D structure of proteins and ligand interaction)
  Chandra VERMA
  Igor BEREZOVSKY
  Peter J. BOND
  FAN Hao

- **Imaging Informatics**
  (automated study of cellular/tissue images with labeled molecules)
  LEE Hwee Kuan
  CHENG Li
  CHIAM Keng Hwee
  LOO Lit Hsin

**Translational Research Division**

- **Natural Product Biology**
  NG Siew Bee

- **Natural Product Chemistry**
  Yoganathan KANAGASUNDARAM

- **Antibody & Product Development**
  Samuel K.E. GAN
Genome location of R-loop-mediated hotspots suggests their critical role in AID-dependent mutations and translocations

Wongsurawat T, Jenjaroenpun P, Kuznetsov VA
Bioinformatics Institute, Agency for Science Technology & Research (A*STAR), Singapore

R-loop structure is RNA-DNA hybrid of nascent RNA transcript hybridized to DNA template, leaving the other DNA strand unpaired. Activation-induced cytidine deaminase (AID) mediates class switch recombination in the immunoglobulin (Ig) genes in the mammalian, where AID acts on ssDNA provided by R-loop. However, our computational predictions (Wongsurawat et al, 2012) and recent experimental data (Aguilera et al, 2012; Ginno et al, 2013) suggest that R-loops and AID targets could be found not only in the Ig genes, but also in many hundreds genes of the genome. It is tempting to speculate that in mammalian genome, genetic lesions generated by AID-dependent mutation/translocation (ADMT) could be initiated by transcriptional R-loop formation mechanism. To test this hypothesis, the R-loop identification on a genomic scale was performed using our in silico R-loop model (Wongsurawat et al, 2012). We predicted a localization of R-loop forming sequence (RLFS) in the human (http://rloop.bii.a-star.edu.sg) and mouse genome. In previous studies, RLFSs have been found in Ig, Myc, Actb, Bcl6, Rhoh and Pim1 genes. In all these cases the RLFS loci predicted by our model were co-localized with experimentally-defined R-loop loci. We found RLFS in demethylated regions of the Dazl and Foxo1 genes that functions are essential in a maintenance of pluripotency (Figure 1) and play a role in this cell differentiation. This finding suggests that R-loop structures might be not only a AID-mediated mutation target, but also a target of AID-mediated stem cell reprogramming. At the genome scale in human, we predicted 3,605 genes, in which RLFSs were co-localized with AID binding sites. Importantly, our model predicts the RLFSs in many known diseases-related complex loci that contain mutations, SNPs and translocations associated with cancer and many other diseases, including diffuse large-cell B-cell lymphoma and leukaemia. Our findings suggest that transcriptional R-looping and association with AID/APOBEC targeting can be common phenomena in mammalian genomes. This analysis supports our hypothesis that R-loops could provide ssDNA segments of RLFS-positive genes and its transcripts for AID/APOBEC-mediated targeting and editing (Wongsurawat et al, submitted). Our data analysis also studies suggest that R-loops could play a casual role in orchestrating ADMT and in initiation of gene alterations and genome instability. The knowledge of RLFS and AID/APOBEC binding regions may provide novel therapeutic approaches to cancer and other genetic diseases.
Role of MECOM/EVI-1 in High-Grade Serous Ovarian Cancer

Arsen O. Batagov1, Surya Pavan Yenamandra1, Emilie A. Bard-Chapeau2, Pironen JENARDENPUN, Anna V. Ivshina1, Vladimir A. Kuznetsov2

1Bioinformatics Institute, A*STAR, Singapore
2IMCB, A*STAR, Singapore

For past 30 years, ovarian cancer mortality rate has remained high, despite considerable efforts directed toward this disease, because most of the patients are diagnosed with high-grade serous ovarian cancer (HG-SOC) at stages when the disease is almost incurable. The present classification system of HG-SOC, which implies the linear progression model of the tumor, fails to be biologically relevant and clinically significant. By contrast, such characteristics as slow asymptomatic progression of HG-SOC, cell shedding with early metastasis into multiple peritoneal organs, along with massive genomic and epigenetic variation, are strong indicators of HG-SOC parallel progression. We discovered that MECOM locus and its product, EVI1 transcription factor, could dictate genome instability and, via a specific pathway, play a key role in generating the clones and their rapid parallel progression. With the help of EVI1, MECOM propagates itself in an explosive onset of copy number variation in the HG-SOC genome. In addition, EVI1 transcriptionally activates a distinct regulatory pathway consisting of six functional branches, which correspond to six genetic programs of tumor development. Two of them are direct reflections of the embryogenic processes, in which EVI1 participates. This pathway is co-amplified with MECOM and thus these loci are fixed in the genotype as a single functional unit, which drives HG-SOC cell progression.

Meta-analysis of transcriptome reveals let-7b as an unfavorable prognostic biomarker and predicts molecular and clinical subclasses in high-grade serous ovarian carcinoma.

Tang Z1, Ow GS, Thiery JP2, Anna V. Ivshina1, Vladimir A. Kuznetsov2

1Bioinformatics Institute, A*STAR, Singapore
2Department of Biochemistry, National University of Singapore, Singapore

High-grade serous ovarian carcinoma (HG-SOC) is a heterogeneous, poorly classified, lethal disease that frequently exhibits altered expressions of microRNAs. Let-7 family members are often reported as tumor suppressors; nonetheless, clinicopathological functions and prognostic values of individual let-7 family members have not been addressed in HG-SOC. In our work, we performed an integrative study to investigate the potential roles, clinicopathological functions and prognostic values of let-7 miRNA family in HG-SOC. Using microarray and clinical data of 1,170 HG-SOC patients, we developed novel survival prediction and system biology methods to analyze prognostic values and functional associations of let-7 miRNAs with global transcriptome and clinicopathological factors. We demonstrated that individual let-7 members exhibit diverse evolutionary history and distinct regulatory characteristics. Statistical tests and network analysis suggest that let-7b could act as a global synergistic interferator and master regulator controlling hundreds of protein-coding genes (Figure 2A-B). The elevated expression of let-7b is associated with poor survival rates, which suggests an unfavorable role of let-7b in treatment response for HG-SOC patients. A novel let-7b-defined 36-gene prognostic survival signature outperforms many clinicopathological parameters, and stratifies HG-SOC patients into three high-confidence, reproducible, clinically distinguishable intermediate- and high-risk, with 5-year overall survival rates of 56-71%, 12-29% and 0-10%, respectively (Figure 2C). Furthermore, the high-risk and low-risk subclasses exhibit strong mesenchymal and proliferative tumor phenotypes concordant with resistance and sensitivity to primary chemotherapy. Our results have led to identification of promising prognostic markers of HG-SOC, which could provide a rationale for gene-based stratification of patients and optimization of treatment regimes.

SELECTED PUBLICATIONS

1. Tang Zhiqun, Ow Ghim Siong, Thiery Jean Paul, Ivshina Anna Vladimirovna, Meta-Yenamandra Surya, EVI1 and let-7b transcriptome reveals let-7b as an unfavorable prognostic biomarker and predicts molecular and clinical subclasses in high-grade serous ovarian carcinoma. International Journal of Cancer, Accepted manuscript online: 3 JUL 2013 13:38AM EST | DOI:10.1002/ijc.28371 Abstract


Over the past decade, numerous cDNA cloning and sequencing projects and genome-tiling array analyses revealed that the mammalian genomes are almost entirely transcribed leading to the generation of the tens of thousands of ncRNAs. Diverse ncRNA species include short miRNAs, piRNAs and much longer ncRNAs (lncRNAs). Few studies performed so far on biological role of lncRNAs suggest extremely diverse mechanisms of action of this class of molecules. Computational prediction of lncRNA function thus faces a serious challenge of decoding the information contained within the sequence of these molecules. We aim to develop a novel computational framework for functional and structural analyses of lncRNAs that integrates high-throughput data related to transcriptional control of lncRNA, secondary structure of these molecules, evolutionary conservation, and functional annotation of co-expressed protein-coding RNAs. The predictions are assumed to be validated through wet-lab experimentation.

In particular, we are interested in exploring the role played by lncRNAs secreted via exosome nanovesicles and their utilization as potential biomarkers for various diseases including cancer, inflammation and metabolic disorders. We are also investigating lncRNAs responsive to environmental stress conditions and their potential link to cellular senescence and cancer.

Characterization of lncRNAs responsive to oxidative stress: link to cellular senescence and cancer

It is still unclear what fraction of thousands of lncRNAs produced by mammalian cells is functional. Cellular response to stress sets a valuable paradigm for revealing functional lncRNAs because during the stress response cells engage essential self-protecting mechanisms and are unlikely to produce unnecessary transcripts. To detect RNAs regulated by oxidative stress, we utilize the next generation RNA sequencing technology (RNA-Seq). RNA-Seq provides unprecedented resolution, allowing us to accurately monitor expression output of each genomic locus including non-protein coding regions. Using RNA-Seq, we uncovered a large number of lncRNAs whose expression is highly responsive to oxidative stress (Figure 1). Oxidative stress is known to be implicated in the development of a number of pathologies, including cancer. In this connection, Ras-induced cell senescence led to down-regulation of a majority of the stress-induced lncRNAs. Several stress-induced lncRNAs were found to have significant impact on the colony formation by semi-immortalized human epithelial cells or inhibit growth of human normal fibroblast cells (Figure 2). Microarray analysis revealed down regulation of genes involved in cell adhesion and integrin binding (Figure 3) as a potential mechanism of stress lncRNAs-mediated cell phenotype. Remarkably, the expression levels of a set of the stress-responsive correlated with breast cancer progression from G1 to G3 stages, with p53 status and could separate basal and
luminal breast cancer cell lines. Thus oxidation-induced lncRNAs might possess the regulatory potential in tumorigenesis and could serve as biomarkers for specific cancer grades. The project was performed in collaboration with Computational Analysis of Genome Complexity, Transcription Regulation and Cellular Phenotypes Team, BII.

Figure 3. Heat map depicting differentially expressed genes upon overexpression of lncRNA J-80 in normal human fibroblasts.

Towards understanding the complete transcriptome of secreted exosome microvesicles

Extracellular vesicles have recently attracted a great interest after the discovery that they contain various RNA species, including mRNA and miRNA. It was demonstrated that mRNA was functional in target cells as it could be translated into proteins. Cellular mRNAs vary in length from 400 nt to 12,000 nt with the average size of transcripts 2,100 nt. However, the majority of RNA present in exosomes has a size between 25 and 700 nt. This suggests that either mRNAs are not major species of RNA delivered by exosomes or that exosomal mRNA is fragmented. Analyzing public microarray data in regard to expression ratio between all probes representing a given transcript for both cellular and exosomal RNAs we demonstrated that mRNAs secreted by human cells transport mostly mRNA fragments and these fragments are largely derived from the 3'-untranslated regions (3'UTR). These calculations have been verified experimentally using qRT-PCR. Our findings suggest the need to reassess the assumption that RNA messages delivered by exosomes are translated into proteins by recipient cells. Instead we propose that RNA fragments delivered by exosomes play regulatory roles. Indeed, 3'UTRs are potentially rich in regulatory activities as they contain elements that confer subcellular localization of mRNA and are enriched in miRNA-binding sites. Thus secreted 3'UTRs may regulate localization, stability and translational activity of mRNAs in target cells.

Tumor cells release large quantities of exosomes into blood and other body fluids. Since exosomes possess characteristic protein and RNA signatures of host tumor cells, analysis of exosomes in various body fluids can be potentially utilized for non-invasive cancer diagnosis. Towards understanding information carried by tumor cell secreted exosomes we have analyzed RNA content of the vesicles secreted by breast cancer cell lines using RNA-Seq. The analysis revealed that exosomes, in addition to mRNAs and miRNAs, transport a number of unannotated long and short noncoding RNAs. Interestingly, the majority of mRNAs was represented by short (80-100 nt) fragments. We observed a number of transcripts in exosomes that were undetectable in the host cells suggesting an existence of a dedicated mechanism for selection of the secreted molecules. The newly identified exosomal RNA species may play regulatory functions in target cells and mediate biological effects and function previously to exosomes. The project is being performed in collaboration with A.O. Batagov, P. Jenjaroenpun and V.A. Kuznetsova, Computational Analysis of Genome Complexity, Transcription Regulation and Cellular Phenotypes Team, BII.

SELECTED PUBLICATIONS

2. Batagov AO, Kurochkin IV. Exosomes secreted by human cells transport large mRNA fragments that are enriched in the 3'-untranslated regions. Biol Direct 2013 Jun 7;8:12.
4. Mizuno Y, et al. Tysnd1 deficiency in mice interferes with secretion (ECER) and their location. The larger is ECER, the stronger is the tendency of the probes to be localized to the 3'-end of the transcripts.

Principal Investigator’s Biography

Igor Kurochkin joined the Bioinformatics Institute in 2009. He earned his Ph.D. in molecular biology from the Institute of Molecular Biology and Genetics in Kiev. After postdoctoral work in the School of Pharmaceutical Sciences at Toho University, Japan (1990-1993), he joined the Holland laboratory of American Red Cross, MD as a visiting research fellow supported by the International Fellowship from the Fogarty International Center, NIH (1993-1995). During 1996-2002, he was a research scientist at Chugai Pharmaceutical Co., Ltd. (now a member of the Roche group). He returned to the academic sector as a research scientist in RIKEN Genomic Sciences Center, Japan (2002-2009).
Families of conserved structural RNAs:

In [2] and [3] we reported the first human genome-wide identification of families of structural RNAs, based primarily on the evolutionary signal across deep multiple alignments of mammalian species. A critical aspect in the success of such cluster analyses is the distance measure between the elements of interest, for which we developed a semi-metric between generative hidden state models of structural RNAs derived from an estimate of Kullback-Leibler divergence by a Monte Carlo approach, but incorporating and controlling for the varying statistical significance of the pairs. Functional hypotheses were generated by enrichments in ENCODE datasets. This study revealed multiple regulatory families of significant interest including the largest new family which consisted of structured ARE elements in many key immunity genes. Many of these elements have recently been independently shown to be controlled by the RNA-binding protein Roquin. Other key families are under ongoing experimental and computational investigation, including a family of presumably regulatory hairpins in the 3'UTR of the key methylation gene MAT2A, for which the structure has been validated by probing experiments (see figure 1). MAT2A is involved in production of the primary methyl donor, SAM, and is known to be post-transcriptionally regulated, and is a potential drug target.

Mechanisms of microRNA targeting:

Using a predictivist approach, we are investigating the determinants and mechanisms of microRNA targeting. We analysed the static sequence–based flanking determinants of microRNA binding and degradation response from combined Argonaute (AGO) PAR-CLIP and expression data using non-parametric predictive models [4]. This revealed that e.g. the determinants of RISC/AGO binding differ from those of the subsequent degradative response (see Hafner et al. Methods 58 (2012) 94105). Moreover, this work led to the first microRNA target prediction model not based on expression data (c.f. TargetScan) but rather on AGO CLIP-seq- the “Antar” predictor leading to improved predictive performance, and a key tool for our subsequent research.

We are also studying the dynamic behaviour of microRNA targeting and, in particular, feed-back mechanisms between regulatory systems. In [5], we studied the time series response of microRNA transfection with the pioneering use of functional data analytic (FDA) approaches to this problem (FDA is a specialised subfield of statistical methods that extends the usual multivariate approaches to continuous functional data). This study incorporated a distance measure based on the functional inner product of the first derivative of the normalised time series curve shape, to focus the study on general features of the expression response following perturbation. Functional PCA revealed a multiphasic response following miRNA transfection which we hypothesised was due to feed–forward regulation. Based on these results, we are currently completing a higher resolution study on miR-9, in collaboration with Copenhagen University, using a more physiologically meaningful locked nucleic acid (LNA)-based microRNA inhibition perturbation. Current results confirm the multiphasic signal as a feed–forward response and has revealed multiple regulatory modules controlled by miR-9, amongst other findings.

Role of m5C RNA methylation in post-transcriptional regulation:

In [6] we reported the discovery of a potential role of 5-methyl cytosine (m5C) RNA methylation in Argonaute (AGO) binding and, presumably, microRNA targeting. This is an emerging field of increasing importance. Currently, we have developed a differential methylation model optimised for the proportion statistics inherent in the bisulphite assay used, and further research in optimising this model is planned. In collaboration with the Australian National University, ongoing studies are using replicated data to understand differential RNA methylation between different tissues and conditions.

Role of histone variants and nucleosome positioning in regulatory mechanisms:

In collaboration with the John Curtin School of Medical Research (JCSMR) at ANU we have been investigating the interaction of histone variants (and epigenetic mechanisms more generally)
with post-transcriptional regulation. We are using functional data analytic techniques (FDA) to model the variance of nucleosome displacements. We have shown that variance of the nucleosome positioning signal can be a sensitive indicator of functional regulatory regions, and have used, uniquely, a functional data analytic approach based on a Fourier basis, to sensitively analyse this signal and its correlation with histone variant and histone modification signals, to identify regions of regulatory importance. Initial results have identified a role of the histone variant H2A.Z in CTCF binding [7]. Another approach, used in [8], is frequency domain space/space-frequency analysis to sensitively detect weak nucleosome signals in degraded DNA samples, although this method has application more generally.

<table>
<thead>
<tr>
<th>SELECTED PUBLICATIONS</th>
</tr>
</thead>
</table>

**Principal Investigator’s Biography**

Brian John Parker completed a Bachelor of Medicine/Bachelor of Surgery (MBBS) and a Bachelor of Science (Mathematics) at the University of Queensland, followed by a PhD (Computer Science) at the University of Sydney, Australia. Postdoctoral studies included research in computational biology/machine learning at Australian National University (ANU) investigating predictive models for structural RNA detection and theoretical analysis of computational validation methods; this was followed by a three year fellowship at the Bioinformatics Centre at Copenhagen University to study hidden state models in computational biology with Prof. Anders Krogh and comparative genomics with Prof. Jakob Pedersen, funded by an Excellence Programme Statistics Network Postdoctoral Fellowship; and, most recently, at the John Curtin School of Medical Research (JCSMR) at the Australian National University investigating the interaction of histone variants (and epigenetic mechanisms more generally) with post-transcriptional regulation, and the role of RNA modifications (esp. RNA methylation) in post-transcriptional regulation.

---

**Figure 1.** Example of a family of conserved hairpins in the human gene MAT2A found by SCFG modeling, demonstrating:

(a) multiple sequence alignment across the six paralogous family members, with substitution evidence for conserved RNA structure across hairpins highlighted in green for compensatory double substitutions and blue for compatible single substitutions.

(b) computationally predicted RNA structure of hairpin A using a phylo-SCFG model; and

(c) Experimental confirmation of RNA structure of hairpin A by inline probing analysis of a MAT2A construct. RNA cleavage products resulting from spontaneous transterfentation (TE) were resolved by denaturing 10% PAGE. Other lanes: (NR) No reaction; (T1) partial digest with RNase T1; (−OH) partial alkaline digest; (Pre) precursor RNA. Selected bands in the T1 lane are labeled with the positions of the respective 39-terminal guanosyl residues, according to the numbering used for hairpin A in panel (b). Filled bars correspond to the positions within hairpin A that are predicted to be largely base-paired, while the open bar corresponds to positions within the putative loop sequence. Arrowheads indicate putative bulged nucleotides C50 and A55.
miRNA regulation of MSC differentiation

miRNAs are increasingly gaining importance as regulators of various cellular functions. Our group has systematically catalogued miRNA expression in differentiating MSC derived from different tissues like human embryonic stem cells, fetal limb stem cells as well as bone marrow. We observe interesting patterns in the miRNA expression and expression of their target genes in MSC from different sources. These patterns in mRNA & miRNA expression explain some of the functional differences observed in these cells in vitro differentiation assays.

We are also investigating the translational state of the MSC transcriptome to further understand the extent of miRNA regulation in MSC. In collaboration with Prabha Sampath at Institute for Medical Biology (IMB), our group has developed an integrated approach to more accurately identify genes regulated at the level of RNA translation (Lee 2012). Combining the translational status of differentiating MSC with their miRNA profiles will enable us to better answer the question: To what extent is miRNA regulation critical for MSC differentiation?

Translational regulation of cell differentiation and function

We are developing and applying tools to analyze and interpret microarray & next generation sequencing data to understand translational regulation of stem cells. Translational regulation appears to play an important role in adipogenic differentiation of MSC. Methods developed to study such processes in MSC are also being applied to other cell types. This approach also results in the identification of miRNAs & other non coding RNAs involved in cellular differentiation. In collaboration with Prabha Sampath, we were able to identify a miRNA specifically expressed in glioblastoma stem cells (Chan, Cell Reports, 2012) while looking for translationally regulated genes. With our help the same group also identified a unique miRNA-mRNA switch that regulates wound healing (Sundaram, Nature 2013). In collaboration with Leah Vardy at IMB we are studying differential translation of alternately spliced isoforms in embryonic stem cells. Translational regulation of stem cells is an exciting area gaining importance as the role of non coding RNA in regulating cellular processes is better understood.

Studying the zebrafish miRnome

In collaboration with S. Mathavan at the Genome Institute of Singapore (GIS), we are also involved in studying the zebrafish miRnome using next generation sequencing. We have successfully cataloged tissue specific expression of miRNAs in zebrafish and identified the targets of these miRNAs. This is the first study to systematically catalog miRNA expression in a tissue & sex specific manner for this important model organism and will yield important insights on the role of miRNAs in tissue development of zebrafish. Further we have also identified novel miRNAs expressed in zebrafish. When complete, this study will more than double the numbers of known miRNAs in zebrafish. These miRNAs can also be useful biomarkers for identification of cells differentiating into these tissues.

Developing embryonic stem cells as cellular models for screening toxic drugs

Understanding cellular differentiation at the signaling level also enables us to predict cellular responses to external stimuli & toxins. In collaboration with Suzanne Kadereit at the University of Konstanz, we identified transcriptional waves of genes as embryonic stem cells differentiate into neurons (Zimmer, Cell Death & Differ 2011) which further led to use of this cell type in assessing developmental neurotoxicity of drugs. This approach was also used to determine suitability of embryonic stem cells as models for assessing toxicity of nanoparticles (Hoelting, Arch Toxicol, 2013).
Vivek Tanavde joined the Bioinformatics Institute, Singapore as a Research Scientist in the Genome & Gene Expression Data Analysis in 2006. Prior to joining BII, he was heading the Hematopoietic Stem Cell Lab at Reliance Life Sciences, Mumbai where his work focused on developing mesenchymal stromal cell based therapies for cardiac and neuronal disorders. In addition, he also set up and managed the clinical flow cytometry laboratory at Reliance Life Sciences. From 1999 to 2002 he was a post doctoral fellow with Dr. Curt Civin at the Sidney Kimmel Cancer Centre, Johns Hopkins University working on expansion of hematopoietic stem cells from umbilical cord blood. Dr. Tanavde obtained his Ph.D from the Cancer Research Institute, Mumbai (1999) in Applied Biology. He is currently the Secretary of the Stem Cell Society Singapore & serves on the Live Education Task Force of the International Society for Advancement of Cytometry (ISAC). In addition, he manages the Intellectual Property portfolio at BII and also liaisons with companies interested in collaborating with BII. He also holds an adjunct appointment at the Institute for Medical Biology (IMB), AU**STAR.

**SELECTED PUBLICATIONS**


**Figure 1.** HNF4A is a common hub for networks derived from alignment data and TargetScan predictions. Gene interaction network in this figure is derived from the dataset of genes with overlapping regions corresponding to peaks from previous mapping.

**Figure 2.** This figure shows the gene interaction network derived from computationally predicted gene targets from TargetScan. A similar topology was observed for gene interaction networks in Figures 1 and 2, with HNF4A as a node amongst the interactions suggesting HNF4A as a possible downstream target for let-7 family miRNAs.
Thus, sequencing is becoming the single most informative research technology in life sciences; consequently, sequence analysis and sequence-based structure and function prediction will be more important than ever. This team invests great effort in maintaining and developing the ANNOTATOR software suite and other tools for biomolecular sequence studies and functional assessments. This technology enables us and other teams in BII to contribute sequence analytic work in a variety of collaborative projects. Due to the small genome of viruses, their genome sequence – protein structures – function – phenotype relationships are more likely be accessible for theoretical approaches. For example, we annotated the Chronic Bee Paralysis Virus (CBPV) genome and a number of other orphan viral genes in collaboration with D. Karlin’s team from Oxford University [3], see Figure. Similar approaches proved powerful in interpreting patient-specific influenza virus mutations (with Sebastian Maurer-Stroh, BII [4,5]. The team contributed sequence-analytic research to publications on granzyme substrates [7], lipid-modified proteins [8] among a few others [9,10]. We provide WWW-service and infrastructural support, for example for the SPACER allostery mode compute server [11].

Frank Eisenhaber co-organized the International Conference on Genome Informatics (GIW) 2013 in Singapore together with Limsoon Wong (NUS) and Wing-Kin Sung (NUS and GIS) [12].

The team was successful in attracting grants and industry collaborations. Birgit Eisenhaber is heading iMaGIN, the computational biology component of A*STAR’s “Biomass-to-Chemicals” research program. Frank Eisenhaber is co-heading the Cat3 “Integrated Genomics Platform” where BII receives support along with GIS. Joint research with Hungarian scientists in sequence analytic tool development and functional interpretation of human mutations is supported by the grant ASTAR-NKTH-007.

The infrastructural backbone of the team, the ANNOTATOR suite of protein sequence analysis tools and databases, received considerable development in the past year. Not only was the cluster hardware completely exchanged with a larger number of more powerful compute nodes. Several new software functionalities were added including recent in-house developments such as the concept of complex/simple TMs, the Zn-metal-binding site predictor HufZinc [13], etc. With the constraints of limited programming capacity and the need for short-term application projects, the current roadmap for ANNOTATOR’s future directions involves consolidating the software suite on protein sequence annotational features while software developments for application tools will use the ANNOTATOR just as annotation compute engine. The additional, application-specific features will be programmed outside the ANNOTATOR framework in a simpler manner. As a key technical proposition, an updated SOAP interface for the satellites has been created to provide access to the ANNOTATOR functionality.

Two satellite web-apps are currently under development. The first satellite will analyze novel microbial proteomes with a focus on identifying proteins with particular functions. This satellite will be a centerpiece of the biomass grant effort. The second satellite provides a compact interactive cartoon that shows sequence features (including multiple sequence alignments and PDB structures) in relation to user specified human mutations.
Birgit Eisenhaber was appointed as a Principal Investigator at the Bioinformatics Institute (BII) Singapore in December 2010. She previously worked as a postdoctoral research fellow at the Institute of Molecular Pathology (IMP) Vienna and at the Experimental Therapeutics Centre (ETC) Singapore. Her research interest is focused on the discovery of molecular functions of previously uncharacterized protein coding genes. The development of accurate prediction tools for proteins’ posttranslational modifications and subcellular localization are one of the key points in her work. The highly pathogenic H7N3 avian influenza strain from July 2012 in Mexico acquired an extended cleavage site through recombination with host 38S rRNA. A new piece in the puzzle of the novel avian influenza virus. " The 24th International Conference on Genome Informatics, GIW2013, in Singapore: Introduction to the systems biology needs and public health." Health Information Science and Systems, 2013, 1, 2.

SELECTED PUBLICATIONS


Our major interest is in the function of the protein complexes ensuring proper chromosomal segregation during eukaryotic cell division. Errors in chromosomal segregation can contribute to the development of cancer and lead to various hereditary diseases. Our experimental approach combines cell and molecular biology methods assisted by data mining and computer modeling of protein structures.

An evolutionarily conserved cohesin complex is comprised of the Smc1, Smc3, and Scc1 proteins and forms a tri-partite ring around sister chromatids holding them together from the replication phase of the cell cycle until mitosis. Paired sisters are thus distinguished from non-related chromosomes and sister chromatid cohesion is critical for their proper segregation to the daughter cells during cell division. Mutations in human genes encoding cohesin subunits cause Cornelia de Lange syndrome, which is believed to be one of the major causes of spontaneous abortions in certain populations. In addition, cohesin is altered in different types of cancer. Cohesin aging in oocytes can lead to chromosomal aberrations in the embryos. Besides advancing our knowledge of the fundamental processes in the living cell, progress in cohesin research helps us to understand the pathology of human diseases. Several aspects of the cohesin function remain a mystery and are a focus of our research.

Figure 1. Cohesin Ring During the Cell Cycle of Budding Yeast.
(A) Cohesin is composed of Smc1, Smc3, and Scc1 subunits, which form a tri-partite ring. In addition, Pds5, Scc3, and Wpl1 proteins interact with the ring and modulate its function.
(B) Cohesin is loaded on the chromosomes and embraces two sister chromatids during DNA replication. The release of cohesin from DNA happens at the beginning of the anaphase and is triggered by the cleavage of its Scc1 subunit by a protease called separase.
(C) Chromosomal spreads were stained with DAPI for DNA and with fluorescent antibodies recognizing Smc3-HA (green) or SMC1-Myc (red) to reveal cohesin bound to DNA.
How is the ring loaded on the DNA?

Several protein complexes involved in DNA metabolism are ring-shaped with DNA passing through the ring. This “topological” mode of protein-DNA interaction makes it possible for the protein complex to slide along the DNA without losing its tight grip at it. This mode of interaction is very stable and very labile at the same time. Since cohesin rings are pre-formed in solution, they have to find the DNA and “import” it inside the ring. The unique feature of the cohesin ring is its enormous size. With a diameter of about 40 nm, cohesin can slide even along the array of nucleosomes. Most of the ring circumference is formed by the long anti-parallel intramolecular coiled coils of the Smc1 and Smc3 proteins. Our research indicates that far from being the mere structural components of the ring, the flexible coiled coil arms play critical role in the ring assembly and its loading on the DNA. Interestingly, some of the mutations that cause Cornelia de Lange syndrome in humans map within the coiled coil region. Combining the reverse genetics with the biochemical and microscopic studies of the novel sets of mutants we endeavor to mechanistically describe the conformational changes within the ring as cohesin is loaded on the DNA.

How is the ring opened?

Another important aspect of the cohesin’s function is the ability of the protein ring to capture two sister chromatids rather than just one. Capturing of the single chromatid has been demonstrated to happen in vivo and might have an important role in chromatin regulation but does not result in the cohesion establishment. The entrapment of both sister chromatids inside the cohesin ring is hampered by the protein called Wpl1, which apparently promotes transient opening of the ring at one of its protein-protein interfaces and escape of the DNA from the ring. This is counteracted by the acetylation of the Smc3 cohesin subunit by the “establishment of cohesion” acetyltransferase Eco1. The discovery of the mechanism of action of Eco1 was greatly facilitated by the careful bioinformatics analysis of its protein sequence. However, no obvious functional insight was obtained from Wpl1 sequence or structure suggesting that it is likely to open the cohesin ring by recruiting additional proteins. We have isolated novel suppressors of eco1 mutation which suggest the potential mechanism of Wpl1 function.

Role of methyltransferases in chromosomal segregation.

In the recent years it became clear that lysine methylation is an abundant post-translational modification on a par with acetylation and is involved in regulation of the diverse cellular processes. Methyltransferases are considered an attractive target of pharmaceutical inhibition. However, incomplete knowledge of the spectrum of their substrates is a major hindrance for the successful development of new drugs. Surprisingly, as of today methylation was reported to regulate only one kinetochore protein, namely the yeast-specific Dam1. We attempt to identify novel targets of methylation among the proteins involved in chromosomal segregation. This work is conducted in close collaboration with Birgit Eisenhaber’s group, which performs the bioinformatics analysis.

SELECTED PUBLICATIONS


Principal Investigator’s Biography

Dmitry Ivanov was appointed as a Principal Investigator at the Bioinformatics Institute A*STAR, Singapore, in July 2012. After graduation from St. Petersburg University, Russia, Dmitry obtained his Ph.D. degree at the laboratory of Prof. Richard Gaynor, University of Texas Southwestern Medical Center at Dallas in 1999. He then received postdoctoral training at the laboratory of Prof. Kim Nasmyth at the Institute of Molecular Pathology in Vienna, Austria. Since November 2006 till July 2012 Dmitry was a Research Group Leader at the Friedrich Miescher Laboratory of the Max Planck Society in Tuebingen, Germany.
Infectious Diseases

One of our strongholds lies in infectious disease research. We have several published and ongoing projects with the WHO CC in Melbourne and other partners relating to influenza drug resistance, viral fitness and antigenic changes. Our FluSurver (http://flusurver.bii.a-star.edu.sg/) is the most complete one-stop influenza mutation analysis tool being used by researchers and surveillance experts globally. We are also the tool of choice for GISAID, the largest influenza database, and this collaboration received greater significance in 2013 with the high profile H7N9 outbreak in China [5] with sequence data being mainly available through GISAID. Besides advancing existing avenues like influenza, we are also adding new success stories with clinical collaborators on HIV [2] and bacterial antibiotics resistance [3,14] for which there is ample potential for further advances in future, especially in combination with next-generation whole genome sequencing. We will remain open to adding more pathogens to this list and are especially vigilant to respond to any emerging new pathogen, such as the novel MERS coronavirus (http://corona.bii.a-star.edu.sg/), for example. We are aware that timely computational analysis at early stages could be the key to detect, control and possibly limit the impact of any new infectious disease outbreaks. Through a co-appointment as visiting scientist to the National Public Health laboratory of the Ministry of Health we contribute our knowledge and computational expertise at the national frontline for infectious disease surveillance.

Human Mutations

We aim at bridging the gap from nucleotide variation to protein structures to interpret effects of human mutations. For example, we have helped clinical collaborators from Duke and Duke-NUS to analyze variants found from patient exome sequencing and tried to mechanistically explain their possible role in hereditary eye diseases [9,13]. Our mutation effect analysis efforts also include a collaboration with I2R and the National University Hospital (NUHS) on personalized cancer medicine where we try to identify genomic variants that can result in a different response to chemotherapy with gemcitabine in non-small cell lung cancer. While these and similar efforts should unveil new biomarkers of drug response and disease, our ongoing industry collaboration project with AITbiotech covers next generation sequencing (NGS) data analysis and the easy linkage of NGS data to the pool of already known and annotated human disease mutations. An interesting new direction we are trying to explore is testing the current abilities of NGS plus bioinformatics for diagnosing a patient’s disease status. We have also participated in developing the 4th version of the SNPeff database with Belgian colleagues which allows analyzing effects of mutations on protein stability and aggregation propensity and are helping our Annotator group to get a foothold in the field with a novel useful Human Mutation Viewer.

Sequence Analytic Tool Development and other Projects

Our fast database search tool, TACHYON (http://tachyon.bii.a-star.edu.sg) jointly developed with the Annotator group, is able...
to retrieve highly similar sequences from the large NR database 2 orders of magnitudes faster than BLAST and continues to receive a lot of positive feedback from the community. We have also been using the same technology to help with our infectious disease research and next generation sequencing data analysis.

Given the easy availability of short reads from genomic sequences, we have extended our in-house-developed method for alignment-free phylogenetic trees to be able to work directly with both assembled full genomes and unassembled short reads as input. This was validated for the case of an E.coli outbreak strain, allowing us to create the correct phylogenetic relationship based on full genomes or short reads from 23 different E.coli and Shigella strains within 2 minutes on a standard desktop computer [4].

The rapid access to pathogen genomes from a new outbreak through Next Generation Sequencing needs to be coupled with computational analysis of the new sequences to help interpreting clinically important features of the outbreak strain quickly. We have developed a workflow that identifies antibiotic-resistant drugs which have a target in the new genome and presence or absence of respective known resistance factors in order to derive a computational drug susceptibility profile of any new bacteria with available genome. This workflow reproduced well the experimentally measured and clinical data for an E.coli outbreak strain from 2011 [14].

Beta-lactamases are often involved in drug resistance by degrading antibacterials. We have developed the BLAC AutoTree webserver (http://blac.bii.a-star.edu.sg) where any new sequence of a beta-lactamase can be plugged into our curated phylogenetic trees of known beta-lactamase families with assigned substrate preferences. We do not just consider normal whole sequence alignments but also create specific trees only considering residues of the substrate binding pocket in order to more precisely predict the drug resistance properties [3].

In our new flagship industry project, Procter & Gamble and BiE are jointly developing Bioinformatics techniques for assessing the allergy potential of proteins using their amino acid sequence and tertiary structure.

SELECTED PUBLICATIONS


Principal Investigator’s Biography

Sebastian Maurer-Stroh studied theoretical biochemistry at the University of Vienna and wrote his master and PhD thesis at the Institute of Molecular Pathology (IMP) in Vienna. After a FEBS and a Marie Curie Postdoc fellowship at the VIB-SWITCH lab in Brussels, he joined the A*STAR Bioinformatics Institute (BiE) in Singapore where he is heading the Protein Sequence Analysis Group in the Biomolecular Function Discovery Division since 2007. He has contributed widely used predictors for posttranslational modifications and catalyzed new biomolecular insights by sequence-based function predictions. Being at the forefront of research during the swine flu pandemic, he is also heading the BiE cross-divisional programme on infectious diseases. He is also adjunct assistant professor at the Nanyang Technological University’s School of Biological Sciences and his expertise in influenza evolution and effects of mutations has earned him an appointment as visiting scientist in the National Public Health Laboratory of the Ministry of Health, Singapore.
Mechanisms underlying biological processes at a molecular level are explored through identifying and/or mapping the characteristics of proteins and their interactions with other proteins, nucleic acids, ligands. The methods/tools used are computational and combine representations at various levels, from the coarse grained to the fully atomistic. The work builds upon foundations that are rooted in rigorous computational biochemistry benchmarked extensively against available experimental data. In particular simulations are combined with detailed experimental information through extensive collaborations with experimental laboratories to provide incisive insights into biology at an atomic level. The group's current research focuses on the p53 pathway, kinases, translation initiation, defensins, antimicrobials and basic structural/computational biophysical chemistry.

The toolbox used consists of: construction of models based on “imagination with a whiff of hand-waving”, homology modeling, molecular dynamics, energy landscapes, reaction paths, ligand-protein/protein-protein dockings including virtual screening. On one hand, the group examines the molecular underpinnings of biomolecular regulation, while on the other, efforts are directed towards ligand/drug discovery and protein/peptide design both from a therapeutic as well as a (bio)technological perspective with a recent focus on tackling resistance.

### The p53 & eIF4E pathways:

An extensive program investigating the relationship between structural-functional aspects of the p53 family has revealed how regulation is orchestrated through multiple subtle and networked interactions. These have guided us in designing a set of novel nanomolar stapled peptides whose ability to enter cells and specifically target the p53-MDM2 axis thus activating p53, has opened a new avenue for designing therapeutics (highlighted in Nature Medicine 19; 120; 2013). The stapled peptides appear to have an advantage that they can inhibit MDM2 mutants that are recalcitrant to small molecule inhibitors. This work is also being extended to explore the reactivation of mutant p53. This highly successful program opens new avenues for targeting the traditionally undruggable universe of protein-protein interactions and is in close collaboration with the p53 laboratory of Prof. Sir David Lane and with collaborators in ICES (A*STAR), National Cancer Centre (SGH), the Universities of Edinburgh, Dundee, Southampton and Harvard.

![Figure 1. Designer peptides – on the top left is a staple peptide (purple ribbon with cyan staple) inhibiting MDM2 (grey surface) with the Met->Ala mutation (blue patch) that is resistant to nutlin. The other 3 figures represent peptides (in yellow ribbon and orange staples) that have been successfully designed to inhibit eIF4E (surface).](image)

In parallel, the effort with the Lane lab is also focused on understanding the mechanisms of the translational initiation cascade and designing aptamers, stapled peptides and small molecules to inhibit a key component in this pathway, eIF4E, which offers opportunities as a major target for therapeutic intervention in several cancers. Recent design efforts combined with extensive biophysical and structural analyses have lead to the design of low nanomolar inhibitory stapled peptides. The work has lead to an agreement with a company to jointly develop some of these peptides. The recent developments and the excitement generated in the p53 field has been outlined in articles in Nature Reviews in Cancer, Nature Reviews in Clinical Oncology that has been published by the joint efforts of teams from Singapore, Karolinska, Cambridge & Harvard Universities. A new project exploring the mechanisms governing cellular uptake of stapled peptides has just been started with generous funding from A*STAR. This has been complemented with funds that have witnessed the setting up of a joint lab in A*STAR by the pioneer of stapled peptides, Prof. Verdine from Harvard.

### Novel antimicrobials:

A highly successful interdisciplinary program involving close collaborations with the group of Prof. Beuerman at the Singapore Eye Research Institute and researchers at Nanyang Technological University, National University of Singapore, Singapore General Hospital and Duke-NUS has resulted in the design of novel antimicrobials. Recent efforts have focused on derivatives of defensins and peptoidic modifications of xanthones, originally derived from the tropical fruit mangosteen. These molecules are membrane targeted (and in contrast to traditional antibiotics, not targeted at a particular biomolecule) with rapid killing times, are non-toxic to human cells and appear to not lead to the emergence of resistance in bacteria compared to conventional antibiotics. The greater anionicity of the bacterial membranes compared to the mammalian appear to be responsible for the rapid adsorption and hence killing of the former and for the non-toxicity of these cationic molecules; the inability of bacteria to easily remodel their membranes coupled to rapid adsorption are possibly the reasons for the lack of resistance. The molecules work against a range of resistant bacteria such as MRSA, can be active against both gram-positive and gram-negative organisms, fungi and more recently have shown activity against MTb and Dengue. A mechanism of action has been developed with technology
disclosures of platforms developed and several patents filed. The project has attracted seed money from ETPL (A*STAR), a TCR grant from A*STAR and the attention of several industries; a small biotech venture is planned.

The EGFR family pathways: In a large translational effort, the group is engaged with experimentalists (Dr. Scaltriti) and clinicians (Prof. Baselga) at Memorial Sloan Kettering and Dr. Daniel Tan (SGH, NCC Singapore) studying the effects of small molecule and antibody based therapies for breast/lung cancers, notably relating to the EGFR pathway. Expanding upon a significant breakthrough in the understanding of molecular mechanisms that underlie the observed synergism between kinase inhibitors and antibody-based therapy for breast cancers characterized by overexpressed HER2 receptors, we are currently developing models to understand the existence of cryptic epitopes, binding sites and interactions with small molecules, antibody. A new program on designing new antibodies is currently under way.

A major focus is directed at unraveling the effects of mutations and silent SNPs in patients and their putative effects on inhibitors. The virtual screening and peptide design efforts of the group are extended to various groups within A*STAR (including the Experimental Therapeutics Centre, IMCB, IMB, SIGN, p53Lab, MEL, ICES), the hospitals in Singapore, universities and hospitals elsewhere, who carry out the synthesis and experimental investigations of the compounds. We are also working with Prof. B Lane of the Institute of Medical Biology (A*STAR) in understanding the role of TGF-β in health/disease. In addition we are combining forces with several groups across A*STAR to develop biofuels over the next few years.

In addition to the translational and clinical focus, the group is also engaged in exploring a variety of problems that underpin fundamental questions such as the role of water in modulating biomolecular function, developments of analytical processes, relationships between flexibility, thermodynamics and function in biomolecules, methods to restabilize mutant proteins to wild type functionality etc. For example, with the increasing power of computations there has been an associated and not unexpected increase in the lengths of atomistic simulations. This immediately results in an enhancement of the conformational ensemble that can be targeted for drugging. To extend the traditional methods of analysis of interactions, we have developed a new scheme using statistical clustering that can separate out the structural features and thermodynamic contributions that arise from distinct conformational families. Applications of these methods are yielding new windows into appreciating the richness that characterizes the landscapes of biomolecular interactions and new opportunities for the design of novel drugs. For example, combining simulations with time dependent mass spec analysis, a novel mode of interaction of the drug nutlin with its target MDM2 has been uncovered that opens a new window into drug design efforts. A large multicenter effort that brings together expertise from various centres in Singapore to combine the spatio-temporal dynamics of biomolecules across multiple space and length scales into a seamless integrated mechanism that may yield biological function has currently begun with a large grant from the Ministry of Education in Singapore.

The success of the group in interfacing the nanoscale with experimental observations has encouraged several experimental groups worldwide to participate with us. A successful working partnership is one in which these groups send a graduate student to our group to learn the art of simulations and subsequently use it as a regular toolkit in their laboratories. These have also resulted in joint graduate programs in various universities including NCBNS, JNU, IISc, Southampton, Manchester, Dundee, Edinburgh.

#### SELECTED PUBLICATIONS


#### Principal Investigator’s Biography

Chandra Verma carried out his undergraduate studies at IIT, Kanpur after which he studied for his D.Phil in York, UK. Subsequently he joined the York Structural Biology lab where he remained until 2003 when he moved to the Bioinformatics Institute, Singapore.
We have developed a model of thermostability in protein complexes [1], which describes stabilization if individual domains and interfaces between them along with strengthening mechanisms for prevention of aberrant assemblies. We have shown that thermophilic trends are universal for obligatory and transient complexes and reflect the major physical mechanisms of stability. Further, we have introduced a generalized concept of protein stability, which includes intermolecular interactions that comprise distinct combinations of stabilizing forces depending on the types of interacting partners [2]. Recently, we have surveyed mechanisms of molecular adaptation based on the physics and evolution of nucleic acids and proteins [3]. DNA, RNA, and proteins are major biological macromolecules that coevolve and adapt to environments as components of the one highly interconnected system. We have explored sequence/structure determinants of mechanisms of adaptation of these molecules, links between them, and results of their mutual evolution.

The whole picture of molecular mechanisms of adaptation and relations between them is far from being complete. Consideration of different environmental factors such as salinity, pressure, etc. will help us to unravel new mechanisms of stability, their sequence/structure determinants, and to understand tradeoffs that Nature embraced en route of the evolution and adaptation.

Evolution of protein function

Enzymes are involved in all processes in living organisms, and their contemporary evolution takes place via mutation and recombination of protein (sub)domains. In our recent studies we have considered an enzymatic function as a combination of elementary units that provide elementary chemical transformations. These units are closed loops, possessing one or few functional residues and bringing them to the active site, Elementary Functional Loops (EFLs) [4,5]. The functions of some EFLs are shared between (super)families of proteins with different biochemical functions and even between different folds. We designed a computational procedure for finding sequence profiles of widely spread EFLs with characteristic functional signatures [4]. Further, we developed a procedure for deriving prototypes, which served as basic units of the first folds/domains with enzymatic functions [5]. Based on the set of prototypes, knowledge of their elementary functions, and structural contexts, it is possible to represent any enzymatic function as a combination of elementary binding/activation chemical reactions provided by the elementary functional loops. Evolutionary connections between different enzymatic functions can also be delineated [6]. Figure 1 exemplifies such connections presenting a snapshot of archaeal domain superfamilies organized in a network by prototypes of elementary functions [4]. Recent advances in genomics and proteomics provided a wealth of sequences and structures, making it possible to unravel intricate evolutionary connections in the realm of protein function. In-depth understanding how first enzymatic domains emerged as combination of prebiotic functional peptides followed by the divergence and recombination of these domains can have important implications for future design desired functions.

Molecular mechanisms of allostery

Protein function depends on the balance between different conformational states, which can be shifted by many external factors that regulate protein activity including localized perturbations such as ligand binding or post-translational modification. When the perturbation site is not directly adjacent to the site of altered activity the regulation is called allosteric. Recently we introduced two concepts of allosteric regulation and communication. The first concept, binding leverage, allows one to measure the ability of a generic ligand to connect to conformational transitions, and thus its potential to have an allosteric effect [7]. Binding leverage concept of allostery is based on the assumption that binding to sites where ligand-protein interactions are connected to important degrees of freedom can affect the conformational equilibrium between active/inactive states. The second concept, leverage coupling, provides a quantitative characteristic of allosteric communication. The major assumption here is that sites that have high binding leverage for the same motion are more likely to be allosterically coupled than sites that only have high binding leverage for motion along
independent degrees of freedom [8]. Figure 2 illustrates work of these concepts by showing functional/allosteric sites and communication between them in phosphofructokinase (PFK), which is allosterically inhibited by phosphoenolpyruvate (PEP) and activated by ADP binding to the same site.

In an era of an exploding number of protein structures, it is important to find catalytic and natural allosteric sites as well as latent binding sites, which can be used as drug targets. Our future major interest, therefore, will be in the developing the computational approaches for predicting allosteric binding sites and design of allosteric effectors.

Collaboration and outreach

We recognize an importance of the outreach for theoretical/computational research. We approach it collaborating with experimental groups and building web-resources available for the public use. In particular, our computational predictions of elementary functions in zinc transporter ZnT-2 (SLC30A2) helped to design experiments that explain transient neonatal zinc deficiency as a result of a dominant negative heterozygous G87R mutation in the transporter [9]. Combination of our methods for predicting allosteric sites with structural/biophysical analysis of the family of matrix metalloproteases allowed us to characterize structure-specific regulatory sites unique for each representative of this family of structural homologues [10]. Our recent web-server SPACER [11] allows researchers with different backgrounds and without any special training to explore allosteric regulation and communication in the protein(s) of interest.

SELECTED PUBLICATIONS


Principal Investigator’s Biography

Igor Berezovsky studied physics and biophysics in the Moscow Engineering Physics Institute (MSc, 1993) and obtained PhD in biophysics from the Moscow Institute of Physics and Technology (1997). Igor’s scientific career started in the Engelhardt Institute of Molecular Biology (Moscow) where he conducted his MSc and PhD research, then worked as a research fellow (until 1998). Igor worked as a postdoctoral researcher at the Weizmann Institute of Science (1999-2002) and Harvard University (2003-2006). In 2007-2012 he was a senior scientist/group leader at the Bergen Center for Computational Science, University of Bergen (Norway), then visited the Weizmann Institute of Science in 2013. Igor joined Bioinformatics Institute in January 2014.
Immune receptor signaling

Toll-like receptors (TLRs) represent the initial gateway to almost all mammalian inflammatory responses to invading microbes. In particular, TLR4 in complex with its MD-2 coreceptor recognizes lipopolysaccharide (LPS) from the outer membranes of invading bacteria, and TLR4 overstimulation is the cause of septic shock, a leading cause of death in intensive care units. Thus TLR4 represents a major therapeutic target, but traditional structure-based drug design approaches have been hampered by the complex nature of both the receptor system and the ligands – these are not small molecules! Large-scale, atomically-detailed simulations of the receptor complex (encompassing ~0.5 million atoms) have recently enabled us to directly observe the switching of the receptor complex between active and inactive states in a ligand-dependent manner. As well as revealing the structural basis for the molecular signaling process, we have thus established an in silico protocol for predicting and refining ligand bioactivities. We are also using computationally-demanding free-energy approaches to accurately estimate the thermodynamic basis for LPS binding to different TLR4 coreceptor signalling states, thus providing a genuinely quantitative means to unravel the complex structure-activity relationships of these molecules. In parallel, we are using simulations to extend our mechanistic understanding of common disease-causing TLR4 SNPs and pathogenic allergens, and to investigate novel immunomodulatory agents, taking advantage of ongoing experimental collaborations at Cambridge and Iowa Universities. Similarly, we are working with collaborators at Baker IDI Heart and Diabetes Institute in Melbourne to establish the connection between lipid metabolic disease and TLR4 regulation.

Taking a broader perspective, we have begun to develop multi-scale simulation approaches to understand how ligand recognition controls the overall TLR4 signaling response. Activation is thought to be associated with receptor dimerization, but the molecular details of this process are unclear. In collaboration with experimental workers at University of Colorado Boulder, we are building coarse-grained membrane models of TLR family members, towards understanding long-timescale complex oligomerization at the cell surface, and subsequent intracellular assembly processes. We are also working with groups at Nanyang Technological University to delineate the mechanisms by which peptides with antimicrobial and anti-inflammatory properties exert their activity within the TLR4 pathway.

The overall aim of our research is to understand the structural basis for key macromolecular recognition and assembly processes associated with essential biochemical functions, including signal transduction, transport, and folding. A particular focus is on the receptor complexes of the innate and adaptive branches of the mammalian immune system. We also have an ongoing interest in biomembranes, and their roles both in carrying information and in compartmentalization. Our main method of choice is molecular simulation, which provides a theoretical basis for studying the structure and dynamics of biomolecules in unprecedented detail. This yields information complementary to data obtained via other biochemical and biophysical approaches, and enables us to formulate and experimentally test new hypotheses with our collaborators. As well as gaining fundamental insights into the mechanisms of action of biomolecules, we help to develop new approaches for pharmaceutical intervention in disease states, and where possible, to extend our observations to underlying principles applicable to (bio)nanotechnology. Our work encompasses several “flavours” of simulation, which enhance the nature and breadth of information we can obtain. On the one hand, we recognize the importance of the thermodynamic landscape, and use rigorous sampling approaches to quantitatively characterize macromolecular conformational changes and assembly processes in atomic detail. On the other, we take advantage of simplified, coarse-grained models to explore biological phenomena over extended time and length scales, enabling us to “fill the gap” between high-resolution structures and low-resolution biophysical data.

---

**Figure 1.** All-atom simulation of bacterial LPS recognition by the mammalian TLR4 receptor / MD-2 co-receptor complex.

**Figure 2.** Simplified models of the oligomerization of membrane-embedded TLR4 in complex with MD-2 co-receptor.
Conserved cavities & the dynamic PDB

With the continuing improvements in high-performance computational hardware and algorithms, the concept of the "dynamic PDB" has become a realistic target for the post-genomic era, in which molecular simulations are used to "screen" the conformational plasticity and ligand binding properties of different sets of proteins in silico. We have taken such an approach to study the binding cavity dynamics of a broad range of lipid-recognition proteins, which differ in their subcellular localization, structure, and function. This study has revealed that apparently diverse proteins can share common conformational dynamics and hence mechanistic strategies to achieve ligand binding. In collaboration with experimental researchers at SiGN, similar approaches are now being extended to understand the ligand binding and exchange properties of the MHC-like CD1 family of proteins, which are specialized for presenting diverse lipid antigen molecules to the T-cell receptor, and represent major targets for vaccine development. In parallel with this, a publically-available program ("trj_cavity") has been developed to enable the high-throughput characterization of protein cavities within computationally or experimentally determined structural ensembles. A comparison of performance with other widely used programs for analysis of protein cavities reveals an orders-of-magnitude improvement in computational cost, making it useful for rapidly analysing novel drug binding sites or catalytic pockets in other proteins.

SELECTED PUBLICATIONS


Compartments in biology and bionanotechnology

A key area of our research seeks to understand the properties of membranes and other compartments, which serve a variety of biomedically relevant functions and are essential to many supramolecular structures in bionanotechnology. The range of associated time and length scales necessitates the use of multi-scale simulation approaches. For example, in the study of lipids, we are combining modeling and simulation approaches with biochemical, biophysical and genetic data from the Pasteur Institute, Paris, to map how a network of secretion machinery proteins interact within the bacterial cell envelope causing virulence and multidrug resistance. In another project, we have made significant progress in the development of a coarse-grained model of what effectively represents a "virtual thylakoid membrane". This heterogeneous system is one of the most complex and realistic biomass membrane models constructed to date, and serves to support our experimental collaborators at the University of Cambridge in understanding the effects of alkane incorporation upon structural and dynamic properties of the membrane and embedded photosynthetic machinery, towards more efficient biofuel harvesting. Finally, we are interested in understanding large-scale assembly processes relevant to nanotechnology. For example, carbon nanotubes (CNTs) are of great biomedical interest in targeted cellular drug delivery, but it is essential that we understand their potential for toxicity and biomembrane disruption. In collaboration with researchers at the Silesian University of Technology, Poland, we are investigating the membrane interaction and translocation properties of functionalized CNTs by combining multiscale and biased-sampling simulation methods with in vitro and in vivo imaging experiments.

Figure 3. Part of the "dynamic PDB" of simulated lipid-recognition proteins, with internal ligand binding surfaces characterized by "trj_cavity".

Figure 4. Translocation of carbon nanotubes across a mammalian membrane model.

Principal Investigator’s Biography

Peter J. Bond has ~12 years of experience in the development and application of computational methods to study biomolecules. Following his graduation in Biochemistry at the University of Oxford, UK in 2001, he moved to the Laboratory of Molecular Biophysics to read for a D. Phil., supported by a Wellcome Trust Prize Studentship. He was awarded an EMBO Long-Term Fellowship in 2007 to carry out research at the Max Planck Institute of Biophysics in Frankfurt, Germany. In 2010, he became a University Lecturer and Group Leader in the Department of Chemistry, University of Cambridge, UK, prior to moving to BII in 2013.
Ligand discovery for proteins in cell signaling pathways

GPCRs and kinases play important roles in cell signaling pathways. Small changes in their activity can have dramatic effects on cellular phenotype. Their dysfunction causes many human developmental and metabolic disorders, as well as certain cancers. The genome of the human body codes for over 900 GPCRs and over 500 kinases, respectively. To date, approximately 120 GPCRs have no ligands identified. Most small molecules that target kinases bind to the highly conserved ATP-binding pocket and often inhibit many kinases simultaneously. The broad goal is to find novel ligands that change protein functions in new ways (e.g. allosteric regulators) for GPCRs and kinases, to contribute to the development of new chemical tools that offer exciting opportunities to understand and regulate the related biological processes such as cell signaling, and to the discovery of new drugs that have better therapeutic properties and more restricted side effects to treat human diseases. GPCRs and kinases show large conformational changes between different functional states, making them ideal platforms for prospective testing of our computational framework described in the other project. In particular, we identify novel ligands (inhibitors, activators) for chemokine receptors. Chemokine receptors are most notable for their role in cell migration. However, inappropriate utilization or regulation of these receptors is implicated in many inflammatory diseases, cancer and HIV, making them important drug targets.

Computational enzymology

With increasing availability of genomic sequences, functional assignment of proteins of unknown function is a major challenge in postgenomic biology that is limiting both understanding...
Principal Investigator's Biography

Fan Hao was appointed Principal Investigator at the Bioinformatics Institute (BII), A*STAR in February 2014. Prior to joining Singapore, he worked as a postdoctoral fellow followed by a research scientist in both Dr. Andrej Sali’s lab and Dr. Brian Shoichet’s lab at University of California, San Francisco (UCSF). He obtained his Ph.D in Biophysical Chemistry in Dr. Alan Mark’s lab at University of Groningen (RUG). He received his undergraduate degree in Biological Sciences in University of Science and Technology of China (USTC).

**SELECTED PUBLICATIONS**


**Computational framework integrating homology modeling, molecular dynamics (MD) simulations, and virtual ligand screening.**

The broad goal is to develop computational techniques to effectively populate and prioritize all available structures of each protein target and to accurately measure protein-ligand interactions, so as to facilitate ligand discovery in pathway/superfamily/genome scale. Many proteins, such as membrane transporters, membrane receptors (e.g. GPCRs), and downstream kinases, can assume multiple conformations. These distinct conformations are often not captured in experimentally determined structures, for a given protein target and even its homologues (templates). There is a pressing need to develop computational methods predicting structures, refining structures, and sampling structures at different states, to provide the basement for ligand discovery. This aim is met by developing a computational framework integrating homology modeling, molecular dynamics (MD) simulations, and virtual ligand screening. This framework significantly extend the applicability of structure-based ligand discovery, and increase the chemical diversity of the ligands.

**Figure 2. The Assignment of Pterin Deaminase Activity to an Enzyme of Unknown Function Guided by Homology Modeling and Docking**
The group of Computer Vision and Pattern Discovery for Bioimages uses advanced computer vision, machine learning and mathematical models to build better machines; for the improvement of health care and discovery of biological knowledge. The group analyses images of tissues, histological slides and 2D/3D live cells assays. These images were acquired using wide-field, confocal and light-sheet microscopes as well as infra-red camera and other kinds of clinical image devices.

In a clinical setting, imaging techniques are becoming important as they are usually non-invasive and advancement of clinical devices has made quantitative analysis of these images an important component for improving health care.

Motivated by the desire to devise better cures for diseases and driven by enabling technologies, biological experiments are becoming more quantitative and generating large amounts of data. These images are then analyzed and used to create new biological hypotheses that are further validated using other experimental means. The group is also working towards the development of better image acquisition protocols to acquire high quality microscopy images.

Computational Meibography

Dry eye disease due to abnormalities in the tear film system is prevalent in the general population. The meibomian glands are located at the rim of the eyelids inside the tarsal plate responsible for the supply of lubrication of the tear film. We use an infra-red imaging technique to analyse the morphological structures of meibomian glands for assessing the level of disease and to evaluate the effectiveness of treatments. For example, healthy glands show distinct zebra-strip like structure (see Figure 1) whereas in unhealthy glands, these structures degenerate into broken fragments. The meibomian glands are automatically segmented by making use of a combination of filters. By performing skeletonization and thinning on the segmented image, the gland and intergland midlines can be extracted. The gland and intergland midlines are used to extract features such as the orientation, gland width, intergland width, and the distribution of midlines. These features are fed to the support vector machine to perform automatic grading for all images acquired in the clinics. The current method is able to achieve 88% accuracy in correctly classifying images into "healthy", "intermediate" and "unhealthy" classes (see Figure 2). We work closely with clinicians from the Singapore General Hospital to develop a computational method for diagnosing the dry eye syndrome.

Breast density assessment using different imaging modalities

Breast density measured using different imaging modalities can be used for (1) early breast cancer risk prediction and (2) radiotherapy planning. However, its clinical potential remains to be fully explored. For (1), we collaborate with Genome Institute of Singapore to evaluate/promote mammographic density as a decision factor for personalized mammography, specifically by refining area-based density via tissue patterns, and by reliably measuring density change, which enable early-stage disease intervention. For (2), we collaborate with National Cancer Center Singapore to automatically extract texture features of dense tissue from 3D computed tomography (CT) that correlate with clinical outcome. This additional information may potentially be used in improving radiotherapy planning.

Digital Pathology

Computer Vision can play an important role in digital pathology. Two projects in our effort to progress digital pathology are on prostate cancer and breast cancer. For the prostate cancer project,
Lee Hwee Kuan obtained his Ph.D. in Theoretical Condensed Matter Physics from Carnegie Mellon University in 2001. He then held a joint postdoctoral position with Oak Ridge National Laboratory (USA) and University of Georgia where he worked on advanced Monte Carlo methods and nanomagnetism. Lee Hwee Kuan moved to Tokyo Metropolitan University in 2003 as a postdoctoral research fellow with the Japan Society for Promotion of Science. In 2005, he returned home to join Data Storage Institute, pioneering the use of ferromagnetic resonance recording method. In 2006, he joined Bioinformatics Institute as a Principal Investigator and became the Head of Imaging Informatics Division in Bioinformatics Institute in 2010.

**SELECTED PUBLICATIONS**


11. T. Dinh, T. Gong et al. Unsupervised Medical Image Classification by Combining Case-Based Classifiers. The 14th World Congress on Medical and Health Informatics 2013.


**Principal Investigator’s Biography**

Lee Hwee Kuan obtained his Ph.D. in Theoretical Condensed Matter Physics from Carnegie Mellon University in 2001. He then held a joint postdoctoral position with Oak Ridge National Laboratory (USA) and University of Georgia where he worked on advanced Monte Carlo methods and nanomagnetism. Lee Hwee Kuan moved to Tokyo Metropolitan University in 2003 as a postdoctoral research fellow with the Japan Society for Promotion of Science. In 2005, he returned home to join Data Storage Institute, pioneering the use of ferromagnetic resonance recording method. In 2006, he joined Bioinformatics Institute as a Principal Investigator and became the Head of Imaging Informatics Division in Bioinformatics Institute in 2010.

**Autonomous Microscopy for Biological Research on Neural Stem/Progenitor Cells**

Neural Stem/Progenitors cells (NSCs/NPs) are cells that give rise to the main cell types of the nervous system; oligodendrocytes, neurons and astrocytes. Studying NSCs/NPs with time-lapse microscopy is critical to the understanding of the biology of these cells. In most of NSC/NP-related experiments, a large number of cells need to be monitored. Consequently, the acquisition of a huge amount of images is required.

In the last year, we presented a novel automated 3D visual tracking of suspension living cells for time-lapse image acquisition using phase-contrast microscopy. This year, we present an automated microscopy system for high content screening. Traditionally, screening specimens on multi wall plates using microscopy is a very labor-intensive task. Moreover, most existing automated exhaustive screening platform take a long time for image acquisition. For example, acquiring images of all cells culturing on a 96 well plate can take several hours, even several days.

We propose a novel high content screening microscopy system for screening living cells on a multi well plate. The system involves computer vision and machine learning approaches for estimating the location and the proper focal level for each of the cells. In our experiments, 100 cells distributed in 10 different wells can be imaged within 250 seconds. The time required for image acquisition includes, moving the stage, switching the objective lens between 4x and 10x, auto focusing and calculating precise locations of the cells. This new high content screening method provides the tool for a fast and automated pipeline for three dimensional cell cultures, image acquisition and screening, paving the way for innovative analysis of NSCs/NPs and the study of neurodegenerative diseases.
Our group aims to develop machine learning approaches for the automated analysis of biological and medical images. Topics of interest include, but not limited to, cellular and tissue image analysis such as segmentation, tracking, and phenotype recognition. Selected Projects are listed below:

### Tracing Filamentary Structured Objects with Crossover

The problem of tracing filamentary structures with crossover has been encountered in various biological and medical image analysis applications, including e.g. neurite tracing, microtubule tracking, retinal vessel tracing, as illustrated in Figure 1. We focus on the application of tracing retinal blood vessels, and our goal is to untangle the vessels such that starting from the optical disk as roots, each of the vessel trees is clearly identified and separated from the rest. After necessary image preprocessing steps, the vessel skeletons are extracted and are further transformed into directed graph representation. The tracing with crossover problem is then formulated as random walks in directed graphs to perform transductive inference, and is addressed as maximizing the accumulative expected number of visits using absorbing Markov chain theory. The approach is shown to bear strong connections to existing works including in particular PageRank used by Google. We also provide complexity and generalization error analyses. Empirically our approach consistently outperform existing methods. Some of the visual results are demonstrated in Figure 2 and Figure 3. This is an on-going research problem. So far we have developed a suite of algorithms, where some have been published (e.g. [3]). To continue, we plan to thoroughly analyze the proposed and related technics, in order to provide as an integrated toolkit our related research works, as well as to further investigate the neurite tracing problem.

### 3D Immersive Bioimage Computing Platform and our in-house Hand Engine

Biology is by nature three dimensional. Currently there is a paradigm shift toward 3D analysis such as 3D cell culture, and 3D computational work.

---

**Figure 1.** Examples of filamentary structures with crossover to be traced: Neurons, microtubules, and retinal vessels.

**Figure 2.** Results of exemplar images of DRIVE dataset: Left: ground-truth. Right: the proposed approach. Black square: error predictions.

**Figure 3.** Results of exemplar images of STARE dataset: Left: ground-truth. Right: the proposed approach. Black square: error predictions.
SELECTED PUBLICATIONS


Development of quantitative assays to measure the biophysical properties of cell migration

Most cell migration studies have been carried out on two-dimensional substrates and are therefore not physiological, even though a lot has been known from such studies about the formation of integrin-mediated cell-matrix adhesion and flat two-dimensional protrusions such as filopodia and lamellipodia. On the other hand, not much is known about how cells migrate in three-dimensional environments. This is particularly important for some tumor cells, which has been known to form bleb-like protrusions and change their shapes in a manner similar to amoebae; the shape changes then allow the tumor cells to squeeze through the pores in the extracellular matrix and invade.

We have been collaborating with wet labs in various A*STAR research institutes to develop a three-dimensional assay that mimics in vivo conditions. We will use this assay to screen for inhibitors of in vivo cell migration. Their effects on cell migration can then be assessed. We will automatically track the cell movement and measure cell speed, directional persistence, as well as traction force magnitudes that the cell exerts on the substrate. These measurements serve as quantitative indicators of the cell’s migratory potential.

Quantitative experiments and modeling of amoeboid migration and chemotaxis

Tumor cells have been observed to migrate using the so-called amoeboid mode, characterized by its independence from integrin-mediated adhesion to the substrate. This is different from the more commonly studied mode of mesenchymal migration, characterized by actin polymerization and focal adhesion assembly. We are studying amoeboid migration and chemotaxis, both experimentally and computationally. For example, we have shown that amoeboid cells can migrate via a “chimneying” mechanism by generating anchoring stresses normal to the substrate and shearing stresses at protrusions that shift the cell body forward resulting in cell movement. In Figure 1, we show examples of the traction stress distribution that amoeboid cells exert on their substrates. Such quantitative calculations allow us to identify new modes of migration.

We will also use inhibitor libraries to screen for and uncover the signaling networks that regulate the amoeboid mode of migration and chemotaxis. Knowledge of such signaling networks will eventually allow us to uncover novel gene functions implicated in migration and chemotaxis as well as identify targets to inhibit tumor invasion during cancer metastasis.

Quantitative modeling of collective cell migration

In multicellular organisms, certain cells such as epithelial cells migrate collectively as a group instead of individually. The skin, for example, is comprised of striated layers of epithelial cells. We are interested in developing quantitative models of how an epithelial cell sheet can maintain its integrity while at the same time accommodating for individual cell movement and rearrangement, cell division, and cell death. Such models require one to understand the molecular mechanisms of cytoskeletal rearrangement and cell-cell adhesion, and how these molecular processes influence the mechanical properties of the cells. Conversely, it is also important to account for the effects of mechanics on cell signaling and hence cytoskeletal rearrangement and regulation of cell-cell adhesion. For example, in Figure 2, we show the results from a simulation of how cells in a piece of tissue can sort themselves in response to an external morphogen while still maintaining tissue integrity.
One specific problem that such a quantitative model can be applied to is to study how an epithelial layer can function as a barrier. This is a problem that is motivated by the consumer care industry. For example, we are interested in using our quantitative models to compute the permeation into the skin of a particular compound such as those found in a cream or lotion that is being applied topically onto the skin. Our quantitative models will allow us to study the effects of perturbing specific signaling pathways so as to modify the ability of a particular chemical to permeate into the skin epithelial layer. We expect such computations to be useful to the consumer care industry where there is a general trend to move away from animal testing to in silico product testing.

**SELECTED PUBLICATIONS**


**Principal Investigator’s Biography**

Chiam Keng Hwee is a theorist working at the interface of physics and biology, collaborating very closely with experimental groups in developing theories and models for a variety of problems in mechanobiology and biological physics, systems biology, and biological fluid mechanics. He received his Ph.D. in physics from the California Institute of Technology in 2003 and his B.S.E. in physics from the University of Michigan in 1997. Prior to joining the Bioinformatics Institute in 2013, he and his group was located at the Institute of High Performance Computing.
Complex Cellular Phenotype Analysis

I. Protein Localization Analysis and Search Tools (PLAST)
Protein subcellular localization is a major determinant of protein function. However, this important protein feature is often described in terms of discrete and qualitative categories of subcellular compartments, and therefore it has limited applications in quantitative protein function analyses. Our group has developed Protein Localization Analysis and Search Tools (PLAST), an automated analysis framework for constructing and comparing quantitative signatures of protein subcellular localization patterns based on microscopy images [1, 2]. PLAST produces human-interpretable protein localization maps that quantitatively describe the similarities in the localization patterns of proteins and major subcellular compartments, without requiring manual assignment or supervised learning of these compartments. Using the budding yeast *Saccharomyces cerevisiae* as a model system, we construct a genome-wide protein localization map, and show that PLAST is more accurate than existing, qualitative protein localization annotations in identifying co-localized proteins (Fig. 1). PLAST can also be used to identify proteins that have similar localization patterns and participate in closely-related biological processes, but do not necessary form stable complexes with each other or localize at the same organelles.

II. Spatial and functional divergences of proteins
Gene duplication is a main source of new genes. A fundamental question in evolutionary biology is how duplicate genes acquired new or altered biological functions. Change in protein subcellular localization, or “protein relocalization”, is a possible mechanism for duplicate genes to achieve functional divergence. Using PLAST, we found an association between spatial and functional divergences of proteins during evolution [1]. Surprisingly, as proteins with common ancestors evolve, they tend to develop more diverged subcellular localization patterns, but still occupy similar numbers of compartments (Fig. 2). This suggests that divergence of protein localization might be more frequently due to development of more specific localization patterns over ancestral compartments than occupation of new compartments.

III. Specific induction of cancer cell apoptosis
Lung cancer is a leading cause of malignancy-related mortality worldwide. Specific induction of cancer cell apoptosis using chemo- or targeted therapeutic agents is a main tool for clinical management of this disease. However, acquired resistance to...
these agents poses a significant challenge to cancer treatment efficacy. Combinatorial approaches to treatment using novel compounds is a possible strategy to alleviate this problem. Magnesium silicate (“talc”) is used clinically for the relief of symptoms associated with chest cavity fluid accumulation frequently observed in cancers. Talc has been found to induce apoptosis of mesothelioma cells, but not non-malignant mesothelial cells. Therefore, we postulate that this mineral could also specifically induce apoptosis of lung cancer cells. In collaboration with Dr. Lee Pyng from the National University Hospital, we use quantitative imaging methods to characterize and study the effects of talc treatment on non-small-cell lung carcinoma cells. Talc could potentially be used as a novel compound in combination with other chemo- or targeted therapeutic agents for treating lung cancer patients.

IV. The cellXpress platform for profiling cellular phenotypes

High-throughput, image-based screens of cellular responses to genetic or chemical perturbations generate huge numbers of images. Automated analysis is required to quantify and compare the effects of these perturbations. Most of our projects are based on the “cellXpress” platform developed by our group for high-throughput, automated analysis of cellular images (Fig. 3) [3]. The platform is written in C/C++ and optimized for modern 64-bit and multi-core CPUs. It supports parallel processing and dynamic job scheduling, which allow it to process terabytes of image data and quantify millions of individual cells under different experimental conditions. We have also designed an intuitive graphical user interface for interactive configuration of the platform and visualization of results. The cellXpress platform is especially useful for managing and analyzing images obtained from high-throughput drug or RNAi screening experiments.

SELECTED PUBLICATIONS


Principal Investigator’s Biography

Loo Lit Hsin studied electrical and computer engineering at Drexel University in Philadelphia, USA, and received his B.S. and M.S. in 2000, and Ph.D. in 2004. To further pursue his interests in systems biology and pharmacology, he became a postdoctoral fellow in the lab of Drs. Steven Altschuler and Lani Wu, which was first located in the Bauer Center for Genomics Research at Harvard University, and then in the Green Center for Systems Biology and the Department of Pharmacology at the University of Texas Southwestern Medical Center, USA. In 2010, Lit Hsin joined the Bioinformatics Institute at Singapore as a Principal Investigator, and has been heading the Complex Cellular Phenotype Analysis Group in the Imaging Informatics Division. He was the recipient of the Award for Excellence in Postdoctoral Research (2010) and the Alfred Gilman Award (2009) by the University of Texas Southwestern Medical Center, and the Doctoral Award in Mathematical Sciences and Engineering (2005) by Drexel University.
Natural Product Biology

The A*STAR Natural Product Library (NPL) was established in 2013 following the acquisition of the Collection from MerLion Pharmaceuticals Pte Ltd (MerLion). The Library has been developed over the past 20 years through sample collection from targeted local habitats, a diverse series of international collaborations and by the strategic acquisition of entire libraries from other companies (GSK, Combinature and Edison). It currently consists of 37,542 plant samples and 123,177 microbial strains (Figure 1). The genetic diversity within the A*STAR NPL collection is exceptional. With 57% of all known cultured fungal genera, over 67% of the world’s plant families and 70% of filamentous bacterial genera represented, the collection has been described as “the most diverse and comprehensive collection of plant and microbial samples in the world” (Prof. Geoffrey A. Cordell, University of Illinois).

In terms of geographic distribution, the collections are truly global with plant and microbial samples from over 100 countries and from all continents including marine samples from the coastal areas off Antarctica. The samples are collected in compliance with the Convention on Biological Diversity. In Singapore, through a long term collaborative arrangement with the National Parks Board, MerLion and the Centre for Natural Product Research had gained access to most of the local environment for the collection of plants and microbes. This includes the Singapore Botanic Gardens, intact native rainforests, and secondary forests surrounding the central catchment area, numerous mangrove systems and the marine environment surrounding many of Singapore’s off-shore islands.

Most of the plants and nearly half of the microbial strains in the Collection have been utilised for the production of organic extracts. Typically each plant sample has one extract generated from a portion of the dried material, whilst the microorganisms have at least four extracts generated through the implementation of specific combinations of media and fermentation conditions. The extract screening library comprises nearly 270,000 un-fractionated and 68,985 HPLC fractionated extracts and is formatted in industry-standard microtiter plates (Figure 1). Approximately 25% of the crude extract library (76,210 extracts; Figure 2) have been chemically fingerprinted using an automated LCMS characterisation system developed by MerLion. The data have been uploaded into a proprietary Fingerprinting (FP) database which allows analysis such as productivity, similarity comparison, as well as chemical diversity analysis. The FP database can be mined to identify novel producers of targeted compounds with indication of relative production yield, definitive measures of extract quality and the selection of productive strains.

The Collection also includes a library of 2,567 purified natural products (Figure 3) with a high proportion of unpublished compounds (18%). 1,843 of them are arrayed in 96-well microtiter plates in a screen-ready format. Most of these compounds have been isolated during the course of diverse bioassay-guided isolation projects and thus represent an excellent source of bioactive natural product chemistry.

Detailed information on the Natural Product Collection is housed in a proprietary database which links information on the plants and strains taxonomy, source information and fermentation conditions to the extracts, biological data and isolated compounds.

High Throughput Screening

Since 1993, the Collection has been used for the discovery of small organic bioactive molecules. This is accomplished through the high

A*STAR Natural Product Library

The main focus of the group is the discovery of bioactive molecules from natural sources such as plants and microbes. This is accomplished by screening extracts derived from plants or microbes against a wide array of biochemical and cellular assays to identify extracts with bioactivity of interest. The subsequent isolation of the compounds present in the extracts that are responsible for the observed biological activity is carried out in collaboration with the Natural Product Chemistry group. The Natural Product Biology group is also responsible for the management and maintenance of the A*STAR Natural Product Library.
throughput screening of the extracts through biochemical or cellular assays to identify biological activities of interest. Once the active extracts have been identified the active principals are purified using bioassay-guided compound isolation. This effort was primarily for drug discovery, and 157 screens against human health targets (covering therapeutic areas such as infectious diseases, oncology, metabolic diseases, inflammation, CNS) had been performed. Approximately 2,000 bioactive compounds have been isolated; ~25% are novel with 16 compounds progressed to animal model studies and 2 lead series were identified (antibacterial and oncology). The Collection was also used to search for food preservatives, natural ingredients such as flavours and nutriceuticals, as well as bioactive molecules for crop protection, animal health, insecticides and cosmeceuticals applications. The discovery and development of new drugs from natural products (NPs) has played a significant role over the last few decades. Over 28% of the new chemical entities and 42% of the anticancer drugs introduced into the market can be traced back to NPs. NPs will continue to play a role in drug discovery as long as there is unmet medical needs. NPL’s huge collection of plant specimens and microbial strains is a rich resource that will continue to be useful for the discovery of novel bioactive compounds.

The Natural Product Biology Lab can perform assays in 96- or 384-well format with absorbance, fluorescence and luminescence readouts. The assays can be cell-free or cell-based assays involving the use of microorganisms or mammalian cells. All screens will be validated for screening against natural product samples before commencement of HTS and strict Quality Control procedures, including scrutiny of screening data using Spotfire visualisation software, are implemented throughout the screen to ensure high quality data. During assay validation a small set of extracts will be tested in the screen to determine the robustness of the assay in terms of day to day assay variation (indicated by assay parameters such as signal, background, signal-to-background ratio, Z' and IC50).

![Figure 3. The purified natural product compounds collection](image)

![Figure 4. Natural product discovery using bioassay-guided compound isolation](image)

**Future Research Strategies**

Traditionally the search for new natural products begins with the growing of microorganisms in the laboratory and testing the fermentation broths or organic extracts for bioactivity. This is then followed by bioassay-guided compound isolation. However, recent research has shown that microorganisms have many natural product biosynthetic gene clusters that do not readily express when grown in the laboratory. Thus, despite decades of fermentation-based screening it is likely that a vast supply of bioactive microbial compounds remains to be discovered. Various genomic and proteomic surveys of microbes have recently been published and demonstrated the fruitfulness of using these newer strategies in the discovery of novel natural products and their biosynthetic pathways. We believe that augmenting the currently available information on the A*STAR Natural Product collection with genome sequence, biosynthetic pathway annotation, and chemical indexing data will greatly increase the utility of the Collection for diverse applications. As a pilot study we will be selecting 100 microbes to develop the methods and work process for chemical indexing and genomic characterisation.

**SELECTED PUBLICATIONS**


**Principal Investigator’s Biography**

Ng Siew Bee joined the Bioinformatics Institute in 2014. She earned her Ph. D in Biochemistry from the Institute of Molecular and Cell Biology (IMCB) in Singapore. Dr. Ng then embarked on her career in natural product research and drug discovery by joining the Centre for Natural Product Research (CNPR), IMCB. Initially, she was involved in the development of new assays and subsequently was promoted to lead the High Throughput Screening group. When CNPR corporatised to form MerLion Pharmaceuticals Pte Ltd she was the Director of Discovery Biology (2002 – 2012), managing all assay development and screening activities, as well as leading the internal antibacterial discovery programme and supporting MerLion antibiotic development programmes. Prior to joining Bil, Dr. Ng was an Adjunct Lecturer at the Singapore Polytechnic.
Discovery Process
Since 1993, the Collection has been used for the discovery of small organic bioactive molecules. This is accomplished through the high throughput screening of the extracts through biochemical or cellular assays to identify biological activities of interest. Once active extracts have been identified, they will be subjected to a two phase chemical characterization process. Phase I entailing dereplication and Phase II comprising of fractionation, compounds isolation and structure elucidation.

Phase I, Dereplication
The purpose of Phase I (dereplication) characterisation is to rapidly identify known compounds which account for the activity observed in a particular extract. This is achieved by simultaneously analysing the extract mixtures by analytical HPLC coupled to a high resolution mass spectrometry and collecting 38 fractions directly in a deep well microtiter format block for biological testing. A schematic of the process and the equipment used at Natural Product Chemistry Lab is shown in Figure 1.

Compounds observed within the active fractions are matched against NPL HRMS/MS/MS database of natural products which have been analysed under the same LC/MS conditions. In conjunction, when these high resolution MS or MS/MS data are “coupled” with the Dictionary of Natural Products or SciFinder search, it may help dereplicate for compounds not represented in the NPL database or identify target compounds (new compounds) for isolation. The NPL HRMS/MS/MS database will be created comprising of 2,567 compounds, including most commonly isolated natural products. Most of these compounds are available at NPL in its pure form. This allows the confirmation of the compound activity in the crude sample from the bioassay activity of the pure compound.

Phase II, Fractionation & Structure Elucidation
Following dereplication, samples will be selected for progression to Phase II analysis, involving fractionation and structure elucidation of the purified active principals. To allow the investigation of the maximum number of samples at this stage, we will develop an efficient process whereby the secondary metabolites of crude extracts are enriched using vacuum-assisted chromatography followed by automated fractionation using preparative reverse phase (C18) HPLC. The gradient can be modified for individual samples to enhance separation of the active regions identified in Phase I. During this initial stage (designated Phase IIa; Figure 2) up to 20 samples can be processed using such a system within a three-day cycle. Aliquots of each fraction are tested for biological activity and can be selected for high resolution mass spectrometer and 1H-NMR analysis if shown to be active. Subsequent sub-fractionation and any further purification steps are performed on the active fractions using methods which are most suited to the chemistry identified within these fractions (designated Phase IIb; Figure 3a). These may comprise of further HPLC separations or orthogonal separation techniques such as normal phase (silica and diol), Sephadex or ion exchange chromatography.

After each separation step, fractions can be tested for biological activity and active fractions will be analysed by high resolution mass spectrometry and 1H-NMR (Figure 3b). Fractions of sufficient purity and amount can then be further analysed using high resolution mass spectrometry and 2D-NMR techniques, as appropriate, to confirm compound identity or to solve novel structures. Novel compounds will be given a unique identifier and will be characterised using the dereplication HPLC system. This adds their retention time and HRMS/MS/MS profile to the A*STAR NPRL MS/MS database, which in turn allows such compounds to be dereplicated from subsequent samples.

Future Research Strategies
1. Chemical & Genomic approach to natural product discovery
We believe that augmenting the currently available information on the A*STAR Natural Product Collection with genome sequence,
biosynthetic pathway annotation, and chemical indexing data will greatly increase the utility of the Collection for diverse applications. As a pilot study, we will be selecting 100 bacteria and 100 fungi to develop the methods and work process for chemical indexing and genomic characterisation. We will use ultra-high performance liquid chromatography coupled to a high resolution mass spectrometry and bioinformatics to measure the chemical diversity of the 200 selected strains by characterising the known and novel natural product compounds they produced. This will help to dereplicate or index for compounds not represented in the NPL database or identify target and new compounds and for isolation and structure elucidation. The chemical indexing fingerprint data generated will be used to identify known and novel compounds and serve as a reference for mapping biosynthetic pathways. In addition, there will be a physical library of novel compounds available for testing. This will increase the value of our purified compound collection by increasing its chemical novelty, and thus offer the possible generation of significant intellectual property (structure leads). In order to achieve this mission, further development of three areas is required: first, a manual chemical assignment strategy for understanding known metabolites (isolation and structure elucidation of such metabolites is important because they are often commercially unavailable); second, a publicly available database of HRMS and MS/MS spectra with precise chemical information including exact mass, elemental composition, and structure to expand the available target information for automatic assignment; and third, an algorithm or platform for automatic assignment using bioinformatics. In addition, we will extend this by developing models to integrate the chemical indexing (chemical fingerprint information) with the genomic sequence and proteomic information in order to develop better models for in silico prediction of secondary metabolites from gene sequences and proteomics. The database will be searchable by chemical structure, genetic sequences and keywords. By building this database based on selected 200 strains collection, we can simply plug in a molecular structure as the search term and return with the possible gene clusters responsible for the compound production.

**Figure 2.** Schematic of Phase IIA: automated large-scale fractionation

**Figure 3.** (a) Schematic of Phase IIB: purification & structure elucidation

We will use the results from the pilot study to assess the feasibility of scaling up to this approach to a larger set of strains.

### 2. Chemical Probes and Drug leads

Natural products have long been recognised as privileged scaffolds because they have evolved specifically to interact with biological macromolecules, especially proteins. Thus, they are ideal tools for exploring biological phenomena and assemble biochemical pathways. Therefore, available biological data of the 2.567 purified compounds and extracts will be mined by bioinformatics to identify potential candidates of chemical probes. In addition, some antibacterial and antitumour lead compounds that were not followed through by MerLion, this will further be worked upon in partnership with other research groups. We will isolate gram quantities of the compounds of interest for medicinal chemistry to synthesize a focused library of analogues around the compounds, thus offering the possible generation of potential drug candidates.

**SELECTED PUBLICATIONS**


**Principal Investigator’s Biography**

Yoganathan joined the Biinformatics Institute in January 2014. He obtained his PhD in Natural Product Chemistry (Organic Chemistry) from the Chemistry Department, University of Malaya in 1997. Following this, he worked as a Senior Lecturer for Chemistry (1997-1999) at the Seapang Institute of Technology (University of Adelaide Twinning Program, Malaysia). Prior to joining BII, he was heading the Natural Product Chemistry Group at MerLion Pharmaceuticals, focusing on natural product chemistry which included anti-bacterial and anti-cancer drug development in an industry based collaborative research, as well as supporting MerLion antibiotic development programmes. From 1999 to 2002, he was Senior Research Scientist at the Centre for Natural Product Research, IMB.
Arising from this concerted effort, we have filed invention focused to understanding HIV drug resistance development. Dr. Chandra Verma, an interdisciplinary effort have been mutational data. Experimental platform for both validating and generating disclosures on a mutation solution software and a novel Tock Seng Hospital, Assoc Prof. Kwoh of Sch of Com Eng (NTU), in collaboration with Dr. Tan of SIMTech (A*STAR), Dr. Ng of Tan Drug Design resulted from these efforts with publications following shortly. Their efficacy and specificity. Several invention disclosures have on which antibody therapeutics could leverage on to improve optimal elements necessary, but also to develop a platform mentioned collaborators, we aim to not only investigate the developments. With the interdisciplinary contribution of the role of these antibody elements for diagnostic or therapeutic antibody genetics and biology activity. Together with the ENT dept of Changi General Hospital, our own Dr. Chandra Verma, Prof. Sir Lane and Prof. Bhakoo of p53Lab and SBIC, respectively (A*STAR), we have secured generous funding from A*STAR JCO (2013) to investigate and evaluate the role of these antibody elements for diagnostic or therapeutic developments. With the interdisciplinary contribution of the mentioned collaborators, we aim to not only investigate the optimal elements necessary, but also to develop a platform on which antibody therapeutics could leverage on to improve their efficacy and specificity. Several invention disclosures have resulted from these efforts with publications following shortly.

**Therapeutic Antibodies**

The use of therapeutic antibodies in the alleviation of many pathological conditions has been gaining traction and promise clinically. Amidst this growing interest, significant knowledge gaps remain in the processes involved to improve such current therapeutics and in the making of the ideal therapeutic antibody. To address this, we do detailed antibody engineering of human immunoglobulins to study the impact of these various elements on expression and biological activity.

Together with the ENT dept of Changi General Hospital, our own Dr. Chandra Verma, Prof. Sir Lane and Prof. Bhakoo of p53Lab and SBIC, respectively (A*STAR), we have secured generous funding from A*STAR JCO (2013) to investigate and evaluate the role of these antibody elements for diagnostic or therapeutic developments. With the interdisciplinary contribution of the mentioned collaborators, we aim to not only investigate the optimal elements necessary, but also to develop a platform on which antibody therapeutics could leverage on to improve their efficacy and specificity. Several invention disclosures have resulted from these efforts with publications following shortly.

**Drug Design**

In collaboration with Dr. Tan of SIMTech (A*STAR), Dr. Ng of Tan Tock Seng Hospital, Assoc Prof. Kwoh of Sch of Com Eng (NTU), and Dr. Chandra Verma, an interdisciplinary effort have been focused to understanding HIV drug resistance development. Arising from this concerted effort, we have filed invention disclosures on a mutation solution software and a novel experimental platform for both validating and generating mutational data.

**Product developments**

As the above mentioned research streams have mid to long term incubation periods before tangible outcomes can be seen, the team also has short term goals to provide commercially viable products as solutions to identified gaps in routine molecular biology processes. On this, we have worked with local SMEs in the SPRING-A*STAR GET-UP grant and service agreements to move into ‘Made in Singapore’ research consumables production such as nucleic extraction kits. On this, we have made our own reagents that are comparable to current commercial gold standards, resulting in invention disclosures and publications (e.g. Chan et al., 2013). To boost our own productivity, we have also developed mobile apps e.g. the DNAApp v1.0, which is the first app that allows the analysis of DNA sequencing files on an android platform (Figure 1) in which licensing talks are underway.

This newly formed team was officially set up from the Biomolecular Modeling and Design Division in Oct 2013 to be part of the new Translational Research division. We work on therapeutic antibodies; drug design; and product development. Working with other groups both within and beyond A*STAR, the interdisciplinary nature of streams spans across molecular cell biology, computational science, complexity sciences, biomolecular modeling, and commercialization of products (with local SMEs).

### SELECTED PUBLICATIONS AND INVENTIONS

2. Gan, SKE. Optimization of QIAprep and QIAquick Miniprep and Gel-PCR extraction kits. ETPL Invention disclosure no: BII/2/08123 (Oct 2013).
3. Gan, SKE. RT-PCR Humanization and Rodentization of Monoclonal Antibodies. ETPL Invention disclosure no: BII/2/07997 (July 2013).
6. Karagiannis SN. Characterisation of an engineered trastuzumab IgE antibody and effector cell mechanisms targeting HER2/ neu-positive tumour cells in Cancer Immunology, Immunotherapy, 2009

### Teamleader’s Biography

Samuel Gan got his Dip Biotech from TP (2001), his BSc (Hons) Mol Cel Bio from UCL, UK (2005), his PhD in Allergy from KCL MRC Center for Asthma and Allergy (2008) alongside philosophical and teaching qualifications, MSc from BBK, UK (2010), PG Cert BA from OU, UK (2010). He was in Shanghai as Technical Director before coming back to Singapore A*STAR JCO in 2010. He joined Bioinformatics Institute in 2011 as a post-doc under Dr. Chandra Verma.

---

**Teamleader:**
Samuel K.E. GAN

**Research Associates:**
NGUYEN Phi Vu, LAU Sein Yan, POH Jun Jie

**Figure 1 DNAApp v1.0 full version.** This is the first android app available that can open ab1 sequencing files with capabilities to copy the sequences and reverse complement them for further analyses.
Bio-Computing Centre and Administration Teams

Scientific Computing, Opensource Computing and Scientific portal team

The role of the team is to provide technical IT support and solution in the areas of general IT services, desktop computing, computational cluster, database, scientific data storage and networking. We are closely aligned with our research community to provide strong support in IT-related resources and services for our BII scientists in their research work.

Major activities

1) AFS file servers - all data has been move to new servers with upgraded storage capacity.
2) Virtual Machine hosting based on proxmox/openvz - implemented for some services and servers. Plans to move entire corporate servers onto this platform.
3) Web-based tool for tracking and understanding A*STAR ejournal access and usage. Saves the need for yearly survey from every RI in A*STAR.

Scientific computational cluster upgrade

Over the past one year, BII had upgraded its scientific computing infrastructure particularly in the areas below

1) New set of compute servers for Annotator cluster - exponentially cuts down the running time for calculations and analysis in the cluster. Addition of a 1TB high-ram node allows for handling of big data-sets.
2) On-going. New GPUs nodes for molecular dynamics simulation.

Administration Team

The administration team supports the institute’s leadership to create conditions for scientific work at BII. It also serves as a link to the BMSI Business Centre (BBC) which is the centralized corporate body of A*STAR’s Biomedical Sciences Institutes (BMSI). BBC administers the procurement, finance and human resource management of majority BMSI’s research institutes. Within this setting, the Administration Team facilitates all auxiliary services and serves as BII’s main liaison body with all parties to ensure that matters relating to students, internships, facilities, logistics, administration, etc. function effectively so that BII scientists can concentrate on their areas of expertise in their research work.

From left to right:
Betty KEE, FONG Chew Peng, Executive Director, Dr. Frank EISENHABER, Noraini SULAIMAN, Christine LOW
Dr. Lim’s laboratory in NUS had and is collaborating with BII’s scientists in using bioinformatics to i) aid in the generation of testable hypotheses to elucidate the mechanisms through which novel oncogenes contribute to cancer (Dr. Sebastian MaurerStroh); ii) characterize and make biological sense out of high-throughput data generated from proteomics of cancer models through gene ontology and pathway mapping (Dr. Vladimir A. Kuznetsov) and iii) to predict the structure of novel protein-protein interactions that have implications in small molecule-based cancer therapy (Mallur Srivatsan MADHUSUDHAN/Chandra Verma).

Dr. Lim’s laboratory in NUS had and is collaborating with BII’s scientists in using bioinformatics to i) aid in the generation of testable hypotheses to elucidate the mechanisms through which novel oncogenes contribute to cancer (Dr. Sebastian MaurerStroh); ii) characterize and make biological sense out of high-throughput data generated from proteomics of cancer models through gene ontology and pathway mapping (Dr. Vladimir A. Kuznetsov) and iii) to predict the structure of novel protein-protein interactions that have implications in small molecule-based cancer therapy (Mallur Srivatsan MADHUSUDHAN/Chandra Verma).
SCE (NTU) and BII (A*STAR) PhD programme in Computational Biology and Bioinformatics

Introduction
The scientific landscape of bioinformatics/computational biology continues to change as the field is still evolving. It is anticipated that new scientific insights in computational biology and bioinformatics will impact the continued growth in biomedical sciences and biotechnology and will have considerable socio-economic impact.

Nevertheless, the number of personnel with major specialization in bioinformatics/computational biology hired by academia and industry remains small; yet, the professional requirements remain high and include a thorough education in (i) mathematics and exact natural sciences (physics/chemistry), (ii) computer science including programming, and (iii) life sciences (especially molecular level life sciences). Since bioinformatics applications in the real world occur at the cutting edge of the field, professionals in this field need to have strong research experience. This is exactly the goal of the SCE-BII PhD Programme in Computational Biology and Bioinformatics.

Background
In March 2010, Bioinformatics Institute (BII) of the Agency for Science, Technology, and Research (A*STAR) entered a Memorandum of Understanding (MOU) with the School of Computer Engineering of the Nanyang Technological University (NTU-SCE) to support this SCE-BII PhD Programme in Computational Biology and Bioinformatics.

Resources
SCE has an established infrastructure for students’ admission, guidance, examination, and awarding of academic degrees. The Bioinformatics Research Centre (BIRC) at SCE has faculty members in bioinformatics with considerable research background and ongoing research efforts in the field.

BII will complement SCE in enhancing the profile of a joint PhD course. Faculty of BII provides specialized contributions to bioinformatics/computational biology teaching. In addition, BII will provide the research capacity of its principal investigators with their teams to dramatically enhance the opportunities for research for the PhD students in this programme and in collaboration with the SCE faculty in bioinformatics.

SCE-BII PhD Programme
There will be multiple award of this scholarship to train graduates/postgraduates towards a PhD in Computational Biology and Bioinformatics. Besides a limited set of coursework aimed at complementing the existing knowledge of the applicant in an interdisciplinary manner, involvement in actual research as a member of a research team at BII or SCE will be the main activity during the training.

Supervision and Mentoring of Students
Each PhD student will be supervised by one SCE staff and one BII staff, where one of the supervisors is the main project supervisor and the other is the advisory supervisor. Additional supervisor(s) may be added to the thesis committee if necessary. The PhD research topic will be jointly identified and agreed upon by the SCE and BII supervisors.

The PhD student should fulfill all the coursework, research, and other requirements as stipulated by NTU-SCE for the award of the PhD degree to be conferred by NTU. The supervisor from SCE carries the responsibility for the academic affairs. The PhD thesis committee of each student shall include both supervisors and at least one senior scientist from either BII or SCE.

To earn the PhD degree, the student is required to satisfy the following:

1. To complete 6 courses (18 AUs) within 18 months with a minimum of two 7xxx series courses (at least one must be a SCE CE7xxx series course and other bioinformatics related courses);
2. To pass a qualifying examination by the end of the second year;
3. To pursue an independent in-depth study of approved research topic, leading to a thesis (dissertation);

Eligibility Criteria
Applicants are expected to have:

• completed a full Master of Science Programme in natural sciences or engineering in a related field such as biology, chemistry, physics, computer science or medicine when the PhD is started;
• excellent results during their university studies;
• possessed a strong interest in computational biology/bioinformatics research;
• good performance in the Bachelor’s and Master’s study.

Applicants will be considered from any country on a purely meritocratic basis.

Only shortlisted applicants will be notified.

Further information on matters such as admissions, application procedures, scholarships, fees, etc. is available at:
http://sce.ntu.edu.sg/CurrentStudents/Graduate/Pages/PhD_CBB_prog.aspx
Scientific Meetings and Outreach

Research Exposure Programmes (REP) 2013

▶ A group of research scientists hosting a batch of 45 REP students invited by Singapore Science Centre on 20 Nov 2013. Several scientists took turns to present their respective teams’ projects.

▶ Vladimir Kuznetsov explaining about some award-winning posters to the students on 10 Sept 2013.

▶ Yong Tai Pang showing and explaining our in-house Data Centre and the Data storage facilities to the visitors on 10 Sept 2013.

▶ Chandra Verma giving a Talk to REP students on 5 June 2013.

▶ Irina Kulemzina conducting demo on experimental research to group of students on 10 Sept 2013.

The X-periment!, Singapore Science Festival 2013

▶ An overview of visitors with staff manning our booth at the X-periment! held at Marina Square from 19-21 July 2013.

▶ A young visitor showing his “masterpiece” using tiles to transfer a paper image as part of Imaging and Pattern Discovery exercise.
The 24th International Conference on Genome Informatics (GIW 2013)

Participants at one of the sessions co-hosted by BII with School of Computing, NUS held at Biopolis in Singapore on 16-18 Dec 2013.

Biopolis 10th Anniversary Carnivale Open House 2013

Samantha Kwah and Cecilia Tan engaging visitors at the BII interactive booth.

Adeline Sim demonstrating some research features to visitors during the Carnivale on 18 Oct 2013.

National Science Challenge 2013

Xu Chi giving a demonstration of 3D-CuBE, a BII-developed 3D immersive system for biological computing, that helps users to directly observe, annotate and analyze biological data.

Participants listening intently to a scientist during Agency visit on 1 Aug 2013.
The FY13 Committee comprises of: Hanafi HARRON, Irina KULEMZINA, Dilraj LAMA, NGUYEN Ngoc Minh and YIP Ai Kia. They organised a total of 7 events throughout FY13.

Recreational Club

On 17 Jan 2014, some staff, students and interns brought 22 elderly folks from the Care Corner Social day Care for the Elderly for an excursion to the River Safari and blessed them each with an “Angpow”. The outing expenses were sponsored through funds raised from the Christmas Charity Market.

Halloween BBQ: Christine, Irina, Dilraj and Yen Ling attired in scary accessories and make-up in competition for the most scary look to identify with Halloween and having skulls as prizes.

Annual Dinner held on 22 Feb 2013 at an idyllic & tranquil venue overlooking Straits of Johor. Staff listening to guest speaker, Prof. Bertil Andersson enlightening scientists on subject of Nobel Prize and winners over the decades.

Movie-Cookie Night on 14 Jun 2013: Staff, students and interns munching on delicious cookies prepared by some colleagues before enjoying either movie “Life of Pi” or “Skyfall” screened inhouse.

Annual Dinner held on 22 Feb 2013 at an idyllic & tranquil venue overlooking Straits of Johor. Staff listening to guest speaker, Prof. Bertil Andersson enlightening scientists on subject of Nobel Prize and winners over the decades.

Movie-Cookie Night on 14 Jun 2013: Staff, students and interns munching on delicious cookies prepared by some colleagues before enjoying either movie “Life of Pi” or “Skyfall” screened inhouse.

Movie-Cookie Night on 14 Jun 2013: Staff, students and interns munching on delicious cookies prepared by some colleagues before enjoying either movie “Life of Pi” or “Skyfall” screened inhouse.

On 17 Jan 2014, some staff, students and interns brought 22 elderly folks from the Care Corner Social day Care for the Elderly for an excursion to the River Safari and blessed them each with an “Angpow”. The outing expenses were sponsored through funds raised from the Christmas Charity Market.
BII Location

Address
Bioinformatics Institute
30 Biopolis Street #07-01 Matrix
Singapore 138671

By Car
For visitors who drive, please park your vehicle at B3 (basement 3) and follow the signage “To Matrix Lift Lobby” to locate lift D. You may take lift D to level 1 and approach the receptionist for the visitor’s pass.

By Bus
The following are the Singapore Bus Service Numbers that stop along North Buona Vista Road: 74, 91, 92, 95, 191, 196, 198, 200

By MRT
Board the East-West line or the Circle line and alight at Buona Vista MRT Station. After alighting, walk towards Matrix building at Biopolis Street.