# Table of Contents

- **Welcome Note** .................................................................................................................. 4
- **About APBioNet** .............................................................................................................. 6
- **About Universitas YARSI** .................................................................................................. 6
- **Conference Schedule** ....................................................................................................... 7
- **Oral Presentation Schedule** .............................................................................................. 11
- **Flash Oral Presentation Schedule** ..................................................................................... 15
- **Poster Presentation Schedule** ........................................................................................... 15
- **Guidelines for Oral and Flash Oral Presentation** ............................................................... 18
- **Guidelines for Poster Presentation** ................................................................................... 18
- **Keynote Talks** .................................................................................................................. 19
- **Plenary Talks** .................................................................................................................... 27
- **Workshops** ....................................................................................................................... 30
- **Breakout Session** ............................................................................................................. 35
- **Highlight Track** ................................................................................................................ 37
- **Oral Presentation Abstracts** .............................................................................................. 39
- **Special Industry Session** .................................................................................................. 88
- **Special Session: Bioinformatics in Indonesia** ................................................................. 92
- **Software Demo** ................................................................................................................. 108
- **Poster Presentation Abstracts** .......................................................................................... 110
Welcome Note

Dear colleagues and guests,

On behalf of Asia-Pacific Bioinformatics Network ExCo members and the active local organising committee (LOC), it gives me a great pleasure to welcome all of you to the 18th International Conference on Bioinformatics, held in Jakarta, Indonesia for the first time. The idea to bring InCoB to Indonesia crystallised two-years ago, with the inauguration of APBioNet’s first International Symposium on Bioinformatics (InSyB) in 2017, also held in Jakarta. We congratulate Universitas Yarsi for hosting both the inaugural InSyB and today’s 18th InCoB. It has been a fruitful partnership.

InCoB2019 has taken on the theme of “Bioinformatics for Precision Medicine”, well suited with the current international push in the direction. This year’s InCoB is also special, in that it is for the first-time co-located with the Global Organisation for Bioinformatics Learning (GOBLET) Annual General Meeting (AGM), Genomic Medicine Conference (GMC), as well as the South East Asian Pharmacogenomics Research Network (SEAPharm) meeting, including complimentary workshops by our distinguished speakers and trainers. We implore our participant to take full advantage of this event to further expand your network circle.

InCoB is one of the leading bioinformatics and allied discipline conference in the Asia-Pacific region. We are proud of our history and the outcome that it brings to our audience. In particular, InCoB supplement publications with BioMed Central journals have been one of the channels for researchers and students alike to publish their articles in various areas of bioinformatics, and we are actively looking into providing alternatives to our audience in the future.

I hope that you will benefit from the scientific program and the workshops; and that future research collaborations can blossom from the network you will gain here. We welcome all of you to attend the plenaries, keynotes, oral and poster presentations by our distinguished speakers.

I would like to thank the local staff helpers, session chairs, keynote speakers for blueprinting this conference program. We hope you have a fruitful and memorable experience; one that you will continue to talk post conference.

Finally, I wish you have a 48 hours of productive and successful conference!

Thank you.

Mohammad Asif Khan, PhD
President,
Asia-Pacific Bioinformatics Network (APBioNet)
Welcome to the 18th International Conference on Bioinformatics (InCoB) in Jakarta, Indonesia. We are very grateful and honoured to host the oldest bioinformatics conference of Asia Pacific for the first time in Indonesia. The conference is co-located with Genomic Medicine Conference (GMC), South East Asian Pharmacogenomics Research Network (SEAPHARM) meeting and the Global Organisation for Bioinformatics Learning, Education and Training (GOBLET) Annual General Meeting 2019.

We are excited to have 10 keynotes and plenary talks in the conference. We also provide four complimentary workshops for all InCoB 2019 participants; two of them are supported by GOBLET. We thank them very much for their generous time and significant contribution to the conference.

InCoB 2019 call-for-papers and call-for-posters attracted more than 200 submissions. The review panel was integrated by more than 70 experts, from 20 different nationalities. Each submissions received at least a double peer review. The rigorous review process gives us confidence in the high scientific standard of the 50 oral presentations, 15 flash oral presentations, 34 posters, 4 workshop, 2 highlights, 2 software demos and 1 breakout session accepted. Accepted full papers are now in consideration for publication in one of the participating journals, namely BMC Genomics, BMC Medical Genomics, BMC Systems Biology, BMC Bioinformatics, GigaScience, PeerJ, and Computational Biology and Chemistry (CBAC). We thank the Program Committee Chairs and Members for all their hard work and continuous support for InCoB.

The conference is organised by Universitas Yarsi, in collaboration with Asia-Pacific Bioinformatics Network (APBioNet), and supported by International Society for Computational Biology (ISCB), GOBLET, Masyarakat Bioinformatika dan Biodiversitas Indonesia (MABBi), Ikatan Dokter Indonesia (IDI), and SEAPHARM. We also would like to thank the sponsors for their financial support, the local organizing committee and the student volunteers for their diligent work and dedication, and all the registrants for their participation, and others who contributed directly or indirectly to the success of the conference.

We wish you a memorable conference experience, full of new learnings, exchange, and inspiring moments. Enjoy the conference and enjoy Jakarta!

Rika Yuliwulandari
General Chair
InCoB 2019
About APBioNet

Founded in January 1998, APBioNet is dedicated to the advancement of the field of bioinformatics, specifically, the development of the bioinformatics network infrastructure, the exchange of data and information, the development of training programs, workshops and symposia and the encouragement of collaborations in the field of bioinformatics with an Asia Pacific focus (1). To further the development of the bioinformatics network infrastructure, APBioNet has formed a partnership with the Asia-Pacific Advanced Network project APAN. This project has been endorsed as a priority status project within APEC’s Telecommunications Working Group (APEC TEL Development Cooperation Steering Group – DCSG) in 1999. APBioNet’s collaboration with APAN has paved the way to link DDBJ, GenomeNet, HGC, NCC, ANGIS, MAFFIN, as well as, allow individuals to exchange data and other information quickly. APBioNet has developed web content and maintains servers dedicated to bioinformatics to assist collaboration and information exchange. APBioNet also has a memorandum of understanding with the Asia-Pacific International Molecular Biology Network (A-IMBN) to work with their members to promote awareness among biologists of the need to acquire bioinformatics skills. APBioNet has also sponsored or co-sponsored or co-organised workshops and conferences in the areas of bioinformatics, molecular phylogenetics, biomolecular structural analysis, protein simulations and others, including the International Conference on Bioinformatics (InCOB’02) in Thailand. APBioNet has pioneered outreach programmes that has involved visits and training workshops in Kuala Lumpur, Penang, Taipei, HongKong, Xian, Beijing, Bogor, Bangalore, Bangkok, Los Banos, etc.

About Universitas YARSI

Universitas YARSI (http://yarsi.ac.id) is an Indonesian premier research university which strives to enhance and strengthen its educational programs and has taken various initiatives to complement its educational excellence. Universitas YARSI started its history more than 50 years ago with medical school for undergraduate students. It now has 6 faculties: medicine, dentistry, law, economy, information technology and psychology. Additionally, it also established several post graduate programs: Master of Management, Master of Notary, and Master of Biomedical Sciences. Setting the goal to be in 500 top universities in the world, it is now committed to research more than ever. YARSI research institute was established in 2012 to support its mission to be a top-notch research university. It consists of six core research centers: genetics, stem cell, herbal, telomere, e-health and halal research centers.
# Conference Schedule

**International Conference on Bioinformatics 2019 and co-located events**  
Universitas YARSI, September 9-12, 2019

<table>
<thead>
<tr>
<th>VENUE</th>
<th>TIME</th>
<th>PROGRAM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Monday, September 9, 2019</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Workshop 5 &amp; Ar-Rahim auditorium</td>
<td>9:00 AM - 6:00 PM</td>
<td>GOBLET Annual General Meeting 2019 (Workshop room 5, 11th floor) SEAPharm Meeting 2019 (Ar-rahim auditorium, 12th floor)</td>
</tr>
<tr>
<td>Al-Qudus auditorium</td>
<td>6:30 PM - 8:00 PM</td>
<td>GOBLET and SEAPharm Dinner</td>
</tr>
</tbody>
</table>

<p>| <strong>Tuesday, September 10, 2019</strong> | | |
| Ar-Rahim Auditorium | 8:00-8:40 AM | Registration |
| <strong>8:40-10:00 AM</strong> | <strong>InCoB 2019 Inauguration</strong> | |
| 8:40-8:45 AM | • Recitation of the Holy Quran (Muhammad Badrul Amali) |
| 8:45-8:50 AM | • Singing National Anthem &quot;Indonesia Raya&quot; (Lead by: Talitha) |
| 8:50-8:55 AM | • Address by Rector of Universitas YARSI (Prof. Fasli Jalal, PhD) |
| 8:55-9:00 AM | • Address by Conference Chair (Rika Yuliwulandari, MD, PhD) |
| 9:00-9:05 AM | • Address by APBioNet President (Dr. Asif M. Khan) |
| 9:05-9:30 AM | • Opening speech by Minister of Research, Technology and Higher Education (Prof. Mohamad Nasir, Ph.D., Ak.) |
| 9:50-10:00 AM | • Special Talk: Dr. Mohammad Dimyati (Directorate General of Research and Development at the Ministry of Research, Technology and Higher Education) |
| 11th storey corridor | 10:00-10:15 AM | Coffee Break |
| Ar-Rahim Auditorium | 10:15-11:00 AM | Keynote talk 1: Prof. Dr. Sir Munir Pirmohammed, &quot;Discovering and implementing drug safety biomarkers&quot; (Chair: Prof. Christian Schoenbach) |
| Ar-Rahim Auditorium | 11:00-11:45 AM | Keynote talk 2: Dr. Eija Korpelainen, “Tackling the bioinformatics skills gap” (GOBLET speaker) (Chair: Dr. Yam Wai Keat) |</p>
<table>
<thead>
<tr>
<th>Time</th>
<th>Location</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.45-1:00 PM</td>
<td>Lunch Break &amp; Poster Viewing</td>
<td>Keynote talk 3: Prof. Dr. Katsushi Tokunaga, “Genomic approach to immune-mediated complex diseases” (Chair: Dr. Rika Yuliwulandari)</td>
</tr>
<tr>
<td>1:00-1:45 PM</td>
<td>Ar-Rahim Auditorium</td>
<td>GOBLET AGM 2019 continues (Workshop room 5, 11th floor)</td>
</tr>
<tr>
<td>1:45-2:45 PM</td>
<td>Ar-rahim Auditorium; Workshop room 1-5</td>
<td>Oral Presentation (Parallel session): GMC-SEAPharm conference 2019 (Ar-rahim auditorium, 12th floor)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Theme I: Sequencing and NGS data analysis (Workshop Room 1) (Chair: Dr. Harpreet Singh)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Theme II: Protein structure, function and interaction (Workshop Room 2) (Chair: Dr. Choi Sy Bing)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Theme III: Immunoinformatics and host pathogen interactions (Workshop Room 3) (Chair: Dr. Christian Schoenbach)</td>
</tr>
<tr>
<td>2:45-3:00 PM</td>
<td>11th storey corridor</td>
<td>Coffee Break</td>
</tr>
<tr>
<td>3:00-4:30 PM</td>
<td>Ar-rahim Auditorium; Workshop room 1-5</td>
<td>Oral presentation (Parallel session): GMC-SEAPharm conference 2019 continues</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Theme IV: Genomics and Evolutionary Biology (Workshop Room 1) (Chair: Dr. Shandar Ahmad)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Theme V: Tools, databases and web services in Bioinformatics (Workshop Room 2) (Chair: Dr. Choi Sy Bing)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Theme VI: Network biology and interaction networks (Workshop Room 3) (Chair: Dr. Christian Schoenbach)</td>
</tr>
<tr>
<td>4:30-5:30 PM</td>
<td>Workshop room 4</td>
<td>APBioNet AGM (Chair: Dr. Asif M. Khan) - all are welcome</td>
</tr>
<tr>
<td>6:00-7.30 PM</td>
<td>Al-Qudus Auditorium</td>
<td>Welcome reception (Dinner) for All</td>
</tr>
<tr>
<td>Time</td>
<td>Location</td>
<td>Event</td>
</tr>
<tr>
<td>-----------------</td>
<td>-------------------------------</td>
<td>-----------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>9:00-9:45 AM</td>
<td>Ar-rahim Auditorium</td>
<td>Keynote talk 4: Prof. Dr. Christine Orengo, “CATH Functional families - insights into impacts of genetic variations” (ISCB Speaker; Chair: Dr. Bruno Gaeta)</td>
</tr>
<tr>
<td>9:45-10:00 AM</td>
<td>11th storey corridor</td>
<td>Coffee Break</td>
</tr>
<tr>
<td>10:00-10:45 AM</td>
<td>Ar-rahim Auditorium</td>
<td>Keynote talk 5: Dr. Denis Bauer, “Cloud computing and artificial intelligence transforms bioinformatics research” (Chair: Prof. Shoba Ranganathan)</td>
</tr>
</tbody>
</table>
| 10:45-11:45 AM  | Ar-rahim and Al-Qudus Auditorium | Special session: Bioinformatics in Indonesia (Chair: Prof. Asif M. Khan)  
Plenary Talk: A National Center for Bioinformatics is Required for Optimizing Biodiversity-Based Research in Indonesia (Dr. Kholis A. Audah)  
Special industry session: Novocraft and Illumina (Chair: Dave Clements)  
Novocraft Talk: NovoAlign (V4) run time reduction and performance benchmarking using Freebayes variant caller  
Illumina Talk: Illumina’s genomics solutions - enabling fast, accurate, cost effective analysis |
| 11:45-12:45 AM  | Workshop room 1-4             | Oral Flash presentation (Workshop room 5 and 6) (Chair: Dr. Afiahayati and Dr. Wisnu Ananta Kusuma)  
Oral presentation (Parallel session):  
Theme VII: Mass spectrometry and nanobioinformatics (Workshop room 2) (Chair: Dr. Shoba Ranganathan)  
Theme VIII: Genome wide association studies (GWAS) and Biomarker discovery (Workshop Room 3) (Chair: Dr. Adaikalavan Ramasamy)  
Theme IX: Machine learning, AI and novel algorithms-I (Workshop Room 4) (Chair: Dr. Harpreet Singh) |
<p>| 12:45-1:30 PM   | 11th storey corridor          | Lunch Break &amp; Poster Viewing                                                           |</p>
<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>1:30-3:00 PM</td>
<td>Workshop: Detecting differentially expressed genes with RNA-seq (WS room 5)</td>
</tr>
<tr>
<td>Workshop room 5-8</td>
<td>Workshop: An introduction to the Galaxy platform for computational biology – with real-world hands-on demonstration (WS room 1)</td>
</tr>
<tr>
<td>Workshop room 5-8</td>
<td>Workshop: Protein Sequence and Structure Analysis Using Google Cloud Engine (WS room 7)</td>
</tr>
<tr>
<td>Workshop room 5-8</td>
<td>Workshop: Train the Trainer (WS room 8)</td>
</tr>
<tr>
<td>3:00-3:20 PM</td>
<td>Coffee Break</td>
</tr>
<tr>
<td>3:20-5:00 PM</td>
<td>Workshop continues</td>
</tr>
<tr>
<td>3:20-5:30 PM</td>
<td>MABBI General Meeting 2019 - <em>All are welcome</em></td>
</tr>
<tr>
<td>5:30 PM - end</td>
<td>Sightseeing (optional)</td>
</tr>
</tbody>
</table>

**Thursday, September 12, 2019**

<table>
<thead>
<tr>
<th>Time</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:00-09:45 AM</td>
<td>Keynote talk 6: Prof. Dr. Heru Suhartanto, “High Performance Computing challenges in in-silico drug design based on Indonesia Medical Plants” (Chair: Dr. Choi Sy Bing)</td>
</tr>
<tr>
<td>Ar-rahim Auditorium</td>
<td>Highlights track (Workshop room 5) (Chair: Dr. Gareth Price)</td>
</tr>
<tr>
<td>Ar-rahim Auditorium</td>
<td>Software Demo (Workshop room 6) (Prof. Christian Schonbach)</td>
</tr>
<tr>
<td>Ar-rahim Auditorium</td>
<td>Highlight 1: A comprehensive map of intron branchpoints and lariat RNAs in plants (Yun Zheng)</td>
</tr>
<tr>
<td>Ar-rahim Auditorium</td>
<td>Software Demo 1: novoWorx: A Genome Data Management &amp; Analytics Platform (Kaamesh Ganisen)</td>
</tr>
<tr>
<td>Ar-rahim Auditorium</td>
<td>Highlight 2: Characterization of cyclophilin gene family in wheat and its implications in heat stress response (Harpreeet Singh)</td>
</tr>
<tr>
<td>Ar-rahim Auditorium</td>
<td>Software Demo 2: abciba: a background correction R package for the illumina bead arrays (Rohmatul Fajriyah)</td>
</tr>
<tr>
<td>10:15-10:30 AM</td>
<td>Coffee Break</td>
</tr>
<tr>
<td>Ar-rahim Auditorium</td>
<td>Oral presentation (Parallel session): Breakout Session: Building a regional Galaxy Community and Platform (Ar-rahim auditorium) (Chair: Dave Clements)</td>
</tr>
<tr>
<td>Ar-rahim Auditorium</td>
<td>Plenary talk: Prof. Shandar Ahmad, “Deep versus conventional learning for biological data through changing paradigms” (Chair: Dr. Gareth Price)</td>
</tr>
<tr>
<td>12.00-1:00 PM</td>
<td>Lunch Break &amp; Poster Viewing</td>
</tr>
<tr>
<td>Workshop room 1-2</td>
<td>Oral presentation (Parallel session): Theme XI: Machine learning, AI and novel algorithms-III (Workshop room 1) (Chair: Dr. Gareth Price)</td>
</tr>
<tr>
<td>Workshop room 1-2</td>
<td>Special Lecture: Population Genetics, Prof. Akihiro Fujimoto (Workshop room 3) (Chair: Dr. Rika Yuliwulandari)</td>
</tr>
<tr>
<td>TIME</td>
<td>Workshop Room 1</td>
</tr>
<tr>
<td>----------</td>
<td>--------------------------------------------------------------------------------</td>
</tr>
<tr>
<td></td>
<td>Theme XI: Sequecing and NGS data analysis</td>
</tr>
<tr>
<td>2:15-2:30 PM</td>
<td>O-32: Younghi Lee: Differential alternative splicing regulation among hepatocellular</td>
</tr>
<tr>
<td>2:30-2:45 PM</td>
<td></td>
</tr>
<tr>
<td>Time</td>
<td>Session</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------------------------------------------------------------------</td>
</tr>
<tr>
<td>3:00-4:30 PM</td>
<td><strong>Theme IV: Genomics and Evolutionary Biology</strong></td>
</tr>
<tr>
<td>3:00-3:15 PM</td>
<td><strong>Theme V: Tools, databases and web services in Bioinformatics</strong></td>
</tr>
<tr>
<td>3:15-3:30 PM</td>
<td><strong>Theme VI: Network biology and interaction networks</strong></td>
</tr>
<tr>
<td>3:30-3:45 PM</td>
<td><strong>Theme IV: Genomics and Evolutionary Biology</strong></td>
</tr>
<tr>
<td>3:45-4:00 PM</td>
<td><strong>Theme IV: Genomics and Evolutionary Biology</strong></td>
</tr>
<tr>
<td>4:00-4:15 PM</td>
<td><strong>Theme IV: Genomics and Evolutionary Biology</strong></td>
</tr>
<tr>
<td>4:15-4:30 PM</td>
<td><strong>Theme IV: Genomics and Evolutionary Biology</strong></td>
</tr>
<tr>
<td>4:30-4:45 PM</td>
<td><strong>Theme IV: Genomics and Evolutionary Biology</strong></td>
</tr>
<tr>
<td>Time</td>
<td>Theme VII: Mass spectrometry and nano-bioinformatics</td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------------------------------------------------</td>
</tr>
<tr>
<td>11:45-12:00 PM</td>
<td>O-03: Zaved Hazarika: Computational analysis of Silver nanoparticle - human serum albumin complex</td>
</tr>
<tr>
<td>12:15-12:30 PM</td>
<td>O-09: Yang Ming Lin: MS2CNN: Predicting MS/MS spectrum based on protein sequence by Deep Convolutional Neural Networks</td>
</tr>
</tbody>
</table>

**Thursday, September 12, 2019**

<table>
<thead>
<tr>
<th>Time</th>
<th>Theme X: Machine learning, AI and novel algorithms-II</th>
<th>NA</th>
<th>NA</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:00-12:00 PM</td>
<td>O-38: Jacob Bradford: Improving CRISPR guide design with consensus approaches</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>11:00-11:15 AM</td>
<td>O-40: Ling Zou: Predicting synergistic drugs using gradient tree boosting based on features extracted from</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>11:15-11:30 AM</td>
<td></td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Time</td>
<td>Session</td>
<td>Title</td>
<td>Abstract</td>
</tr>
<tr>
<td>--------------</td>
<td>--------------------------------------------------------------------------</td>
<td>-----------------------------------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>1:00-1:45 PM</td>
<td>Theme XI: Machine learning, AI and novel algorithms-III</td>
<td>Theme XII: Disease data modeling and integrative Biology</td>
<td>NA</td>
</tr>
<tr>
<td>1:00-1:15 PM</td>
<td>O-21: Lun Li: A novel constrained reconstruction model towards high-resolution sub-tomogram averaging</td>
<td>O-24: Shobana Sundar: Rv0807, a putative phospholipase A2 of Mycobacterium tuberculosis; Elucidation through sequence analysis, homology modeling, molecular docking and molecular dynamics studies of potential substrates and inhibitors</td>
<td></td>
</tr>
<tr>
<td>1:30-1:45 PM</td>
<td></td>
<td>O-29: Vivitri Dewi Prasastry: Structure-based Discovery of Novel Inhibitors of Mycobacterium tuberculosis CYP121 from Indonesian Natural Products</td>
<td></td>
</tr>
<tr>
<td>1:45-2:00 PM</td>
<td></td>
<td>O-51: Asif M. Khan: A systematic bioinformatics approach for large-scale identification and characterization of host-pathogen shared sequences</td>
<td></td>
</tr>
</tbody>
</table>
Flash Oral Presentation Schedule

Room: Workshop 5

<table>
<thead>
<tr>
<th>Time</th>
<th>ID</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.45-11.52 AM</td>
<td>OF-01</td>
<td>Wahyu Hidayati</td>
</tr>
<tr>
<td>11.52-11.59 AM</td>
<td>OF-04</td>
<td>Akhmad Kamal Nasution</td>
</tr>
<tr>
<td>11.59-12.06 PM</td>
<td>OF-05</td>
<td>Nur Imaniat Sumantri</td>
</tr>
<tr>
<td>12.06-12.13 PM</td>
<td>OF-11</td>
<td>Larasati</td>
</tr>
<tr>
<td>12.13-12.20 PM</td>
<td>OF-12</td>
<td>Alfiatun Hasanah</td>
</tr>
<tr>
<td>12.20-12.27 PM</td>
<td>OF-13</td>
<td>Wisnu Ananta Kusuma</td>
</tr>
<tr>
<td>12.27-12.34 PM</td>
<td>OF-15</td>
<td>Muhammad Arba</td>
</tr>
</tbody>
</table>

Room: Workshop 6

<table>
<thead>
<tr>
<th>Time</th>
<th>ID</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.45-11.52 AM</td>
<td>OF-02</td>
<td>Marsia Gustiananda</td>
</tr>
<tr>
<td>11.52-11.59 AM</td>
<td>OF-03</td>
<td>Fransisca Srioetami Tanochardjo</td>
</tr>
<tr>
<td>11.59-12.06 PM</td>
<td>OF-06</td>
<td>Devy Apriansyah</td>
</tr>
<tr>
<td>12.06-12.13 PM</td>
<td>OF-07</td>
<td>Fadhil Khaliq Surado</td>
</tr>
<tr>
<td>12.13-12.20 PM</td>
<td>OF-08</td>
<td>Nabila Sekar Ramadhanti</td>
</tr>
<tr>
<td>12.20-12.27 PM</td>
<td>OF-09</td>
<td>Maria Susan Anggreainy</td>
</tr>
<tr>
<td>12.27-12.34 PM</td>
<td>OF-10</td>
<td>Marta Nisita Dewanggana</td>
</tr>
<tr>
<td>11.45-11.52 AM</td>
<td>OF-14</td>
<td>Kindi Adam</td>
</tr>
</tbody>
</table>

Poster Presentation Schedule

⇒ Odd serial numbered posters will be presented on the first day of the conference (September 10, 2019).
⇒ Even serial numbered posters will be presented on the second day of the conference (September 11, 2019).

<table>
<thead>
<tr>
<th>Poster Presentation Number</th>
<th>Corresponding Authors</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>P-03</td>
<td>Koel Mukherjee</td>
<td>Studying – analyzing - evaluating and reusing of old drug molecules against new target protein causing human cervical cancer- an in silico approach</td>
</tr>
<tr>
<td>P-05</td>
<td>Hendrick Gao-Min Lim</td>
<td>A Comparative Study between Transcript-level and Gene-level Classification on Multiple Affymetrix Platforms of Colorectal Cancer Microarray Studies</td>
</tr>
<tr>
<td>P-06</td>
<td>Lokesh Thangamani</td>
<td>Rv0807, a putative phospholipase A2 of Mycobacterium tuberculosis; Elucidation through sequence analysis, homology modeling, molecular docking and molecular dynamics studies of potential substrates and inhibitors</td>
</tr>
<tr>
<td>P-07</td>
<td>Jagath Rajapakse</td>
<td>A deep neural network approach to predicting clinical outcomes of neuroblastoma patients</td>
</tr>
<tr>
<td>P-08</td>
<td>Seema Mishra</td>
<td>Target identification for Multi-Target Drug Designing for Mycobacterium tuberculosis active infection by systems biology approach</td>
</tr>
<tr>
<td>P-09</td>
<td>Adil El-Hussain</td>
<td>In Silico Simulation of Bacteriophages and Salmonella typhi Population Dynamics</td>
</tr>
<tr>
<td>P-11</td>
<td>Hiroyuki Toh</td>
<td>Alignment Viewer with a Function to Collect Annotation of Proteins from LOD</td>
</tr>
<tr>
<td>P-12</td>
<td>Abhirupa Ghosh</td>
<td>Prediction of drug resistance in MTB using Machine learning algorithms</td>
</tr>
<tr>
<td>P-13</td>
<td>Sudipto Bhattacharjee</td>
<td>Prediction of Lung Diseases with Pulmonary Function Test (PFT) Features using Machine Learning</td>
</tr>
<tr>
<td>P-14</td>
<td>Steven G. Rozen</td>
<td>Evaluation of software for mutational signature analysis based on realistic synthetic data</td>
</tr>
<tr>
<td>P-16</td>
<td>Muhammad Arba</td>
<td>Identification of Phosphatidylinositol 3-kinase δ (PI3Kδ) Inhibitor: Pharmacophore-based Virtual Screening and Molecular Dynamics Simulation</td>
</tr>
<tr>
<td>P-17</td>
<td>Nur Imaniaty Sumantri</td>
<td>Structural Insight of Alteration Intronic Beta Globin Mutations in Beta Thalassemia Major via mRNA Folding</td>
</tr>
<tr>
<td>P-18</td>
<td>Fadilah Fadilah</td>
<td>Ligand-based drug design of Pinostrobin derivatives by pharmacophore modelling as wild-type and mutant her2 inhibitor</td>
</tr>
<tr>
<td>P-19</td>
<td>Jamil Ahmad</td>
<td>3D structure determination and validation of mTORC2 using ab initio and homology modeling techniques</td>
</tr>
<tr>
<td>P-20</td>
<td>Linda Erlina</td>
<td>Molecular Docking and Molecular Dynamics Simulation Analysis For Finding HDAC Class IIa Inhibitor as New Antidiabetic Drug</td>
</tr>
<tr>
<td>P-21</td>
<td>Hui San Ong</td>
<td>Advancing pathogenicity profiling through genomics big data analytics</td>
</tr>
<tr>
<td>P-22</td>
<td>Shubha Sharma</td>
<td>Study on Activation of Constitutive Androstane Receptor(CAR) by Herbs for Prevention of Gallstone –A Computational Approach</td>
</tr>
<tr>
<td>P-23</td>
<td>Lautaro Soler</td>
<td>Image processing on the brain: 3D printing MRI images for further observation</td>
</tr>
<tr>
<td>P-24</td>
<td>Kanupriya Tiwari</td>
<td>Mapping Key Alternative Splicing events during Neural Crest Cell Delamination</td>
</tr>
<tr>
<td>P-25</td>
<td>Fadel Berakdar</td>
<td>RNA-seq Analysis Benchmark: Datasets &amp; Tools Resources</td>
</tr>
<tr>
<td>P-26</td>
<td>Afiahayati</td>
<td>Classification for Buccal Mucosa’s DNA Damage Measurement with Super Tiny Dataset Using Transfer Learning</td>
</tr>
<tr>
<td>P-27</td>
<td>Edward Daniel Gomez</td>
<td>Molecular Profiling of the Nonconopeptides in the Envenomation Strategy of Conus lenavati through High-Throughput Sequencing Platforms</td>
</tr>
<tr>
<td>P-28</td>
<td>Priya Kumari</td>
<td>Comparative study of microbial nitrilase enzymes: A computational approach</td>
</tr>
<tr>
<td>P-29</td>
<td>Rika Yuliwulandari</td>
<td>HLA-B*15:02 As Predisposition Allele to Carbamazepine Induced SJS/TEN is Frequently Observed in Indonesia</td>
</tr>
<tr>
<td>P-30</td>
<td>Restu Syamsul Hadi</td>
<td>Platelet Rich Plasma Stimulate Human Dermal Fibroblast Viability, Proliferation and Migration in High Glucose Condition</td>
</tr>
<tr>
<td>-----</td>
<td>-------------------</td>
<td>---------------------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>P-31</td>
<td>Ibnul Asifuzzaman</td>
<td>A Robust Missing Value Imputation Approach for Microarray Gene Expression Data Analysis</td>
</tr>
<tr>
<td>P-32</td>
<td>Retno Wilujeng Susilowati, Kinasih Prayuni</td>
<td>Isoniazid Metabolizing Enzyme Genotypes Distribution in Indonesian Populations: Awareness to Anti-Tuberculosis Drug Induced Liver Injury (AT-DILI)</td>
</tr>
<tr>
<td>P-33</td>
<td>Ari Yuniastuti</td>
<td>Inulin of Lesser Yam (Dioscorea Esculenta) Tuber As An Colorectal Anticancer Agent In Silico</td>
</tr>
</tbody>
</table>
**Guidelines for Oral and Flash Oral Presentation**

**PRESENTATION TIME**
- All speakers will be allotted 15 minutes (for oral presentation) or 7 minutes (for flash oral presentation) to speak. This includes setup and questions.
- Going overtime is not permitted.
- Presentation slides must cover the materials as cited in the abstract.

**MATERIAL**
- Speakers are encouraged to have their completed presentation on a USB/Flash drive in 16:9 (widescreen) format. The USB/Flash drive should be presented to the operator during registration.
- In order have seamless transitions between presentations, speakers are not allowed to use their own laptops. All slideshows will be copied onto main presentation device.
- Only fonts that are included in the basic installation of MS-Windows will be available (English version of Windows). Use of other fonts can impact the layout and visual of the presentation. If speakers insist on using different fonts, these fonts must be embedded into presentation.

**Guidelines for Poster Presentation**
- Poster should be in an A1 size. The Orientation of your poster must be in portrait style.
- The poster should be easily readable at a distance of one meter
- Avoid using a mixture of type/font styles.
- Poster pasting material will be provided by the conference organizing committee at the registration desk.
- There is no printing service on the conference venue. Please bring your printed copy with you to avoid inconvenience. However, shops with printing services within 500 m distance from the conference site may be available for which authors may explore the details on their own.
- The poster must be presented by one of the co-authors as mentioned in your original submission. Proxy presentations are not allowed.
- You are requested to set up your poster in the morning of the first day of the conference i.e. September 10, 2019.
- Each poster will be assigned a poster number. Please watch out this space for your poster number.
- Please ensure that your poster is displayed exactly on the poster board carrying the number assigned to you.
- Posters will remain on display throughout the conference and should be removed after the Valedictory program.
- Presenting authors should be available on their poster display position during the poster session allotted to them (first and second day) for discussion with conference participants and judges who will view and optionally discuss their poster contents at their convenience.
- Each poster will require that at least one author has registered for the conference exclusively representing that poster. If none of the authors of the poster has registered for the conference, the poster boards will not be issued to the poster.
Keynote Talks

Keynote Talk 1

Professor Sir Munir Pirmohamed MB ChB (Hons), MRCP, PhD, FRCP, FRCP(E), FBPharmacolS (University of Liverpool, UK)

Biography

Professor Sir Munir Pirmohamed is currently David Weatherall Chair in Medicine at the University of Liverpool, and a Consultant Physician at the Royal Liverpool University Hospital. He is also the Associate Executive Pro Vice Chancellor for Clinical Research and Head of Department of Molecular and Clinical Pharmacology. He also holds the only NHS Chair of Pharmacogenetics in the UK, and is Director of the M.R.C. Centre for Drug Safety Sciences, and Director of the Wolfson Centre for Personalised Medicine. He is also an inaugural NIHR Senior Investigator, and Fellow of the Academy of Medical Sciences in the UK. He has authored over 350 peer-reviewed publications.

Professor Pirmohamed’s research focuses on individual variability in drug response (including anti-cancer drugs), both safety and efficacy, with a view to evaluating the mechanisms, and identifying strategies to personalise healthcare in order to optimise drug efficacy and minimise toxicity. The work spans the whole spectrum from discovery to implementation with laboratory based studies being linked translationally to patient studies, with the aim being to develop the evidence base that can move discoveries from the lab to the clinic, and from clinic to application. Professor Pirmohamed has received a number of honours including most recently, the William Withering Medal from the Royal College of Physicians.

Abstract

DISCOVERING AND IMPLEMENTING DRUG SAFETY BIOMARKERS

Adverse drug reactions (ADRs) are a major clinical problem accounting for a great deal of morbidity, mortality and are a drain on healthcare resources. ADRs can generally be divided into on-target and off target reactions. Both types of ADRs have a genetic predisposition, but the quantitative contribution of genetic vs. non-genetic factors varies with the type of reaction, the drug implicated and the patient’s clinical co-morbidities. Genomic research in this area has progressed from discovery, using candidate gene approaches, to genome-wide approaches. The latter has met with a lot of success because the effect size of pharmacogenomic biomarkers is greater than the effect size observed for complex diseases. A notable area has been the identification of different HLA alleles predisposing to a variety of serious immune mediated
adverse drug reactions. The effect size in such cases can be so high that it allows for implementation into practice, which has prevented ADRs in patients. An example of this is the use of HLA-B*57:01 to prevent abacavir hypersensitivity, and HLA-B*15:02 to prevent carbamazepine-induced Stevens-Johnson Syndrome and toxic epidermal necrolysis in South East Asian patients. Genetic predisposing loci outside the HLA region are also being discovered which is providing new insights into the mechanisms of the ADRs. Although the general framework for implementation has focused on the use of genetic tests to predict and prevent ADRs, genetic tests can be used in more novel ways to improve patient outcomes, including for diagnosis and stratification – such approaches will be discussed in the presentation.

**Keynote 2: GOBLET Speaker**

![Dr. Eija Korpelainen](image)

### Dr. Eija Korpelainen
(CSC-IT Center for Science, Finland)

**Biography**

Dr. Eija Korpelainen works currently as a Product manager at CSC – IT Center for Science. Being the national supercomputing centre, CSC provides bioinformatics services for all the universities and research institutes in Finland. Eija has long experience in enabling life scientists to analyze high-throughput data. She has run over 100 training courses in Finland and abroad, and her team has established a popular YouTube channel with tutorial videos. The book “RNA-seq data analysis – practical approach” by Korpelainen et al has been one of the most popular bioinformatics books in recent years. In order to enable biologists to analyze data more efficiently, her team has developed the open source analysis software Chipster which is used worldwide. Eija has participated in several international consortia including the EU funded EMBRACE, AllBio and SeqAhead, and she is currently the national training coordinator of the European ELIXIR initiative. Eija has been involved in GOBLET from the start. Eija holds a PhD from the Medical faculty of University of Adelaide, Australia, and she completed her post doctoral research as an EMBO long term fellow in the University of Helsinki, Finland. Eija has over 40 peer-reviewed publications and an H-index of 21.

**Abstract**

**TACKLING THE BIOINFORMATICS SKILLS GAP**

Life sciences have been revolutionized by high-throughput technologies, and new applications of sequencing are emerging constantly. In order to fully exploit the potential of these new technologies, researchers need sufficient data analysis skills and computing resources. While Big Data and cloud solutions have been actively developed in order to solve the computing challenge, data analysis skills of life scientists are still lagging behind.
In this talk I discuss tackling the bioinformatics skills gap at two levels. Firstly, I introduce GOBLET, the Global Organization for Bioinformatics Learning, Education and Training. Secondly, I discuss the different approaches that our national bioinformatics core facility in Finland has used in order to enable life scientists to analyse data. These two topics are of course interconnected, as international collaboration and sharing experiences benefits everybody. GOBLET is a not-for-profit Foundation, established in 2012 to harmonise worldwide bioinformatics training activities. Its members share materials and best practices, set standards, and provide high-quality resources to support learning, education and training. Together with its partner organizations, GOBLET has co-organized more than 40 training workshops and events around the world for different audiences ranging from bioinformatics trainers and end-users to biology teachers. In order to promote bioinformatics training and to increase its visibility, GOBLET has published several articles and established the Education Community of Special Interest as part of the ISMB conference together with ISCB.

CSC, the national supercomputing centre in Finland, has been a member of GOBLET from the start. We provide bioinformatics services for all the universities in the country, and the majority of our users have biomedical background. As the users are numerous, geographically scattered, and lack bioinformatics education, we need to find scalable ways to enable them to perform data analysis. I will discuss the combination of bioinformatics training approaches and software development that has helped us to fulfill our task.

Keynote 3

Prof. Katsushi Tokunaga, PhD
(Research Institute, National Center for Global Health and Medicine, Japan)

Biography

Prof. Katsushi Tokunaga is a Professor of the Department of Human Genetics, Graduate School of Medicine, at the University of Tokyo in Japan since 1995. He received his Ph.D. from Graduate School of Science at the University of Tokyo. He was awarded a number of honors and scientific prizes including Young Investigator Awards of The Japan Society of Human Genetics and of The Japan Society of Blood Transfusion, and Society Award of The Japan Society of Human Genetics. He is a member of Science Council of Japan, President of Japanese Society for Histocompatibility and Immunogenetics, and Director of Japan Society of Human Genetics. He also serves as the Editor-in-Chief for Human Genome Variation and Advisory Editor for Journal of Human Genetics, HLA, and International Journal of Immunogenetics. His research interests focus on genome-wide search for genetic factors to various complex diseases, with special reference to HLA- disease associations.
Abstract

GENOMIC APPROACH TO IMMUNE-MEDIATED COMPLEX DISEASES

We have been performing genome-wide association studies (GWAS) for a variety of complex diseases. Utilization of population-specific SNP array and genome-wide imputation based on WGS of thousands Japanese enabled effective identification of disease susceptibility genes. Above all, *HLA* genes have been identified as the most important genetic factor for many immune-mediated diseases. Major findings in our GWAS, *HLA* genotyping, and highly accurate *HLA* imputation are presented. Narcolepsy, type I diabetes, rheumatoid arthritis, primary biliary cholangitis, autoimmune hepatitis, childhood nephrotic syndrome, autoimmune hepatitis, childhood nephrotic syndrome, and peach and shrimp allergies were most strongly associated with the *HLA-DR* and *HLA-DQ* genes, while hepatitis B virus (HBV) related diseases and cold medicine-related Stevens-Johnson syndrome were most strongly associated with the *HLA-DP* and *HLA-A* genes, respectively. Recent progress of *HLA* and *KIR* typing using next-generation sequencer (NGS) and their application to disease association studies are also presented. Moreover, in infectious diseases including hepatitis and tuberculosis, interactions between host and pathogen genome variations were detected.

Keynote 4: ISCB Speaker

**Prof. Christine Orengo, PhD, FRS**

*(University College London, UK)*

Biography

Christine Orengo is a computational biologist, whose core research has been the development of robust algorithms to capture relationships between protein structures, sequences and functions. She has built one of the most comprehensive protein classifications, CATH, used worldwide by tens of thousands of biologists, and central to many pioneering structural and evolutionary studies.

CATH structural and functional data for hundreds of millions of proteins has enabled studies that revealed essential universal proteins and their biological roles, and extended characterisation of biological systems implicated in disease e.g. in cell division, cancer and ageing. CATH functional sites have revealed protein residues implicated in enzyme efficiency and bacterial antibiotic resistance. This data also identified genetic variations likely to be driving human diseases and the drugs that can be repurposed to offset the pathogenic effects. Christine is a Vice President of the International Society of Computational Biology (ISCB). She is a Fellow of the Royal Society of Biology and Elected member of EMBO since 2014, and a Fellow of ISCB since 2016. She is a founder of ELIXIR 3DBioInfo.
Abstract

**CATH FUNCTIONAL FAMILIES - INSIGHTS INTO IMPACTS OF GENETIC VARIATIONS**

Powerful tools for comparing protein structures and protein sequences have allowed us to analyse proteins from more than 20,000 completed genomes and identify 5500 evolutionary domain superfamilies, comprising a total of ~90 million domains. These superfamilies cover nearly 70% of domains from all kingdoms of life and are captured in our CATH resource. Some structural frameworks seem particularly suited to supporting different residue arrangements in the active sites and structural variations on the surfaces of the domains which can modify protein functions. Sub-classification of CATH superfamilies into functional families (FunFams) allows us to examine the structural mechanisms of function evolution in these superfamilies.

We have used the CATH-FunFams to analyse the impacts of genetic variations. For example, we observe that a particular mode of alternative splicing – Mutually Exclusive Exons (MXE) – is typically associated with variation in a small subset of residues on the surface of the protein and close to known or predicted functional sites for that protein. We analysed these effects using publicly available MXE data from 5 model organisms. Some compelling examples of MXE events in glycolytic proteins have been explored in more detail. The CATH-FunFams have also be used to determine whether genetic variations linked to human disease e.g. cancer, result in changes in residues close to functional sites, thereby modifying the functions of the proteins and affecting specific pathways and processes.

Keynote 5

Dr. Denis C. Bauer

(Transformational Bioinformatics Team, CSIRO, Australia)

**Biography**

Dr. Denis Bauer is an internationally recognised expert in machine learning, specifically in processing big genomic data to help unlock the secrets in human DNA – secrets that could change the course of human history. Her achievements include developing an open-source, artificial intelligence-based search engine in the cloud that helps researchers pinpoint the exact genes they need to study or edit to cure disease.

As CSIRO’s transformational bioinformatics leader, Denis is involved in Australian and international initiatives to integrate genomics into medical practice. She is frequently invited as a keynote at heavyweight medical and IT conferences including Amazon Web Services Summit.
Denis holds a BSc from Germany and PhD in Bioinformatics from the University of Queensland, and has completed postdoctoral research in both biological machine learning and high-throughput genetics. She has 33 peer-reviewed publications (14 as first or senior author) and an H-index 14.

Abstract

CLOUD COMPUTING AND ARTIFICIAL INTELLIGENCE TRANSFORMS BIOINFORMATICS RESEARCH

Genomic data is outpacing traditional Big Data disciplines, producing more information than Astronomy, twitter, and YouTube combined. As such, Genomic research has leapfrogged to the forefront of Big Data and Cloud solutions. We developed software platforms using the latest in cloud architecture, artificial intelligence and machine learning to support every aspect genome medicine; from disease gene detection through to validation and personalized medicine.

This talk outlines how we find disease genes for complex genetic diseases, such as ALS, using VariantSpark, which is a custom machine learning implementation capable of dealing with Whole Genome Sequencing data of 80 million common and rare variants. To support disease gene validation, we created GT-Scan, which is an innovative web application, which we think of it as the “search engine for the genome”. It enables researchers to identify the optimal editing spot to create animal models efficiently. The talk concludes by demonstrating a novel cloud-based architecture that enables a decision support framework capable of processing genomic and medical data at a speed fit for point-of-care application.

Keynote 6

Prof. Heru Suhartanto, PhD
(University of Indonesia, Indonesia)

Biography

Prof. Heru Suhartanto a Professor in Faculty of Computer Science, Universitas Indonesia He holds a BSc from Department of Mathematics, University of Indonesia. He holds Master of Science, from Department of Computer Science, The University of Toronto, Canada since 1990. He also holds PhD in Parallel Computing from Department of Mathematics, The
University of Queensland since 1998. His main research interests are Numerical, Parallel, Cloud and Grid computing. He is also a member of reviewer of several referred international journal such as journal of Computational and Applied Mathematics, International Journal of Computer Mathematics, and Journal of Universal Computer Science. Furthermore, he has supervised some Master and PhD students; he has won some research grants; holds several software copyrights; published a number of books in Indonesian and international papers in proceeding and journal. Some of the awards that he has received include Honorary Professor, School of ITEE, The University of Queensland, (2014 – 2017); Adjunct Professor, School of ITEE, The University of Queensland, (2017-2019); Universitas Indonesia Best Researcher, 2007 and The University of Queensland Alumni Award, 2010.

Abstract

HIGH PERFORMANCE COMPUTING CHALLENGES IN INSILICO DRUG DESIGN BASED ON INDONESIA MEDICAL PLANTS

Molecular Dynamics is one part of In silico drug design processes, and it is a computer simulation method for studying the physical movements of atoms and molecules. The atoms and molecules are allowed to interact for a fixed period, giving a view of the dynamic evolution of the system*. Assuming that there are N atoms in a molecule, the time complexity is O(N²). Usually, one needs to do a great number of timesteps simulation such as in 100 Nano Second (NS) and large computing time as the number of atoms (N) are huge due to the complexity of the molecule. For example, the Gromacs results in 90 NS which shows the development of veskel DPPC (dipalmitoyl phosphatidyl choline) [De Vries2008], required 3750 days of single CPU processes, or 117.2 days of 32 CPUs processes. It is becoming more challenges as not many Research institution own HPC infrastructure, and many of the users are not in Computer Science background. In this talk, the authors will present their experience in developing a feasible Cloud Computing environment for users in developing countries, such as Indonesia. Some findings on the computing aspects and some results on Indonesia medical plants simulation will be given.

Keynote 7

Prof. Adrian V.S. Hill, MD, PhD
(University of Oxford, UK)

Biography

Professor Hill is Director of the Jenner Institute, which focuses on designing and developing vaccines for infectious diseases prevalent in developing countries, such as HIV/AIDS, malaria and tuberculosis. He also heads a group at the Wellcome Trust Centre for Human Genetics
which studies genetic susceptibility factors for common bacterial diseases. He is a passionate believer in the power of molecular medicine to design and deliver new health care interventions that will improve the lives of the poorest billion in sub-Saharan Africa and elsewhere. His own vaccine research programme has developed one of the most promising potential vaccines for malaria which is currently in large scale trials in infants in sub-Saharan Africa. In 2014 his group led the first clinical trial of an Ebola virus vaccine targeting the outbreak of Ebola in West Africa.

His expertise includes vaccines for Ebola, vaccines for malaria, immunology and vaccinology of infectious diseases, human genetics of susceptibility to tuberculosis, leprosy, sepsis and other bacterial diseases and vaccines against intracellular pathogens. He received several honours and awards, including 1999 Fellow of the Royal College of Physicians, 1999 Fellow of the Academy of Medical Sciences, and 2005 Director of the Jenner Institute.

Abstract

THE HUMAN GENETICS OF INFECTION AND VACCINATION: GETTING TO SCALE

Human genetic factors influencing infectious disease susceptibility were first identified over sixty years ago with several striking examples now well-defined and understood. These include CCR5 and globin gene variants impacting on HIV/AIDS and malaria susceptibility respectively. Modern genomic analyses now provide immensely greater power to interrogate genomic variation in detail and large numbers of susceptibility loci have been identified by genome-wide association scans. However, the number of reproducibly defined loci is relatively small compared to many other disease types. I will discuss some of the challenges in this area and potential solutions offered by the advent of large cohort and biobank repositories and new high throughput immunoassays.

The study of human genetic variation impacting responses to vaccination is a much newer endeavour but one with several advantages compared to infectious disease genetics. Emerging findings indicate a major role of HLA, immunoglobulin gene and other immunogenetic loci in determining vaccine potency, efficacy and durability.

In addition to translational utility these two related areas provide tantalising insights into the selective pressures that infectious pathogens have exerted on our genetic make-up.
Plenary Talks

Prof. Shandar Ahmad, PhD
(Jawaharlal Nehru University, India)

Biography

Shandar Ahmad is a Professor of Bioinformatics and Dean at the School of Computational and Integrative Sciences, Jawaharlal Nehru since 2016. Earlier, he worked in National Institute of Biomedical Innovation, Health and Nutrition, Japan (2007-2016), where he still holds a Visiting Scientist honorary position. He was also an adjunct Associate Professor Osaka University (2008-2016).

Shandar Ahmad is interested in developing data-driven algorithms and applications for biological data with an integrative perspective. He has made major contributions to the study of protein-DNA interactions using classical as well as deep machine learning methods. He developed first neural network-based method for predicting DNA-binding sites and has extended his work to many aspects of transcription factor dynamics and basic understanding of protein-DNA recognition. His group aims to leverage on his past and ongoing works in the field of sequence and structural analysis of protein-DNA and other biomolecular interactions, particularly their modeling through conventional and deep machine learning algorithms. He has experience in studying patterns of global and condition-specific gene expression, single cell transcriptome analysis, miRNA and NGS data analysis for important Biological systems such as Influenza vaccination response, Uni-parental disomy and other critical problems in the field of Biomedical research. His group has also developed deep learning algorithms to interpret ECG, EEG and medical images data using deep learning models such as LSTM, CNN and deep neural network. For more, visit his website www.sciwhylab.org.

Shandar Ahmad has also interest in Poetry. He writes in English, Urdu and Hindi and his recent collection of English Ghazal was launched as part of the book “Dialogues with a caged parrot and Ghazal of an eavesdropper”, published by Studera Press (https://www.amazon.in/Dialogues-Caged-Parrot-Ghazals-Eavesdropper-Rabinarayan/dp/9385883747). He blogs his thoughts and poetry on www.shandarthought.net.
Abstract

DEEP VERSUS CONVENTIONAL LEARNING FOR BIOLOGICAL DATA THROUGH CHANGING PARADIGMS

Biological data analytics have gone through the paradigms of basic statistics and many different forms of supervised and unsupervised learning as well as feature selection. Deep learning has revolutionized all fields of scientific discovery including biomedical science. Many problems which were earlier solved poorly by conventional statistical and machine learning methods have been dramatically outperformed by deep learning models such as Convolutional neural networks (CNN), multi-layer auto-encoders and long short-term memory (LSTM) neural networks. Many barriers, which were outside the reach of conventional machine learning, have been overcome, partly because of better training systems and GPU computing. However, recent benchmarks have also brought in a word of caution against blindly applying deep learning to all the scientific problems. In this talk, I will discuss some of the contexts in which choice of deep versus conventional learning has to be made carefully. I will share some key findings of our own lab to identify factors that contribute to the success of one over the other in solving a given biological or clinical problem such as fMRI scans, function annotation, transcriptome analysis, protein-DNA interactions and conformational dynamics of DNA.

Plenary Talk 2

Prof. Akihiro Fujimoto, PhD
(The University of Tokyo, Japan)

Biography

Dr. Akihiro Fujimoto is a Professor at Department of Human Genetics, Graduate School of Medicine, The University of Tokyo. He was also an associate professor of Department of Drug Discovery Medicine, Graduate School of Medicine, Kyoto University. He received his Ph.D from the University of Tokyo in 2008, where he worked to reveal genes that were related to human adaptations. He joined RIKEN institute and carried out analysis of next generation sequence data. He analyzed the first Japanese whole genome and liver cancer genomes. His current interests are identification of genetic variations and somatic mutations, and interpretation of their functional impacts.
Abstract

IDENTIFICATION OF STRUCTURAL VARIATIONS AND ANALYSIS OF THEIR FUNCTIONAL ROLES

Structural variations (insertions, deletions, copy number variations, inversions and translocations) are functionally important types of genetic variations. Structural variations can influence structure, copy number and expression level of genes. However, identification and interpretation of their functional roles are still difficult tasks in genetics studies. We consider that there are large numbers of structural variations in the human and cancer genomes, but most of them have not been identified. We developed methods to detect structural variations and analyzed their functional impacts.

In this presentation, we would like to talk about analysis of structural variations in 300 liver cancer genomes, intermediate-size deletions in a Japanese population, and application of a long-reads sequencing technology to detect structural variations. Analysis of the cancer genomes showed that large number of structural variations exists in the cancer genomes and they can influence the gene expression level. Analysis of intermediate-size deletions revealed previously unreported deletions in a Japanese population and a part of them were associated with gene expression level. We also developed a method to detect structural variations by analyzing data from a long-reads sequencing technology. The analysis detected 2-3 times larger number of structural variations than those from a current short-reads sequencing technology. We consider that our study can contribute to the understanding of the landscape of structural variations in the human genome, and interpretation of the importance of them.
Workshop 1

Detecting differentially expressed genes with RNA-seq

Instructor: Dr. Eija Korpelainen
(CSC-IT Center for Science, Finland)

Workshop details

This workshop introduces the participants to RNA-seq data analysis methods, tools and file formats. It covers the whole workflow from quality control and alignment to quantification and differential gene expression analysis. The workshop consists of lectures and practical exercises. The free and user-friendly Chipster software is used in the exercises, so no previous knowledge of Unix or R is required, and the workshop is thus suitable for everybody.

The lectures are available as videos, and the participants are requested to view them prior to the workshop. This gives you more time to reflect on the concepts so that you can use the workshop time more efficiently. The lectures are summarized, and questions answered during the workshop.

In this workshop, the participants will learn how to:

1. check the quality of reads with FastQC
2. infer strandedness with RseQC
3. align RNA-seq reads to the reference genome with HISAT2 and STAR
4. perform alignment level quality control using RseQC
5. quantify expression by counting reads per genes using HTSeq
6. check the experiment level quality with PCA plots and heatmaps
7. analyze differential expression with DESeq2 and edgeR
8. take multiple factors (including batch effects) into account in differential expression analysis
Workshop details

Overview: Galaxy is a scientific workflow platform that aims to make computational biology accessible to research scientists that do not have computer programming or systems administration experience. By providing a graphical, web-based interface for running bioinformatics tools and pipelines, Galaxy removes the need to learn complex command-line syntax or the need to interact with computational job management systems. Galaxy also makes your analyses reproducible by maintaining a replicable history of analysis steps, and makes your research more transparent by allowing you to share your analyses with others. Here we will be showcasing the features of Galaxy on the Galaxy Australia service, usegalaxy.org.au. Login is open to all and only requires an Correspondence to access the full resources of Galaxy Australia.

Workshop content: The workshop will be run in two halves, an overview of the Galaxy platform, using Galaxy Australia as the example, then a practical demonstration of DNA sequence variant identification. We will start with the basics of Galaxy, by providing a run through of the interface, demonstrate how to get private and public data into Galaxy, and perform a common type of analysis using popular software tools. We will then build on this base to create a more complex workflow, publish this workflow and perform a more complex analysis, using an exemplar variant calling pipeline. We will also discuss the options available for installing and using Galaxy, including freely available public servers.

Target audience: System Administrators and researchers with an interest in carrying out bioinformatics analysis, either as a core service for an institute and for individual research questions. Attendees should bring a laptop to participate in the variant calling demonstration. The only required software is a web browser (Chrome preferred), both Mac and Windows supported.

Learning Outcomes: Attendees will learn how to:
1. Access, login and use Galaxy to perform an analysis
2. Import and Export of data
3. Design and execution of chained tools / workflows
4. Import / Export
5. Sharing of an analysis with a collaborator
6. Perform a series of analyses on a research topic – here will be DNA sequence variant identification
Workshop 3

**Protein Sequence and Structure Analysis Using Google Cloud Engine**

**Instructor:** Prof. Shandar Ahmad and Ajay Arya  
(Jawaharlal Nehru University, India)

**Workshop details**

Overview: The computational needs of bioinformatics are constantly increasing. Although sophisticated ready-made tools are increasingly available, in order to fully control their methods, bioinformaticians and other data scientists will need to write or modify their own software. Recently, there has also been a shift in computational architectures, from single-core desktop and laptop computers to multicore and distributed systems such as cloud computing. This shift necessitates a change in the way that we approach programming and think about algorithms in general and also specifically in bioinformatics. In this workshop we will introduce Google cloud computing techniques and tools to perform basic sequence and structure analysis of proteins.

Objectives: This course will give a theoretical background as well as hands-on experience in the following topics.

- Scalable computation with Google Cloud Platform
- High-performance, concurrent algorithms for data analysis on Google compute engine
- Basic Protein Sequence and Structure Analysis using Google Cloud

Requirement for signing up to Google Cloud Platform (GCP):
This workshop will need Google Cloud Services for all. Every participant is required to sign up for an account in Google Cloud Service to use GCP. This entails following steps.

1. If you do not already have a Google account (e.g. gmail), you need to create one for free.
2. After having created a Google account, login to your Google account and specifically sign up for Cloud account. To do so, navigate to https://cloud.google.com and click, on “Try it free!”. This will start your GCP sign up process. Complete the process. You will need to enter your billing information but the free services can be used and are enough to complete all the activities in your workshop. GCP free trial gives you approximately $300 worth of cloud services valid for 12 months from the date of signing up.
3. If you have used up your GCP free account allocation before or during the workshop, you may have to go to paid option and complete the payment.

Workshop will consists of three one hour lectures. Throughout the workshop, lectures will start with a lecture and theoretical introduction to each topic, followed by hands-on exercises.
Lecture 1. Scalable programming basics
Basic ideas of parallel and scalable programming will be introduced. General tools for parallelization in standard software such as R will be reviewed.

Lecture 2. Google Cloud Engine
Basic introduction on cloud computing will be given. This will follow as specific implementations in Google cloud. MapReduce versus Apache Spark cloud computing frameworks will be introduced.

Basic sequence analysis tasks such as amino acid composition calculation, amino acid propensity in a pair of sequences, information contents from multiple alignments etc. Will be demonstrated and hands on exercises will be provided. For structure analysis, structures from PDB will be taken as examples and secondary structure, solvent accessibility and binding sites will be computed.

Who can attend: Participants are expected to have at least some minimal prior experience in programming, using any language (R, Python, or Perl would suffice). A basic working knowledge of molecular biology is helpful but not essential.

Maximum intake: A maximum of 50 participants will be accommodated. Selection (in case of excess requests) will be made by the course coordinator and teaching faculty based on the compatibility and usefulness of the course and also considering regional, social and gender diversity.

Requirements: Users will be required to subscribe to basic Cloud services in Google. This may amount to about USD 50 per person and will be done by users themselves. No financial transactions is needed with organizers or hosts.
Workshop details

This interactive Train-the-Trainer workshop aims to provide individuals new to training, or those wishing to develop their skills, with guidance and tips for developing and delivering training in bioinformatics, exploring a range of methods appropriate to different learning styles, examining the requirements for a successful course and acquiring appropriate feedback.

The workshop will:

• provide guidance on general training techniques and appropriate use of methods based on learner needs, including some general do’s and do not’s for successful training
• provide a framework for successful curriculum design and further development, to enable trainers to build training appropriate to their learner’s needs

By the end of this workshop, trainers should be able to:

• Recognise the elements of good and bad training
• Describe the role they play in providing an engaging and comfortable learning environment for their trainees
• Construct appropriate session / curriculum design for their target audience
• Explore ideas and inspiration for developing training materials and apply these appropriately
• Seek and act upon feedback to improve their training through reflective practice

Intended audience: This workshop is suitable for anyone wishing to develop their skills in bioinformatics training.
**Breakout Session**

Overview: Galaxy ([galaxyproject.org](http://galaxyproject.org)) is a widely used and deployed open-source platform that integrates thousands of analysis tools, and supports data integration and analysis, workflow creation, data visualization, and sharing and publishing your analyses and pipelines. Galaxy is also a global community of researchers, programmers, trainers, and users all doing and supporting data intensive research. Galaxy supports regional communities and resources including a large Australian community ([usegalaxy.org.au](http://usegalaxy.org.au)). Galaxy Australia already supports users regionally from: China, India, Indonesia, Japan, Malaysia, New Zealand, Philippines, Singapore, South Korea, Sri Lanka, Thailand and Vietnam. Galaxy Australia operates a nationally distributed federation of compute resources, where individual researcher analysis requests are assigned according to a sophisticated rules system optimised for rapid result generation.

This breakout will present a vision for a regional Galaxy community and platform for Asia, Australia, and Oceania that proposes broadening the compute infrastructure underpinning the service to include regional resources. Furthermore the vision will highlight the control afforded to each contributing country in fair-use of contributing resources.

The session will run for one hour with three presentations from:

- Dave Clements - Global Galaxy communities and services, 10 minutes
- Gareth Price - Galaxy Australia as a usegalaxy.* public service, 10 minutes
- Nuwan Goonasekera - Future Galaxy developments for distributed data access, 10 minutes

Followed by 30 minutes of moderated discussion on the benefits of growing a Galaxy service to support a greater number and diversity of researchers regionally. Topics covered will include but are not limited to:

- What could a regional service look like?
- The benefits for contributing institute / organisation
- How does each institute control access to resources
- Can staff time be contributed as well as or instead of computer infrastructure
- What is the sustainability model
- Regional branding on the Galaxy service

The session will conclude with a commitment to grow a community of interested parties and agree on a time frame to resourcing the regional Galaxy service.

**Short description of prior experiences in organizing a workshop session**

Target audience and relevance to InCoB 2019

Galaxy Project has over 120 public services globally and in recent years moved to synergise efforts amongst Galaxy instances to maximise the efficient operation of the services. Each Galaxy instance benefits from the community efforts of the other Galaxy services. The usegalaxy.* public services exist in Australia, Europe and the United States. A managed Galaxy service regionally would benefit local researchers. As such, this BoF will be of interest to system administrators, researchers and students with an interest in carrying out bioinformatics analysis, either as a core service for an institute and for individual research questions.
Highlight Track

Highlight 1

A COMPREHENSIVE MAP OF INTRON BRANCHPOINTS AND LARIAT RNAs IN PLANTS

Xiaotuo Zhang1,2,3, Yong Zhang4,5, Taiyun Wang1, Zhiwei Li1, Jinping Cheng1, Haoran Ge1, Qi Tang1, Kun Chen4, Li Liu4, Junqiang Guo2,3, Chenyu Lu1, Binglian Zheng1,6, & Yun Zheng2,3,6

1 State Key Laboratory of Genetic Engineering, Ministry of Education Key Laboratory of Biodiversity Sciences and Ecological Engineering, Institute of Plant Biology, School of Life Sciences, Fudan University, Shanghai 200438, China
2 Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan 650500, China
3 Yunnan Key Laboratory of Primate Biomedical Research; Institute of Primate Translational Medicine, Kunming University of Science and Technology, Kunming, Yunnan 650500, China
4 Faculty of Life Science and Technology, Kunming University of Science and Technology, Kunming, Yunnan 650500, China
5 These authors contributed equally to the work
6 Correspondence: zhengyun5488@gmail.com and zhengbl@fudan.edu.cn

Lariats are formed by excised introns, in which the 5’ splice site joints with the branchpoint (BP) as the first step during splicing. Lariat RNAs are traditionally thought to be quickly degraded by DBR1 (RNA debranching enzyme 1). However, recent findings in animals showed that a substantial fraction of lariats were widely detected as circular forms, and that BP recognition is critical for lariat metabolism. In contrast, the features of BPs and to what extent lariat RNAs escape debranching are largely unexplored in plants. Here, we analyzed more than 900 RNA sequencing datasets to document plant BPs and lariat-derived circular RNAs on a genome-wide scale. In total, we identified 13872 BPs in Arabidopsis, 5199 BPs in tomato, 29582 BPs in rice, and 13487 BPs in maize. Features of plant BPs are highly similar to those in yeast and human, in that BPs are adenine-preferred, closely distributed to the 3’ splice site, and flanked by uracil- enriched sequences. Intriguingly, more than 60% of introns harbor multiple BPs, and BP usage is obviously tissue-specific. Furthermore, we performed circular RNA sequencing from both wild type Arabidopsis plants and the *dbr1* mutant, and found that 10580 lariat RNAs accumulate under physiological conditions. Moreover, most of these lariat RNAs originate from longer or retroelement-depleted introns, and the expression of these lariat RNAs significantly correlated with the incidence of back-splicing of parent exons. Collectively, our results provide the first comprehensive map of intron BPs and lariat RNAs in plants, and uncover a novel link between lariat turnover and splicing.
Cyclophilins are ubiquitous proteins present in a wide range of organisms with several members possessing PPIase activity. *In silico* analysis of the recently available wheat genome (IWGSC RefSeq 1.0) revealed 83 different wheat cyclophilin genes that encode 85 different proteins ranging between 17 kDa to 92 kDa. 58 TaCYPs possess only a single cyclophilin-like domain, while others also possess additional functional domains (Figure 1). The TaCYPs were predicted to be localized to cytosol (28), chloroplast (22), nucleus (19) mitochondria (9), secretory system (4) and ER (3). Gene structure analysis showed the presence of introns in the ORFs as well as 5′ and 3′ UTRs. These genes are distributed unevenly on the different chromosomes and have evolved as a result of whole genome, segmental, ectopic and tandem duplications (Figure 2, 3). The expression of different TaCYPs is regulated differently in the wheat seedlings in response to heat stress, suggesting their role in thermotolerance.
Oral Presentation Abstracts

Oral ID: O-01

BIGRAM-PGK: PHOSPHOGLYCERYLATION PREDICTION USING THE TECHNIQUE OF BIGRAM PROBABILITIES OF POSITION SPECIFIC SCORING MATRIX

Abel Chandra, Alok Sharma, Abdollah Dehzangi, Daichi Shigemizu, and Tatsuhiko Tsunoda

1 Institute for Integrated and Intelligent Systems, Griffith University, Brisbane, QLD-4111, Australia
2 Department of Medical Science Mathematics, Medical Research Institute, Tokyo Medical and Dental University (TMDU), Tokyo, 113-8510, Japan
3 Laboratory for Medical Science Mathematics, RIKEN Center for Integrative Medical Sciences, Yokohama, 230-0045, Kanagawa, Japan
4 Department of Computer Science, Morgan State University, Baltimore, Maryland, USA
5 School of Engineering and Physics, Faculty of Science Technology and Environment, University of the South Pacific, Suva, Fiji
6 Medical Genome Center, National Center for Geriatrics and Gerontology, Obu, Aichi 474-8511, Japan
7 CREST, JST, Tokyo, 102-8666, Japan
8 Laboratory for Medical Science Mathematics, Department of Biological Sciences, Graduate School of Science, The University of Tokyo, Tokyo 108-8639, Japan

Correspondence: abelavit@gmail.com, alok.sharma@griffith.edu.au

The biological process known as post-translational modification (PTM) is a condition whereby proteomes are modified that affects normal cell biology, and hence the pathogenesis. A number of PTMs have been discovered in the recent years and lysine phosphoglycerylation is one of the fairly recent developments. Even with a large number of proteins being sequenced in the post-genomic era, the identification of phosphoglycerylation remains a big challenge due to factors such as cost, time consumption and inefficiency involved in the experimental efforts. To overcome this issue, computational techniques have emerged to accurately identify phosphoglycerylated lysine residues. However, the computational techniques proposed so far hold limitations to correctly predict this covalent modification.

We propose a new predictor in this paper called Bigram-PGK which uses evolutionary information of amino acids to try and predict phosphoglycerylated sites. The benchmark dataset which contains experimentally labelled sites is employed for this purpose and profile bigram occurrences is calculated from position specific scoring matrices of amino acids in the protein sequences. The statistical measures of this work, such as sensitivity, specificity, precision, accuracy, Mathews correlation coefficient and area under ROC curve have been reported to be 0.9642, 0.8973, 0.8253, 0.9193, 0.8330, 0.9306, respectively.

The proposed predictor, based on the feature of evolutionary information and support vector machine classifier, has shown great potential to effectively predict phosphoglycerylated and non-phosphoglycerylated lysine residues when compared against the existing predictors. The data and software of this work can be acquired from https://github.com/abelavit/Bigram-PGK.
Single Molecule Real-Time (SMRT) sequencing is a recent advancement of Next Gen technology developed by Pacific Bio (PacBio). It comes with an explosion of long and noisy reads demanding cutting edge research to get most out of it. To deal with the high error probability of SMRT data, a novel contextual Locality Sensitive Hashing (conLSH) based algorithm is proposed in this article, which can effectively align the noisy SMRT reads to the reference genome. Here, sequences are hashed together based not only on their closeness, but also on similarity of context. The algorithm has $O(n^{\rho+1})$ space requirement, where $n$ is the number of sequences in the corpus and $\rho$ is a constant. The indexing time and querying time are bounded by $O\left(\frac{n^{\rho+1}ln\ n}{ln\ P_2}\right)$ and $O(n^\rho)$ respectively, where $P_2 > 0$, is a probability value. This algorithm is particularly useful for retrieving similar sequences, a widely used task in biology. The proposed conLSH based aligner is compared with rHAT, popularly used for aligning SMRT reads, and is found to comprehensively beat it in speed as well as in memory requirements. In particular, it takes approximately 24.2% less processing time, while saving about 70.3% in peak memory requirement for H.sapiens PacBio dataset.
Drug delivery in excess concentration and at not-specified positions inside the human adversely affects the body and gives rise to another disease. Several methods have been developed to deliver the drug molecules in required amount and at specific target. Nanoparticle mediated drug delivery is one such process and has gained some success at primarily level. The effect of nanoparticles on human body is an important concern and it has been unravelled by measuring the protein-nanoparticle interactions. Human Serum Albumin (HSA) is one of the most abundant blood plasma protein and a potential drug carrier. Silver nanoparticle (AgNP) has been selected to investigate its suitability for biomedical application. Here, we have measured the impact of AgNP on HSA structure and function with the help of all-atom molecular dynamics simulations (MDS).

An extensive MDS on protein, AgNP and AgNP adjuvant protein in five different orientations has provided a dynamic picture of HSA-AgNP interactions. The number of non-bonded contacts and interaction energy between protein and AgNP during the 100ns simulations showed that the NP is interacting with HSA and conjugated system got stabilized with time evolution of MDS trajectories. Serum albumin is a transport protein and any change in the structure may obstruct its function. The comparative analysis (RMSDs and RMSFs) between protein-NP conjugate and bare protein simulations have not shown any major fluctuations. Further, the relative changes in the protein secondary structure has again revealed that HSA structure remain unaltered. Additional two parameters have been utilized to probe the alternation at individual amino acid level and this scrutiny also supported our results. We have shown the dynamical interaction between HSA-AgNP and no negative influence on the protein structure and function by a computational approach.

The present MDS based investigation confirms that the AgNP is interacting with HSA without affecting its tertiary and secondary structure in turn the protein function as well. Also, it recommends the AgNP application to transport conjugated drug molecules because it has no adverse effect on a serum protein. Since HSA is present in the circulatory system, therefore, it may open various applications of AgNP in biomedical field.
Oral ID :O-04

CLASSIFICATION OF ADAPTOR PROTEINS USING RECURRENT NEURAL NETWORKS AND PSSM PROFILES

Nguyen Quoc Khanh Le¹, Quang H. Nguyen², Xuan Chen³, Susanto Rahardja⁴ and Binh P. Nguyen⁵

¹ School of Humanities, Nanyang Technological University, 50 Nanyang Ave, 639798 Singapore, Singapore.
² School of Information and Communication Technology, Hanoi University of Science and Technology, 1 Dai Co Viet, 10000 Hanoi, Vietnam.
³ Beijing Genomics Institute, 21 Hongan 3rd Street, 518083 Shenzhen, China.
⁴ School of Marine Science and Technology, Northwestern Polytechnical University, 127 West Youyi Road, 710072 Xi’an, China.
⁵ School of Mathematics and Statistics, Victoria University of Wellington, Gate 7, Kelburn Parade, 6140 Wellington, New Zealand

Correspondence: binh.p.nguyen@vuw.ac.nz

Adaptor proteins are a broad category of carrier proteins that play a role in signal transduction. They generally contain several modular domains, each of which having its own binding activity, and act by forming complexes with other intracellular-signaling molecules. Many studies determined that the adaptor proteins had been implicated in a variety of human diseases. Therefore, creating a precise model to predict the function of adaptor proteins is one of the vital task in bioinformatics and computational biology. Few computational biology studies have been conducted to predict the protein functions, and in most of those studies, they could not use original position specific scoring matrix (PSSM) profiles to feed into the neural network. This leads to the lack of information and the neural network consequently could not achieve the optimum result. In this paper, we present an innovative approach by incorporating deep recurrent neural networks (RNN) and PSSM profiles to resolve this problem.

Compared to other state-of-the-arts, our approach achieved an evidential improvement in all of the common measurement metrics. The area under the receiver operating characteristic curve (AUC) metric in prediction of adaptor proteins in the cross-validation and independent datasets were 0.893 and 0.853, respectively.

The findings of this study could provide a basis for further research that can use the combination of RNN and PSSM profiles in bioinformatics. Moreover, scientists can use our approach to solve various protein function prediction problems in the future.
RULE-BASED META-ANALYSIS REVEALS THE MAJOR ROLE OF PB2 IN INFLUENCING INFLUENZA A VIRUS VIRULENCE IN MICE

Fransiskus Xaverius Ivan, Chee Keong Kwoh
Biomedical Informatics Lab, School of Computer Science and Engineering, Nanyang Technological University, Singapore

Correspondence: fivan@ntu.edu.sg, asckkwoh@ntu.edu.sg

Influenza A virus (IAV) poses threats to human health and life. Many individual studies have been carried out in mice to uncover the viral factors responsible for the virulence of IAV infections. Nonetheless, a single study may not provide enough confident about virulence factors, hence combining several studies for a meta-analysis is desired to provide better views. For this, we documented more than 500 records of IAV infections in mice, whose viral proteins could be retrieved and the mouse lethal dose 50 or alternatively, weight loss and/or survival data, was/were available for virulence classification.

IAV virulence models were learned from various datasets containing aligned IAV proteins and the corresponding two virulence classes (avirulent and virulent) or three virulence classes (low, intermediate and high virulence). Three proven rule-based learning approaches, i.e., OneR, JRip and PART, and additionally random forest were used for modelling. PART models achieved the best performance, with moderate average model accuracies ranged from 65.0% to 84.4% and from 54.0% to 66.6% for the two-class and three-class problems, respectively. PART models were comparable to or even better than random forest models and should be preferred based on the Occam’s razor principle. Interestingly, the average accuracy of the models was improved when host information was taken into account. For model interpretation, we observed that although many sites in HA were highly correlated with virulence, PART models based on sites in PB2 could compete against and were often better than PART models based on sites in HA. Moreover, PART had a high preference to include sites in PB2 when models were learned from datasets containing the concatenated alignments of all IAV proteins. Several sites with a known contribution to virulence were found as the top protein sites, and site pairs that may synergistically influence virulence were also uncovered.

Modelling IAV virulence is a challenging problem. Rule-based models generated using viral proteins are useful for its advantage in interpretation, but only achieve moderate performance. Development of more advanced approaches that learn models from features extracted from both viral and host proteins shall be considered for future works.
Oral ID: O-06

GALAXY AUSTRALIA - A TRULY NATIONAL OPEN-SOURCE BIOINFORMATICS PLATFORM

Gareth Price1, Derek Benson 2, Simon Gladman 3, Igor Makunin 2, Anna Syme 4, Grahame Bowland 5, Helen van de Pol1, Christina Hall 3, Nuwan Goonasekera 3, Nigel Ward 6, Jeffrey H Christiansen 7, Andrew Lonie 3

1 Queensland Facility for Advanced Bioinformatics, Queensland Cyber Infrastructure Foundation, 306 Carmody Rd, St Lucia, Queensland 4067, Australia, Phone +61 7 3346 2624, org.price@qfab.org
2 Research Computing Centre, Level 5, Axon Building, The University of Queensland, St Lucia Queensland 4072, Australia, Phone: +61 7 3365 8350, {d.benson, i.makunin} @uq.edu.au
3 Melbourne Bioinformatics, University of Melbourne, 187 Grattan Street, Victoria 3010, Australia, Phone +61 3 9035 5822, {alonie, crhall, helen.vandepol, n.goonasekera, simon.gladman}@unimelb.edu.au
4 Royal Botanic Gardens Private Bag 2000, Birdwood Avenue, South Yarra, Victoria 3141, Australia, Phone: +61 3 9252 2300, anna.syme@rbg.vic.gov.au
5 Centre for Comparative Genomics, Murdoch University, 90 South Street, Murdoch, Western Australia 6150, Australia, Phone: +61 8 9360 6088, gbowland@ccg.murdoch.edu.au
6 Queensland Cyber Infrastructure Foundation, Level 5, Axon Building, The University of Queensland, St Lucia QLD 4072, Australia, Phone: +61 7 3365 7534, nigel.ward@qcif.edu.au
7 Queensland Cyber Infrastructure Foundation and Research Computing Centre, Level 5, Axon Building, The University of Queensland, St Lucia QLD 4072, Australia, Phone: +61 7 33657534.

Correspondence name: g.price@qfab.org

The Galaxy Project provides an installable environment with easy user access to managed computational resources, designed to support the data-intensive endeavours of researchers generating data from the latest generation of high-throughput assays. Australia nationalised multiple Galaxy services into Galaxy Australia (https://usegalaxy.org.au/) and has seen annualised user growth of 40%. The Galaxy Australia service is configured with local and remote servers in Queensland, Victoria and the Australian Capital Territory, respectively. The service has been configured to distribute jobs to allow for training, in conjunction with EMBL Australia Bioinformatics Resource (EMBL-ABR) network, and to process long run time high-resource jobs whilst maintaining the general analysis service. For trusted third parties, Galaxy Australia allows automatic user creation, data ingest and can trigger analysis workflows. Galaxy Australia is browser agnostic and freely available with no installation required or registration required.
MADOKA: AN ULTRA-FAST APPROACH FOR LARGE-SCALE PROTEIN STRUCTURE SIMILARITY SEARCHING

Lei Deng¹, Guolun Zhong¹, Chenzhe Liu¹ and Hui Liu²

¹School of Computer Science and Engineering, Central South University, 410075, Changsha, China
²Lab of Information Management, Changzhou University, 213164, Changzhou, China

Correspondence name: hliu@cczu.edu.cn

Protein structure comparative analysis and similarity searches play essential roles in structural bioinformatics. A couple of algorithms for protein structure alignments have been developed in recent years. However, facing the rapid growth of protein structure data, improving overall comparison performance and running efficiency with massive sequences is still challenging.

Here, we propose MADOKA, an ultra-fast approach for massive structural neighbor searching using a novel two-phase algorithm. Initially, we apply a fast alignment between pairwise structures. Then, we employ a score to select pairs with more similarity to carry out a more accurate fragment-based residue-level alignment. MADOKA performs about 6-100 times faster than some existing methods, including TM-align and SAL. Moreover, the quality of structure alignment of MADOKA is better than the existing algorithms in terms of TM-score and number of aligned residues. We also develop a web server to search structural neighbors in PDB database (About 135,000 PDB files contains around 360,000 protein chains in total), as well as additional features that include 3D structure alignment visualization and download links.

Conclusions: MADOKA is an efficient tool to search protein structure similarity.
STATISTICAL CONSIDERATIONS AND MACHINE LEARNING APPROACHES FOR RAPID STRAIN TYPING OF STAPHYLOCOCCUS HAEMOLYTICUS BASED ON MATRIX-ASSISTED LASER DESORPTION IONIZATION-TIME-OF FLIGHT MASS SPECTROMETRY

Chia-Ru Chung¹, Hsin-Yao Wang²,³, Frank Lien², Yi-Ju Tseng²,⁴, Chun-Hsien Chen⁴, Tzong-Yi Lee⁵,⁶, Jorng-Tzong Horng¹, and Jang-Jih Lu²,⁷,⁸

¹Department of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan,
²Department of Laboratory Medicine, Chang Gung Memorial Hospital at Linkou, Taoyuan, Taiwan,
³Ph.D. Program in Biomedical Engineering, Chang Gung University, Taoyuan, Taiwan,
⁴Department of Information Management, Chang Gung University, Taoyuan, Taiwan,
⁵School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, China,
⁶Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen, China,
⁷School of Medicine, Chang Gung University, Taoyuan, Taiwan,
⁸Department of Medical Biotechnology and Laboratory Science, Chang Gung University, Taoyuan, Taiwan

Staphylococcus haemolyticus is one of the most significant coagulase-negative staphylococci (CoNS), and often causes severe infections. Rapid strain typing of pathogenic S. haemolyticus is an important part of modern public health infectious disease control and critical to outbreak response, since it facilitates the identification of the origin of infection. Such information results in rapid infection control, which benefits tremendously critically-ill patients. However, currently strain typing methods, such as multi-locus sequencing, are relatively expensive and have a relatively long turn-around-time. A method for the rapid strain typing of pathogens that is suitable for routine use in clinics and hospitals is still not available. Analysis of matrix-assisted laser desorption ionization-time of flight mass spectrometry (MALDI-TOF MS) with machine learning approaches is a promising way of providing rapid strain typing. In this study, we first developed a statistical test-based alignment method for dealing with the shifting problem on mass spectrometry, and constructed machine learning-based classifiers for classifying different strains of S. haemolyticus. The multi-class area under the receiver operating characteristic curve (AUC) and accuracy were 0.848 and 0.866, respectively. Additionally, we further employed a variety of statistical tests and different feature selection strategies to investigate the specific informative peaks. In conclusion, the approach proposed in this study not only developed a cost effective and rapid identification method for the strain typing of S. haemolyticus, but provided a specific strategy for dealing with the shifting problem on mass spectrometry.
Tandem mass spectrometry allows biologists to identify and quantify protein samples in the form of digested peptide sequences. When performing peptide identification, spectral library search is more sensitive than traditional database search but is subject to the peptides that have been previously identified. An accurate tandem mass spectrum prediction tool is thus crucial in expanding the peptide space and increasing the coverage of spectral library search.

We proposed a non-linear regression model, MS2CNN, based on a deep learning algorithm - deep convolutional neural network. The features for our model are amino acid composition, predicted secondary structure, and physical-chemical features such as isoelectric point, aromaticity, helicity, hydrophobicity and basicity. MS2CNN was trained with 5-fold cross validation by three-way data split on the large scale human HCD MS2 dataset of Orbitrap LC-MS/MS downloaded from National Institute of Standards and Technology. It was then evaluated on a publicly available independent test dataset of human HeLa cell lysate from LC-MS experiment. On average, our model shows better Cosine Similarity and Pearson Correlation Coefficient (0.690 and 0.632) than those of MS2PIP (0.647 and 0.601) and are comparable with those of pDeep (0.692 and 0.642). Notably, for the more complex MS2 spectra of 3+ peptides, MS2PIP is significantly better than both MS2PIP and pDeep.

We showed that MS2CNN outperformed MS2PIP for 2+ and 3+ peptides and pDeep for 3+ peptides. It implies our convolutional neural network model, MS2CNN, is capable of generating highly accurate MS2 spectra for LC-MS/MS experiments using Orbitrap machines, which can be of great help in protein and peptide identifications. The results suggest incorporating more data for deep learning model can potentially improve the performance.
GENOME-WIDE IDENTIFICATION, PHYLOGENY, AND EXPRESSION ANALYSIS OF THE SBP-BOX GENE FAMILY IN EUPHORBIACEAE

Jing Li¹,², Xiaoyang Gao¹, Shiye Sang¹,², Changning Liu¹,³

¹CAS Key Laboratory of Tropical Plant Resources and Sustainable Use, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Kunming 650223, China
²University of Chinese Academy of Sciences, Beijing 100049, China
³The Innovative Academy of Seed Design, Chinese Academy of Sciences, Menglun, Mengla, Yunnan 666303, China

Correspondence: lijing2@xtbg.ac.cn

Euphorbiaceae is one of the largest families of flowering plants. For the exceptional diversity of growth forms and nearcosmopolitan distribution, it has attracted human interest since ancient times. SBP-box (SBP) genes encode a type of plant-specific transcription factors that play critical roles in numerous biological processes, especially in flower development. We performed a genome-wide identification and characterization of SBP genes from four high economic Euphorbiaceae species.

In total, 77 SBP genes were identified in four Euphorbiaceae genomes. All these SBP proteins are belong to 3 length ranges and 10 groups. Group-6 is absent in Arabidopsis thaliana, but it is conserved existing in Euphorbiaceae. Segmental duplication played the most important roles in the expansion processes of Euphorbiaceae SBP genes, and all these duplicated genes were subjected to positive selection. In addition, about two thirds of Euphorbiaceae SBP genes are the potential targets of miR156, and some miR-regulated SBP genes show high intensity and tissue contrasting expression profiles. The expression profiles of different stress treatments demonstrated broad involvement of Euphorbiaceae SBP genes in response to various abiotic and hormonal treatments and functional divergence.

In this research, 77 SBP genes were identified in four Euphorbiaceae species, and their phylogenetic relationships, protein physicochemical characteristics, duplication, tissue and stress response expression and potential roles in Euphorbiaceae development were studied. This study lays foundation of further studies of Euphorbiaceae SBP genes, which provide valuable information for future functional exploration of Euphorbiaceae SBP genes.
INSTANCE-BASED LEARNING FOR PERSONALIZED CANCER DIAGNOSIS
AND TREATMENT PLANNING

Md. Sarwar Kamal\textsuperscript{1}, Daniel Catchpoole\textsuperscript{1,2,3}, Jinyan Li\textsuperscript{1}

\textsuperscript{1}University of Technology Sydney, \\
\textsuperscript{2}Westmead Children Hospitals, \\
\textsuperscript{3}University of Sydney

Correspondence: jinyan.li@uts.edu.au*

Existing “global” learning approaches (e.g., decision trees, support vector machines) have been widely used for \textit{in-silicon} disease diagnosis. With a specific training data set, these global learning approaches generate a decision model of fixed parameters, regardless of the change from one test instance (patient) to another. Such a fixed model applied for all test samples is not consistent with the principle of the personalized cancer diagnosis which requires that only the most relevant knowledge from the training data should be tailored for the accurate diagnosis of each test sample. Instance-based learning (IBL), or called “lazy learning”, consider a different set of knowledge from the training data to reach a decision for each different test instance, suitable for precision medicine.

This work proposes a novel instance-based learning approach. Each time for a new test instance, our method generates a relevance matrix specifically for that instance through the scanning of the whole training data. This relevance matrix may change when a different test instance is applied. Then, a Bipartite graph and K-connected sub-graphs are constructed from the relevance matrix to ensure only dominant features (genes) of the test instance are extracted responsible for cancer. Finally, Bayesian classification and maximum a posterior (MAP) are used to make the accurate diagnosis decision. Our rigorous experimental analysis confirms that instance-based learning can outperform global learning techniques in accuracy, precision, recall, and F-measures. Moreover, IBL can identify patient relevant sub-cohort and gene subset for an individual instance as shown in our case studies on colon, breast, bladder, and adrenal cancer, which are useful for disease understanding and personalized treatment planning.
Drug-drug interactions (DDIs) are a major concern in patients' medication. It's unfeasible to identify all potential DDIs using experimental methods which are time-consuming and expensive. Computational methods provide an effective strategy, however, facing challenges due to the lack of experimentally verified negative samples. To address this problem, we propose a novel positive-unlabeled learning method named DDI-PULearn for large-scale drug-drug interaction predictions. DDI-PULearn first generates seeds of reliable negatives via OCSVM (one-class support vector machine) under a high-recall constraint and via the cosine-similarity based KNN (k-nearest neighbors) as well. Then trained with all the labeled positives (i.e., the validated DDIs) and the generated seed negatives, DDI-PULearn employs an iterative SVM to identify a set of entire reliable negatives from the unlabeled samples (i.e., the unobserved DDIs). Following that, DDI-PULearn represents all the labeled positives and the identified negatives as vectors of abundant drug properties by a similarity-based method. Finally, DDI-PULearn transforms these vectors into a lower-dimensional space via PCA (principal component analysis) and utilizes the compressed vectors as input for binary classifications. The performance of DDI-PULearn is evaluated on simulative prediction for 149,878 possible interactions between 548 drugs, comparing with two baseline methods and five state-of-the-art methods. Related experiment results show that the proposed method for the representation of DDIs characterizes them accurately. DDI-PULearn achieves superior performance owing to the identified reliable negatives, outperforming all other methods significantly. In addition, the predicted novel DDIs suggest that DDI-PULearn is capable to identify novel DDIs. The results demonstrate that positive-unlabeled learning paves a new way to tackle the problem caused by the lack of experimentally verified negatives in the computational prediction of DDIs.
Differences in cell-type composition across subjects and conditions often carry biological significance. Recent advancements in single cell sequencing technologies enable cell-types to be identified at the single cell level, and as a result, cell-type composition of tissues can now be studied in exquisite detail. However, a number of challenges remain with cell-type composition analysis – none of the existing methods provides perfect cell-type identification and variability related to cell sampling exists in any single cell experiment. This necessitates the development of method for estimating uncertainty in cell-type composition.

We developed a novel single cell differential composition (scDC) analysis method that performs differential cell-type composition analysis via bootstrap resampling which captures the uncertainty associated with cell-type proportions of each subject via bias-corrected and accelerated bootstrap confidence intervals. We assess the performance of our method using a number of simulated datasets and synthetic datasets curated from publicly available single cell datasets. In simulated datasets, scDC correctly recovers the true cell-type proportions. In synthetic datasets, the cell-type composition output from scDC has very high concordance with reference cell compositions from the original data. Since the majority of datasets tested in this study have only 2 to 5 subjects per condition, the addition of confidence intervals enables better comparisons of compositional differences between subjects and across conditions.

Finally, we have made our method available to the scientific community as part of the scdney package (Single Cell Data Integrative Analysis) R package, available from https://github.com/SydneyBioX/scdney.
Molecular characterization of individual cancer patients is important because cancer is a complex and heterogeneous disease with many possible genetic and environmental causes. Many studies have been conducted to identify diagnostic or prognostic gene signatures for cancer from gene expression profiles. However, some gene signatures may fail to serve as diagnostic or prognostic biomarkers and gene signatures may not be found in gene expression profiles.

In this study, we developed a general method for constructing patient-specific gene correlation networks and for identifying prognostic gene pairs from the networks. The main difference of our method from previous ones includes (1) it is focused on finding prognostic gene pairs rather than prognostic genes and (2) it can identify prognostic gene pairs from gene expression profiles even when no significant prognostic genes exist.

Evaluation of our method with extensive data sets of three cancer types (breast invasive carcinoma, colon adenocarcinoma, and lung adenocarcinoma) showed that our approach is general and that gene pairs can serve as more reliable prognostic signatures for cancer than genes. Our study also revealed that prognosis of individual cancer patients is associated with the existence of prognostic gene pairs in the patient specific network and by the size of a subnetwork of the prognostic gene pairs in the patient-specific network. Although preliminary, our approach will be useful for finding gene pairs to predict survival time of patients and to tailor treatments to individual characteristics. The program for dynamically constructing patient-specific gene networks and for finding prognostic gene pairs is available at http://bclab.inha.ac.kr/pancancer.
Oral ID: O-15

PREDICTING SYNTHETIC LETHAL INTERACTIONS IN HUMAN CANCERS USING GRAPH REGULARIZED SELF-REPRESENTATIVE MATRIX FACTORIZATION

Jiang Huang1, Min Wu2, Fan Lu3, Le Ou-Yang3,4* and Zexuan Zhu*1

1 College of Computer Science and Software Engineering, Shenzhen University, Nanhai Ave 3688, 518060 Shenzhen, China.
2 Institute for Infocomm Research (I2R), A*STAR, 1 Fusionopolis Way, Singapore, Singapore.
3 Guangdong Key Laboratory of Intelligent Information Processing and Shenzhen Key Laboratory of Media Security, College of Electronics and Information Engineering, Shenzhen University, Nanhai Ave 3688, 518060 Shenzhen, China.
4 Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen, China.

Correspondence: leouyang@szu.edu.cn

Synthetic lethality has attracted a lot of attentions in cancer therapeutics due to its utility in identifying new anticancer drug targets. Identifying synthetic lethal (SL) interactions is the key step towards the exploration of synthetic lethality in cancer treatment. However, biological experiments are faced with many challenges when identifying synthetic lethal interactions. Thus, it is necessary to develop computational methods which could serve as useful complements to biological experiments.

In this paper, we propose a novel graph regularized self-representative matrix factorization (GRSMF) algorithm for synthetic lethal interaction prediction. GRSMF first learns the self-representations from the known SL interactions and further integrates the functional similarities among genes derived from Gene Ontology (GO). It can then effectively predict potential SL interactions by leveraging the information provided by known SL interactions and functional annotations of genes. Extensive experiments on the synthetic lethal interaction data downloaded from SynLethDB database demonstrate the superiority of our GRSMF in predicting potential synthetic lethal interactions, compared with other competing methods. Moreover, case studies of novel interactions are conducted in this paper for further evaluating the effectiveness of GRSMF in synthetic lethal interaction prediction.

In this paper, we demonstrate that by adaptively exploiting the self-representation of original SL interaction data, and utilizing functional similarities among genes to enhance the learning of self-representation matrix, our GRSMF could predict potential SL interactions more accurately than other state-of-the-art SL interaction prediction methods.
Oral ID: O-16

**scReClassify: POST HOC CELL TYPE CLASSIFICATION OF SINGLE-CELL RNA-SEQ DATA**

Taiyun Kim¹, Kitty Lo¹, Thomas Geddes¹,²,³, Hani Jieun Kim¹,²,³, Jean Yee Hwa Yang¹,³, Pengyi Yang¹,³,⁴

¹School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia; ²School of Life and Environmental Science, University of Sydney, NSW 2006, Australia; ³Charles Perkins Centre, University of Sydney, NSW 2006, Australia; ⁴Children’s Medical Research Institute, Faculty of Medicine and Health, University of Sydney, NSW 2145, Australia

Correspondence: pengyi.yang@sydney.edu.au

Single-cell RNA-sequencing (scRNA-seq) is a fast emerging technology allowing global transcriptome profiling on the single cell level. Cell type identification from scRNA-seq data is a critical task in a variety of research such as developmental biology, cell reprogramming, and cancers. Typically, cell type identification relies on human inspection using a combination of prior biological knowledge (e.g. marker genes and morphology) and computational techniques (e.g. PCA and clustering). Due to the incompleteness of our current knowledge and the subjectivity involved in this process, a small amount of cells may be subject to mislabelling.

Here, we propose a semi-supervised learning framework, named scReClassify, for ‘post hoc’ cell type identification from scRNA-seq datasets. Starting from an initial cell type annotation with potentially mislabelled cells, scReClassify first performs dimension reduction using PCA and next applies a semisupervised learning method to learn and subsequently reclassify cells that are likely mislabelled initially to the most probable cell types. By using both simulated and real-world experimental datasets that profiled various tissues and biological systems, we demonstrate that scReClassify is able to accurately identify and reclassify misclassified cells to their correct cell types. We implemented scReClassify as an R package. It can be used for scRNA-seq data as a post hoc cell type classification tool to fine-tune cell type annotations generated by any cell type classification procedure.

**Availability:** scReClassify is implemented in the scdney (Single Cell Data Integrative Analysis) R package and is freely available from [https://github.com/SydneyBioX/scdney](https://github.com/SydneyBioX/scdney)
Oral ID: O-17

AUTOENCODER-BASED CLUSTER ENSEMBLES FOR SINGLE-CELL RNA-SEQ DATA ANALYSIS

Thomas Geddes¹,², Taiyun Kim¹, Lihao Nan², James Burcheld³, Jean Yee Hwa Yang¹, Dacheng Tao² and Pengyi Yang¹,⁴.

¹Charles Perkins Centre, School of Mathematics and Statistics, Faculty of Science, The University of Sydney, NSW 2006, Australia and
²UBTECH Sydney Artificial Intelligence Centre and the School of Computer Science, Faculty of Engineering and Information Technologies, The University of Sydney, NSW 2008, Australia and
³Charles Perkins Centre, School of Life and Environmental Sciences, Faculty of Science, The University of Sydney, NSW 2006, Australia and
⁴Children’s Medical Research Institute, Faculty of Medicine and Health, The University of Sydney, NSW 2145, Australia

Correspondence: pengyi.yang@sydney.edu.au

Single-cell RNA-sequencing (scRNA-seq) is a transformative technology, allowing global transcriptomes of individual cells to be profiled with high accuracy. An essential task in scRNA-seq data analysis is the identification of cell types from complex samples or tissues profiled in an experiment. To this end, clustering has become a key computational technique for grouping cells based on their transcriptome profiles, enabling subsequent cell type identification from each cluster of cells. Due to the high feature-dimensionality of the transcriptome (i.e. the large number of measured genes in each cell) and because only a small fraction of genes are cell type-specific and therefore informative for generating cell type-specific clusters, clustering directly on the original feature/gene dimension may lead to uninformative clusters and hinder correct cell type identification.

Here, we propose an autoencoder-based cluster ensemble framework in which we first take random subspace projections from the data, then compress each random projection to a low-dimensional space using an autoencoder artificial neural network, and finally apply ensemble clustering across all encoded datasets for generating clusters of cells. We employ four evaluation metrics to benchmark clustering performance and our experiments demonstrate that the proposed autoencoder-based cluster ensemble can lead to substantially improved cell type specific clusters when applied with both the standard k-means clustering algorithm and a state-of-the-art kernel-based clustering algorithm (SIMLR) designed specifically for scRNA-seq data. Compared to directly using these clustering algorithms on the original datasets, the performance improvement in some cases is up to 100%, depending on the evaluation metrics used. These results suggest that the proposed framework can facilitate more accurate cell type identification as well as other downstream analyses.

Availability: The code for creating the proposed autoencoder-based cluster ensemble framework is freely available from https://github.com/gedcom/autoencoder_cluster_ensemble
IMPARO: INFERRING MICROBIAL INTERACTIONS THROUGH PARAMETER OPTIMISATION

Rajith Vidanaarachchi¹, Marnie Shaw¹, Sen-Lin Tang², Saman Halgamuge¹

¹ Australian National University, Australia
² Biodiversity Research Center, Academia Sinica, Taiwan

Correspondence:rajith.v@anu.edu.au

Microbial Interaction Networks (MINs) provide important information for understanding bacterial communities. MINs can be inferred by examining microbial abundance profiles. Abundance profiles are often interpreted with the Lotka Volterra model in research. However existing research fails to consider a biologically meaningful underlying mathematical model for MINs or to address the possibility of multiple solutions.

In this paper we present IMPARO, a method for inferring microbial interactions through parameter optimisation. We use biologically meaningful models for both the abundance profile, as well as the MIN. We show how multiple MINs could be inferred with similar reconstructed abundance profile accuracy, and argue that a singular solution is not always satisfactory. Using our method, we successfully inferred clear interactions in the gut microbiome which have been previously observed in in-vitro experiments.

IMPARO was used to successfully infer microbial interactions in human microbiome samples as well as in a varied set of simulated data. The work also highlights the importance of considering multiple solutions for MINs.
Module detection algorithms relying on modularity maximization suffer from an inherent resolution limit that hinders detection of small topological modules, especially in molecular networks where most biological processes are believed to form small and compact communities. We propose a novel cluster refinement approach that helps finding functionally significant modules of molecular networks.

The module refinement algorithm improves the quality of topological modules in protein-protein interaction networks by finding biologically functionally significant modules. The algorithm is based on the fact that functional modules in biology do not necessarily correspond to those of maximum modularity. Larger modules corresponding to maximal modularity are incrementally modularized again under specific constraints so that smaller yet topologically and biologically valid modules are achieved. We show improvement in quality and functional coverage of modules using experiments on synthetic and real protein-protein interaction networks. We also compare our results with the existing methods.

The proposed algorithm finds smaller but functionally relevant modules that are undetected by classical quality based methods for modular detection. The refinement procedure helps to detect more functionally enriched modules in protein-protein interaction networks, which are also more coherent with functionally characterised gene sets.
A NOVEL CONSTRAINED RECONSTRUCTION MODEL TOWARDS HIGH-RESOLUTION SUB-TOMOGRAM AVERAGING

Renmin Han¹, Lun Li², Peng Yang, Fa Zhang², Xin Gao¹

¹King Abdullah University of Science and Technology (KAUST), Computational Bioscience Research Center (CBRC), CEMSE Division, Thuwal, 23955-6900, Saudi Arabia.
²High Performance Computer Research Center, Chinese Academy of Sciences, Beijing 100190, China.
³University of Chinese Academy of Sciences, Beijing, China.

Correspondence: renmin.han@kaust.edu.sa

Electron tomography (ET) offers unique capacity to image biological structures in situ. However, the resolution of ET reconstructed tomograms is not comparable to that of the single particle cryo-EM. If many copies of the object of interest are present in the tomograms, their structures can be reconstructed separately and averaged together to increase the signal-to-noise ratio and improve the resolution, which is known as the sub-tomogram averaging. To date, the resolution improvement of the sub-tomogram averaging is still limited because each reconstructed sub-tomogram is of low reconstruction quality due to the missing wedge issue in the tilt series images.

In this paper, we propose a novel computational model, the constrained reconstruction model (CRM), to better recover the information from the multiple sub-tomograms to compensate for the missing wedge issue in each of them. We first formulate the traditional averaging method and our CRM as linear systems, and prove that the solution space of CRM is no larger, and in practice much smaller, than that of the traditional averaging method. We then propose a sparse Kaczmarz algorithm to solve the formulated CRM, and further extend the solution to the simultaneous algebraic reconstruction technique (SART). Experimental results demonstrate that CRM can significantly alleviate the missing wedge issue and improve the final reconstruction quality. In addition, our model is robust to the number of images in each tilt series, the tilt range, and the noise level.

Availability: The Matlab code is available at https://github.com/icthrm/constrained-reconstruction and the C/C++ codes of CRMSIRT and CRM-SART is available upon acceptance.
Oral ID: O-22

DESIGNING OF PRECISE VACCINE CONSTRUCT AGAINST VISCERAL LEISHMANIASIS THROUGH PREDICTED EPITOPE ENSEMBLE: A CONTEMPORARY APPROACH

Garima Singh1, Manisha Pritam1, Monisha Banerjee2, Akhilesh Kumar Singh1 and Satarudra Prakash Singh1*

1Amity Institute of Biotechnology, Amity University Uttar Pradesh, Lucknow Campus, Lucknow-226028, India
2Molecular and Human Genetics Lab, Department of Zoology, University of Lucknow, Lucknow- 226007, India

Correspondence: spsingh5@amity.edu

Visceral leishmaniasis (VL) caused by *Leishmania donovani* is a fatal parasitic disease affecting primarily the poor population in endemic countries. Increasing number of deaths as well as resistant to existing drugs necessitates the development of an effective vaccine for successful treatment of VL. The present study employed a combinatorial approach for designing monomer vaccine construct against *L. donovani* by applying forecasted B- and T-cell epitopes from 4 genome derived antigenic proteins having secretory signal peptides and glycoprophatidylinositol (GPI) anchors with ≤ 1 transmembrane helix. The forecasted population coverage of chosen T cell epitope ensemble (combined HLA class I and II) cover 99.14% of world-wide human population. The predicted 3D structure of vaccine constructs (VC1/VC2) were modeled using homology modeling approach and docked to innate immune receptors TLR-2 and TLR-4 with docking energies -1231.4/-910.3 and -1119.4/-1476, correspondingly. Overall, the aforementioned designed vaccine constructs were found appropriate for including in self-assembly protein nanoparticles (SAPN) for further study in developing cutting-edge precision vaccine against VL in short duration with cost-effective manner.
DNA methylation is a crucial epigenomic mechanism in various biological processes. Using whole genome bisulfite sequencing (WGBS) technology, the methylated cytosine sites can be revealed at single nucleotide level. However, the WGBS data analysis process is usually complicated and challenging.

To alleviate the associated difficulties, we integrated the WGBS data processing steps and downstream analysis into a two-phase approach. First, we developed a Docker container DocMethyl to deal with read mapping, methylation calling and scoring to a methylation status summary, the mtable. Next, the mtable files are uploaded to the web server EpiMOLAS_web for linking with gene annotation databases that enable rapid data retrieval and analyses.

To our knowledge, the EpiMOLAS framework consisting of DocMethyl and EpiMOLAS_web is the first approach which attempts to include Docker containerization technology and web-based system for WGBS data analysis from raw data processing to downstream analysis. EpiMOLAS will help users cope with their WGBS data and also conduct reproducible analysis of publicly available data to gain insights into the mechanisms underlying complex biological phenomenon. The Galaxy Docker image DocMethyl is available at https://hub.docker.com/r/lsbnb/docmethyl/. EpiMOLAS_web is publicly accessible at http://symbiosis.iis.sinica.edu.tw/epimolas/.
RV0807, A PUTATIVE PHOSPHOLIPASE A2 OF MYCOBACTERIUM TUBERCULOSIS; ELUCIDATION THROUGH SEQUENCE ANALYSIS, HOMOLOGY MODELING, MOLECULAR DOCKING AND MOLECULAR DYNAMICS STUDIES OF POTENTIAL SUBSTRATES AND INHIBITORS

Shobana Sundar*, Lokesh Thangamani, Gowdham Manivel, Praveen Kumar and Shanmughavel Piramanayagam
Department of Bioinformatics, Bharathiar University, Coimbatore.

*Correspondence: sundar.shobana@gmail.com

Mycobacterium tuberculosis has the ability to scrounge off the host macrophages and create a cordial environment for its survival. Identification of mechanisms favoring this purpose leads to novel treatment strategies for tuberculosis. Rv0807, a homolog of MSMEG_5817 from M. smegmatis may be crucial for mycobacterium survival within the host macrophages. In this study through insilico approaches, we intend to identify the putative role for Rv0807 from M. tuberculosis and its essentiality for mycobacterium survival within the macrophages. We have designed motifs for Rv0807, MSMEG_5817 and related phospholipase sequences from different strains. From sequence analysis, we suggest that Rv0807 could be a Phospholipase A2 of Mtb. We have predicted some important residues to be a part of the catalytic process of the Rv0807 homodimer and we have also performed insilico mutation studies of these residues to understand their role in catalysis. Rv0807 could be a potential drug target as it binds phosphatidylinositol-3- phosphate (PI3P) and could be involved in processing the host cell PI3Ps, thereby blocking the phagosomal maturation. A pharmacophore hypothesis was generated based on the ligand binding site and a set of Pretomanid related compounds were screened against the Rv0807 homodimer. The top five compounds having better docking scores and good ADME properties were selected as best inhibitors and analyzed further. Molecular dynamics studies of Rv0807 homodimer with PI3P, demonstrated a lot of conformational changes in the protein structure as it gets occluded through the course of simulation. The movement of a loop atop the ligand binding site, suggests of a lid-like region as seen in many other phospholipases. Thus, the insights gained through this study will provide some novel ideas to understand Rv0807 catalysis and its role in Mycobacterium survival within the macrophage.
CHARACTERIZATION OF RISK GENES OF AUTISM SPECTRUM DISORDERS USING GENE EXPRESSION PROFILES

Yenching Lin¹, Srinivasulu Yerukala Sathipati², and Shinn-Ying Ho¹,²,³,⁴

¹Interdisciplinary Neuroscience Ph.D. Program, National Chiao Tung University, Hsinchu, Taiwan
²Institute of Bioinformatics and Systems Biology, National Chiao Tung University, Hsinchu, Taiwan
³Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, Taiwan
⁴Center For Intelligent Drug Systems and Smart Bio-devices (IDS2B), National Chiao Tung University, Hsinchu, Taiwan

Correspondence: syho@mail.nctu.edu.tw

Autism spectrum disorders (ASD) refers to a wide spectrum of neurodevelopmental disorders, which emerges during infancy and continues throughout a lifespan. Although, substantial efforts have been made towards therapeutic approaches, core symptoms persist lifelong in many individuals with ASD. This work aims to identify risk genes associated with to temporospatial regions of individuals with ASD for exploring the specificity and early detection of ASD.

This work proposes a support vector machine (SVM) based method, called SVM-ASD, to identify risk genes of ASD from expression profiles. SVM-ASD used an efficient feature selection algorithm to identify 19 temporospatial features for predicting ASD risk genes. SVM-ASD achieved a 10-fold cross-validation accuracy, sensitivity, specificity, area under a receiver operating characteristic curve, and test accuracy of 81.83%, 0.84, 0.79, 0.84, and 72.27%, respectively. The top 10 temporospatial features according to their contribution to the prediction accuracy were Posteroventral (inferior) parietal cortex-13pcw, Primary visual cortex (striate cortex area V1/17)-8yrs, Posterior superior temporal cortex (area S1, area 3,1,2)-16pcw, Striatum-13pcw, Orbital frontal cortex-40yrs, Anterior (rostral) cingulate (medial prefrontal cortex)-8pcw, Dorsal thalamus-12pcw, Amygdaloid complex-8yrs, Primary auditory cortex (core)-8yrs, and Hippocampus (hippocampal formation)-11yrs. Furthermore, significance of these top 10 features in ASD were analysed.

Analysis of the identified temporospatial features revealed their important roles in ASD etiology. The identified temporospatial features would be helpful to explore gene expression profiles that are implicated in different regions of the brain in individuals with ASD.
Myc is a crucial player in cellular proliferation and a known regulator of cancer pathobiology. Modulation of Myc expression targeting the Myc Protein-Protein Interactors (PPIs) like Myc-Max has till now been the most explored approach. However, this approach threatens the normal cells where Myc expression is required for proliferation. This demands the need for a new strategy to indirectly modulate Myc expression. Indirect modulation can be achieved by regulating Myc turnover. FBXW7 mediates the ubiquitination and subsequent degradation of Myc which is reversed by USP28. In this study, the interaction of USP28 with FBXW7 as well as with its substrate, Ubiquitin (Ub) were used as targets. Computation based high-throughput screening of bioactive small chemicals using molecular docking method was implemented to predict USP28 inhibitors. For the two regions, docking study with AutoDock Vina gave top 10 best scoring drugs which were identified and tabulated. The two regions defined in the study as FBXW7 binding and Ub binding also encompass the areas in which USP28 differed from USP25, a homologue with a different role. Out of these the best scoring drugs were explored for their role in cancer, if any. This study was performed keeping in mind re-purposing of these known drugs for possible alternative anti-Myc cancer therapy.
Enhancers are non-coding DNA fragments which are crucial in gene regulation (e.g., transcription and translation). Due to the great difference in the locating site among these enhancers, identifying their locations is, therefore, more complicated than other genetic factors. To address this biological issue, several in silico studies have been done to identify and classify enhancer sequences among a myriad of DNA sequences using computational advances. Although recent studies have come up with improved performance, shortfalls in these learning models still remain. To overcome the current limitations of previous learning models, we introduce iEnhancer-ECNN - a novel prediction framework using One-Hot Encoding (OHE) for data transformation and ensemble Convolutional Neural Network (CNN) for model construction to identify enhancers as well as classify their strength. The dataset from Liu et al.’s study was used to develop and evaluate the models. Comparative analysis between iEnhancer-ECNN and other state-of-the-arts was done to fairly assess the model performance.

Our results demonstrated that iEnhancer-ECNN has better performance compared to other state-of-the-art predictors. The accuracy of the ensembled training model for enhancer identification and classification are 0.765 and 0.695, respectively. In both enhancer identification and classification models, iEnhancer-ECNN achieves slightly lower specificity values compared to those of other methods introduced in previous studies. However, the improvement in the area under the receiver operating characteristics curve (AUC), sensitivity, and Matthews’s correlation coefficient (MCC) is remarkable, especially with respective to the enhancer classification model where the increases are about 13%, 55.5%, and 84%, respectively. Our source code is available at http://homepages.ecs.vuw.ac.nz/~nguyenb5/bioinformatics/enhancers

According to all the evaluation metrics, iEnhancer-ECNN outperforms other previously proposed methods with significant improvement in most of the metrics.
Logicome Profiler: EXHAUSTIVE DETECTION OF STATISTICALLY SIGNIFICANT LOGIC RELATIONSHIPS FROM COMPARATIVE OMICS DATA

Tsukasa Fukunaga$^1$ and Wataru Iwasaki$^{2,3,4,5,6}$

$^1$Department of Computer Science, Graduate School of Information Science and Technology, The University of Tokyo, 113-0033 Tokyo, Japan.
$^2$Department of Biological Sciences, Graduate School of Science, The University of Tokyo, 113-0032, Tokyo, Japan.
$^3$Department of Computational Biology and Medical Science, Graduate School of Frontier Sciences, The University of Tokyo, 277-8568 Chiba, Japan.
$^4$Atmosphere and Ocean Research Institute, The University of Tokyo, 277-8564 Chiba, Japan.
$^5$Institute for Quantitative Biosciences, The University of Tokyo, 113-0032 Tokyo, Japan.
$^6$Collaborative Research Institute for Innovative Microbiology, The University of Tokyo, 113-0032 Tokyo, Japan.

*Correspondence: fukunaga@is.s.u-tokyo.ac.jp

Logic relationship analysis is a data mining method that comprehensively detects item triplets that satisfy logic relationships from a binary matrix dataset, such as an ortholog table in comparative genomics. Thanks to recent technological advancements, many binary matrix datasets are now being produced in genomics, transcriptomics, epigenomics, metagenomics, and many other fields for comparative purposes. However, regardless of presumed interpretability and importance of logic relationships, existing data mining methods are not based on the framework of statistical hypothesis testing. That means, the type-1 and type-2 error rates are neither controlled nor estimated.

Here, we developed Logicome Profiler, which exhaustively detects statistically significant triplet logic relationships from a binary matrix dataset (Logicome means ome of logics). To improve the statistical power in multiple testing correction, Logicome Profiler adjusts the significance level by a two-stage Bonferroni method instead of direct application of the Bonferroni method. Its application to an ocean metagenomic dataset showed that Logicome Profiler can effectively detect statistically significant triplet logic relationships among environmental microbes and genes, which include those among urea transporter, urease, and photosynthesis-related genes.

In this study, we proposed Logicome Profiler, which is the first method for logic relationship analysis based on the framework of statistical hypothesis testing. Beyond omics data analysis, Logicome Profiler is applicable to various binary matrix datasets in general for finding significant triplet logic relationships. The source code is available at https://github.com/fukunagatsu/LogicomeProfiler.
Oral ID: O-29

STRUCTURE-BASED DISCOVERY OF NOVEL INHIBITORS OF MYCOBACTERIUM TUBERCULOSIS CYP121 FROM INDONESIAN NATURAL PRODUCTS

Vivitri Dewi Prasasty1*, Sandra Cindana2, Fransiskus Xaverius Ivan2, Hilyatuz Zahroh3, Ernawati Sinaga4

1 Faculty of Technobiology, Atma Jaya Catholic University, Jakarta 12930, DKI Jakarta, Indonesia
2 Department of Biotechnology, Faculty of Life Sciences, Surya University, Tangerang 15810, Banten, Indonesia
3 Genetics Research Centre, Universitas Yarsi, Jakarta 10510, DKI Jakarta, Indonesia
4 Faculty of Biology, Universitas Nasional, Jakarta 12520, Indonesia

*Correspondence: vivitri.dewi@atmajaya.ac.id

Tuberculosis (TB) continues to be a serious global health threat with the emergence of multidrug-resistant tuberculosis (MDR-TB) and extremely drug-resistant tuberculosis (XDR-TB). There is an urgent need to discover new drugs to deal with the advent of drug resistant TB variants. This study aims to find new M. tuberculosis CYP121 inhibitors by screening of Indonesian natural products using the principle of structure-based drug design and discovery. In this work, eight natural compounds isolated from Rhoeo spathacea and Pluchea indica were selected based on their antimycobacterial activity. Derivatives compound were virtually designed from these natural molecules to improve the interaction of ligands with CYP121. Virtual screening of ligands was carried out using AutoDock Vina followed by 50 ns molecular dynamics simulation using YASARA to study the inhibition mechanism of the ligands. Two ligands, i.e. kaempferol (KAE) and its benzyl derivative (KAE3) are identified as the best CYP121 inhibitors based on their binding affinities and adherence to the Lipinski’s rule. Results of molecular dynamics simulation indicate that KAE and KAE3 possess a unique inhibitory mechanism against CYP121 that is different from GGJ (control ligand). The control ligand alters the overall dynamics of the receptor, which is indicated by changes in residue flexibility away from CYP121 binding site. Meanwhile, the dynamic changes caused by the binding of KAE and KAE3 are isolated around the binding site of CYP121. These ligands can be developed for further biological activity tests.
GO2Vec: TRANSFORMING GO TERMS AND PROTEINS TO VECTOR REPRESENTATIONS USING GRAPH EMBEDDINGS

Xiaoshi Zhong*, Rama Kaalia and Jagath C Rajapakse

School of Computer Science and Engineering, Nanyang Technological University, Singapore

*Correspondence: xszhong@ntu.edu.sg

Semantic similarity of gene ontology (GO) terms is a fundamental measure for many bioinformatics applications, such as determining functional similarity of genes or proteins. Most previous research exploit the information content to estimate the semantic similarity of GO terms; recently some research exploit the word embeddings to learn vector representations from corpus for the similarity of GO terms. In this paper, we introduce a new method named GO2Vec, which exploit the graph embeddings to learn vector representations from GO graph for the similarity of GO terms. Our method combines information from both GO graph and GO annotation data, and its learned vectors can be applied to a variety of bioinformatics applications, such as the similarity of proteins and protein-protein interaction prediction. To evaluate the quality of GO2Vec, we use its learned vectors to conduct two kinds of experiments: (1) semantic similarity of proteins on the Collaborative Evaluation of GO-based Semantic Similarity Measures (CESSM) dataset and (2) protein-protein interaction prediction on the Yeast and Human datasets from the STRING database. Experimental results on the two kinds of experiments demonstrate superior performance over the information content based similarity measures and the word embedding based similarity measures.
AN IMPROVED \textit{DE NOVO} GENOME ASSEMBLY OF THE COMMON MAMMOSET GENOME YIELDS IMPROVED CONTIGUITY AND INCREASED MAPPING RATES OF SEQUENCE DATA

Vasanthan Jayakumar\textsuperscript{1}, Ishii Hiromi\textsuperscript{1}, Misato Seki\textsuperscript{1}, Wakako Kumita\textsuperscript{2}, Takashi Inoue\textsuperscript{2}, Sumitaka Hase\textsuperscript{1}, Kengo Sato\textsuperscript{1}, Hideyuki Okano\textsuperscript{3,4}, Erika Sasaki\textsuperscript{2}, Yasubumi Sakakibara\textsuperscript{1*}

\textsuperscript{1}Department of Biosciences and Informatics, Keio University, Yokohama, Kanagawa, 223-8522, Japan.
\textsuperscript{2}Department of Marmoset Biology and Medicine, Central Institute for Experimental Animals, Kawasaki, Kanagawa, 210-0821, Japan.
\textsuperscript{3}Department of Physiology, Keio University School of Medicine, Japan.
\textsuperscript{4}Laboratory for Marmoset Neural Architecture, RIKEN Center for Brain Science, Wako-shi, Saitama, 351-0198, Japan.

Correspondence: yasu@dna.bio.keio.ac.jp

The common marmoset (\textit{Callithrix jacchus}) is one of the most studied primate model organisms. However, the marmoset genomes available in the public databases are highly fragmented and filled with sequence gaps, hindering research advances related to marmoset genomics and transcriptomics.

Here we utilize single-molecule, long-read sequence data to improve and update the existing genome assembly and report a near-complete genome of the common marmoset. The assembly is of 2.79 Gb size, with a contig N50 length of 6.37 Mb and a chromosomal scaffold N50 length of 143.91 Mb, representing the most contiguous and high-quality Marmoset genome up to date. Approximately 90\% of the assembled genome was represented in contigs longer than 1 Mb, with approximately 104 fold improvement in contiguity over the previously published marmoset genome. More than 98\% of the gaps from the previously published genomes were filled successfully, which improved the mapping rates of genomic and transcriptomic data on to the assembled genome.

Altogether the updated, high-quality common marmoset genome assembly provide improvements at various levels over the previous marmoset assembly versions. This will allow researchers working on primate genomics to apply the genome more efficiently for their genomic and transcriptomic sequence data.
**DIFFERENTIAL ALTERNATIVE SPLICING REGULATION AMONG HEPATOCELLULAR CARCINOMA WITH DIFFERENT RISK FACTORS**

Young-Joo Jin¹², Seyoun Byun¹, Seonggyun Han¹, John Chamberlin¹, Dongwook Kim¹, Min Jung⁴ Kim¹³, and Younghee Lee¹⁴*

1 Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, Utah, USA; 2 Division of Gastroenterology, Department of Internal Medicine, Inha University School of Medicine, Incheon, South Korea; 3 Pharmacy program, Massachusetts College of Pharmacy and Health Sciences, Worcester Massachusetts; 4 Huntsman Cancer Institute, University of Utah School of Medicine, Salt Lake City, Utah, USA

Correspondence: younghee.lee@utah.edu

Hepatitis B virus (HBV), hepatitis C virus (HCV), and alcohol consumption are predominant causes of hepatocellular carcinoma (HCC). However, the molecular mechanisms underlying how differently these causes are implicated in HCC development are not fully understood. Therefore, we investigated differential alternative splicing (AS) regulation among HCC patients with these risk factors. We conducted a genome-wide survey of AS events associated with HCCs among HBV (n=95), HCV (n=47), or alcohol (n=76) using RNA-sequencing data obtained from The Cancer Genome Atlas.

In three group comparisons of HBV vs. HCV, HCV vs. alcohol, and HBV vs. alcohol for RNA seq (ΔPSI>0.05, FDR<0.05), 133, 29, and 93 differential AS events (143 genes) were identified, respectively. Of 143 AS genes, eight and one gene were alternatively spliced specific to HBV and HCV, respectively. Through functional analysis over the canonical pathways and gene ontologies, we identified significantly enriched pathways in 143 AS genes including immune system, mRNA splicing-major pathway, and nonsense-mediated decay, which may be important to carcinogenesis in HCC risk factors. Among eight genes with HBV-specific splicing events, HLA-A, HLA-C, and IP6K2 exhibited more differential expression of AS events (ΔPSI>0.1). Intron retention of HLA-A was observed more frequently in HBV-associated HCC than HCV- or alcohol-associated HCC, and intron retention of HLA-C showed vice versa. Exon 6 of IP6K2 was less skipped in HBV-associated in HCC compared to HCV- or alcohol-associated HCC.

AS may play an important role in regulating transcription differences implicated in HBV-, HCV-, and alcohol-related HCC development.
LePrimAlign: LOCAL ENTROPY-BASED ALIGNMENT OF PPI NETWORKS TO PREDICT CONSERVED MODULES

Sawal Maskey\(^1\) and Young-Rae Cho\(^{1,2,*}\)

\(^1\) Department of Computer Science, Baylor University, One Bear Place #97141, Waco, TX, 76798 USA
\(^2\) Bioinformatics Program, Baylor University, One Bear Place #97141, Waco, TX, 76798 USA

*Correspondence: Young-Rae Cho@baylor.edu

Cross-species analysis of protein-protein interaction (PPI) networks provides an effective means of detecting conserved interaction patterns. Identifying such conserved substructures between PPI networks of different species increases our understanding of the principles deriving evolution of cellular organizations and their functions in a systematic level. In recent years, network alignment techniques have been applied to PPI networks to predict evolutionary conserved modules. Although a wide variety of network alignment algorithms have been introduced, developing a scalable local network alignment algorithm with high accuracy is still challenging.

We present a novel pairwise local network alignment algorithm, called LePrimAlign, to predict conserved modules between PPI networks of two different species. The proposed algorithm exploits the results of a pairwise global alignment algorithm with many-to-many node mappings. It also applies the concept of graph entropy to detect initial cluster pairs from two networks. Finally, the initial clusters are expanded to increase the local alignment score that is formulated by a combination of intra-network and inter-network scores. The performance comparison with state-of-the-art approaches demonstrates that the proposed algorithm outperforms in terms of accuracy of predicted protein complexes and quality of alignments.

The proposed method produces local network alignments of higher accuracy in predicting conserved modules even with large biological networks at a reduced computational cost.
The 18th International Conference on Bioinformatics

Oral ID: O-34

THE TRANSCRIPTOME VARIATIONS OF PANAX NOTOGINSENG ROOTS TREATED WITH DIFFERENT FORMS OF NITROGEN FERTILIZERS

Xiaohong Ou¹³⁴, Shipeng Li¹², Peiran Liao¹, Xiuming Xiu¹, Binglian Zheng⁵, Ye Yang¹, Dahui Liu⁴ and Yun Zheng²

¹ Kunming Key Laboratory of Sustainable Development and Utilization of Famous-Region Drug, Key Laboratory of Panax notoginseng Resources Sustainable Development and Utilization of State Administration of Traditional Chinese Medicine, Faculty of Life Science and Technology, Kunming University of Science and Technology, Kunming, 650500, China.
² Yunnan Key Lab of Primate Biomedicine Research; Institute of Primate Translational Medicine, Kunming University of Science and Technology, Kunming, 650500, China.
³ Guizhou University of Traditional Chinese Medicine, Guiyang, 550025, China.
⁴ Institute of Medicinal Plants, Yunnan Academy of Agricultural Sciences, Kunming, 650500, China.
⁵ State Key Laboratory of Genetic Engineering, Collaborative Innovation Center for Genetics and Development, Institute of Plant Biology, School of Life Sciences, Fudan University, Shanghai, 200438, China.

*Correspondence: zhengyun5488@gmail.com

The sensitivity of plants to ammonia is a worldwide problem that limits crop production. Excessive use of ammonium as the sole nitrogen source results in morphological and physiological disorders, and retarded plant growth.

In this study we found that the root growth of Panax notoginseng was inhibited when only adding ammonium nitrogen fertilizer, but the supplement of nitrate fertilizer recovered the integrity, activity and growth of root. Twelve RNA-seq profiles in four sample groups were produced and analyzed to identify deregulated genes in samples with different treatments. In comparisons to NH₄⁺ treated samples, ACLA-3 gene is up-regulated in samples treated with NO₃⁻ and with both NH₄⁺ and NO₃⁻, which is further validated by qRT-PCR in another set of samples. Subsequently, we show that the some key metabolites in the TCA cycle are also significantly enhanced when introducing NO₃⁻. These potentially enhance the integrity and recover the growth of Panax notoginseng roots.

These results suggest that the activated TCA cycle contributes to the increased Panax notoginseng root yield when applying both ammonium and nitrate fertilizer.
EFFECTS OF ORDERED MUTATIONS ON DYNAMICS IN SIGNALING NETWORKS

Maulida Mazaya¹, Hung-Cuong Trinh², and Yung-Keun Kwon¹*

¹ School of IT Convergence, University of Ulsan, 93 Daehak-ro, Nam-gu, Ulsan 44610, Republic of Korea.
² Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam.

*Correspondence: kwonyk@ulsan.ac.kr (YKK)

Many previous clinical studies have found that accumulated sequential mutations are statistically related to tumorigenesis. However, they are limited in fully elucidating the significance of the ordered-mutation because they did not focus on the network dynamics. Therefore, there is a pressing need to investigate the dynamics characteristics induced by ordered-mutations.

To quantify the ordered-mutation-inducing dynamics, we defined the mutation-sensitivity and the order-specificity that represent if the network is sensitive against a double knockout mutation and if mutation-sensitivity is specific to the mutation order, respectively, using a Boolean network model. Through intensive investigations, we found that a signaling network is more sensitive when a double-mutation occurs in the direction order inducing a longer path and a smaller number of paths than in the reverse order. In addition, feedback loops involving a gene pair decreased both the mutation-sensitivity and the order-specificity. Next, we investigated relationships of functionally important genes with ordered-mutation-inducing dynamics. The network is more sensitive to mutations subject to drug-targets, whereas it is less specific to the mutation order. Both the sensitivity and specificity are increased when different-drug-targeted genes are mutated. Further, we found that tumor suppressors can efficiently suppress the amplification of oncogenes when the former are mutated earlier than the latter.

Taken together, our results help to understand the importance of the order of mutations with respect to the dynamical effects in complex biological systems.
IDENTIFICATION OF LUNG CANCER GENE MARKERS THROUGH KERNEL MAXIMUM MEAN DISCREPANCY AND INFORMATION ENTROPY

Zhixun Zhao¹, Hui Peng¹, Xiaocai Zhang¹, Yi Zheng¹, Fang Chen², Liang Fang³ and Jinyan Li*¹

¹ Advanced Analytics Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Australia.
² Faculty of Engineering and Information Technology, University of Technology Sydney, PO Box 123, Broadway, NSW 2007, Australia.
³ School of Computer, National University of Defense Technology, 410073, Changsha, China.

*Correspondence: Jinyan.Li@uts.edu.au

The early diagnosis of lung cancer has been a critical problem in clinical practice for a long time, and discovering differential expressed genes as disease markers through high-throughput sequencing data is one of the promising solutions. However, the current gene differential expression analysis (DEA) methods have two main drawbacks: First, these methods based on fixed statistical hypotheses are not always effective; Second, these methods are not able to identify a certain expression boundary when there is no obvious expression level gap between control and experiment groups.

This paper proposed a novel way to identify marker genes and gene expression level boundary for lung cancer. By calculating a kernel maximum mean discrepancy, our method can evaluate the expression differences between normal-normal adjacent to the tumor (NAT), normal-tumor and NAT-tumor groups. Compared with two conventional methods t-test and fold change, the top average ranked genes selected by our method achieve the best performance under all metrics in 10-fold cross-validation. Then GO and KEGG enrichment analysis are conducted to analyze the biological function of the top 100 ranked genes. At last, we choose the top 10 average ranked genes as lung cancer markers and identify the gene expression boundary in normal, NAT and tumor samples as the classification criterion.

The proposed kernel maximum mean discrepancy and expression boundary detection method is useful to identify gene markers for lung cancer. Our method is not only more efficient than conventional DEA methods but also provides a reliable way to identify the gene expression level boundary.
AN AUTOMATED 3D MODELLING PIPELINE FOR CONSTRUCTING 3D MODELS OF MONogenean HARDPART USING MACHINE LEARNING TECHNIQUES

1,2 Bee Guan Teo and 2 Sarinder Kaur Dhillon*

1 School of Engineering, Monash University Malaysia, Kuala Lumpur, Malaysia
2 Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia

*Correspondence: sarinder@um.edu.my

Studying structural and functional morphology of small organisms such as monogenean, is difficult due to the lack of visualization in three dimensions. One possible way to resolve this visualization issue is to create digital 3D models which may aid researchers in studying morphology and function of the monogenean. However, the development of 3D models is a tedious procedure as one will have to repeat entire complicated modelling process for every new target 3D shape using a comprehensive 3D modelling software. This study was designed to develop an alternative 3D modelling approach to build 3D models of monogenean anchors which can be used to understand these morphological structures in three dimensions. The aim of this alternative 3D modelling approach is to avoid repeating the tedious modelling procedure for every single target 3D model from scratch.

An automated 3D modelling pipeline empowered by an Artificial Neural Network (ANN) was developed. This automated 3D modelling pipeline enables automated deformation of a generic 3D model of monogenean anchor into another target 3D anchor. The 3D modelling pipeline empowered by ANN has managed to automate the generation of the 8 target 3D models (representing 8 species: Dactylogyrus primaries, Pellucidhaptor merus, Dactylogyrus falcatus, Dactylogyrus vastator, Dactylogyrus pterocleidus, Dactylogyrus falcunguis, Chauhanellus auriculatum and Chauhanellus caelatus) of monogenean anchor from the respective 2D illustrations input without repeating the tedious modelling procedure.

Despite some constraints and limitation, the automated 3D modelling pipeline developed in this study has demonstrated a working idea of application of machine learning approach in a 3D modelling work. This study has not only developed an automated 3D modelling pipeline but also has demonstrated a cross-disciplinary research design that integrates machine learning into a specific domain of study such as 3D modelling of the biological structures.
CRISPR-based systems are playing an important role in modern genome engineering. A large number of computational methods have been developed to assist in the identification of suitable guides. However, there is only limited overlap between the guides that each tool identifies. This can motivate further development, but also raises the question of whether it is possible to combine existing tools to improve guide design.

We considered 10 leading guide design tools, and their output for a set of guides for which experimental validation data is available. We found that consensus approaches were able to outperform all individual tools. The best performance (with a precision of 0.924) was obtained when combining five of the tools and accepting all guides selected by at least four of them.

These results can be used to improve CRISPR-based studies, but also to guide further tool development. However, they only provide a short-term solution as the time and computational resources required to run five tools may be impractical in certain applications.
JCDB: A COMPREHENSIVE KNOWLEDGE DATABASE FOR JATROPHA CURCAS, AN EMERGING MODEL FOR WOODY ENERGY PLANTS

Xuan Zhang\textsuperscript{1,2}, Bang-Zhen Pan\textsuperscript{1}, Maosheng Chen\textsuperscript{1}, Wen Chen\textsuperscript{1}, Jing Li\textsuperscript{1}, Zeng-Fu Xu\textsuperscript{1,*} and Changning Liu\textsuperscript{1,*}

\textsuperscript{1} CAS Key Laboratory of Tropical Plant Resources and Sustainable Use, Xishuangbanna Tropical Botanical Garden, The Innovative Academy of Seed Design, Chinese Academy of Sciences, Kunming 650223, China
\textsuperscript{2} University of Chinese Academy of Sciences, Beijing 100049, China

*Correspondences: liuchangning@xtbg.ac.cn; zfxu@xtbg.ac.cn

Jatropha curcas is an oil-bearing plant, and has seeds with high oil content (~40%). Several advantages, such as easy genetic transformation and short generation duration, implement J. curcas as an emerging model for woody energy plants. With the development of high-throughput sequencing, the genome of J. curcas has been sequenced and a mass of transcriptome data have been produced. How to integrate and analyze these omics data is crucial for functional genomics research on J. curcas. Results: By establishing pipelines for processing novel gene identification, gene function annotation, and gene network construction, we systematically integrated and analyzed a series of J. curcas transcriptome data. Based on these data, we constructed a Jatropha curcas database named JCDB. The JCDB database not only includes general gene information, gene functional annotation, gene interaction network, and gene expression matrix, but also provides the tools for browsing, searching and downloading against all data, as well as online BLAST, JBrowse genome browser, ID conversion, and gene network analysis tools. Conclusions: JCDB is the most comprehensive and well annotated database for J. curcas. We believe it will make valuable contributions to the functional genomics study of J. curcas. The Jatropha curcas database is accessible at http://jcdb.xtbg.ac.cn.
PREDICTING SYNERGISTIC DRUGS USING GRADIENT TREE BOOSTING BASED ON FEATURES EXTRACTED FROM DRUG-PROTEIN HETEROGENEOUS NETWORK

Hui Liu¹,², Wenhao Zhang¹, Lixia Nie², Xiancheng Ding³, Judong Luo⁴ and Ling Zou⁵

¹ Lab of Information Management, Changzhou University, Jiangsu, China.
² School of Information Science and Engineering, Changzhou University, Jiangsu, China.
³ Information Center, Changzhou University, Jiangsu, China., Jiangsu, Jiangsu 213164, China.
⁴ Department of Oncology, the Affiliated Changzhou No.2 People’s Hospital of Nanjing Medical University, Changzhou, China.

*Correspondence: zouling@cczu.edu.cn

Although targeted drugs have led to remarkable advances in the treatment of cancer patients, their clinical benefits to tumor therapies are greatly limited due to intrinsic and acquired resistance of cancer cells against such drugs. Drug combinations synergistically interfere with protein networks to inhibit more effectively the activity level of carcinogenic genes, and thus play an increasingly important role in the treatment of complex disease. In this paper, we combined the drug similarity network, protein similarity network and known drug-protein associations into a drug-protein heterogenous network. Next, we run random walk with restart (RWR) on the heterogenous network using the combinatorial drug targets as the initial probability, and obtained the probability distribution after convergence as the feature vector of each drug combination. Taking these feature vectors as input, we trained a gradient tree boosting (GTB) classifier to predict new drug combinations. We have conducted performance evaluation on the widely used data sets of drug combinations came from DCDB database. The experimental results show that our method achieved higher performance than seven typical classifiers and traditional boosting algorithms. From the perspective of network pharmacology, our method effectively exploits the topological attributes and interactions of drug targets in the overall biological network, which has been shown to be a systematic and reliable approach for drug discovery.
DECONVOLUTION OF BULK GENE EXPRESSION PROFILES FROM COMPLEX TISSUES TO QUANTIFY SUBSETS OF IMMUNE CELLS

Yen-Jung Chiu, Yi-Hsuan Hsieh, Yen-Hua Huang*
Institute of Biomedical Informatics, Yang-Ming University, Taiwan
Correspondence: yhhuang@ym.edu.tw

To facilitate the investigation of the pathogenic roles played by different types of immune cells in complex tissues like tumors, a few computational methods for deconvoluting the bulk gene expression profiles to predict the cell composition have been created. However, available methods were usually developed along with a set of reference gene expression profiles consisting of imbalanced replicates across different cell types. Methods and results: Therefore, the objective of this study was to create a new deconvolution method and also build a new set of reference gene expression profiles that incorporate more microarray replicates of the immune cells that have been frequently implicated with the poor prognosis of cancers, such as T helper cells, Regulatory T cells and macrophage M1/M2 cells. Our deconvolution method was developed by choosing $\varepsilon$-Support Vector Regression ($\varepsilon$-SVR) as the core algorithm assigned with a loss function subject to the L1-norm penalty. The benchmark revealed that the performance of our method was at least comparable to that of a state-of-art tool, CIBERSORT, by using the test data sets of in silico cell mixtures, simulated bulk tissues, and real human samples with known immune cell fractions.

We developed a new cell composition deconvolution method and the implementation of this method was all based on the publicly available R and Python packages. Besides, we compiled a new set of reference gene expression profiles for analyzing 7 types of immune cells by recruiting 135 public-domain microarray samples. This new reference gene expression signature matrix might allow a more robust prediction of the cell fractions for tumor associated immune cells.
RAPID CLASSIFICATION OF GROUP B STREPTOCOCCUS SEROTYPES BASED ON MATRIX-ASSISTED LASER DESORPTION IONIZATION-TIME OF FLIGHT MASS SPECTROMETRY AND MACHINE LEARNING TECHNIQUES

Hsin-Yao Wang¹,², Wen-Chi Li³, Kai-Yao Huang³, Chia-Ru Chung⁴, Jorng-Tzong Horng⁴,⁵, Jen-Fu Hsu⁶,⁷,*, Jang-Jih Lu¹,⁸,*, and Tzong-Yi Lee³,¹⁰,*

¹ Department of Laboratory Medicine, Chang Gung Memorial Hospital at Linkou, Taoyuan 33305, Taiwan
² Ph.D. Program in Biomedical Engineering, Chang Gung University, Taoyuan City, Taiwan
³ Warshel Institute for Computational Biology, The Chinese University of Hong Kong, Shenzhen 518172, China
⁴ Department of Computer Science and Information Engineering, National Central University, Taoyuan 32001, Taiwan
⁵ Department of Bioinformatics and Medical Engineering, Asia University, Taoyuan City, Taiwan
⁶ Division of Pediatric Neonatology, Department of Pediatrics, Chang Gung Memorial Hospital, Linkou, Taoyuan 33305, Taiwan
⁷ School of Traditional Chinese Medicine, College of Medicine, Chang Gung University, Taoyuan 33302, Taiwan
⁸ Department of Medical Biotechnology and Laboratory Science, Chang Gung University, Taoyuan, Taiwan
⁹ Department of Medicine, College of Medicine, Chang Gung University, Taoyuan, Taiwan
¹⁰ School of Life and Health Sciences, The Chinese University of Hong Kong, Shenzhen 518172, China

Correspondence: leetzongyi@cuhk.edu.cn; 29 janglu45@gmail.com and hsujianfu@cgmh.org.tw

Group B streptococcus (GBS) is an important pathogen causing invasive infection including sepsis or meningitis. GBS serotyping is an essential means for the investigation of possible infection outbreak and can identify possible source of infectious origin. Although it is possible to determine GBS serotypes by either immuno-serotyping or geno-serotyping, both traditional methods are time-consuming and labor-intensive. In recent years, the matrix-assisted laser desorption ionization-time of flight mass spectrometer (MALDI-TOF MS) has been reported to be an effective tool to determine GBS serotypes in a more rapid and accurate manner. Thus, this work aims to incorporate the machine learning techniques with MALDI-TOF MS to carry out the identification of GBS serotypes.

In this study, totally 787 GBS isolates, obtained from three research and teaching hospitals, were analyzed by MALDI-TOF MS, and the serotype of the GBS was determined by geno-serotyping experiment. The peaks of mass-to-charge ratios were regarded as the attributes to characterize the various serotypes of GBS. Machine learning algorithms, such as support vector machine (SVM) and random forest (RF), were then used to construct predictive models for the five different serotypes (Types Ia, Ib, III, V, and VI). After the optimizations on feature selection and model generation based on training datasets, the accuracies of the selected models attained 54.9%-87.1% for various serotypes based on independent testing data. Specifically, for the major serotypes, namely type III and type VI, the accuracies were 73.9% and 70.4%, respectively.

The proposed models have been adopted to implement a web-based tool (GBSTyper), which is now freely accessible at http://140.138.77.239/~Joy105/gbs_web/, for providing an efficient and effective detection of GBS serotypes based on MALDI-TOF MS spectrum. Overall, this work has demonstrated the combination of MALDI-TOF MS and machine intelligence could provide a practical means of clinical pathogen testing.
MetaVelvet-DL: a MetaVelvet DEEP LEARNING EXTENSION FOR DE NOVO METAGENOMICS ASSEMBLY

Kuo-ching Liang¹ and Yasubumi Sakakibara¹,*

¹Department of Biosciences and Informatics, Keio University, 3-14-1 Hiyoshi, Kohoku-ku Yokohama, 223- 8522, Japan

*Correspondence: kuoching@dna.bio.keio.ac.jp

The increasing use of whole genome sequencing technology in the sequencing of metagenomic samples has spurred a need for improvement in de novo assemblers to facilitate the discover of unknown species and their genomic functions in metagenomic samples. MetaVelvet is a metagenome assembly algorithm proposed as an extension to the de Bruijn graph-based de novo assembler Velvet for the assembly of whole genome short sequence read data.

In this work, we improve upon the performance of the metagenome assembler MetaVelvet by using a deep learning-based model, and present that the recent advancements in machine learning due to deep learning approaches offer the opportunity to better exploit sequence information to better differentiate between the sequenced genomes of different species in a metagenomic sample of a mixture of bacteria species. Assembly of the Critical Assessment of Metagenome Interpretation (CAMI) dataset showed that our deep learning-based metagenome assembler MetaVelvet-DL produced longer contigs with less chimeric assemblies compared to those of MetaVelvet and state-of-the-art algorithms.

The deep-learning-based extension for MetaVelvet has been shown to provide more accurate de novo assemblies of whole genome metagenomic data, with lower chimeric assembly rates and longer correctly assembled single species contigs. The authors believe that this improvement can help in the further the understanding of microbiome through metagenomics by providing a more accurate description of metagenomic samples under analysis.
IDENTIFICATION OF HIGHLY CONSERVED, SEROTYPE-SPECIFIC DENVIRUS SEQUENCES: IMPLICATIONS FOR VACCINE DESIGN

Li Chuin Chong¹ and Asif M. Khan¹, *

¹ Centre for Bioinformatics, School of Data Sciences, Perdana University, Serdang, Selangor 43400, Malaysia

*Correspondence: asif@perdanauniversity.edu.my

The sequence diversity of dengue virus (DENV) is one of the challenges in developing an effective vaccine against the virus. Highly conserved, serotype-specific (HCSS), immune-relevant DENV sequences are attractive candidates for vaccine design, and represent an alternative to the approach of selecting pan-DENV conserved sequences. The former aims to limit the number of possible cross-reactive epitope variants in the population, while the latter aims to limit the cross-reactivity between the serotypes to favour a serotype-specific response. Herein, we performed a large-scale systematic study to map and characterise HCSS sequences in the DENV proteome.

All reported DENV protein sequence data for each serotype was retrieved from the NCBI Entrez Protein (nr) Database (txid: 12637). The downloaded sequences were then separated according to the individual serotype proteins by use of BLASTp search, and subsequently removed for duplicates and co-aligned across the serotypes. Shannon’s entropy and mutual information (MI) analyses, by use of AVANA, were performed to measure the diversity within and between the serotype proteins to identify HCSS nonamers. The sequences were evaluated for the presence of promiscuous T-cell epitopes by use of NetCTLpan 1.1 and NetMHCIIpan 3.2 server for human leukocyte antigen (HLA) class I and class II supertypes, respectively. The predicted epitopes were matched to reported epitopes in the Immune Epitope Database.

A total of 2,321 nonamers met the HCSS selection criteria of entropy < 0.25 and MI > 0.8. Concatenating these resulted in a total of 337 HCSS sequences. DENV4 had the most number of HCSS nonamers; NS5, NS3 and E proteins had among the highest, with none in the C and only one in prM. The HCSS sequences were immune-relevant; 87 HCSS sequences were both reported T-cell epitopes/ligands in human and predicted epitopes, supporting the accuracy of the predictions. A number of the HCSS clustered as immunological hotspots and exhibited putative promiscuity beyond a single HLA supertype. The HCSS sequences represented, on average, ~40% of the proteome length for each serotype; more than double of pan-DENV sequences (conserved across the four serotypes), and thus offer a larger choice of sequences for vaccine target selection. HCSS sequences of a given serotype showed significant amino acid difference to all the variants of the other serotypes, supporting the notion of serotype-specificity.

This work provides a catalogue of HCSS sequences in the DENV proteome, as candidates for vaccine target selection. The methodology described herein provides a framework for similar application to other pathogens.
Predicting protein-protein interactions by using the sequences of amino acids has two significant problems. The first is representing each sequence as a feature vector and the second is designing a model that can identify the protein interactions. Thus, effective feature extraction methods can lead to improved model performance. Methods: In this study, we used two types of feature extraction methods—global encoding and pseudo-substitution matrix representation (PseudoSMR)—for representing the sequences of amino acids in human proteins and Human Immunodeficiency Virus type 1 (HIV-1) to address the classification problem of predicting protein-protein interactions. We also compared principal component analysis (PCA) with independent principal component analysis (IPCA) as methods for transforming Rotation Forest.

Feature extraction using global encoding with Rotation Forest (PCA) obtained accuracy, sensitivity, specificity, and precision of 79.77, 79.91, 79.07, and 79.77%, respectively. Rotation Forest (IPCA) obtained accuracy, sensitivity, specificity, and precision of 77.20, 76.65, 77.8, and 79.40%, respectively. The results were slightly more mixed for PseudoSMR, with Rotation Forest (PCA) achieving an accuracy, sensitivity, specificity, and precision of 80.23, 81.25, 79.35, and 79.28%, respectively. Rotation Forest (IPCA) achieved 77.83, 76.87, 79.92, and 79.26%, respectively, for the same measures.

Both global encoding and PseudoSMR can represent the sequences of amino acids quite well. Rotation Forest (PCA) performed better than Rotation Forest (IPCA) in terms of predicting protein-protein interactions between HIV-1 and human protein.
Toll-like receptor 9 is a key innate immune receptor involved in detecting infectious diseases and cancer. TLR9 activates the innate immune system via recognition of foreign single stranded DNA containing unmethylated cytosine-guanine (CpG) motifs. Due to the considerable number of rotatable bonds in DNA, high-throughput in silico screening of CpG oligodeoxynucleotides (ODNs) is challenging via traditional structure-based virtual screening approaches. In the current study, we present a machine learning based method for predicting novel mouse TLR9 (mTLR9) agonists based on features including count and position of motifs, the distance between the motifs and graphically derived feature such as the radius of gyration and moment of Inertia. We employed an in-house experimentally validated dataset of 396 single-stranded synthetic CpG-ODNs, to compare the results of five machine learning algorithms. Since the dataset was highly imbalanced, we used an ensemble learning approach based on repeated random down-sampling.

Using in-house experimental TLR9 activity data we found that random forest algorithm outperformed other algorithms for our dataset. Therefore, we developed a cross validated ensemble classifier of 20 random forest models. The average Matthews Correlation Coefficient and balanced accuracy of our ensemble classifier in test samples was 0.61 and 80.0%, respectively with the maximum balanced accuracy and Matthews correlation coefficient of 87.0% and 0.75, respectively. We confirmed common sequence motifs including ‘CC’, ‘GG’, ‘AG’, ‘CCCG’ and ‘CGGC’ were overrepresented in mTLR9 agonists. Overall, we showed that the random forest algorithm outperformed other machine learning algorithms including support vector machines, shrinkage discriminant analysis, gradient boosting machine and neural networks. Predictions on randomly generated ODNs were ranked and the top 100 ODNs were experimentally tested for TLR9 activity.

Due to its predictive performance and simplicity, the random forest technique is a useful method for prediction of mTLR9 ODN agonists. We combined repeated random down-sampling with random forest to overcome the class imbalance problem and achieved promising results. Experimental validation of the top 100 predictions confirmed that 91 ODNs showed high activity.
PHYLOGENETIC ANALYSIS OF TYPE IX SECRETION SYSTEM (T9SS) PROTEIN COMPONENTS REVEALED THAT PORR UNDERGOES HORIZONTAL GENE TRANSFER

Reeki Emrizal, Nor Azlan Nor Mohammad

Institute of Systems Biology, Universiti Kebangsaan Malaysia

Author Correspondence: norazlanm@ukm.edu.my

Porphyromonas gingivalis is one of the major bacteria that causes periodontitis. Chronic periodontitis is a severe form of periodontal disease that occurs due to prolong inflammatory conditions. If left untreated, deterioration of the supporting structures such as gingiva, bone, and ligament can ultimately lead to tooth loss. Virulence factors produce by P. gingivalis that are responsible for the pathophysiology of periodontitis are secreted by Type IX Secretion System (T9SS). T9SS-acquiring bacteria have been linked to several systemic diseases such as atherosclerosis, aspiration pneumonia, cancer, rheumatoid arthritis, and diabetes mellitus. This study aims to investigate the phylogenetic relationship and taxonomic distribution between putative members of T9SS component protein families. There are 20 protein components of T9SS being investigated in this study. We have constructed multiple sequence alignments for each component using homologs of those components. Then we proceed to phylogenetic analysis by constructing the maximum-likelihood (ML) trees. ML trees for 19 protein components of T9SS exhibit clustering of terminal nodes based on their respective classes under Bacteroidetes phylum. The ML tree of PorR, which is an aminotransferase that involved in Wbp pathway that produces structural sugar of A-LPS, exhibits different clustering pattern of terminal nodes where the nodes do not cluster based on their respective classes. Hence, PorR might evolve independently from the other T9SS protein components which might suggest that PorR is acquired by T9SS-acquiring bacteria through horizontal gene transfer. The part of P. gingivalis strain ATCC 33277 genome that contains porR gene has been extracted to support the possibility that porR gene has been horizontally transferred. Through homology searching using NCBI blastx, we found that seven genes (including porR) that involved in the biosynthesis of A-LPS that anchored the virulence factor secreted by T9SS to bacterial cell surface are flanked by insertion sequences (ISs) that encode IS5 family transposase. The IS5 transposons contain a single open reading frame that encodes for the transposase that will cleave the 12 bp inverted repeats that flanked the transposons. Consequently, this can mobilise the intervening DNA segment that contains porR gene and subsequently contributes to the possibility that porR gene is subjected to conjugative transfer. The taxonomic distribution of T9SS protein components revealed that they can be found across all classes under Bacteroidetes phylum. Additionally, we have identified species under Chitinophagia, Saprospiria, and unclassified that acquired homologs of T9SS protein components that, to our knowledge, have not been reported. In conclusion, this study can provide a better understanding about the phylogeny and taxonomic distribution of T9SS protein components.
Euphorbiaceae is one of the largest families of flowering plants. For the exceptional diversity of growth forms and nearcosmopolitan distribution, it has attracted human interest since ancient times. SBP-box (SBP) genes encode a type of plant-specific transcription factors that play critical roles in numerous biological processes, especially in flower development. We performed a genome-wide identification and characterization of SBP genes from four high economic Euphorbiaceae species.

In total, 77 SBP genes were identified in four Euphorbiaceae genomes. All these SBP proteins are belong to 3 length ranges and 10 groups. Group-6 is absent in Arabidopsis thaliana, but it is conserved existing in Euphorbiaceae. Segmental duplication played the most important roles in the expansion processes of Euphorbiaceae SBP genes, and all these duplicated genes were subjected to positive selection. In addition, about two thirds of Euphorbiaceae SBP genes are the potential targets of miR156, and some miR-regulated SBP genes show high intensity and tissue contrasting expression profiles. The expression profiles of different stress treatments demonstrated broad involvement of Euphorbiaceae SBP genes in response to various abiotic and hormonal treatments and functional divergence.

In this research, 77 SBP genes were identified in four Euphorbiaceae species, and their phylogenetic relationships, protein physiochemical characteristics, duplication, tissue and stress response expression and potential roles in Euphorbiaceae development were studied. This study lays foundation of further studies of Euphorbiaceae SBP genes, which provide valuable information for future functional exploration of Euphorbiaceae SBP genes.
Accurate classification of diffuse gliomas, the most common tumors of the central nervous system in adults, is important for appropriate treatment. However, detection of isocitrate dehydrogenase (IDH) mutation and chromosome 1p/19q codeletion, biomarkers to classify gliomas, is time- and cost-intensive and diagnostic discordance remains an issue. RNA editing has emerged as a novel cancer diagnostic and prognostic marker, but the diagnostic and prognostic value for gliomas remains largely unexamined.

We aim to (1) unravel the relationship between RNA editing and IDH mutation and 1p/19q codeletion and to (2) predict presence/absence of IDH mutation and 1p/19q codeletion using machine learning algorithms. By characterizing genome-wide adenosine to inosine RNA editing signatures of 638 gliomas, we found that tumors with wildtype IDH, on average, had higher total editing level compared with those carrying IDH mutation (Kolmogorov-Smirnov test, p<0.0001). When tumor grade was considered, however, only grade I-V tumors with wildtype IDH exhibited higher total editing level. According to 10-fold cross-validation, support vector machines (SVM) outperformed random forest and AdaBoost (DeLong test, p <0.05). The area under the receiver operating characteristic curve (AUC) of SVM in predicting IDH mutation and 1p/19q codeletion were 0.989 and 0.990, respectively. After performing feature selection, AUCs of SVM and AdaBoost in predicting IDH mutation were higher than that of random forest (0.985 and 0.983 vs. 0.977; DeLong test, p <0.05), but AUCs of the three algorithms in predicting 1p/19q codeletion were similar (0.976-0.982). Furthermore, 67% of the six continuously misclassified samples by our 1p/19q codeletion prediction models were misclassifications in the original labelling after inspection of 1p/19q status and/or pathology report, demonstrating the accuracy and clinical utility of our models.

This study represents the first genome-wide RNA editing analysis to date of adult diffuse glioma and identifies RNA editing as a novel diagnostic and prognostic biomarker for gliomas. Our prediction model provides standardized, accurate, reproducible and objective classification of gliomas. Our model is not only useful in clinical decision-making, but also able to identify editing events that have the potential to serve as biomarkers and therapeutic targets in glioma management and treatment.
Biology has entered the era of big data with the advent of high-throughput omics technologies. Biological databases provide public access to petabytes of data and information facilitating knowledge discovery. Over the years, sequence data of pathogens has seen a large increase in number of records given the relatively small genome size and their important role as infectious and symbiotic agents. Humans are host to numerous pathogenic diseases such as that by viruses, many of which are responsible for high mortality and morbidity. The interaction between pathogens and humans over the evolutionary history has resulted in sharing of sequences, with important biological and evolutionary implications. This study describes a large-scale, systematic bioinformatics approach for identification and characterization of shared sequences between the host and pathogen. Mapping of the host-pathogen share-ome has important implications for the development of vaccines, disease surveillance and the discovery of new pathogens. The generic workflow is applicable to a variety of pathogens, of viral, bacterial or parasitic origin. Herein, an application of the workflow is demonstrated to map the Flaviviridae-human share-ome.
Novocraft Technologies

NOVOALIGN (V4) RUN TIME REDUCTION AND PERFORMANCE BENCHMARKING USING FREEBAYES VARIANT CALLER

Colin Hercus*, Akzam Saidin1

1Novocraft Technologies Sdn. Bhd., Petaling Jaya, Selangor, Malaysia

Correspondence: akzam@novocraft.com; colin@novocraft.com

NovoAlign is an NGS sequence alignment software for accurate mapping of short reads onto a reference genome. We have recently updated the software from the previous version (V3.09.02) to version 4 (V4) which includes runtime reduction, revision of default options for alignment scoring, removal of obsolete functions and changes in command line options. To assess the changes in run time, we ran a comparison between V4 versus V3.09.02 and BWA-MEM on several publically available NA12878 sequence datasets; (i) NextSeq 500 v2: Nextera Rapid Capture Exomes, (ii) WGS HiSeqX (~17X) and (iii) WGS MGISEQ2000 (~35X). For dataset (i), we observed that V4 has significantly reduced the runtime required to 14 minutes from 605 minutes (V3.09.02), bwa-mem (0.7.12-r1039) runtime was at 6 minutes. The result has also shown a reduction of 13% in false positive variant calls in V4-freebayes compared to bwa mem-freebayes aligner-variant caller combination. In WGS datasets (ii) and (iii), we noted that V4 has a runtime of 2.8 and 6.7 hours, for each respective datasets. Improvements were also observed in variant calling with freebayes, where an increase of 1.5-3.5% true positive indel recalls were recorded, with approximately 1-2% increase in precision for both SNP and indel as compared to bwa-mem.
Empowering Research Through Smart Solutions

**Proprietary Software**

- **novoAlign**
  - DNA sequence alignment software designed to map short reads onto a reference genome.

- **novoSort**
  - Multithreaded merge and sort tool for BAM files.

- **novoLR Package**
  - A complete toolkit for hybrid genome assembly with mixed long and short reads.

**Data Analysis Services**

- **Types of Services:**
  1. Result Assessment & Quality Control
  2. NGS Data Analysis
     - Whole Genome, RNAseq, Exome, Amplicons, etc.
  3. NGS de novo Assembly
     - (from data pre-processing, assembly to annotation)
  4. Custom Analysis

**Analysis Platforms**

- Powered by in-house algorithms, these platforms are designed for biological scientists, clinical and diagnostics researchers without background in bioinformatics, and/or those who do not wish to run complex bioinformatics algorithm for sequence analysis.

- **novoWorX**
  - NGS genome data management and analytics platform.

- **novoClinic**
  - A clinical bioinformatics solution in a single platform.

Over 200 licensees globally...

**Professional Bioinformatics Specialists**

Software Solutions | Analysis Services | Consultation

**Novocraft Technologies Sdn Bhd**
C-23A-05, 3 Two Square, Section 19, 46300 Petaling Jaya, Selangor, Malaysia

office@novocraft.com  +603 7960 0541
http://www.novocraft.com  +603 7960 0540
The landscape of genomics is trending towards high-intensity sequencing, increasingly driven by large initiatives at a national scale. As scientists work with ever-growing datasets, there is a need for bioinformatics analysis tools and solutions that drive towards speed and scalability, delivering high quality results while efficiently using resources and minimizing costs. To this end, Illumina has developed a number of solutions to allow rapid and accurate analysis to enable simplified and reliable scalability across a range of applications. Join us and learn more about our latest developments for optimized analysis tools and achieving success at scale.
Transforming the future of genomics, together
Your partner in providing best-in-class solutions and services

At Illumina we strive to provide innovative technologies to unlock the power of the genome. To discover how illumina can help reveal new connections with next-generation sequencing (NGS), array and informatic solutions visit the Illumina & PT Pandu Biosains booth at Incob 2019.
Biography
In 2009, he moved to Saudi Arabia to establish a Biochemistry Department and Laboratory at College of Medicine, University of Hail, Saudi Arabia (2009-2012). He was the leading scientist in the establishment of the biggest Indonesian Medical Education and Research Institute at Faculty of Medicine, University of Indonesia (2011-2012). Before he joined Swiss German University in 2015, he held various positions at different institutions. His notable findings include but not limited to elucidation of protein-protein interactions and flavin transfer mechanism in a two-component enzyme system in alkanesulfonate monooxygenase in E. coli. He also characterized of a novel protein/enzyme involved in flagella assembly and activity. His current research interests are in the area of antibiotics resistant and drug discovery from Indonesian biodiversity, particularly for antimicrobial and anticancer agents. His works had been published in highly reputable journals, proceedings and book chapter. He initiated the establishment of Indonesia Natural Products Extracts Library (INPEL).

In addition to academics’ activity, Dr. Audah is heavily engaged in various organizations namely Indonesian Society for Bioinformatics and Biodiversity (Founding President) (2016-present), Indonesian Society for Biotechnology Study Program (Founding Member), and Indonesian Society for Biochemistry and Molecular Biology (Executive Member) (2017-present). As for community services engagement, Dr. Audah is involved in An Annual National Conference on Community Services and Social Corporate Responsibility since 2015 with various roles and as the Chairman (2019). He is the chairman of a non-profit organization that run a school for underprivileged students (about 400 students) (2015-present).

Abstract

A National Center for Bioinformatics is Required for Optimizing Biodiversity-Based Research in Indonesia

Biodiversity-based research generates big amount of data acquired (high throughput) at different organization level from variety of organisms. The more organisms studied, the more data collected and need to be stored, analysed and interpreted to become more meaningful biological information. Therefore, application of bioinformatics has become necessity to fulfil those needs. In the context of Indonesia, the need of bioinformatics in various aspects even become more important due to the fact that Indonesia is considered as one of countries with mega-biodiversity.
As the awareness of the importance of biodiversity-based research increases, capacity building as well as better coordination, direction and organization of such activities are urgently required. The latter efforts can be done by establishing a national center for bioinformatics in the country. In order to realize this agenda, collaboration, cooperation and communication among related government, private institutions and communities are needed.

Indonesian Society for Bioinformatics and Biodiversity (ISBB) which is also known as Masyarakat Bioinformatika dan Biodiversitas Indonesia (MABBI) together with other organizations play an important role in catalysing the establishment of the Indonesia National Center for Bioinformatics. Its role includes but not limited to connecting different institutions both government and privates to discuss together this novel cause. This International Conference of Bioinformatics (InCob 2019) held by YARSI University and its organizing partners such as APBioNet, Goblet and MABBI itself is a golden opportunity to move the idea forward.
The term of endophytic microbes has been become familiar for more than a decade. The term is used for groups of microbes which live inside of plants tissues whether bacteria or fungus. There is a huge number of plants widely used as medicinal plants to cure diseases, for instance infectious diseases, cancer, and metabolic syndromes. In this research we aimed to get endophytic fungi isolates from fruits of *Averrhoa bilimbi* Linn. and have antidiabetic activity through an alpha- amylase inhibition assay. The research conducted by several steps, endophytic fungi isolation, macroscopic identification, isolates screening for antidiabetic activity by alpha-amylase inhibition assay, and molecular identification of potential endophytic fungus, respectively. There were 12 isolates and among those isolates, BW-10 isolate had higher activity compare to positive control. The most potential endophyte identified as *Trametes elegans* by molecular identification. The scaling-up fermentation, phytochemistry identification and in vitro modelling of antidiabetics activity will be the further experiments in the next project.
T cells are the primary effector cells of tumor immunity. T cells recognized peptides derived from tumor antigens which is presented as a complex with HLA (histocompatibility leukocyte antigen - HLA) on the surface of target cells. HLA-A*24:07 is the major HLA alleles in Indonesia. HLA A*24:07 is considered unique and not a member of the HLA A*24 supertype, which is represented by HLA A*24:02. The two molecules, A*24:02 and A*24:07, share high homology except for one amino acid residue number 70. Residue number 70 in A*24:02 is H and in A*24:07 is Q. This residue happens to be in the peptide binding groove of the HLA allele, and affecting at least 3 out of 6 peptide binding pockets. HLA-A*24:07 is a less characterized allele, and at the moment, there is only one T-cell epitopes reported to be restricted by this HLA allele (IEDB website). The paper will present the peptide binding motif of HLA A*24:07, the benchmark of online T cell epitope prediction server, and the use of prediction server to predict peptide derived from cancer antigens that will bind to HLA A*24:07.
**MYCOBACTERIUM TUBERCULOSIS BEIJING AND EAI LINEAGE MUTATION PATTERNS OF TB PREVALENCE SURVEY SPECIMENS (2013)**

Kindi Adam\(^1\), Yuni Rukminiati\(^1\), Fauzul Muna\(^1\), Francisca Srioetami Tanoerahardjo\(^1\)* \(^1\)National Institute of Health Research and Development, Ministry of Health Republic of Indonesia

Correspondence: kindiad@yahoo.co.id; ftanoerahardjo@gmail.com

*M.tuberculosis* strain lineage Beijing and EAI (East-African India) are dominant strains that circulate in Indonesia. Beijing Lineage can replicate faster than laboratory strains, thus the tendency for mutations in Beijing lineage to be higher. EAI lineage in Indonesia because it is more adaptive to the host. The purpose of this study was to determine mutation patterns associated with anti-tuberculosis drug resistance by genotype in the TB Lineage Beijing and EAI strains that circulate in Indonesia. Four Beijing TB strain specimens (sample code Beijing-1, Beijing-2, Beijing 3 and Beijing-4) and 3 EAI strains (code EAI-1, EAI-2, EAI-3) were sequenced using Illumina Miseq. Sequence data quality was analyzed by FastQC and Qualimap and cleaned with trimmomatic. Mapping analysis is carried out with BWA. Analysis of variance was performed using SAMtools and GATK 2.5. based on analysis of total genome sequences conducted, Beijing-1 strain has a mutation of 2079 (whereas, drug-resistant mutations in the Gyr A, rpsL, rrs, rpsA, katG) genes, Beijing-2 totaling 2289 (Gyr A, rpsL, rrs, rrl), Beijing-3 was 2112 (GyrA, rpsL, rrs, eis) and Beijing-4 were 2392 (GyrA, rpsL, rrs, eis). Whereas strain EAI-1 had mutations of 2392 (GyrA, mshA, rpsL, rrs, ahpC, thyA, embA), EAI-2 as many as 3114 (GyrA, mshA, rpsL, rrs, ahpC, embA) and EAI-3 as many as 2879 (Ea-2) GyrA, rpsL, rrs, ahpC, embA). EAI mutation strain (on average 2955) larger than Beijing strain (on average 2218) and mutations positions related to drug resistance were also more numerous (on average 6 compared to 4.25 respectively).
Jamu is one of herbal medicine from Indonesia. Jamu is formulated from herbs that are known to have efficacy in curing diseases. Indonesian Jamu Herbs (IJAH) webserver is a useful tool in the development of Jamu. IJAH uses compounds, proteins, and compound-protein interactions data which is obtained from open access databases. The data from internet were not enough, so IJAH needs a prediction model to complete the compound-protein interactions. This study focused on prediction of Jamu formulas using Bipartite Local Model Network Interaction-Profile Inferring (BLM-NII) and analyzing the synergistic effect of compounds in Jamu formula using Network Target-Based Identification of Multicomponent Synergy (NIMS). The incomplete interactions between compound and protein were predicted by BLM-NII, before predicting the Jamu formulas using Support Vector Machine. In conclusion, the area under precision-recall (AUPR) curve of BLM-NII model is 0.70. Moreover, a module to evaluate the synergistic effect between compounds has been implemented in IJAH webserver with the best combination of compound for Diabetes Mellitus type 2 is reached by Glipizide and Mitiglinide with 0.45 score synergy.
PRELIMINARY STUDY OF MYCOBACTERIUM TUBERCULOSIS VIRULENCE FACTORS AS BIOMARKER CANDIDATE FOR DIAGNOSTIC TOOLS

Nur Imaniati Sumantri*

Faculty of Engineering, Universitas Indonesia, Depok, Indonesia

Correspondence: nur.imaniati@gmail.com

Tuberculosis (TB) is an infectious disease caused by Mycobacterium tuberculosis which has capability in deceiving human immune response and is multi-drug resistance pathogen. Accurate and safe detection methods is required to be developed to minimize contact and transmission of the pathogen to health workers. Finding of several novel virulence factors of MTB bring through antigenic candidates for diagnostic tools development. This study aim to analyze structural attributes of Rv0310c and Rv1255c as potential marker. Sequences were obtained from Mycobrowser (http://mycobrowser.epfl.ch), 3D protein modelling was performed by SWISS-MODEL (https://swissmodel.expasy.org), and protein-protein interaction was analyzed by STRING (https://string-db.org). The structural elucidation showed Ramachandran favoured of Rv0310c and Rv1255c were 91.28 % and 93.58%, respectively. Protein interaction revealed those virulence factors weave specific and meaningful functional association. This study exposed structural and functional regions in disease-causing effector of TB that enables them to be considered as biomarker in safe and simple diagnostic tools.
ASSOCIATION OF SINGLE NUCLEOTIDE POLYMORPHISM AND PHENOTYPES IN TYPE 2 DIABETES MELLITUS USING STEPWISE REGRESSION

Devy Apriansyah¹, Wisnu Ananta Kusuma²,*

¹Department of Computer Science, IPB University, Bogor, Indonesia
²Tropical Biopharmaca Research Center, IPB University, Bogor, Indonesia

Correspondence: devyapriansyah@gmail.com; ananta@apps.ipb.ac.id

Precision medicine is the application of medical science to profile groups of individuals based on the unique biological characteristics of each individual. Currently precision medicine focuses on complex diseases such as type 2 diabetes mellitus. The study of associations in determining unique biological characteristics is one that supports precision medicine. In this study, the biological characteristics of type 2 diabetes mellitus were obtained from DNA markers and phenotypes of rat. The DNA marker used is single nucleotide polymorphisms (SNPs). SNPs data were taken from Mouse Phenome Database based on 22 protein candidates regarding type 2 diabetes mellitus. The pre-processing stage was required by filling the missing value and coding in SNPs data. SNPs selection as a feature was conducted by the Stepwise Selection method to select SNPs that are statistically significant. The phenotype data used is insulin tolerance. The association of SNPs and insulin tolerance phenotype was conducted using multiple linear regression method. The association results obtained 8 SNPs that were significantly associated with insulin tolerance phenotype based on t-test analysis. The resulting regression model is acceptable because it has passed the multicollinearity, heterogeneity, and autocorrelation test.
A SNPs (single nucleotide polymorphisms) identification and validation process which is conducted using wet lab is costly and time consuming. Thus, computer programs can be used to generate more efficient process. SNPs information could be obtained through some preliminary stages that is necessary to be processed. Each stage would be processed by a computer programs or tools. In this research, some stages and tools were used to identify SNPs including aligning sequence using Bowtie 2, preprocessing the result of sequence alignment using GATK, identifying and filtering SNPs using SAMtools/BCFtools, and predicting SNPs effect using SnpEff. A pipeline was developed to manage the order of execution in indentifying SNPs, since there are many stages to be processed. Using pipeline, the output of each stage could be forwarded into the next stage as an input. This research also used the SNPs information to construct a phylogenetic tree. This pipeline, that has been developed, successfully identified SNPs from reads sequences input, either paired or unpaired, and could show the detailed information of each potential SNPs. The phylogenetic tree was also successfully constructed from four soybean varieties using UPGMA and Neigbor-Joining method using VCF-kit.
IMBALANCED DATA OPTIMIZATION FOR DRUG-TARGET INTERACTION WITH DIFFERENT CONTRIBUTION SAMPLING BASED ON SVM ENSEMBLE

Nabila Sekar Ramadhanti¹, Wisnu Ananta Kusuma¹²⁺, Annisa¹

¹Department of Computer Science, Bogor Agricultural University, Bogor, West Java, Indonesia
²Tropical Biopharmaca Research Center, Bogor Agricultural University, Bogor, West Java, Indonesia

Correspondence: nabila_skr@apps.ipb.ac.id; ananta@apps.ipb.ac.id

Imbalanced data had been one of the problems that arise in processing data. This research will focus on handling imbalanced data problem for drug-target interaction data. There is many targets or protein and even much more drug compound existed in many databases, but many of their interactions are unknown yet. This unknown interaction is what led drug target interaction to imbalanced data. The unknown interaction has already been worked on as drug testing has been conducted even long before the data science era. But conducting drug testing requires a long duration and even bigger cost. This drug-target interaction data may be used for further analysis such as drug prediction that may help ease the selection process for drug testing. There are many ways of handling imbalanced data, the naïve way like sampling may also work, there is even a deep learning solution to solve the imbalanced data, but this research will focus on sampling problem, combining some methods such as BSVM, SMOTE, and RUS with SVM ensemble which written as Different Contribution Sampling (DCS) method. This method had been used before for some imbalanced data problem, but not on the drug-target interaction data. As this method proved superior to some other sampling method, this research will use the basics of DCS to improve the imbalanced drug-target interaction data.
JENSEN SHANNON DIVERGENCE AND QUANTUM RELATIVE ENTROPY FOR LOCUS SELECTION IN STR-DNA BASED TRIBAL INFERENCE SYSTEM

Maria Susan Anggreainy\textsuperscript{1}, M. Rahmat Widyanto\textsuperscript{2}, Belawati H. Widjaja\textsuperscript{2}, Nurtami Soedarsono\textsuperscript{3}

\textsuperscript{1}Informatics Management, Vocational School, IPB University, Bogor, Indonesia
\textsuperscript{2}Faculty of Computer Science, Universitas Indonesia, Depok, Indonesia
\textsuperscript{3}Faculty of Dentristy, Universitas Indonesia, Depok, Indonesia

Correspondence: mariasusananggreainy@apps.ipb.ac.id

The selection of a locus that is significant in distinguishing the uniqueness between tribes is expected to improve the accuracy of the tribal inference system that has been developed. The proposed research was developed using fuzzy logic by utilizing the probability distribution of allel markers as fuzzy membership functions. The combination of uniqueness between tribes is analyzed with Quantum Relative Entropy (QRE) and illustrated in two distribution models namely independent allel models and paired allel models. Paired allel models show better results with the percentage of recall precessions and accuracy above 99%. Statistical tests with McNemar show that Quantum Relative Entropy (QRE) is better than Jensen-Shanon Divergence (JSD) in terms of choosing the right combination of loci in a tribal inference system.
GENOMIC DIVERSITY OF ESKAPE-INFECTING BACTERIOPHAGE AND LYSIN PROTEIN

Marta Nisita Dewanggana¹, Adi Yulandi¹, Yogiara²*

¹Department of Biology, Faculty of Biotechnology Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia
²Master in Biotechnology Program, Faculty of Biotechnology Atma Jaya Catholic University of Indonesia, Jakarta, Indonesia

Correspondence: martanisita@gmail.com; yogiara@atmajaya.ac.id

A group of pathogenic bacteria such as Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, and Enterobacter as known as ESKAPE bacteria is antibiotic-resistant pathogens and the primary cause of nosocomial infections. Bacteriophage therapy is used as an alternative therapy to antibiotics for multi-drug resistant (MDR) bacteria, including the ESKAPE group. A bacteriophage produces a lysin enzyme during lytic phase to destroy the peptidoglycan structure that is present in the cell wall. However, the information regarding diversity and phylogenetic relationship among bacteriophage is still not explored yet. So, This study aimed to obtain phylogenetic relationship information between ESKAPE bacteria-infected bacteriophage and its lysin protein. MEGA X and MEGA CC were used to construct a phylogenetic tree. BRIGG and MAUVE were used to compare particular bacteriophage genomes. Modeller and GROMACS were used to build protein model and molecular dynamics analysis, respectively. According to phylogenetic and comparative genomic analysis, bacteriophage and its lysin were tend to group based on their host. Six lysin protein sequences were used to model lysin protein, but only two protein sequences were appropriate for further study. As shown by RMSD graphs, lysin was reasonably stable on the temperature range (300, 310, and 313 K) with different optimal temperatures for each model. RMSF graph analysis followed by ESPript3 analysis showed predicted point mutations on all three protein sequences. This study demonstrated that both bacteriophages infection and lysin activity are host specific.
Drug repositioning is the reuse of existing drugs to treat a new indication other than the initial one. Research on drug repositioning can be done by observing the interaction of drug compounds with disease proteins that react positively. One of the challenges in predicting the interaction of compounds and proteins is imbalanced data. Deep semi-supervised learning has proven to be effective in handling prediction models with imbalanced data. The unsupervised based pre-training in deep semi-supervised learning can represent input from unlabeled data (majority data) properly and optimize weights initialization in the classification model. Pre-training can act as an optimizer and regularizer to improve the performance of the classification model with imbalanced data. This study implements the Deep Belief Network (DBN) as a pre-training with Deep Neural Network (DNN) as a classifier thus compares it with Stacked Auto Encoder (SAE) pre-training. The data used in this study are ion channel, GPCR, and nuclear reactor dataset sourced from KEGG BRITE, BRENDA, SuperTarget, and DrugBank databases. The results of this study indicate that on the dataset, pre-training using DBN for feature extraction has a better optimization effect than SAE. This can be seen from DNN performance improvement in accuracy (3-4.5%), AUC (4.5%), precision (5.9-6%), and F-measure (3.8%).
ADVERSE EFFECT OF TRANSFLUTHRIN ON SEXUAL-RELATED HORMONES THROUGH MOLECULAR DOCKING STUDY

Alfiatun Hasanah¹, Didik H. Utomo¹,*, Djoko Soeatmadji²

¹Research And Education Center For Bioinformatics, INBIO Indonesia, Malang, Indonesia
²Department Of Internal Medicine Laboratory, Faculty Of Medicine, Brawijaya University, Malang, Indonesia

Correspondence: alfiatun.hasanah1@gmail.com; didik.huswo@gmail.com

Transfluthrin is a synthetic pyrethroid used as the main component of mosquito repellent. Its adverse effect is still debatable, but it is known to affect sexual-related hormones after prolonged exposure. The study aimed to evaluate the toxicity of transfluthrin and its interaction with androgen receptor (AR) and estrogen receptor (ESR). Transfluthrin (PubChem ID: 656612) was first analyzed using admetSAR to identify its cellular bioavailability. It then docked using PyRx, an automated virtual screening tool, to find out its molecular interaction with AR (PDB ID: 1E3G) and ESR (PDB ID: 1A52) compared to each control ligands which are metribolone and estradiol, respectively on active site of the receptor. Docking results then visualized using PyMol. admetSAR analysis reported that transfluthrin can enter the cell, pointed by its molecular weight (371.16 g/mol), AlogP value (4.88), and H-bond acceptor (2). It was also showed that transfluthrin can bind to ESR and AR with predicted score 0.7305 and 0.6494 respectively. These predicted scores represent the potential ability of transfluthrin to alter these sexual-related hormones, resulted in their distortion. Interestingly, docking results showed a relatively high binding affinity score of transfluthrin-AR (-9.0 Kcal/mol), while transfluthrin-ER complex scored -8.1 Kcal/mol. Additionally, transfluthrin binds specifically to the active sites of each receptor's native ligand, acting as a competitive inhibitor supported by visualization data. Thus, it can be concluded that computational data revealed the hypothesis that transfluthrin has adverse effect on sexual-related hormones through competitive binding.
IJAH ANALYTICS: HERBAL MEDICINE FORMULATION SYSTEM BASED ON NETWORK PHARMACOLOGY AND GRAPH MINING

Wisnu Ananta Kusuma¹, Rudi Heryanto², Farit Mochamad Afendi³

¹Department of Computer Science, IPB University
²Department of Chemistry, IPB University
³Department of Statistics, IPB University

Correspondence: ananta@ipb.ac.id

Herbal medicine is a typical Indonesian herbal medicine which is formulated from plants which are considered to have properties in curing diseases. The development of herbal medicine formulas in Indonesia is very important because herbal medicine has lower side effects compared to conventional medicine. We developed an application to predict the formula for herbal medicine namely IJAH Analytics. IJAH Analytics is a computer program for predicting herbal formulas developed based on network pharmacology and machine learning approach. Ijah initially had the hypothesis that the herbal formula consisted of at least plants that had four pharmacological activities: analgesic, antibacterial, anti-inflammatory, and plants with certain efficacy targets. This system utilizes data of compounds, proteins, and interactions between plant-compound-protein-diseases obtained from public database on the Internet. However the data of interactions between compound and protein are very limited so it is required a mechanism to predict those interactions in order to yield a good formulation of herbal medicine. IJAH Analytics aims to perform prediction model of herbal formula and analyze the synergy effect of the compounds making up herbal formula and obtain new herbal formulations based on pharmacological activities. Bipartite Local Model Network Interaction-Profile Infering (BLM-NII) is used to predict the interactions between compounds and proteins in the herbal medicine formula network. The Network Target-based Identification of Multicomponent Synergy (NIMS) is used to analyze synergistic effects between compounds. From this model, a network was produced which described the relationship between plants, herbal compounds, compounds, proteins and diseases. The main features of IJAH Analytics could be described as follows (i) finding the best combination of plants as a candidate for herbal formulas for certain diseases, (ii) calculating synergy score from a combination of two compound candidates for herbal formulas, (iii) visualizing the relationship between plants, compounds, proteins, and diseases, (iv) providing the detailed information on plants, compounds, proteins, and diseases.
Phosphatidylinositol 3-kinase δ (PI3Kδ) is a validated drug target for the treatment of cancer. The present study aims to search for new inhibitors of PI3Kδ by employing pharmacophore modelling using LigandScout Advanced 4.3 software. The three hydrogen bond acceptors and two hydrophobic features were proposed as a pharmacophore model using LASW1976 structure. The model was then validated using Area Under Curve (AUC) of Receiver Operating Characteristic (ROC) and GH score. It was used to screen new molecules in the ZINC database, which resulted in 599 hits. All 599 hits were then docked into PI3Kδ protein, and five best hits were submitted to 50-ns molecular dynamics simulations. Each hit complexed with PI3Kδ underwent minor conformational changes as indicated by the values of Root Mean Square Deviation (RMSD) and Root Mean Square Fluctuation (RMSF). Furthermore, prediction of the binding free energy using Molecular Mechanics-Poisson Boltzmann Surface Area (MM-PBSA) method showed that five hits, i.e., Lig25/ZINC253496376, Lig682/ZINC98047241, Lig449/ZINC85878047, Lig554/ZINC253389510, and Lig199/ZINC12638303, had lower binding energy compared to LASW1976. This result indicated their potentials as new inhibitors of PI3Kδ.
Software Demo

Software Demo 1

novoWorx: A GENOME DATA MANAGEMENT & ANALYTICS PLATFORM

Kaamesh Ganisen¹, Muhammad Farhan Mohd Tahir¹, Muhammad Sukeri¹, Abdul Malik Ahmad¹, Chong Su Wei¹, Hamidah Ghani¹, Akzam Saidin*¹

¹Novocraft Technologies Sdn. Bhd., Petaling Jaya, Selangor, Malaysia

Correspondence: kaamesh@novocraft.com; akzam@novocraft.com

NovoWorx is a comprehensive multi proprietary platform designed to accelerate research via a fully-automated analysis platform that employs the best practices in running millions of reads to analyze next generation sequencing data. It is a highly customizable, modular system, designed with portability in mind to allow users to process large datasets consisting of multiple samples without having to run complex bioinformatics tools. NovoWorx is primarily built on a PHP and EXTJS front end while the back end uses a Unix operating system with multi language support to operate the tools used for protocols. The whole platform is enclosed in a virtual machine and can be run on any operating system capable of running a virtual machine manager. The software is scalable where users can opt to run it on a single machine, a high performance server or a server cluster. NovoWorx features include project management, protocol management, reference genome management, user management, data repository, statistical analysis integration, and an integrated genome browser among others. The protocols in novoWorx are highly optimized and fully customizable according to users’ analysis requirements, making it a flexible system that is able to run any compatible analysis pipeline with ease. The system is also equipped with standard protocols such as basic alignment and sorting with mark duplicates, whole genome sequence analysis, whole exomes sequence analysis, targeted amplicon sequence analysis and a few others. All the modules for the protocols are individually built in docker containers that allow for high portability and version control.

Availability and Requirements

- Project name: novoWorx¹
- Project home page: http://dev.novocraft.com/demo/NovoWorx/NovoWORX/
- Operating system(s): Linux, Windows, Mac OS X¹
- Programming language: PHP, EXTJS, Python, Perl, Java
ABCIVA: A BACKGROUND CORRECTION R PACKAGE FOR THE ILLUMINA BEAD ARRAYS

Rohmatul Fajriyah¹, Istvan Berkes²

¹Department of Statistics, Universitas Islam Indonesia, Yogyakarta, Indonesia ²Institute of Statistics, Graz University of Technology, Austria ³Department of Statistics, Universitas Islam Indonesia, Yogyakarta, Indonesia

Correspondence: rfajriyah@uii.ac.id; rfajriyah@uii.ac.id

In microarray data, various processes and technologies have been implemented and developed to produce it, where the noise may contribute to the data measurement. In order to compare and analyse data, pre-processing is applied before further analysis. Background correction is the first step in the pre-processing of microarray data analysis and it represents the critical step. Several methods have been proposed mainly for the Affymetrix platform. In this proposed demo, we will demonstrate the use of abciba, an R package for the background correction for Illumina platform, based on Fajriyah [1] and Fajriyah [2] model. The models consider an additive noise, rather than multiplicative and devise a background correction at the level of bead expression. We show how the benchmarking data is used to evaluate the proposed model empirically. We also provide a simulation study to determine the best fit for some microarray data.

Availability and Requirements

* Project name: abciba
* Project home page: -
* Operating system(s): Linux, Windows, Mac OS X * Programming language: R
IDENTIFICATION OF NOVEL AND POTENT MITOGEN-ACTIVATED PROTEIN KINASE KINASE INHIBITORS: A STRUCTURE-BASED DRUG DESIGN APPROACH.

Geet Madhukar, Dr. Naidu Subbarao*
School Of Computational and Integrative Sciences, Jawaharlal Nehru University New Delhi, India
Ph: (+91) 9687016818
Correspondence: geet85_sit@jnu.ac.in , nsrao@mail.jnu.ac.in*

Head and neck squamous cell carcinoma (HNSCC), one of the most common causes of deaths due to cancers in Asian Countries and the sixth most common cause of cancer globally is a heterogeneous group of upper aerodigestive tract malignancies. [1] MEK1 and MEK2, the 'gatekeepers' of ERK1/2 activity participates in the RAS-RAF-MEK-ERK signal transduction cascade. This cascade participates in the regulation of a large variety of processes including cell cycle progression, cell migration, differentiation, metabolism, and proliferation, apoptosis. Activation of MEK1 is brought about by phosphorylation of S218 and S222 in its activation segment. The kinase activity of KSR is also required for MEK activation. [2] The present study aims to find out the potent and novel inhibitors of MEK1. Extensive database screening from NCI, ChEMBL (KinaseSARfari), Chemdivision (Anticancer and Stock) were performed. Computational approaches were employed to carry out a Structure-Based Drug Design study. Screened compounds from NCI, KinaseSARfari and Chemdivision (Anticancer) showed superior results and surpassed the benchmark compounds with huge differences. Top compounds of Chemdivision (Anticancer) (D146-0335) and KinaseSARfari (ChEMBL43781) reported GLIDE Gscore of -11.05 kcal/mol and -14.68 kcal/mol respectively whereas benchmark compounds had maximum GLIDE Gscore of -9.37 kcal/mol. Results were also validated using Xscore and were found consistent. A total of 10 compounds from all the screened libraries were selected and are currently under in-vitro investigation, for further validation of our findings.

Figure 1: Top ligands from NCI, KinaseSARfari and Chemdiv occupy same pocket
Human Papilloma Virus (HPV) is the main cause of cervical cancer that encodes six non-structural proteins (E1 To E6) and two structural proteins (L1 and L2). The drive of this study is to recycle or reuse old ovarian drug against new targets using the available soft information tools and software’s. In order to accomplish the above said aim, atomic-level models of HPV group of protein (HPV-16 E6) was modelled. Drug molecules against HPV proteins from drug databases (Drug databank, Zinc database, FDA) were identified. Each of the screened drugs were docked to the active sites of the target protein using the Cdocker algorithm by the software Discovery Studio and Auto Dock Vina. Based on the docking scores of different protein-drug complexes, the drug molecule with the best Dock score was noted as Rucaparib. All the protein-drug complexes were optimized and simulated using GROMACS. Binding of Rucaparib to the DNA binding domain of HPV-16 E2 prevents it to inhibit p53, which otherwise may abrogate the cell cycle events and thereby ceases uncontrolled cell division. Therefore, Rucaparib can be established as one of the repurposed drugs for Cervical Cancer treatment.

Keywords: FDA approved drugs, Lipinski rule, MD simulation, HPV protein
ASSOCIATION BETWEEN DRUG-GENE-INTERACTION, DRUG-DRUG-INTERACTION, AND DRUG-DRUG-GENE-INTERACTION AND SWITCHING, DOSE ADJUSTMENT AND EARLY DISCONTINUATION OF (ES)CITALOPRAM: AN EXPLORATIVE STUDY

Muh. Akbar Bahar1,2*, Pauline Lanting3, Jens H.J. Bos1, Rolf H. Sijmons3, Eelko Hak1, Bob Wilffert1,4 1Dept. of PharmacoTherapy, -Epidemiology & -Economics, University of Groningen, Groningen, the Netherlands. 2Faculty of Pharmacy, Hasanuddin University, Makassar, Indonesia. 3Dept. of Genetics, University Medical Center Groningen, Groningen, the Netherlands. 4Dept. of Clinical Pharmacy and Pharmacology, University Medical Center Groningen, Groningen, The Netherlands.

Correspondence: akbarbahar@unhas.ac.id; akbarbahar@unhas.ac.id

(Es)citalopram users commonly fail to achieve remission in their first treatment episode. Drug-gene-interaction (DGI), drug-drug-interaction (DDI), and drug-drug-gene-interaction (DDGI) may undermine the efficacy of (es)citalopram. Therefore, we aimed to explore the association between DGI, DDI, and DDGI and the effects on (es)citalopram dispensing. As part of the PharmLines Initiative, adult Caucasians (≥18 years old) from the Lifelines cohort (167,729 participants) with linked dispensing data from the University of Groningen prescription database IADB.nl (730,000 patients) and with genetic information on CYP2C19/3A4 genotypes were studied. Exposure groups were first time users of (es)citalopram with (1) DGI (CYP2C19/3A4 deviating phenotype), (2) DDI (CYP2C19/3A4/2D6 inhibitors/inducers), and (3) DDGI (co-presence of DDI and DGI). Outcomes were drug switching or dose adjustment, and early discontinuation within 90 days after the start of (es)citalopram. We applied logistic regression modeling to estimate adjusted odd ratios with corresponding 95% confidence interval. Overall, 316 (es)citalopram starters (median 45 years, 62% women) were included. DGI between (es)citalopram and decreased function of CYP2C19 tended to increase odds of switching and dose reduction, but seemed to reduce risk of discontinuation regardless of CYP3A4 phenotypes. These associations might be modified by CYP2C19 intermediate metabolizer. Meanwhile, DDI and DDGI showed an indication towards dose reduction and switching, respectively. As conclusions, DGI involving decreased function of CYP2C19, regardless of CYP3A4 phenotypic status, may increase the risk of switching or dose reduction of (es)citalopram. For DDI and DDGI, trends towards varying directions of effects were found, but larger studies are needed to confirm these findings.

A scientific statement:
Drug-Gene-Interaction (DGI) involving decreased metabolic function of CYP2C19, regardless of CYP3A4 phenotypic status, may increase the risk of switching or dose reduction of (es)citalopram.
A COMPARATIVE STUDY BETWEEN TRANSCRIPT-LEVEL AND GENE-LEVEL CLASSIFICATION ON MULTIPLE AFFYMETRIX PLATFORMS OF COLORECTAL CANCER MICROARRAY STUDIES

Hendrick Gao-Min Lim¹,* and Yuan-Chii Gladys Lee¹

¹ Graduate Institute of Biomedical Informatics, Taipei Medical University, 11031 Taipei, Taiwan

Several classification algorithms have been applied into microarray studies for detection of the colorectal cancer. Algorithms such as naïve bayes, random forest, logistic regression, support vector machine, and deep learning have been used in previous studies. The accuracy of these algorithms shown promising result through n-fold validation. However, there is lack of comparison between transcript-level and gene-level classification across multiple platforms microarray gene expression result. Meanwhile, the comparison is needed to validate the classification performance across platforms to make sure there is no imbalance result. Therefore, we compared the classification performance of transcript-level with gene-level using multiple Affymetrix microarray platforms through different classification algorithms including: naïve Bayes, random forest, logistic regression, support vector machine, and deep learning. We evaluated the performance using several parameters including: accuracy, area under ROC curve, precision and recall. As the result, we found the imbalance performance in transcript-level classification from multiple studies can be solved through gene-level classification by applying annotation and merging. Furthermore, by applying batch effect removal method can make gene-level classification performance slightly improved.
RV0807, A PUTATIVE PHOSPHOLIPASE A2 OF MYCOBACTERIUM TUBERCULOSIS; ELUCIDATION THROUGH SEQUENCE ANALYSIS, HOMOLOGY MODELING, MOLECULAR DOCKING AND MOLECULAR DYNAMICS STUDIES OF POTENTIAL SUBSTRATES AND INHIBITORS

Shobana Sundar*, Lokesh Thangamani, Gowdha Manivel, Praveen Kumar and Shanmughavel Piramanayagam

Department of Bioinformatics, Bharathiar University, Coimbatore.

*Corresponding Author Dr. Shobana Sundar Correspondence id: sundar.shobana@gmail.com

Mycobacterium tuberculosis has the ability to scrounge off the host macrophages and create a cordial environment for its survival. Identification of mechanisms favoring this purpose leads to novel treatment strategies for tuberculosis. Rv0807, a homolog of MSMEG_5817 from M. smegmatis may be crucial for mycobacterium survival within the host macrophages. In this study through insilico approaches, we intend to identify the putative role for Rv0807 from M. tuberculosis and its essentiality for mycobacterium survival within the macrophages. We have designed motifs for Rv0807, MSMEG_5817 and related phospholipase sequences from different strains. From sequence analysis, we suggest that Rv0807 could be a Phospholipase A2 of Mtb. We have predicted some important residues to be a part of the catalytic process of the Rv0807 homodimer and we have also performed insilico mutation studies of these residues to understand their role in catalysis. Rv0807 could be a potential drug target as it binds phosphatidylinositol-3- phosphate (PI3P) and could be involved in processing the host cell PI3Ps, thereby blocking the phagosomal maturation. A pharmacophore hypothesis was generated based on the ligand binding site and a set of Pretomanid related compounds were screened against the Rv0807 homodimer. The top five compounds having better docking scores and good ADME properties were selected as best inhibitors and analyzed further. Molecular dynamics studies of Rv0807 homodimer with PI3P, demonstrated a lot of conformational changes in the protein structure as it gets occluded through the course of simulation. The movement of a loop atop the ligand binding site, suggests of a lid-like region as seen in many other phospholipases. Thus, the insights gained through this study will provide some novel ideas to understand Rv0807 catalysis and its role in Mycobacterium survival within the macrophage.

Keywords: Rv0807, Mycobacterium tuberculosis, homology modeling, molecular docking, insilico mutation, Molecular dynamics
 INSTANCE-BASED LEARNING FOR PERSONALIZED CANCER DIAGNOSIS AND TREATMENT PLANNING

Md. Sarwar Kamal¹, Daniel Catchpoole¹,²,³, Jinyan Li¹*

¹University of Technology Sydney, ²Westmead Children Hospitals, ³University of Sydney
MdSarwar.Kamal@student.uts.edu.au, daniel.catchpoole@health.nsw.gov.au, jinyan.li@uts.edu.au*

Existing “global” learning approaches (e.g., decision trees, support vector machines) have been widely used for in-silicon disease diagnosis. With a specific training data set, these global learning approaches generate a decision model of fixed parameters, regardless of the change from one test instance (patient) to another. Such a fixed model applied for all test samples is not consistent with the principle of the personalized cancer diagnosis which requires that only the most relevant knowledge from the training data should be tailored for the accurate diagnosis of each test sample. Instance-based learning (IBL), or called “lazy learning”, consider a different set of knowledge from the training data to reach a decision for each different test instance, suitable for precision medicine.

This work proposes a novel instance-based learning approach. Each time for a new test instance, our method generates a relevance matrix specifically for that instance through the scanning of the whole training data. This relevance matrix may change when a different test instance is applied. Then, a Bipartite graph and K-connected sub-graphs are constructed from the relevance matrix to ensure only dominant features (genes) of the test instance are extracted responsible for cancer. Finally, Bayesian classification and maximum a posterior (MAP) are used to make the accurate diagnosis decision. Our rigorous experimental analysis confirms that instance-based learning can outperform global learning techniques in accuracy, precision, recall, and F-measures. Moreover, IBL can identify patient relevant sub-cohort and gene subset for an individual instance as shown in our case studies on colon, breast, bladder, and adrenal cancer, which are useful for disease understanding and personalized treatment planning.
TARGET IDENTIFICATION FOR MULTI-TARGET DRUG DESIGNING FOR MYCOBACTERIUM TUBERCULOSIS ACTIVE INFECTION BY SYSTEMS BIOLOGY APPROACH

Praveena Koyyada, Dr Seema Mishra*

Department of Biochemistry, School of Life Sciences, University of Hyderabad, Hyderabad, India

Correspondence: Praveena Koyyada: praveena.bt94@gmail.com; Dr Seema Mishra: seema_uoh@yahoo.com

As the world progresses in discovering the realms of the universe and in technology, bacteria have evolved, developing tolerance and resistance towards drug and antibiotics. One such microbe is Mycobacterium tuberculosis H37Rv, causative agent of tuberculosis, which owes 1.3 million deaths annually. The genomic architecture of Mycobacterium comprises genes responsible for persistence and survival in harsh conditions of host environment [1] and thus, establishing infection. For instance, the most recent discovery shows MarP gene of bacteria involved in acid tolerance that enables the survival and replication of bacteria in host cell lysozymes [2]. In such a scenario, it is essential to deduce an efficacious, shorter drug regimen that evades the persistence of bacteria and further, works towards elimination of MDR Mycobacterium tuberculosis. In order to achieve our aim, we initially enlisted the major transcription factors/coregulators and their corresponding target genes that are crucial in developing bacterial infection. Our computational network building and, statistical understanding on TFOE gene sequencing [3] data propose novel transcription regulators Rv0022c, Rv0054, Rv0081, Rv0452, Rv0827c, Rv0880, Rv1027c, Rv1176c, Rv1221, Rv3286c, Rv3676, Rv3862c to be amongst major regulators regulating target genes involved in multiple major pathways such as oxidoreductase respiratory ETC, response to oxygen levels, host cell cytoplasm component metabolism, and cellular external encapsulation. We have also identified that decisive virulence, adaptation genes of PE_PGRS family are regulated by four major regulators Rv1176c, Rv0880, Rv0081 and Rv3862c. An insightful study on their corresponding regulatory pathway analysis reveals key metabolic pathways and their respective genes to be targeted.
IN SILICO SIMULATION OF BACTERIOPHAGES AND SALMONELLA TYPHI POPULATION DYNAMICS

Ayman Elshayeb\textsuperscript{a}, A. A. Ahmed\textsuperscript{a}, M. El Siddig\textsuperscript{a} and A.A. EL Hussien\textsuperscript{a}

\textsuperscript{a} Faculty of Science, University of Khartoum, Sudan

Salmonella typhi is a causative agent of food poisoning and enteric fever. The Simulation of bacteriophages’ efficiency in bacteria treatment is the integration of their lytic cycle in destroying bacteria with those found in the disciplines of mathematics, statistics and computational systems in order to find the exact numbers of phages that can totally lysate bacteria. This work was conducted to create a simulative model for the development of dynamic interactions between phages and their bacterial host (Salmonella typhi).

The model was developed from calculations of bacteriophage replication to solve the distinct mechanisms of their efficiency including changes in lifecycles, therapeutic dose and mortality rates. Simulated data are compared with data obtained \textit{in vitro} to assess the suitability of the model for multiplicity of infection.

\textit{In vitro} observations showed that the strength and mechanisms of bacteriophage can alter the determination of Salmonella typhi as antimicrobial therapy. The exponential growth curves solved the interactions of bacteriophage with their host in certain time decay, the changes in concentrations over time was solved by differential equations used to determine the therapeutic outcome.

\textit{In silico} predicting of the potentiality of phages in lysing Salmonella typhi was estimated by using this model due the experiments conditions (\textit{in vitro}). For more accurate estimations the model was programmed MS Excel sheet and simulated as a simple computer program.

The predicting of the potentiality of lytic phages in lysing Salmonella typhi can be estimated by using this mathematical model due the experiments environment conditions. Therefore, it is likely that the mathematical model could be made to work computationally by changing their values according to the laboratorial experiment conditions.
Sequence data have been rapidly accumulated in accordance with recent progress of sequencing technology. The huge volume of the sequence data is a rich source of biological knowledge. At the same time, however, the large quantity hinders the analysis of a multiple sequence alignment with the user-interactive manner. One of the problems is difficulty in assigning annotation to each sequence of a large alignment. In this study, we have developed an alignment viewer named ASHViewer [1], which has a function to collect annotation data from linked open data (LOD). Previously, one of the authors (A.Y.) had developed a search system called LOD Surfer for general use in life sciences [2]. Based on the system, a version specified for multiple amino acid sequence alignment was developed and implemented in ASHViewer [1]. Previous version can collect three types of GO terms, molecular function, cellular components and biological process, and visualize the collected GO terms in concentric circles surrounding an unrooted phylogenetic tree. The new version can collect the source organisms and active sites. The source organisms are visualized in the concentric circles like GO terms. The active sites are shown in the conservation profiles calculated from the alignment. Other functions of the viewer will be also discussed.
Poster ID: P-12

PREDICTION OF DRUG RESISTANCE IN MTB USING MACHINE LEARNING ALGORITHMS
Abhirupa Ghosh, Sudipto Saha*

Division of Bioinformatics, Bose Institute, Kolkata, India

Correspondence: AG: abhirupa@jcbose.ac.in; SS: ssaha4@jcbose.ac.in

Drug-resistant (DR) tuberculosis (TB) is one of the major global healthcare concern including India as faster and accurate detection of drug resistance(s) in *Mycobacterium tuberculosis* (MTB) clinical isolates pose a great challenge [1,2]. In this study, the publicly available WGS sequences of MTB with known drug resistance phenotypes were processed to obtain single nucleotide polymorphisms (SNPs) in specific DR genes [3]. Further, the DR gene mutations were used to develop Naïve Bayes model(s) for predicting the susceptibility and resistance of MTB. Here, the prediction model(s) were trained/tested on gene mutations in rifampicin and isoniazid associated genes namely, rpoB, inhA and katG. Three different values were used to represent point mutations such as no mutations as zero, heterogeneous mutation as 0.5, and homogenous mutations as 1 for generating vectors. The final score from Naïve Bayes model for each MTB isolates gene sequence data is computed as log likelihood ratio of drug resistance vs drug susceptible posterior probabilities. The developed prediction model(s) were tested in fivefold cross-validation for mono and multiple drugs resistance with reasonable accuracy. The performance of the model(s) were also assessed on a blind data set. In summary, the prediction model(s) were developed using Naïve Bayes for predicting drug susceptible/resistant TB isolates using SNP data from WGS sequences.

Figure: The Workflow to predict drug resistance in TB using in silico approach. 1-3: Development of prediction model using public domain data. 4: validation of the model using a blind dataset. 5: The performance evaluation of the model.
Asthma and Chronic Obstructive Pulmonary Disease (COPD) are airway inflammatory diseases that exhibit similar symptoms like wheezing, breathlessness and chest tightness [1]. Although there are similarities in symptoms, but the treatment regimen are different. So, it is a challenge to diagnose these diseases accurately at the early stage. Here, Pulmonary Function Test (PFT) reports of patients were used as features for supervised machine learning and attempt was made to classify these diseases using Support Vector Machines (SVMs) [2]. Feature selection algorithms were used to improve the performance measures. In this study, COPD was chosen as positive dataset whereas asthma and chronic cough were selected as negative dataset. An accuracy of 71.8% was achieved by Radial Basis Function (RBF) kernel using 10-fold cross validation. Similar result was achieved for blind dataset as well (accuracy of 75.88%). In summary, SVM was used for supervised learning for prediction of most common pulmonary diseases like COPD and asthma with PFT features.

**Figure 1: Flowchart**
EVALUATION OF SOFTWARE FOR MUTATIONAL SIGNATURE ANALYSIS BASED ON REALISTIC SYNTHETIC DATA

Steven G Rozen*, Yang Wu, S M Ashiqul Islam, Ludmil B Alexandrov, Michael R Stratton

Duke-NUS Centre for Computational Biology and Programme in Cancer & Stem Cell Biology, Duke-NUS Medical School, Singapore. Department of Cellular and Molecular Medicine and Department of Bioengineering and Moores Cancer Center, University of California, La Jolla, San Diego, CA, USA. Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK

Correspondence: SGR: steve.rozen@duke-nus.edu.sg

Mutational signature analysis has become a mainstay in the study of cancer genomes [1,2]. While ≥16 approaches to signature discovery and/or attribution have been implemented, there has been negligible evaluation of their usability, utility, or accuracy on substantial sets of realistic synthetic data. Here we present a suite of synthetic data sets, an R package for generating such synthetic data sets, and evaluations of multiple software packages on these synthetic data sets. We have developed a freely available, open-source R package, SynSig, for generating synthetic data sets for mutational signature analysis (Figure 1, https://github.com/steverozen/SynSig). SynSig emphasizes generation of realistic synthetic data, based on the field's best understanding of known mutational signature profiles and their contributions to mutational spectra in various tumor types. We have used SynSig to generate two collections of synthetic data, and have evaluated mutational signature software on these collections. Study 1 is based on a suite of 8 realistic synthetic data sets designed to probe challenges in signature analysis (data and results at https://doi.org/10.7303/syn18497223). Study 2 is based on synthetic mutational spectra with correlated mixtures of two signatures (data set available at https://doi.org/10.5281/zenodo.2636980). Evaluation of software in these studies shows that these synthetic data provide an important resource for assessing, and subsequently improving, approaches for discovering and attributing mutational signatures. Moreover, the tests confirmed that use of current approaches to discover signatures is not a purely automatic process. Instead, signature discovery requires human judgement that considers all available data, evidence, and reasonable priors.

Figure 1: Visualization of a synthetic mutational spectrum resembling that of a pancreatic adenocarcinoma. The height of each bar represents the number of single base substitutions in the context of the preceding and following bases. For example, the highest blue bar is the number of TCT to TAT mutations.
IDENTIFICATION OF PHOSPHATIDYL INOSITOL 3-KINASE δ (PI3Kδ) INHIBITOR: PHARMACOPHORE-BASED VIRTUAL SCREENING AND MOLECULAR DYNAMICS SIMULATION

Muhammad Arba1*, Malindo Sufriadin1, Ruslin1, Ida Usman2

1Faculty of Pharmacy, Halu Oleo University, Kendari, Indonesia, 93231 2Department of Physics, Halu Oleo University, Kendari, Indonesia, 93231

Correspondence: muh.arba@uho.ac.id

Phosphatidylinositol 3-kinase δ (PI3Kδ) is a validated drug target for the treatment of cancer. The present study was aimed to search for new inhibitor of PI3Kδ by employing pharmacophore modelling using LigandScout software. The three hydrogen bond acceptors and two hydrophobic features were proposed as a pharmacophore model using LASW1976 structure. The model was then validated using Area Under Curve (AUC) of Receiver Operating Characteristic (ROC) and GH score, and it was used to screen new molecules in the ZINC database, which resulted in 599 hits. All 599 hits were then docked into PI3Kδ protein, and five best hits were submitted to 50-ns molecular dynamics simulations. Each hit complexed with PI3Kδ, underwent minor conformational changes as indicated by the values of Root Mean Square Deviation (RMSD) and Root Mean Square Fluctuation (RMSF). Furthermore, prediction of the binding free energy using Molecular Mechanics-Poisson Boltzmann Surface Area (MM-PBSA) method showed that four hits i.e. Lig25, Lig199, Lig554, and Lig682 had much lower binding energy as compared to LASW1976, which indicated their potential as new inhibitor of PI3Kδ.

Figure 1: interaction of Lig199 with PI3Kδ.
β-thalassemia major is a severe prevalent blood disorder caused by mutations interfere the primary mRNA transcription. Thus, eventually dissolve regular splicing and produce β\(^0\)-thalassemia result in β-thalassemia major phenotype. Mutations at position IVS-1 nt 1 and IVS-1 nt 5 recently reported as two of the most prevalent mutation of thalassemia major. RNAfold WebServer was used to predict thermodynamic stability of mRNA sequences carrying 4 types of intron mutation in beta globin gene. The lowest centroid secondary structure MFE levels was performed by IVS1nt5 G>C (-334.86 kcal/mol), followed by IVS1nt5 G>T (-330.86 kcal/mol), IVS1nt1 G>T (-322.16 kcal/mol), and IVS1nt1 G>A (-316.16 kcal/mol). Our study presented that intronic mutation effect structural alteration, which result functional failure in globin chain protein of β-thalassemia mayor.
Designing of effective and save drug candidates for breast cancer Her2+ was important due to high number of the overexpression in breast cancer. Its overexpression will lead to aggressiveness and worse prognosis. Mutations also showed as a fascinating topics in Her2 protein. Few of HER2 mutation is likely to result in resistance to HER2 tyrosine kinase inhibitors. We used 6 missense mutation generated from SNP Entrez of NCBI. We did the homology modelling to create the 3D structures of mutated Her2 protein. The ligand-based drug design was used to design a new Her2 inhibitor. We designed Pinostrobin – Mannich Base derivatives because of their structures showed similarity with best pharmacophore model. Based on molecular docking result, we could see that ΔG of Lapatinib in all mutant Her2 was higher than in wild-type Her2. Differ from Lapatinib, all of pinostrobin derivatives show good activity against wild-type and mutant Her2, especially 6-(indole-1-ilmethyl)pinostrobin (8), that showed as the best binding energy in both wild-type and mutant Her2. From the results, we can find out that there are candidates for Her2-targeted inhibitor that sensitive to both wild-type and mutant. So, breast cancer patients with Her2 therapy (with or without mutant of Her2 status) could be given the HER2-targeted inhibitor without any resistance issues.
As the mammalian target of rapamycin (mTOR) regulated pathway has been found most commonly activated and deregulated in cancer, so there is a need to understand the underlying mechanism of its key interacting proteins. Structural insights of proteins so far undiscovered will unravel these phenomena. 3D structure of mTORC2 is built using several ab initio protein structure prediction tools and homology modelling algorithm of Modeller. The quality and validation of the obtained models were accessed using Errat, PROSA and Qmean softwares while the Ramachandran plot was used to access the overall stereochemical properties of the proteins. The individual protein sub-units were finally docked using HADDOCK webserver. This approach will aid in understanding the function of individual proteins and a step forward towards drug designing against cancer.
MOLECULAR DOCKING AND MOLECULAR DYNAMICS SIMULATION ANALYSIS FOR FINDING HDAC CLASS IIA INHIBITOR AS NEW ANTIDIABETIC DRUG

1,2* Linda Erlina, 3 Arry Yanuar, 1,2 Fadilah, 1,2 Rafika Indah Paramita

1 Department of Medical Chemistry, Faculty of Medicine, Universitas Indonesia, Jakarta – Indonesia 10430
2 Bioinformatic Cluster, Indonesian Medical Education and Research Institute (IMERI) – Faculty of Medicine, Universitas Indonesia, Jakarta – Indonesia 10430
3 Faculty of Pharmacy, Universitas Indonesia, Depok, West Java – Indonesia 16424

*Correspondence: lindaerlina22@gmail.com

Epigenetics is related to modifications of gene expression and predicted can give a potential implication in some diseases, including metabolic disorders. Among the epigenetic mechanisms currently known, histone modifications such as acetylation and deacetylation have been reported as important regulators of gene transcription, which influence physiological processes involved in several diseases, including diabetes. HDAC Class IIA is one of important epigenetic factor for glucose metabolism, liver, adipogenesis, myogenesis and muscle metabolism. In this research, the aim is to find the new candidates for epigenetic (HDAC Class IIa) drug, especially from herbal compounds. The pharmacophore-based virtual screening has already done in the previous study by Linda Erlina, et al (2017). Based on virtual screening result, there are six hit compounds (artocarpesin, dimboa glucosidase, eriodictin, avicularin, mirabijalone c and luteolin) that have high pharmacophore fit score then continue to molecular docking and molecular dynamic simulation for 10ns. The evaluation parameters of molecular docking and molecular dynamic simulation include: amino acids residue interaction, RSMD and RMSF. Based on molecular docking and molecular dynamic simulation analysis, three of herbal compounds (artocarpesin, mirabijalone c and eriodictin) have a high potency as new antidiabetic drug.
ADVANCING PATHOGENICITY PROFILING THROUGH GENOMICS BIG DATA ANALYTICS

Shrmilita Mahendran¹, Maurice H.T. Ling¹, Sy Bing Choi¹, Effirul Ikhwan Ramlan², Mohd Firdaus- Raih³, Asif M. Khan¹, 4* and Hui San Ong¹*

¹ Centre for Bioinformatics, School of Data Sciences, Perdana University, Serdang, Malaysia.
² School of Computing and Intelligent Systems, Ulster University
³ Faculty of Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Malaysia
⁴ Graduate School of Medicine, Perdana University, Serdang, Malaysia

Correspondence: asif@perdanauniversity.edu.my, huisan.ong@perdanauniversity.edu.my

Rapid emergence and advancement of next generation sequencing technologies over the last decade has generated a growing wealth of digital footprint, with genomic data alone predicted to be scaling 40 Exabytes by 2025. Harvesting and harnessing these streams of big data through integrative analytics for the control of pathogen outbreaks and rise in antimicrobial resistance have always been one of the promises of the post-genomic era. The identification and characterization of virulence and pathogenicity factors from microbial genomes has allowed for a precise profiling of pathogenicity. This paper highlights the genomic developments around pathogenicity profiling, in particular big data meta-analytics of microbial pathogens’ volatile signatures. Machine learning has been pivotal in advancing pathogenicity, paving the way for AI-powered next generation tools. Digital organisms are becoming valuable for high throughput simulations to better understand the origin and evolution of pathogenicity.
The 18th International Conference on Bioinformatics

Poster ID: P-22

STUDY ON ACTIVATION OF CONSTITUTIVE ANDROSTANE RECEPTOR (CAR) BY HERBS FOR PREVENTION OF GALLSTONE – A COMPUTATIONAL APPROACH

1 Bhavya Sethi, 1 Shashank Raj, 2 Rajani Sharma, 1* Shubha Rani Sharma & 1 Raju Poddar 1 Department of Bio-Engineering, Birla Institute of Technology, Mesra, Ranchi, Jharkhand - 835215, India
2 Amity University, Ranchi, Jharkhand- 834002

*Correspondence: srsharma@bitmesra.ac.in

The gallstone disease is one of the most common gastrointestinal diseases. It has been found that the activation of CAR (Constitutive Androstane Receptors) mice prevents the formation of gallstones. The identification of 6,7-Dimethylesculetin as a novel herbal activator of the human Constitutive Androstane Receptor and its subsequent verification as an agonist by conducting a docking of the molecule with the hCAR is an important part of this study. Thus, this particular molecule can be used for the treatment of the gallstones- both cholesterol and pigmented. As to how, the molecule derived from the herb Artemisia capillaris can be used as an effective drug for actual use is subject of further research. Also, it will be extremely beneficial to determine the preventive cure for the gallstones by decreased cholesterol secretion and increased bilirubin clearance. This research work identifies the potential agonists of CAR from different Indian herbs both for human and murine. By docking of the human CAR with each of the agonist and using other bioinformatics tools, we can narrow down the search for an agonist in the herbs or plants. The agonists which were considered for study were 6-(4-chlorophenyl) imidazo[2,1-b][1,3]thiazole-5-carbaldehyde O-(3,4-dichlorobenzyl)oxime (CITCO) and 6,7-Dimethylesculetin. Of these, the latter was derived from the herb Artemisia capillaris and its docking with the hCAR along with further analysis confirmed that it is a potential CAR agonist. There is a decreased cholesterol release and increased bilirubin clearance which prevents both, cholesterol and pigmented gallstones.
IMAGE PROCESSING ON THE BRAIN: 3D PRINTING MRI IMAGES FOR FURTHER OBSERVATION

Lautaro Franco Soler
Madison High School, United States

Correspondence: lsoleracademix@gmail.com

The purpose of this project was to take partially processed MRI images of the brain and further process them in order to create a 3D model for further, physical analysis once printed. This approach attempted to mirror the processes medical officials use on a daily basis for reconstructing 3D models of organs to determine if any abnormalities are present, for use prior to surgery or any other medical intervention. In this case, the goal of this project was to use publicly available MRI data to recreate a 3D model of the brain in order to be able to 3D print and then inspect it for evidence of any irregularities. In order to do this, the MRI image slices needed to be cleaned up, normalized and filtered. Histograms were used to determine brightness cutoff ranges and to set a specific brightness threshold. Blurring techniques and edge detection were used to determine the most likely edges of the images of the brain. In the end, edge detection allowed for a 3D model to be generated and exported for 3D printing.
MAPPING KEY ALTERNATIVE SPlicing EVENTS DURING NEURAL CREST CELL DELAMINATION

Kanupriya Tiwari\textsuperscript{1}, Anushree Balachandran\textsuperscript{1}, Ernst J. Wolvetang\textsuperscript{1}, Lars K. Nielsen\textsuperscript{1,2*}

\textsuperscript{1}Australian Institute for Bioengineering and Nanotechnology (AIBN), The University of Queensland, QLD 4072, Australia
\textsuperscript{2}The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Kgs. Lyngby, DK-2800, Denmark

Correspondence: lars.nielsen@uq.edu.au; k.tiwari@uq.edu.au

Neural crest cells (NCCs) are multipotent stem cells induced during neurulation in the developing nervous system of a vertebrate embryo. Given the correct transcriptional/epigenetic cues, they detach from the neural tube and migrate extensively generating a variety of cell types - facial cartilage and bones, skin pigment cells, particular cell types of the heart, eye and inner ear etc. Delamination from the neural tube is the most crucial step during NCC differentiation as they undergo epithelial to mesenchymal transition (EMT) to attain their migratory phenotype. In this work, we used a custom alternative splicing detection and differential analysis pipeline to contrast bulk RNA-seq data from SOX10\textsuperscript{-} (pre-migratory) and SOX10\textsuperscript{+} (migratory) NCCs. We detected several and nervous system development. Several splice factors such as CELF2, RBM38 were differentially expressed between the two populations indicating a fundamental role of alternative splicing in promoting NCC delamination and migration.
RNA-SEQ ANALYSIS BENCHMARK: DATASETS & TOOLS RESOURCES

Fadel Berakdar 1, Colin Hercus 1, Su Wei Chong 1, Akzam Saidin 1,*

1 Novocraft Technologies Sdn. Bhd., Petaling Jaya, Selangor, Malaysia

Correspondence: fadel@novocraft.com; akzam@novocraft.com

Building upon previous work done by SimBA[1], which focuses on simulated human RNA-seq (100bp paired-end) for normal and somatic mutation for benchmarking RNA-seq reads mapping to genome, splice junction alignment, single nucleotide variants and indels, and fusion. We adopted their methodology and tools and extended the scope to include additional RNA-seq aligners and simulated RNA-seq reads (100bp and 150bp paired-end) from multiple species (Homo sapiens, Mus musculus, Danio rerio, Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana and Saccharomyces cerevisiae) for both normal and/or somatic condition. Currently we are performing benchmarking on five aligners; novoSplice, novoAlign, STAR, HISAT2 and GSNAP. In our preliminary benchmarking using human simulated dataset, we observed that novoSplice, novoAlign and STAR produces alignments that are high in sensitivity and precision for both genome and splice junction mapping (Figure 1 & 2). This study aims to provide an up to date RNA-seq analysis tools benchmarking information and resources for the bioinformatics community in the near future.

References

1- SimBA: A methodology and tools for evaluating the performance of RNA-Seq bioinformatic pipelines. BMC Bioinformatics 18 (2017) 428

Figure 1: Results for GRCh38-101bp-160M-normal-0.014 data-set

Figure 2: Results for GRCh38-101bp-160M-somatic-0.016 data-set
Comet assay or single cell gel electrophoresis assay (SCGE) is a method which is frequently used to measure the damage of DNA. The results of comet assay is a set of comet images, then the comet images are classified to measure the level of DNA damage. Currently, there are several softwares for comet assay image analysis, both free and commercial. Commercial software is quite expensive, while free software is limited, especially for buccal mucosa cell and super tiny comet image dataset. In this research, we propose a classification model for comet assay with super tiny image dataset which is taken from buccal mucosa cells. We propose a transfer learning based convolutional neural network (CNN) model. We have compared the transfer learning model with CNN-support vector machine (SVM) and ordinary CNN. In our experiments, we use super tiny dataset consisting of 73 images. Our transfer learning model gives an accuracy 70.5%, while CNN-SVM gives 62.3% and ordinary CNN gives 63.5%. We also compare our transfer learning model with most frequently used, free comet assay analysis software, OpenComet. Open-Comet gives an accuracy 11.5%. Our transfer learning model is promising for comet assay for buccal mucosa cell and super tiny dataset.
MOLECULAR PROFILING OF THE NONCONOPEPTIDES IN THE ENVENOMATION STRATEGY OF *Conus Lenavati* THROUGH HIGH-THROUGHPUT SEQUENCING PLATFORMS

Gomez, Edward Daniel D.¹*, Constantino, Elcid V.¹, Dela Cruz, Maverick D.¹, Calugay, Ronie J.¹, and Lluisma, Arturo O.²

¹Department of Biology, College of Science, University of the Philippines Baguio, Baguio City, Philippines

²Marine Genomics and Molecular Genetics Laboratory, Marine Science Institute, University of the Philippines Diliman, Quezon City, Philippines

Correspondence: edgomez4@up.edu.ph; edgomez4@up.edu.ph

Cone snails (genus: *Conus*) are marine molluscs which utilize a sophisticated venom production and delivery system that shoots harpoon-like teeth to subdue prey. For decades, *Conus* research was largely limited to this genus’ venom, collectively known as conopeptides—complex cocktail of neurotoxic peptides. Hence, the nonconopeptides, proteins within the venom duct not classified as conopeptides, has been overlooked even though these components play critical roles in a wide spectrum of processes such as physiology of venom delivery, post-translational modification of conopeptides, and envenomation. In this study, transcriptome analysis via high-throughput sequencing (HTS) technologies provided the first comprehensive view of the nonconopeptide diversity of a deep-sea cone snail. Notably, the nonconopeptide-mediated molecular envenomation strategies of *Conus lenavati*, were elucidated in this study. HTS revealed that there are 3,749 nonconopeptides where there are 61 different envenoming factors. These toxins were then further categorized to six clusters according to their mode of action. Results show that majority of the envenomation proteins belong to the hemotoxic cluster which attacks the haemostatic system of the prey. However, the genes encoding for neurotoxins were the most expressed envenoming proteins (see Figure 1). Interestingly, these hemostasis-impairing toxins are usually reported in snake species hence, this may be the first to report the presence of these toxins in the *Conus* genus. Lastly, nonconopeptides such as insulin and veficolin found within the venom duct has pharmacological potentials which could be a source of alternative medicine.

Figure 1. The relative expression of peptides in log2(TPM) clustered based on envenomation function of the three transcriptomes of *C. lenavati* (lena1, lena2, and lena3).
COMPARATIVE STUDY OF MICROBIAL NITRILASE ENZYMES: A COMPUTATIONAL APPROACH

Priya Kumari and Raju Poddar

Birla Institute of Technology, Mesra

Correspondence: Priyasinghchas@gmail.com

Nitriles are becoming a major precursor of many industries which is a very toxic and carcinogenic that is harmful to biodiversity. Apart from the toxicity, it is being rigorously used in synthetic rubber, agriculture, cosmetics, dyes, pharmaceuticals, plastics and textile industry. This toxic compound gets accumulated in the environment after its usage leading to pollution. Nitrilases (‘green catalysts’) are industrially important enzymes expressed widely in both prokaryotes and eukaryotes that belongs to nitrilase superfamily that consists of 13 branches which involved in the hydrolysis of nitriles to amides and acids depending on the nature of the enzyme. Commercially, nitrilases, nitrile hydratases and amidase are major nitriles metabolizing hydrolytic enzymes employed for the biotransformation of nitriles into corresponding useful carboxylic acid as product and in bioremediation of nitriles polluted sources.

A large ensemble-based dataset was utilized from bacteria, fungi, plants and animals. We studied the coding gene sequences, amino acid sequences of the nitrilases from different species and discovered conserved motif linked with related other species. The inferred evolutionary tree shows nitrilase gene clusters are shared among bacteria, fungi and plants. Structural analysis (molecular dynamics simulation, principal component analysis (PCA), dynamic cross correlation (DCCM), root mean squared inner product (RMSIP), and free energy surface (FES)) revealed that the folding of catalytic sites is similar among species; however, the loop region varies. We provide evidence-based on PCA that the nitrilases are clustered into different clades due to variation inside chains. The results are consistent with the hypothesis that bacterial nitrilases are in ecological and evolutionary relationships with fungi and plants during plant-pathogen interaction to a large extent. This compact and detail results also open new dimensions for further studying and improvement of industrially important nitrilase enzymes.
The human leukocyte antigen (HLA) is the most polymorphic gene in the human genome. It is located on the short arm of chromosome 6 encompasses 239 antigenic loci that about 40% of them are immunogenic. HLA-B has been used as pharmacogenomics marker for predicting drug associated with adverse reaction. There are more than 300 ethnics in Indonesia but the study for HLA polymorphism in Indonesia is still limited. The HLA-B genotypes of 36 Javanese ethnic, 201 Sundanese ethnic, and 58 Malay ethnic unrelated individuals with 3 generation back of ethnic origin was carried out using Polymerase Chain Reaction-sequence specific oligonucleotides probe (PCR-SSOP) method. Our result showed that HLA-B*15:02 is the most frequent allele in Javanese (16.67%) and Malay (15.52%) ethnic populations. HLA-B*15:02 (10.95%) in Sundanese ethnics is second frequent allele after HLA-B*15:13 (10.95%). HLA- B*15:02 is known to have association with Carbamazepine-induced Steven Johnson Syndrome / toxic epidermal necrolysis (SJS/TEN) in Asian population including Han Chinese, Taiwanese, and Southeast Asian population, including Indonesia. Our previous study showed that HLA- B*15:02 is strongly associated with Carbamazepine-induced SJS/TEN (OR= 6.5) in Javanese population. Our data will support further case-control study in other ethnics. Additionally, our study can helps establishing pharmacogenomic-based individualization of drug treatment in Indonesia. However future study to evaluate the medical utility and cost-effectiveness of genetic testing in Indonesia also should be considered to predict and prevent the serious condition associated with adverse drug reaction, leading to patient-friendly personalized medicine.
PLATELET RICH PLASMA STIMULATE HUMAN DERMAL FIBROBLAST VIABILITY, PROLIFERATION AND MIGRATION IN HIGH GLUCOSE CONDITION

Restu Syamsul Hadi¹, Indra Kusuma² and Ratih Rinendyaputri³

¹Department of Biology-Anatomy, Faculty of Medicine YARSI University
²Department of Physiology, Faculty of Medicine YARSI University
³Center for Research and Development of Biomedical and Basic Health Technology, Ministry of Health

Human dermal fibroblasts (HDF) are adult stem cells isolated from the dermis layer of the skin. HDF has a potential role in the process of cutaneous wound repair. Use for the treatment of wound healing based on regenerative medicine. The objective of this study was conducted to examine the effect of PRP on viability, proliferation, and migration of HDF as a model of in vitro wound healing under high glucose condition. In this experimental study, fibroblasts were isolated from foreskin then grown in medium Dulbecco's Minimal Essential Medium (DMEM) complete with serum 10%. HDF is characterized using flow cytometer. Cell viability was evaluated using the TC20 Automatic cell counter. To prove the role of PRP on cell proliferation HDF carried out with the CCK-8 test on the 5 treatment groups. To evaluate the ability of migration were used scratch assay. The result showed that HDF expressed CD70, CD90, and CD105 positive as the mesenchymal marker. These results demonstrated that PRP induces HDF viability, proliferation, and migration under high glucose conditions. Our findings provide a basis for the development of agents that can promote wound healing under diabetic conditions.
A ROBUST MISSING VALUE IMPUTATION APPROACH FOR MICROARRAY GENE EXPRESSION DATA ANALYSIS

Md. Ibnul Asifuzzaman  Department of Statistics Begum Rokeya University Rangpur, Bangladesh cmasif.nil@gmail.com, Md. Shahjaman Department of Statistics Begum Rokeya University Rangpur, Bangladesh shahjaman_brur@yahoo.com, Habiba AkterDepartment of Statistics Begum Rokeya University Rangpur, Bangladesh habiba23bru@gmail.com

Nowadays researchers can measure expression profiles of thousands of genes simultaneously with low costs by the advent of high-throughput technologies. Gene expression data (GED) often contain missing values or outliers due to various reasons of data generating process. Although, missing imputation and outliers handling both is equally important for downstream analysis using GED, most of the methods conduct this two task separately. A number of methods have been developed for missing imputation and they are widely used in GED. However, they cannot deal with outlier problem. As a result, in presence of outliers with the missing imputed datasets by these traditional methods the performance of the subsequent analysis can be hampered and may produce misleading results. Therefore, in this paper, an attempt is made to develop a robust approach using the minimum beta divergence method, which simultaneously handle the outliers and impute the missing values. We investigate the performance of the proposed method in a comparison of some popular missing value imputation methods such as Zero, KNN, robust SVD, EM and random forests through feature selection using both simulated and real GED. We used three performance indices such as Frobenius norm (FOBN), area under the ROC curve (AUC) and misclassification error rate (MER) to choose the optimal method. Both simulated and real data analysis results confirm the superiority of the proposed method than other methods in presence of outliers (1%, 5%, and 10%). Whereas, in absence of outliers, the proposed method performs slightly better than the other competing methods.
ISONIAZID METABOLIZING ENZYME GENOTYPES DISTRIBUTION IN INDONESIAN POPULATIONS: AWARENESS TO ANTI-TUBERCULOSIS DRUG INDUCED LIVER INJURY (AT-DILI)

Retno Wilujeng Susilowati, Kinasih Prayuni, Intan Razari, Abdul Salam M Sofro and Rika Yuliwulandari

1 Department of Histology, Faculty of Medicine, YARSI University, Jakarta, Indonesia
2 Genetic Research Center, YARSI University, Jakarta, Indonesia
3 Medical Genetics Section, Department of Medical Sciences, Ministry of Public Health, Nonthaburi, Thailand
4 Department of Pharmacology, Faculty of Medicine, YARSI University, Jakarta, Indonesia
5 The Pharmacogenomics Working Group, Indonesia

Tuberculosis (TB) is still remains as a major health problem in the world. Isoniazid (INH) is one of the most important drugs used in TB treatment that also the major drug involved in anti-tuberculosis drug induced liver injury (AT-DILI) incidence. INH is metabolized by arylamine N-acetyltransferase2 (NAT2), cytochrome P450 2E1 (CYP2E1) and glutathione S-transferase (GST). Our purpose was to examine the distribution in these enzymes genotypes as susceptibility factors to AT-DILI in Indonesian population. Our study showed that the frequency of slow acetylator (SA) status of NAT2 gene was high in these populations. Previous studies reported significant association of SA status with AT-DILI, with NAT2*6A/*6A genotype is frequently observed in Javanese-Sundanese (12.7%) and Buginese (18%) ethnics, while NAT2*6A/7B is the most frequent genotypes in Malay (14%) ethnic. Our previous study also showed that NAT2*6A allele was significantly associated with susceptibility to AT-DILI. CYP2E1 PstI/RsaI polymorphisms with a c1/c1 genotype are frequent in Indonesia population (83% in Javanese-Sundanese ethnics, 89% in Malay ethnics, and 79% in Buginese ethnics) than c1/c2 and c2/c2 genotypes. The GSTM1-null and GSTT1-null genotype was more frequent among Javanese-Sundanese ethnics (99%, 67%) than among the Indonesian Malay (67.2%, 36%) and Buginese (71%, 53%). The combined distribution of the GSTM1 and GSTT1 genes revealed that 66.7% of the individuals from the Javanese-Sundanese population lack both the genes, whereas only 21.1% in the Indonesian Malay and 36% in Buginese population. Our results will be helpful for future epidemiological or clinical studies and for understanding the genetic basis of INH metabolizing genes polymorphisms in Indonesian. This result also can help effective treatment decisions based on individual genotypes for AT-DILI prevention in TB treatment.
INULIN OF LESSER YAM (*Dioscorea Esculenta*) TUBER AS AN COLORECTAL ANTICANCER AGENT *IN SILICO*

Yuniastuti A\(^1\), Iswari RS\(^1\), Yanuarita T\(^2\)

\(^1\)Biology Department, Faculty of Mathematic and Natural Science Universitas Negeri Semarang, Semarang, Indonesia
\(^2\)Faculty of Medicine, University of Muhammadiyah Semarang

*Correspondence: ariyuniastuti@mail.unnes.ac.id*

Inulin is a fructant polymer compound whose main constituent consists of fructose units with β bonds (2 → 1) and has a degree of polymerization (DP) between 2-60. Inulin is thought to have antitumor activity and has been tested in vitro to induce apoptosis in WiDr colorectal cancer cells. The objective of this study was to determine the affinity and mechanism of gembili tuber (*Dioscorea esculenta*) inulin compound to the target protein of Caspase-3 as a colorectal anticancer in silico with molecular docking. The study was carried out exploratively with the stages of preparing the database of inulin 3D structure COX-2 and Caspase-3 protein, 3D inulin structure optimization, protein preparation, molecular docking methods validation, and inulin docking on the protein. Docking results were assessed from the bond energy and hydrogen bonds produced between inulin and protein COX-2 and Caspase-3. The smaller value of the binding energy, the bond between inulin and protein is stronger and more stable. The result showed that inulin has activity as a colorectal anticancer because it is able to inhibit COX-2 and induce Caspase-3 with a bond energy value of \(-10.56\) and \(-5.48\). This results indicate that inulin has the potential to induce apoptosis in colorectal cancer.
The 18th International Conference on Bioinformatics

Co-organised by:

UNIVERSITAS YARSI

APBioNET