## Workshop Notes

IJCAI-17
MELBOURNE

The 3rd international workshop on

# Advances in Bioinformatics and Artificial Intelligence: Bridging the Gap (BAI)

Melbourne, Australia, August 20, 2017
http://bioinfo.uqam.ca/IJCAI_BAI2017/

**Editors:**

| **Name** | **Affiliation** |
|---|---|
| Wajdi Dhifli | University of Lille<br>wajdi.dhifli@univ-lille2.fr<br>https://sites.google.com/site/wajdidhifli/ |
| Abdoulaye Baniré Diallo | University of Quebec at Montreal (Canada)<br>diallo.abdoulaye@uqam.ca<br>http://labo.bioinfo.uqam.ca |
| Engelbert Mephu Nguifo | University Clermont Auvergne (France)<br>mephu@isima.fr<br>http://www.isima.fr/mephu |
| Mohammed Javeed Zaki | Rensselaer Polytechnic Institute, NY (USA)<br>zaki@cs.rpi.edu<br>http://www.cs.rpi.edu/~zaki/ |

# Preface

The goal of this workshop called Bioinformatics and Artificial Intelligence (**BAI**) is to bring together active scholars and practitioners at the frontiers of Artificial Intelligence (AI) and Bioinformatics. AI holds a tremendous repertoire of algorithms and methods that constitute the core of different topics of bioinformatics and computational biology research. BAI goals are twofolds :

- How can AI techniques contribute to bioinformatics research?, and
- How can bioinformatics research raise new fundamental questions in AI?

Contributions clearly points out answers to one of these goals focusing on AI techniques as well as focusing on biological problems.

## Aims and Scope:

AI has played an increasingly important role in the analysis of sequence, structure and functional patterns or models from sequence databases. Bioinformatics aims to store, organize, explore, extract, analyze, interpret, and utilize information from biological data. The main outcome of this workshop is to present latest results in this exciting area at the intersection of biology and AI.

AI approaches can revolutionize new age of bioinformatics and computational biology with discoveries in basic biology, evolution, metagenomics, system biology, regulatory genomics, population genomics and diseases, structural bioinformatics, protein docking, next-generation sequencing (NGS) data processing, chemoinformatics, etc.

Bioinformatics provides opportunities for developing novel AI methods. Some of the grand challenges in bioinformatics include protein structure prediction, homology search, epigenetics, multiple alignment and phylogeny construction, genomic sequence analysis, gene finding and gene mapping, as well as applications in gene expression data analysis, drug discovery in pharmaceutical industry, etc.

Two questions were at the heart of this workshop :

- How can AI techniques contribute to Bioinformatics research, and in particular dealing with biological problems?
- How can Bioinformatics raise new fundamental research problem for AI research?

This one-day workshop aims at bringing together scholars and practitioners active in Artificial Intelligence driven Bioinformatics, to present and discuss their research, share their knowledge and experiences, and discuss the current state of the art and the future improvements to advance the ***intelligent*** practice of computational biology.

## Workshop Topics:

Topics of interest lie at the intersection of AI and Bioinformatics. They include, but are not limited to, the following inter-linked topics:

Artificial Intelligence :

- Constraints, satisfiability and search
- Knowledge representation, reasoning and logic
- Machine learning and data mining
- Planning and scheduling
- Agent-based and multi-agent systems
- Web and knowledge-based information systems
- Natural language processing
- Uncertainty

Bioinformatics :

- Comparative genomics
- Evolution and phylogenetics
- Epigenetics
- Functional genomics
- Genome organization and annotation
- Genetic variation analysis
- Metagenomics
- Pathogen informatics
- Population genetics, variation and evolution
- Protein structure and function prediction and analysis
- Proteomics
- Sequence analysis
- Systems biology and networks

## Workshop Contributions:

This year, the papers submitted to the workshop were carefully peer-reviewed by at least three members of the program committee and among the 10 submissions, 5 papers with the highest scores were selected. We would like to thank all the PC members and the reviewers for their reviews, as well as all the authors for their contributions. The workshop was a one day format with two keynote speakers and five oral presentations.

## Keynote Speakers:

The first keynote speaker was **Dr. Saman Halgamuge**, Professor and Director of Research School of Engineering at the Australian National University, Canberra, Australia. His talk was entitled : « *Unsupervised Deep Learning: Applications in Metagenomics, Metabolomics and Drug Characterisation* ». Most of the existing Deep Learning methods rely on the assumption that all possible class labels sufficient to apply Supervised Learning are available. Although these types of learning algorithms can be generalized, their predictive power will be heavily constrained in the presence of partial information of a problem. For example, the classes that are available to a classifier are assumed to be ground truth, and their correctness is not generally questioned. In contrast to this approach, we propose a learning framework where the number of classes within a dataset do not need to be known a priori, and more specifically, the entire set of class labels are not required at the time of training. Instead, we propose to develop a method that will be able to infer the number of classes based only on the data and generate a more representative set of classes to train a robust classifier. Furthermore, we will also relax the assumption that these class labels are ground truth, and allow a degree of uncertainty in their correctness. An interesting solution for a subclass of these problems is Positive Unlabelled Learning. Applying data analytics to microbial ecology has direct benefits to the design of vaccines and treatments to emerging pathogens, such as the Zika virus. In Metagenomic applications, very little may be known since we have only curated information pertaining to less than 2% of microbial diversity, and far less for novel variants of viruses. It is therefore not a realistic assumption that one can access all the true and underlying (organism) classes of any available data when analysing these organisms. Moreover, if we also consider the different and unknown number of effects that viral mutants can have on different hosts, and that these mutations could be linked to several environmental or geographical factors, we arrive at a complex, heterogeneous data set where labels are mostly unavailable, or any pre-existing labels available may be incorrect or not applicable to emerging viral strains. Even so, all these different types of data are essential to building a near-complete picture of the problem and understanding these pathogens at a deeper, more intimate level. Statistical analysis of DNA sequence data has previously assisted us in identification of features that may further be used to discriminate species in a sample of multiple organisms using unsupervised learning methods. Methods for increasing the resolution in realising the microbial population structure in a metagenomic sample is being worked on and coupling known data with unsupervised learning is found to be useful. Repositioning of existing drugs as appropriate medication for previously not associated medical conditions can reduce the time, costs and risks of drug development by identifying new therapeutic effects. Investigating and understanding the interactions between drugs as well as how they work on our body is important in improving the effectiveness of clinical care. A method based on Positive Unlabelled Learning and Growing Self Organising Maps is used on data available in DrugBank database. It was possible to infer 589 drug pairs that are likely to not interact with each other. Unsupervised Deep Learning also contributes in working with multielectrode array data.

The Second keynote speaker was **Dr. Shoba Ranganathan**, Professor of Bioinformatics at the Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, Australia. Her talk was entitled : « A protocol for finding missing proteins ». In the quest to uncover the entire human proteome, finding "missing proteins" remains the Holy Grail of scientists. In order to capture existing information, in addition to high-stringency MS data, we have launched the MissingProteinPedia (MPP; missingproteins.org), as an integrative biological database. While MPP incorporates automated data collection, novel tools for functional annotation and collated publications, there is an urgent need to identify a protocol for evaluating MPP data, to facilitate missing protein annotation jamborees. We will present how best to evaluate "extraordinary evidence" for missing proteins, with some exciting data confirming successful uncovering of some missing proteins.

## Oral Presentations:

The accepted papers were presented during the workshop.

## Workshop Program:

| Time | Event |
|---|---|
| **8:00 - 9:00** | **Registration and opening** |
| **9:00 - 10:00** | **Keynote speaker 1:** Prof. Saman Halgamuge<br>*Unsupervised Deep Learning: Applications in Metagenomics, Metabolomics and Drug Characterisation.* |
| **10:00 - 10:30** | Coffee Break |
| **10:30 - 11:00** | **Oral presentation:** Qingyu Chen, Xiuzhen Zhang, Yu Wan, Justin Zobel and Karin Verspoor.<br>*Sequence Clustering Methods and Completeness of Biological Database Search* |
| **11:00 - 11:30** | **Oral presentation:** Aidan O'Brien, Piotr Szul, Oscar Luo, Andrew George, Robert Dunne and Denis Bauer.<br>*Breaking the curse of dimensionality for machine learning on genomic data.* |
| **11:30 - 12:00** | **Oral presentation:** Alexandre Bazin, Didier Debroas and Engelbert Mephu Nguifo.<br>*A De Novo Robust Clustering Approach for Amplicon-Based Sequence Data.* |
| **12:00 - 14:00** | Lunch |
| **14:00 - 15:00** | **Keynote speaker 2:** Prof. Shoba Ranganathan<br>*A protocol for finding missing proteins.* |
| **15:00 - 15:30** | **Oral presentation:** Borut Budna, Martin Gjoreski, Anton Gradišek and Matjaz Gams.<br>*JSI Sound – a machine-learning tool in Orange for simple biosound classification.* |
| **15:30 - 16:00** | **Oral presentation:** Isabelle Bichindaritz and Thomas Quinn.<br>*Feature Selection and Deep Learning for Survival Analysis.* |
| **16:00 - 16:30** | Coffee Break |
| **16:30 - 17:00** | **Concluding remarks** |

## Program Committee:

Sabeur Aridhi, University of Lorraine (France)
Enrico Coiera, Macquarie University (Australia)
Annie Chateau, University of Montpellier 2 (France)
Sergio Vale Aguiar Campos, Federal University of Minas Gerais (Brasil)
Elisabetta De Maria, University of Nice Sophia-Antipoliss (France)
Wajdi Dhifli, University of Lille (France)
Abdoulaye Baniré Diallo, University of Quebec at Montreal (Canada)
Mohamed Elati, University of Evry-Val-d'Essonne (France)
Anna Gambin, University of Warsaw (Poland)
Simon de Givry, INRA Toulouse (France)
Tu-Bao Ho, JAIST School of Knowledge Science (Japan)
Attila Kertesz-Farkas, National Research University Higher School of Economics (Russia)
Chung-Shou Liao, National Tsing Hua University (Taiwan)
Fréderique Lisacek, Swiss Institute of Bioinformatics, Geneva (Switzerland)
Mondher Maddouri, Taibah University (Saudi Arabia)
Osamu Maruyama, Kyushu University (Japan)
Engelbert Mephu Nguifo, University Clermont Auvergne (France)
Claire Nedellec, INRA Jouy-en-josas (France)
Dave Ritchie, INRIA-LORIA, University Henry Poincaré, Nancy (France)
Sushmita Roy, University of Wisconsin, Madison (USA)
Marcilio de Souto, University of Orleans (France)
Dechang Xu, Harbin Institute of Technology (China)
Mohammed J. Zaki, Rensselaer Polytechnic Institute, NY (USA)

## Acknowledgements :

We would also like to thank all authors for contributing to our workshop and for their great presentation at the workshop. Furthermore, we thank all reviewers and subreviewers for their time and efforts in helping us build an interesting program.

*The Workshop chairs*
Wajdi Dhifli
Abdoulaye Banier Diallo
Engelbert Mephu Nguifo
Mohammed Javeed Zaki