ECCB 2020

Virtual!

# 19th

# European Conference on Computational Biology

## Planetary Health and Biodiversity

## BOOK OF ABSTRACTS

31st August - 8th September, 2020

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

INB Spanish National Bioinformatics Institute

elixir SPAIN

eccb20.org

FEDER
Fondo Europeo de Desarrollo Regional
UNIÓN EUROPEA
"Una manera de hacer Europa"

GOBIERNO DE ESPAÑA

MINISTERIO DE CIENCIA E INNOVACIÓN

Instituto de Salud Carlos III

# Index

# Presentation

Dear colleague,

We are delighted to welcome you to the **19th European Conference on Computational Biology (ECCB2020)** under the motto Planetary Health and Biodiversity that will take place **from August 31st until September 8th, 2020**. Due to the current circumstances surrounding COVID-19, the ECCB2020 will offer a reduced and virtual event full of scientifically attractive talks for nine days. The **ECCB2020 main conference on September 7th and 8th** will offer a two-day programme with 6 distinguished keynote speakers and 41 insightful themed talks selected from delegates submissions as well as the newly session entitled "A glimpse into Global Bioinformatics Communities" in collaboration with SolBio (Iberoamerican Society for Bioinformatics) and ELIXIR as Latin American and European representatives, respectively. The week prior to the main event, delegates will be invited to join the **New Trends in Bioinformatics by ECCB**, a newly designed format to virtually host a selection of ECCB2020 workshops and tutorials. The 6th European Student Council Symposium (ESCS 2020) will take place virtually on September 6th.

ECCB2020 welcomes scientists working in a variety of disciplines, including bioinformatics, computational biology, systems biology, artificial intelligence, biology, medicine, environmental sciences, and many more. Participating in ECCB2020 will be the perfect opportunity to keep pace with cutting edge research while interacting with a diverse and broad representation of the research community across many domains in Life Sciences and beyond. We are sure that the conference will facilitate knowledge dissemination of great interest and benefit, vivid scientific debates and collaborations, enjoyable interactions through virtual platforms and will stimulate a creative exchange of ideas.

We would also like to express our warm thanks to our kind sponsors and exhibitors for their generous support on the ECCB2020.

For sure it will be a unique event, so we invite you to join us at the ECCB2020!

With best regards,

**Alfonso Valencia & Salvador Capella-Gutierrez on behalf of the ECCB2020**

# Committees

## Steering Committee

- **Alfonso Valencia** | Barcelona Supercomputing Center (BSC), INB/ELIXIR-ES, Spain
- **Salvador Capella-Gutierrez** | Barcelona Supercomputing Center (BSC), INB/ELIXIR-ES, Spain
- **Arcadi Navarro** | University Pompeu Fabra (UPF), INB/ELIXIR-ES, Spain
- **Fatima Al-Shahrour** | Spanish National Cancer Research Center (CNIO), INB/ELIXIR-ES, Spain
- **Christine Orengo** | University College London (UCL), ELIXIR-UK, United Kingdom

## Organising Committee

- **Salvador Capella-Gutierrez** | Barcelona Supercomputing Center (BSC), INB/ELIXIR-ES, Spain
- **Eva Alloza** | Barcelona Supercomputing Center (BSC), INB/ELIXIR-ES, Spain
- **Edurne Gallastegui** | Barcelona Supercomputing Center (BSC), INB/ELIXIR-ES, Spain
- **Jen Harrow** | ELIXIR, United Kingdom
- **Ioannis Kavakiotis** | University of Thessaly, Hellenic Pasteur Institute, ECCB2018 representative, Greece
- **Alberto Langtry Yáñez** | Spanish National Cancer Research Center (CNIO), ISCB-SC RSG Spain representative, Spain

## Workshops/SIGs

- **Ferran Sanz** | Hospital del Mar Medical Research Institute (IMIM), Pompeu Fabra University (UPF), INB/ELIXIR-ES, Spain (Chair)
- **Rita Casadio** | University of Bologna, Italy
- **Ioannis Xenarios** | University of Lausanne, Switzerland
- **Chris Evelo** | Maastricht University, The Netherlands
- **Hedi Peterson** | University of Tartu, Estonia
- **Olivier Taboureau** | Paris Diderot University, France

## Tutorials

- **Oswaldo Trelles** | University of Malaga (UMA), INB/ELIXIR-ES, Spain (Chair)
- **Pedro L. Fernandes** | Instituto Gulbenkian de Ciência (IGC), Portugal
- **Celia van Gelder** | Dutch Techcentre for Life Sciences (DTL), The Netherlands
- **Fotis Psomopoulos** | Center for Research and Technology Hellas (CERTH), Greece
- **Gonzalo Claros** | University of Malaga (UMA), Spain
- **Sonia Tarazona** | Polytechnic University of Valencia (UPV), Spain
- **Francisco García García** | Príncipe Felipe Research Center (CIPF), Spain

# Committees

## Topics Chairs & co-Chairs

### Topic: DATA

- **Josep Lluís Gelpí** | University of Barcelona, Barcelona Supercomputing Center (BSC), INB/ELIXIR-ES, Spain (Chair)

### Topic: GENES

- **Ana Conesa** | University of Florida, United States (Chair)
- **Mark Robinson** | University of Zurich, Switzerland
- **Jacques van Helden** | Institut Français de Bioinformatique (FIB), Aix-Marseille University, France
- **Artemis Hatzigeorgiou** | University of Thessaly, Hellenic Pasteur Institute, Greece

### Topic: GENOME

- **Toni Gabaldón** | Barcelona Supercomputing Center (BSC), Institute for Research in Biomedicine (IRB Barcelona), INB/ELIXIR-ES, Spain (Chair)
- **Stephane Rombauts** | Vlaams Instituut voor Biotechnologie (VIB), University of Gent, Belgium

### Topic: PROTEINS

- **Modesto Orozco** | Institute for Research in Biomedicine (IRB Barcelona), INB/ELIXIR-ES, Spain (Chair)
- **Mark Wass** | University of Kent, United Kingdom

### Topic: SYSTEMS

- **Patrick Aloy** | Institute for Research in Biomedicine (IRB Barcelona), INB/ELIXIR-ES, Spain (Chair)
- **Amelie Stein** | University of Copenhagen, Denmark
- **Pedro Beltrao** | EMBL-European Bioinformatics Institute (EBI), United Kingdom

## Programme Committee Reviewers

### Topic: DATA

Animesh Acharjee
Yeşim Aydin Son
Michael Bada
Nina Baumgarten
Fatemeh Behjati Ardakani
Takis Benos
Olivier Bodenreider
Valentina Boeva
Kristina Buschur
Wei Chen
Maria Chikina
Dennis Hecker
Robert Hoehndorf
Andreas Karwath
Ioannis Kavakiotis
Kiavash Kianfar
Maxat Kulmanov
Ranjan Kumar Maji
Dimitrios Manatakis
Robert A. Mcdougal
Natasa Miskov-Zivanov
Arzucan Ozgur
Athanasia Pavlopoulou
Natapol Pornputtapong
Vineet Raghu
Paul Schofield
Marcel Schulz
Luke Slater
Oznur Tastan
Margarita C. Theodoropoulou
Georgios Tsaousis
Sophia Tsok
Xiaoqian Wang
Xuefeng Wang

# Committees

## Topic: GENES

Sara Aibar
Nikolaos Alachiotis
Claudia Angelini
Philipp Bucher
Domenica D'Elia
Dick De Ridder
Slavica Dimitrieva
Alexandre Ferreira-Ramos
Tomas Flouri
Mudassar Iqbal
Hosna Jabbari
Zeynep Kalender-Atak
Dimitra Karagkouni
Despina Kontos
Anshul Kundaje
Harri Lähdesmäki
Ahmad Mahfouz
Tereza Manousaki
Manolis Maragkakis
Fabio Marroni
Maria Paraskevopoulou
Chris Penfold
Cécile Pereira
Teresa Przytycka
Yanjun Qi
Davide Risso
Łukasz Roguski
Roland Schwarz
Rainer Spang
Terence Speed
Camille Stephan-Otto-Attolini
Simone Tiberi
Bartek Wilczynski

## Topic: GENOME

Mohamed Abouelhoda
Tyler Alioto
Jasmijn Baaijens
Rolf Backofen
Endre Barta
Markus Bauer
Philipp Benner
Sebastian Böcker
Michael Charleston
Spyridon Chavlis
Rayan Chikhi
Panagiotis Chouvardas
Thomas Derrien
Van Hoan Do
Kai Dührkop
Nadia El-Mabrouk
Markus Fleischauer
Caroline Friedel
Erik Garrison
George Giannakopoulos
Stefan Haas
Bernhard Haubold
David Heller
Ralf Herwig
Steve Hoffmann
Manuel Holtgrewe
Wolfgang Huber
Daniel Huson
Panagiotis Ioannidis
Katharina Jahn
Irene Julca
Prabhav Kalaghatgi
Sivarajan Karunanithi
Christophe Klopp
Helene Kretzmer
Anastasia Krithara
Matthias Lienhard
Antoine Limasset
Gen Lin
Marina Marcet
Burkhard Morgenstern
Alena Mysickova
Christoforos Nikolaou
Nikolaos Panoussis
Anthony Papenfuss
Cinta Pegueroles
Nico Pfeifer
Solon Pissis
Christopher Pockrandt
Dimitris Polychronopoulos
Bui Quang-Minh
René Rahn
Thomas Rattei
Knut Reinert
Stephane Rombauts
Marcel Schulz
Alexander Schönhuth
Karel Sedlar
Fritz Sedlazeck
Philip Stevens
Jens Stoye
Wing-Kin Sung
Krister Swenson
Morgane Thomas-Chollier
Tanya Vavouri
Martin Vingron
Tim Warwick
Emanuel Weitschek
Sebastian Will
Ralf Zimmer
Matthias Zytnick

# Committees

### Topic: PROTEINS

Xavier Barril
Asa Ben-Hur
Anne-Claude Camproux
Alessandra Carbone
Rita Casadio
Chakra Chennubhotla
Evangelia Chrysina
Murat Can Cobanoglu
Thomas Dandekar
Xavier Daura
Narcis Fernandez-Fuentes
Juan Fernandez-Recio
Oriol Fornés Crespo
Franca Fraternali
Anthony Gitter
Nicholas M. Glykos
F. Xavier Gomis-Rüth
Attila Gursoy
Des Higgins
Liisa Holm
Vassiliki Iconomidou
Hosna Jabbari
David Juan
Daisuke Kihara
Andrzej Kloczkowski
David Koes
Anália Lourenço
Lars Malmstroem
Henry Martell
Liam McGuffin
Markus Nebel
Irilenia Nobeli
Baldo Oliva
Sandra Orchard
Florencio Pazos
Andreas Prlic
Defne Surujon
Tsung-Heng Tsai
Wim Vranken

### Topic: SYSTEMS

Mohammed Alquraishi
Xavier Barril
Anais Baudot
Richard Bonneau
Christine Brun
Daria Bunina
Eugenio Cinquemani
Davide Cirillo
Diego Di Bernardo
Nadezhda T. Doncheva
Miquel Duran-Frigola
Mirjana Efremova
Juan Fernandez-Recio
Adria Fernandez-Torras
Laura I. Furlong
Atilla Gabor
Anthony Gitter
Traver Hart
Ioannis Kavakiotis
Katja Luck
Julien Martinelli
Chad Myers
Kay Nieselt
Baldo Oliva
Carles Pons
Delphine Ropers
Colm Ryan
Julio Saez-Rodríguez
Sylvain Solyman
Denis Thieffry
Andreas Zanzoni

# Sponsors

## Organising Institutions

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

INB Spanish National Bioinformatics Institute

elixir SPAIN

FEDER Fondo Europeo de Desarrollo Regional
UNIÓN EUROPEA
"Una manera de hacer Europa"
GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN
isciii Instituto de Salud Carlos III

## Our Kind Sponsors

### Gold Sponsor

EMBL-EBI

### Bronze Sponsor

GOBLET

### Strategic Sponsor

elixir

### Institutional Sponsors

iSCB INTERNATIONAL SOCIETY FOR COMPUTATIONAL BIOLOGY

RES RED ESPAÑOLA DE SUPERCOMPUTACIÓN

St. Jude Children's® Research Hospital
Finding cures. Saving children.
ALSAC • DANNY THOMAS, FOUNDER

BIO INFO 4WOMEN

### Official Sponsors

GCAT TACG GCAT genes
an Open Access Journal by MDPI

CAMBRIDGE UNIVERSITY PRESS

GRIFOLS

# Programme at a glance

| CEST | MONDAY, 7 September | | |
|---|---|---|---|
| 09.00 - 10.00 | Keynote by **Modesto Orozco** on **Simulation of DNA. from the atom to the chromatin**<br>Chaired by **Alfonso Valencia** | | |
| 10.00 - 11.00 | **Parallel Track #01 Genomes**<br>Chaired by **Stephane Rombauts** | **Parallel Track #02 Systems**<br>Chaiered by **Patrick Aloy** | **Parallel Track #03 Data**<br>Chaired by **Josep Lluís Gelpí** |
| | **Huan Shi** \| A general near-exact k-mer counting method with low memory consumption enables de novo assembly of 106x human sequence data in 2.7 hours<br>**Maor Asif** \| DeepSELEX: Inferring DNA-binding preferences from HT-SELEX data using multi- class CNNs<br>**Emilio Dorigatti** \| Joint epitope selection and spacer design for string-of-beads vaccines<br>**Adelme Bazin** \| panRGP: a pangenome-based method to predict genomic islands and explore their diversity | **Shohag Barman** \| A Neuro-Evolution Approach to Infer a Boolean Network from Time-Series Gene Expressions<br>**Taiki Fuji** \| Feasible-Metabolic-Pathway-Exploration Technique using Chemical Latent Space<br>**Yahui Long** \| Ensembling graph attention networks for human microbe-drug association prediction<br>**Gian Marco Messa** \| A Siamese Neural Network model for theprioritization of metabolic disorders by integratingreal and simulated data | **Dong-gi Lee** \| Dementia Key Gene Identification with Multi- Layered SNP-Gene-Disease Network<br>**Thomas Gumbsch** \| Enhancing statistical power in temporal biomarker discovery through representative shapelet mining<br>**Tamim Abdelaal** \| SCHNEL: Scalable clustering of high dimensional single-cell data<br>**Derrick Blakely** \| FastSK: Fast Sequence Analysis with Gapped String Kernels |
| 11.00 - 11.20 | BioBreak | | |
| 11.20 - 12.10 | Keynote by **Geneviève Almouzni** on **Chromatin plasticity, cell fate and identity**<br>Chaired by **Christine Orengo** | | |
| 12.10 - 13.10 | **Parallel Track #04 Genomes**<br>Chaired by **Toni Gabaldón** | **Parallel Track #05 Systems**<br>Chaired by **Patrick Aloy** | **Parallel Track #06 Data**<br>Chaired by **Josep Lluís Gelpí** |
| | **Jack Lanchantin** \| Graph Convolutional Networks for Epigenetic State Prediction Using Both Sequence and 3D Genome Data<br>**Huy Nguyen** \| Finding Orthologous Gene Blocks in Bacteria: The Computational Hardness of the Problem and Novel Methods to Address it<br>**Sayaka Miura** \| PathFinder: Bayesian inference of clone migration histories in cancer<br>**Sarah Christensen** \| Detecting Evolutionary Patterns of Cancers using Consensus Trees | **Salvador Casani** \| Padhoc: A computational pipeline for Pathway Reconstruction On the Fly<br>**Sergio Doria-Belenguer** \| Probabilistic Graphlets Capture Biological Function in Probabilistic Molecular Networks<br>**Alina Renz** \| FBA reveals guanylate kinase as a potential target for antiviral therapies against SARS-CoV-2<br>**David Merrell** \| Inferring Signaling Pathways with Probabilistic Programming | **Jason Fan** \| Matrix (Factorization) Reloaded: Flexible Methods for Imputing Genetic Interactions with Cross-Species and Side Information<br>**Wesley Qian** \| Batch Equalization with a Generative Adversarial Network<br>**Jose Barba** \| Using a GTR+Γ substitution model for dating sequence divergence when stationarity and time-reversibility assumptions are violated<br>**Elliot Layne** \| Supervised learning on phylogenetically distributed data<br>**Kerem Ayoz** \| The Effect of Kinship in Re-identification Attacks Against Genomic Data Sharing Beacons |
| 13.10 - 13.30 | **EMBL-EBI**<br>**European Bioinformatics Institute** | | |
| 13.30 - 15.00 | Break | | |
| 15.00 - 16.20 | **A glimpse into Global Bioinformatics Communities: Latin America - SolBio**<br>Chaired by **Javier De Las Rivas** | | |
| | **Benilton Carvalho** \| The Brazilian Initiative on Precision Medicine: Strategies and Findings<br>**Gregorio Iraola** \| Building city-scale genomic cartographies for improved response to emerging infectious diseases<br>**Wendy González Diaz** \| Molecular Modeling of Ion Channels-Associated Diseases<br>**Alejandra Medina Rivera** \| Logical modeling of dendritic cells in vitro differentiation from human monocytes unravels novel transcriptional regulatory interactions | | |
| 16.20 - 16.30 | BioBreak | | |
| 16.30 - 17.20 | Keynote by **Deborah Marks** on **Prediction and design of biological sequences with neural machines**<br>Chaired by **Baldo Oliva** | | |
| 17.20 - 17.30 | **Announcements** | | |
| 17.30 - 19.00 | Network & Connect | | |

# Programme at a glance

| CEST | TUESDAY, 8 September | | | |
|---|---|---|---|---|
| 09.00 - 09.10 | Announcements | | | |
| 09.10 - 10.00 | Keynote by **Fabian Theis** on **Modeling cellular state and dynamics in single cell genomics**<br>Chaired by **Marc Martí-Renom** | | | |
| 10.00 - 11.00 | **Parallel Track #07 -** Chaired by **Toni Gabaldón** | | **Parallel Track #08 -** Chaired by **Ana Conesa** | |
| | Data | Genomes | Genes | Protein |
| | **PhD. Qiao Liu** \| DeepCDR: a hybrid graph convolutional network for predicting cancer drug response<br>**Joanna Ficek** \| SCIM: Universal Single-Cell Matching with Unpaired Feature Sets | | **Ayse Dincer** \| Adversarial Deconfounding Autoencoder for Learning Robust Gene Expression Embeddings<br>**Vu Viet Hoang Pham** \| DriverGroup: A novel method for identifying driver gene groups | |
| | **Marleen Nieboer** \| svMIL: Predicting the pathogenic effect of TAD boundary-disrupting somatic structural variants through multiple instance learning<br>**Luca Nanni** \| Exploring Chromatin Conformation and Gene Co-Expression through Graph Embedding | | **Wenjing Xuan** \| CLPred: A sequence-based protein crystallization predictor using BLSTM neural network<br>**Janani Durairaj** \| Geometricus Represents Protein Structures as Shape-mers Derived from Moment Invariants | |
| 11.00 - 11.20 | BioBreak | | | |
| 11.20 - 12.10 | Keynote by **Bissan Al-Lazikani** on **More than the sum of parts: Multidiciplinary big data in cancer therapy**<br>Chaired by **Patrick Aloy** | | | |
| 12.10 - 13.10 | **Parallel Track #09 Genes**<br>Chaired by **Artemis Hatzigeorgiou** | | **Parallel Track #10 Proteins**<br>Chaired by **Mark Wass** | |
| | **Milad Mokhtaridoost** \| An efficient framework to identify key miRNA-mRNA regulatory modules in cancer<br>**Joske Ubels** \| RAINFOREST: A random forest approach to predict treatment benefit in data from (failed) clinical drug trials<br>**Mohammad Lotfollahi** \| Conditional out-of-sample generation for un-paired data using transfer VAE<br>**Guillem Ylla** \| MirCure: A tool for quality control, filter, and cu-ration of microRNAs of animals and plants | | **Sabrina de Azevedo Silveira** \| GRaSP: a graph-based residue neighborhood strategy to predict binding sites<br>**Lukasz Kurgan** \| PROBselect: accurate prediction of protein- binding residues from proteins sequences via dynamic predictor selection<br>**Yisu Peng** \| New mixture models for decoy-free false discovery rate estimation in mass-spectrometry proteomics<br>**Zhenling Peng** \| APOD: accurate sequence-based predictor of disordered flexible linkers | |
| 13.10 - 13.30 | **RES**<br>**Spanish Supercomputing Network** | | | |
| 13.30 - 15.00 | Break | | | |
| 15.00 - 16.20 | **A glimpse into Global Bioinformatics Communities: ELIXIR**<br>Chaired by **Jennifer Harrow** | | | |
| | **Sameer Velankar \|** 3D-Beacons: An integrative, distributed platform for FAIR access to experimental and predicted macromolecular structures<br>**Jose Ramon Macias Gonzalez \|** 3DBionotes-COVID19 Edition: bringing together structural and functional information on SARS-CoV-2 proteome<br>**Deborah Caucheteur \| COVoc:** a COVID-19 ontology to support literature triage<br>**Bjoern Gruening \|** The ELIXIR Tools Platform - solutions for COVID-19 research<br>**Azab Abdulrahman \|** Nordic development on federating the EGA<br>**Flora D'Anna \|** FAIR data by design | | | |
| 16.20 - 16.30 | BioBreak | | | |
| 16.30 - 17.20 | Keynote by **Londa Schiebinger** on **Gendered Innovations in Biomedicine, Machine Learning, and Robotics**<br>Chaired by **Niklas Blomberg** | | | |
| 17.20 - 17.30 | ECCB2020 - Virtual Closing | | | |

# Scientific Programme

## Monday, September 7, 2020

**Time indicated in CEST**

09:00 - 10:00 **Keynote speaker** - **Modesto Orozco**
Chaired by **Alfonso Valencia**
*Barcelona Supercomputing Center (BSC), Spain*

09:10 - 10:00 Simulation of DNA. from the atom to the chromatin
**Modesto Orozco**
*Institute for Research in Biomedicine (IRB Barcelona), Spain*

10:00 - 11:00 **Parallel Track #01**
Chaired by **Stephane Rombauts**
*Center for Plant System Biology, VIB-UGent, Belgium*

10:00 - 10:15 A general near-exact k-mer counting method with low memory consumption enables de novo assembly of 106x human sequence data in 2.7 hours
**Christina Huan Shi**
*The Chinese University of Hong Kong, Hong Kong*

10:15 - 10:30 DeepSELEX: Inferring DNA-binding preferences from HT-SELEX data using multi-class CNNs
**Maor Asif**
*Ben-Gurion University, Israel*

10:30 - 10:45 Joint epitope selection and spacer design for string-of-beads vaccines
**Emilio Dorigatti**
*Ludwig Maximilian Universitaet Muenchen, Germany*

10:45 - 11:00 panRGP: a pangenome-based method to predict genomic islands and explore their diversity
**Adelme BAZIN**
*LABGeM - UMR8030, Genoscope, France*

10:00 - 11:00 **Parallel Track #02**
Chaired by **Patrick Aloy**
Institute for Research in Biomedicine (IRB Barcelona), Spain

10:00 - 10:15 A Neuro-Evolution Approach to Infer a Boolean Network from Time-Series Gene Expressions
**Shohag Barman**
*American International University-Bangladesh, Bangladesh*

10:15 - 10:30 Feasible-Metabolic-Pathway-Exploration Technique using Chemical Latent Space
**Taiki Fuji**
*Hitachi Ltd., Japan*

10:30 - 10:45 Ensembling graph attention networks for human microbe-drug association prediction
**Yahui Long**
*Nanyang Technological University, China*

10:45 - 11:00 A Siamese Neural Network model for theprioritization of metabolic disorders by integratingreal and simulated data
**Gian Marco Messa**
*King Abdullah University of Science and Technology (KAUST), Saudi Arabia*

10:00 - 11:00 **Parallel Track #03**
Chaired by **Josep Ll. Gelpi**
*University of Barcelona - Barcelona Supercomputing Center (BSC), Spain*

10:00 - 10:15 Dementia Key Gene Identification with Multi-Layered SNP-Gene-Disease Network
**Dong-gi Lee**
*Ajou University, South Korea*

10:15 - 10:30 Enhancing statistical power in temporal biomarker discovery through representative shapelet mining
**Thomas Gumbsch**
*ETH Zürich, Switzerland*

10:30 - 10:45 SCHNEL: Scalable clustering of high dimensional single-cell data
**Tamim Abdelaal**
*Delft University of Technology, Netherlands*

10:45 - 11:00 FastSK: Fast Sequence Analysis with Gapped String Kernels
**Yanjun Qi**
*University of Virginia, United States*

11:00 - 11:20 BioBreak

Monday, September 7, 2020

11:20 - 12:10 **Keynote speaker** - **Geneviève Almouzni**
Chaired by **Christine Orengo**
*University College London, United Kingdom*

11:20 - 12:10 Chromatin plasticity, cell fate and identity
**Geneviève Almouzni**
*Chromatin Dynamics team at Institut Curie, Science Academy in France, LifeTime Initiative, France*

12:10 - 13:10 **Parallel Track #04**
Chaired by **Toni Gabaldón**
*Barcelona Supercomputing Center (BSC), Spain*

12:10 - 12:25 Graph Convolutional Networks for Epigenetic State Prediction Using Both Sequence and 3D Genome Data

**Jack Lanchantin**
*University of Virginia, United States*

12:25 - 12:40 Finding Orthologous Gene Blocks in Bacteria: The Computational Hardness of the Problem and Novel Methods to Address it

**Huy Nguyen**
*Iowa State University, United States*

12:40 - 12:55 PathFinder: Bayesian inference of clone migration histories in cancer

**Sayaka Miura**
*Temple University, United States*

12:55 - 13:10 Detecting Evolutionary Patterns of Cancers using Consensus Trees

**Sarah Christensen**
*University of Illinois at Urbana-Champaign, United States*

12:10 - 13:10 **Parallel Track #05**

Chaired by **Patrick Aloy**
*Institute for Research in Biomedicine (IRB Barcelona), Spain*

12:10 - 12:25 Padhoc: A computational pipeline for Pathway Reconstruction On the Fly

**Salvador Casani**
*BioBam Bioinformatics, Spain*

12:25 - 12:40 Probabilistic Graphlets Capture Biological Function in Probabilistic Molecular Networks

**Sergio Doria-Belenguer**
*Barcelona Supercomputing Center (BSC), Spain*

Monday, September 7, 2020

12:40 - 12:55 FBA reveals guanylate kinase as a potential target for antiviral therapies against SARS-CoV-2

**Alina Renz**
*Computational Systems Biology, University of Tübingen, Germany*

12:55 - 13:10 Inferring Signaling Pathways with Probabilistic Programming

**David Merrell**
*University of Wisconsin - Madison, United States*

12:10 - 13:10 **Parallel Track #06**

Chaired by **Josep Ll. Gelpi**
*University of Barcelona - Barcelona Supercomputing Center (BSC), Spain*

12:10 - 12:25 Matrix (Factorization) Reloaded: Flexible Methods for Imputing Genetic Interactions with Cross-Species and Side Information
**Jason Fan**
*University of Maryland, United States*

12:25 - 12:40 Batch Equalization with a Generative Adversarial Network
**Wesley Qian**
*University of Illinois at Urbana-Champaign, United States*

12:40 - 12:55 Using a GTR+Γsubstitution model for dating sequence divergence when stationarity and time-reversibility assumptions are violated
**Jose Barba-Montoya**
*Temple University, United States*

12:55 - 13:10 Supervised learning on phylogenetically distributed data
**Elliot Layne**
*McGill University, Canada*

13:10 - 13:25 The Effect of Kinship in Re-identification Attacks Against Genomic Data Sharing Beacons
**Kerem Ayoz**
*Bilkent University, Turkey*

13:10 - 13:30 **Meet our sponsor: EMBL-EBI- European Bioinformatics Institute**

> 13:10 - 13:30 What's new at EMBL-EBI?
> **Amonida Zadissa**
> **Sarah Morgan**
> *EMBL-EBI - European Bioinformatics Institute, United Kingdom*

13:30 - 15:00 Break || Meet our exhibitors

Monday, September 7, 2020

15:00 - 16:20 **A glimpse into Global Bioinformatics Communities: Latin America - SolBio**
Chaired by **Javier De Las Rivas**
*Spanish National Research Council (CSIC), Universtity of Salamanca (USAL), Spain*

15:00 - 15:20 The Brazilian Initiative on Precision Medicine: Strategies and Findings
**Benilton Carvalho**
*University of Campinas, Brazil*

15:20 - 15:40 Building city-scale genomic cartographies for improved response to emerging infectious diseases
**Gregorio Iraola**
*Microbial Genomics Laboratory, Institut Pasteur de Montevideo, Uruguay*

15:40 - 16:00 Molecular Modeling of Ion Channels-Associated Diseases

**Wendy González Diaz**

*Center for Bioinformatics and Molecular Simulations (CBSM), University of Talca, Chile*

16:00 - 16:20 Logical modeling of dendritic cells in vitro differentiation from human monocytes unravels novel transcriptional regulatory interactions

**Alejandra Medina Rivera**

*Universidad Nacional Autonoma de Mexico, Mexico*

16:20 - 16:30 BioBreak

16:30 - 17:30 **Keynote speaker - Debora Marks**

Chaired by **Baldo Oliva**

*Pompeu Fabra University, Spain*

16:30 - 17:20 Prediction and design of biological sequences with neural machines

**Debora Marks**

*Department of Systems Biology, Harvard Medical School, United States*

17:30 - 19:00 Network & Connect || Meet our exhibitors

## Tuesday, September 8, 2020

Tuesday, September 8, 2020

09:00 - 10:00 **Keynote speaker - Fabian Theis**
Chaired by **Marc Marti-Renom**
CNAG-CRG, *Spain*

09:00 - 10:00 Modeling cellular state and dynamics in single cell genomics
**Fabian Theis**
*Institute of Computational Biology, Helmholtz Zentrum München, Germany*

10:00 - 11:00 **Parallel Track #07**
Chaired by **Toni Gabaldón**
*Barcelona Supercomputing Center (BSC), Spain*

10:00 - 10:15 DeepCDR: a hybrid graph convolutional network for predicting cancer drug response
**Qiao Liu**
*Tsinghua University, China*

10:15 - 10:30 SCIM: Universal Single-Cell Matching with Unpaired Feature Sets
**Joanna Ficek**
*ETH Zürich, Switzerland*

10:30 - 10:45 svMIL: Predicting the pathogenic effect of TAD boundary-disrupting somatic structural variants through multiple instance learning
**Marleen Nieboer**
*UMC Utrecht, Netherlands*

10:45 - 11:00 Exploring Chromatin Conformation and Gene Co-Expression through Graph Embedding
**Luca Nanni**
*Politecnico di Milano, Italy*

10:00 - 11:00 **Parallel Track #08**
Chaired by **Ana Conesa**
*University of Florida, United States*

10:00 - 10:15 Adversarial Deconfounding Autoencoder for Learning Robust Gene Expression Embeddings
**Ayse Dincer**
*University of Washington, United States*

10:15 - 10:30 DriverGroup: A novel method for identifying driver gene groups
**Vu Viet Hoang Pham**
*University of South Australia, Australia*

Tuesday, September 8, 2020

10:30 - 10:45 CLPred: A sequence-based protein crystallization predictor using BLSTM neural network
**Wenjing Xuan**
*Central South University, China*

10:45 - 11:00 Geometricus Represents Protein Structures as Shape-mers Derived from Moment Invariants
**Janani Durairaj**
*Wageningen University, Netherlands*

11:00 - 11:20 BioBreak

11:20 - 12:10 **Keynote speaker - Bissan Al-Lazikani**
Chaired by **Patrick Aloy**
*Institute for Research in Biomedicine (IRB Barcelona), Spain*

11:20 - 12:10 More than the sum of parts: Multidiciplinary big data in cancer therapy
**Bissan Al-Lazikani**
*The Institute of Cancer Research, United Kingdom*

12:10 - 13:10 **Parallel Track #09**
Chaired by **Artemis Hatzigeorgiou**
*University of Thessaly, Hellenic Pasteur Institute, Greece*

12:10 - 12:25 An efficient framework to identify key miRNA-mRNA regulatory modules in cancer
**Milad Mokhtaridoost**
*Koç University, Turkey*

12:25 - 12:40 RAINFOREST: A random forest approach to predict treatment benefit in data from (failed) clinical drug trials
**Joske Ubels**
*UMC Utrecht, Netherlands*
r
12:40 - 12:55 Conditional out-of-sample generation for un-paired data using transfer VAE
**Mohammad Lotfollahi**
*Helmholtz Zentrum München, Germany*

12:55 - 13:10 MirCure: A tool for quality control, filter, and cu-ration of microRNAs of animals and plants
**Guillem Ylla**
*Harvard University, United States*
Tuesday, September 8, 2020

12:10 - 13:10 **Parallel Track #10**
Chaired by **Mark Wass**
*University of Kent, United Kingdom*

12:10 - 12:25 GRaSP: a graph-based residue neighborhood strategy to predict binding sites
**Sabrina de Azevedo Silveira**
*Universidade Federal de Viçosa, Brazil*

12:25 - 12:40 PROBselect: accurate prediction of protein-binding residues from proteins sequences via dynamic predictor selection
**Lukasz Kurgan**
*Virginia Commonwealth University, United States*

12:40 - 12:55 New mixture models for decoy-free false discovery rate estimation in mass-spectrometry proteomics
**Yisu Peng**
*Northeastern University, United States*

12:55 - 13:10 APOD: accurate sequence-based predictor of disordered flexible linkers
**Zhenling Peng**
*Tianjin University, China*

13:10 - 13:30 **Meet our sponsor: RES - Spanish Supercomputing Network**

13:10 - 13:30 The Spanish Supercomputing Network offers HPC resources to the scientific community
**Jordi Mas**
*Spanish Supercomputing Network (Red Española de Supercomputación RES), Spain*

13:30 - 15:00 Break || Meet our exhibitors

15:00 - 16:20 **A glimpse into Global Bioinformatics Communities: Europe - ELIXIR**
Chaired by **Jen Harrow**
*ELIXIR, United Kingdom*

15:00 - 15:13 3D-Beacons: An integrative, distributed platform for FAIR access to experimental and predicted macromolecular structures
**Sameer Velankar**
*Protein Data Bank In Europe, EMBL-EBI, United Kingdom*

15:13 - 15:26 3DBionotes-COVID19 Edition: bringing together structural and functional information on SARS-CoV-2 proteome
**Jose Ramon Macias Gonzalez**
*Biocomputing Unit, CNB-CSIC, Spain*

15:26 - 15:39 COVoc: a COVID-19 ontology to support literature triage
**Déborah Caucheteur**
*HES-SO Genève / Swiss Institute of Bioinformatics (SIB), Switzerland*

15:39 - 15:52 The ELIXIR Tools Platform - solutions for COVID-19 research
**Bjoern Gruening**
*University of Freiburg, Germany*

15:52 - 16:05 Nordic development on federating the EGA
**Abdulrahman Azab Mohamed**
*University Center for Information Technology, University of Oslo, Norway*

16:05 - 16:18 FAIR data by design
**Flora D'Anna**
*VIB, Belgium*

16:20 - 16:30 BioBreak

16:30 - 17:30 **Keynote speaker - Londa Schiebinger || ECCB2020-Virtual Closing**
Chaired by **Niklas Blomberg**
*ELIXIR, United Kingdom*

16:30 - 17:20 Gendered Innovations in Biomedicine, Machine Learning, and Robotics
**Londa Schiebinger**
*Stanford University, United States*

# Pre-meetings

New Trends in Bioinformatics by ECCB | 31st August - 4th September 2020

ECCB 2020
Virtual!
BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación
INB Spanish National Bioinformatics Institute
elixir SPAIN
FEDER
UNIÓN EUROPEA
GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN
Instituto de Salud Carlos III

**New Trends in Bioinformatics by ECCB** is the new format designed to virtually host a selection of the ECCB2020 workshops and tutorials. These satellite events will run the week before the main event during the afternoon to promote broad participation at both European and International level.

The **workshops** will provide participants with an informal setting to discuss technical issues, exchange research ideas, and to share practical experiences on a range of focused or emerging topics in bioinformatics. Each workshop will provide an interesting perspective on the cutting edge of a selected research field. Workshops may include any form of presentation such as talks or panel discussions.

The **tutorials** will provide participants with lectures and hands-on training covering topics relevant to the field of bioinformatics. Each tutorial will offer participants an opportunity to learn about new areas of bioinformatics research, to get an introduction to important established topics, or to develop advanced skills in areas they are already familiar with.

## Mon, 31 Aug 2020

| Code | Title | Time (CEST) |
|---|---|---|
| NTB-T01 | Machine Learning and Omics data: Opportunities for advancing biomedical data analysis in Galaxy | 13:30 - 16:30 |
| NTB-T02 | Powerful Presentations Tutorial - How to prepare, design and deliver high-impact presentations | 13:30 - 16:30 |
| NTB-T03 | Using Deep Learning For Image and Sequence Analysis | 17:00 - 20:00 |

## Tue, 1st Sep 2020

| Code | Title | Time (CEST) |
|---|---|---|
| NTB-T04 | Keeping up with epigenomic analysis: theory and practice | 13:30 - 16:30 |
| NTB-T05 | Computational modelling of cellular processes: regulatory vs metabolic systems | 13:30 - 16:30 |
| NTB-W01 | CRISPR Informatics for Functional Genomics, Cancer Targeting, and Beyond | 17:00 - 20:00 |
| NTB-T06 | Reconstruction, analysis, and visualization of phylogenomic data with the ETE Toolkit | 17:00 - 20:00 |

# Pre-meetings

## Wed, 2nd Sept 2020

| Code | Title | Time (CEST) |
|---|---|---|
| NTB-T07 | Deep dive into metagenomic data using metagenome-atlas and MMseqs2 | 13:30 - 16:30 |
| NTB-EW01 | ELIXIR \| Workshop on FAIR Computational Workflows | 13:30 - 16:30 |
| NTB-T08 | Full-Length RNA-Seq Analysis using PacBio long reads: from reads to functional interpretation | 17:00 - 20:00 |
| NTB-EW02 | ELIXIR::GA4GH: Advancing genomics through expedited data access enabled by standards and ontologies | 17:00 - 20:00 |

## Thu, 3rd Sep 2020

| Code | Title | Time (CEST) |
|---|---|---|
| NTB-W02 | Annual European Bioinformatics Core Community (AEBC2) Workshop 2020 | 13:30 - 16:30 |
| NTB-ET01 | ELIXIR \| 3D-Bioinfo: Integrating structural and functional data to support in silico predictions in drug design | 13:30 - 16:30 |
| NTB-T09 | Introduction to structural bioinformatics for evolutionary analysis | 17:00 - 20:00 |
| NTB-EW03 | ELIXIR \| Biological Data Analysis Using InterMine | 17:00 - 20:00 |

## Fri, 4th Sep 2020

| Code | Title | Time (CEST) |
|---|---|---|
| NTB-T10 | Biomedical Data and Text Processing using Shell Scripting | 13:30 - 16:30 |
| NTB-W03 | BioNetVisA: biological network reconstruction, data visualization and analysis in biology and medicine | 13:30 - 16:30 |
| NTB-W04 | Advances in computational modelling of cellular processes and high-performance computing | 17:00 - 20:00 |
| NTB-W05 | Computational Pangenomics: Algorithms & Applications | 17:00 - 20:00 |

# Pre-meetings

ESCS2020
ISCB Student COUNCIL
6th EUROPEAN Student Council Symposium
September 6, virtual conference
ISCB
ISCB Student COUNCIL

The **6th European Student Council Symposium (ESCS 2020)** will take place on Sunday 6th September 2020. The ESCS provides a forum for students and young researchers in the fields of Bioinformatics and Computational Biology to meet, present their work in front of an international audience, build a network within the computational biology community and obtain opportunities that can contribute to the development of their scientific career.

The primary goal of the event is to stimulate interactions between students coming from different research institutes, providing them with a platform to socialize with their peers and senior researchers at an international level. A number of distinguished keynote speakers are invited to present important advances in the field, share their experiences in academia and provide stimulating career advice.

More info: www.escs2020.iscbsc.org

## Modesto Orozco

*Institute for Research in Biomedicine (IRB Barcelona), Spain*

## Simulation of DNA. From the Atom to the Chromatin

**Monday, 7 September | 09:00 - 10:00h (CEST)**

Chaired by **Alfonso Valencia**
*Barcelona Supercomputing Center (BSC), INB/ELIXIR-ES, ICREA, Spain*

### Abstract

DNA is the paradigm of a multiscale system, where sub-Angstrom details can affect the structure and the properties of a meter-long fiber. Such complex systems need to be tackled from multiphysics approaches, combining fine representation of specific details with an overall low-resolution picture of the entire chromatin. I will summarize recent advances from our work on the development of a continuum of methodologies, which starting from accurate physical models allowed us to claim until a full representation of chromatin.

### Biography

Modesto Orozco was born in Barcelona in 1962, obtained BS in Chemistry (1985), MS in Biochemistry (1988) and Ph.D. in Chemistry (1990) at the University of Barcelona. Assistant Professor (1989), Professor (1991) and Full Professor (2001) at the Department of Biochemistry at the University of Barcelona. He was a visiting professor at Yale University 1991-1993. Since 2004 he is Group Leader at the Institute for Research in Biomedicine, since 2005 Director of the Life Sciences Department at the Barcelona Supercomputing Center, from 2006-2013 was the director of the Joint IRB-BSC Program on Computational Biology, and since 2014 director of the Joint BSC-CRG-IRB Program on Computational Biology. Director Integrative Research Nodes. IRB Barcelona. INTERESTS Our main interest is to understand living organisms by means of the basic rules of physics. We use the tools of theoretical physics and computational chemistry and biology to understand the basic principles of life. Areas were we are especially active include: i) mining of genomic information to link genotypic changes and pathologies for personalized medicine, ii) connection of physical properties of DNA with chromatin structure and genomic regulation, and iii) study of the dynamic of proteins and the mechanism of information transfer in macromolecules.

## Geneviève Almouzni

*Chromatin Dynamics team at Institut Curie, Science Academy in France, LifeTime Initiative, France.*

## Chromatin plasticity, cell fate and identity

**Monday, 7 September 2020 | 11:20 - 12:10h (CEST)**

Chaired by **Christine Orengo**
*University College London (UCL), ELIXIR-UK, United Kingdom.*

**Abstract**
During development and throughout life, a variety of specialized cells must be generated to ensure the proper function of each tissue and organ. Chromatin dynamics plays a key role in determining cellular states, whether totipotent, pluripotent, multipotent, or differentiated, therefore influencing cell fate decisions and reprogramming. In my laboratory we focus on the capacity of histone variants, chaperones, modifications, and heterochromatin factors to influence cell identity and its plasticity during development but also during disease onset and progression in particular cancer. Recent advances in single-cell technologies now allow to access to cellular trajectories both in healthy tissues and disease states with an unprecedented level of detail. With the LifeTime Initiative https://lifetime-fetflagship.eu, we have mobilized an interdisciplinary community at European level to exploit these approaches to understand how cells transition into a diseased state in order to detect and treat diseases before major tissue damage has occurred. By generating new knowledge and integrating it with new technological solutions in healthcare, we aim to intercept disease onset or progression based on a patient's particular molecular and cellular signatures. LifeTime vision is based on research programmes directed to the patient and the implementation of a cell-based and interceptive medicine in Europe, leading to a significant improvement of citizens' living quality.

**Biography**
Geneviève Almouzni, PhD (EMBO member, Member of the French Academy of Sciences, fellow of the American Association for the Advancement of Sciences, Director of the Research Center of the Institut Curie from sept. 2013 to sept. 2018 and honorary director since then) is director of research exceptional class at the CNRS. She is Principal Investigator of the Chromatin dynamics team in the Nuclear dynamics research Unit (UMR3664 CNRS/Institut Curie) since 1999. She is a world leader in understanding genome organization and function during development and disease in particular in cancer. She has combined biochemistry, cell biology and physical approaches with advanced imaging to explore chromatin dynamics. Active in the field of epigenetics and European actions, she coordinated the EpiGeneSys Network of Excellence to move epigenetics towards systems biology. She is highly engaged in promoting young scientist careers. She received prestigious grants (ERC Advanced Grants) and awards including Woman in Sciences FEBS / EMBO (2013) and the grand prix FRM (2014). She served on the EMBO Council (Vice-chair in 2014), ERC Council (2019), chair of the alliance EU-LIFE and co-chairs European FETFlagship initiative LifeTime.

## Debora Marks

*Department of Systems Biology, Harvard Medical School, Boston, MA, United States.*

### Prediction and design of biological sequences with neural machines

**Monday, 7 September | 16:30 - 17:20h (CEST)**

Chaired by **Baldo Oliva**
*Structural Bioinformatics lab (GRIB-IMIM), Spain.*
*Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Spain.*

**Abstract**

What can we do with a million or a billion genomes? Understanding how variation across genomes shapes the properties of biomolecules, cells, and organisms is a foundational question in biology. I will demonstrate how probabilistic generative modeling of genetic variation can give surprisingly direct answers to questions about 3D structures, dynamics, the effects of mutations and the design of biological systems. Our new work extends from the undirected models of genetic variation to deep directed models using newly developed variational autoencoders, and autoregressive models that do not depend on sequence alignments. From purely unsupervised learning, we have improved of prior-art for predicting the effects of mutations and successfully designed optimized antibody libraries. I will introduce challenges for extending these methods to diverse biomedical and engineering applications, with specific examples of successful probabilistic models to generate novel functional biotherapeutics
https://marks.hms.harvard.edu/
https://github.com/debbiemarkslab

**Biography**

Debora Marks established her laboratory 5 years ago after a career in industry and more recent degrees in mathematics and computational biology, aiming to accelerate fundamental discoveries in biomedicine. Developing robust statistical methods including unsupervised machine learning, Debora's lab was able predict 3-dimensional protein structures from sequence alone, predict the fitness effects of human genetic variation, make robust generative models for therapeutic and antibody design and design deimmunization. Most recently she has extended these new tools to apply in dimension reduction and multimodal modelling of diverse biological and clinical data, including benchmarked approaches to combinations of RNA, protein and image data.
Mission to develop AI for design of biological interventions for human health and social justice.

## Fabian J. Theis

*Institute of Computational Biology ICB, Helmholtz Munich.*

## Modeling cellular state and dynamics in single cell genomics

**Tuesday, 8 September | 09:10 - 10:00 (CEST)**

Chaired by **Marc Marti-Renom**
*Centre Nacional d'Anàlisi Genòmica - Centre de Regulació Genòmica (CNAG-CRG).*
*The Barcelona Institute of Science and Technology (BIST).*

### Abstract

Modeling cellular state as well as dynamics e.g. during differentiation or in response to perturbations is a central goal of computational biology. Single-cell technologies now give us easy and large-scale access to state observations on the transcriptomic and more recently also epigenomic level. In particular, they allow resolving potential heterogeneities due to asynchronicity of differentiating or responding cells, and profiles across multiple conditions such as time points, space and replicates are being generated.

In this talk I will shortly review scVelo, our recent model for dynamic RNA velocity, allowing estimation of gene-specific transcription and splicing rates, and illustrate its use to estimate a shared latent time in pancreatic endocrinogenesis. I will then show CellRank, a probabilistic model based on Markov chains which makes use of both transcriptomic similarities as well as RNA velocity to infer developmental start- and endpoints and assign lineages in a probabilistic manner. It allows users to gain insights into the timing of endocrine lineage commitment and recapitulates gene expression trends towards developmental endpoints.

While the above approaches focus on individual gene expression models, recently latent space modeling and manifold learning have become a popular tool to learn overall variation in single cell gene expression. I will wrap by briefly discussing how these tools can be used to integrate single cell RNA-seq data sets across multiple labs in a privacy aware manner.

### Biography

Fabian Theis studied Mathematics and Physics and has PhDs in Physics and Computer Science. After different research stays abroad he was a Bernstein fellow leading a junior research group at the Bernstein Center for Computational Neuroscience, located at the Max Planck Institute for Dynamics and Self Organisation at Göttingen. In 2007 he became junior group leader at the Helmholtz Center Munich and associate professor at the Technical University of Munich. In 2013 he founded the Institute of Computational Biology at Helmholtz Munich and full professor at the math department at TUM. Since 2019 he is associate faculty at the Wellcome Trust Sanger Institute in Hinxton, UK. Also he is scientific director of the Helmholtz Artificial Intelligence Cooperation Unit and coordinates the Munich School for Data Science, founded in 2019.

Some of his major achievements was an ERC starting grant in 2010 and the Erwin-Schrödinger prize for interdisciplinary research in 2017.

He has a long-standing interest in Computational Biology, with specific expertise in Machine Learning in the context of single cell biology. By developing and adapting inference methods to integrate information across scales, he contributes to answering complex biological and medical questions such as stem cell decision making and impact of cellular heterogeneity in systems medicine.

## Bissan Al-Lazikani

*The Institute of Cancer Research (ICR), United Kingdom.*

## More than the sum of parts: Multidiciplinary big data in cancer therapy

**Tuesday, 8 September | 11:20 - 12:10 (CEST)**

Chaired by **Patrick Aloy**
*Institute for Research in Biomedicine (IRB Barcelona), Spain.*

### Abstract

The past two decades have witnessed a revolution in the creation and use of large scale data from genomics to imaging. This has had a transformational impact on precision cancer therapy. But after the initial gains, we appear to be slowing down in terms of the clinical impact. In this talk I will discuss the wins and challenges in large cancer data. I will then exemplify how integrating these large multidisciplinary data is already helping power the next revolution in cancer drug discovery, and precision therapy. I will discuss the technical challenges in bringing disparate multidisciplinary data together, and illustrate successes in both preclinical cancer drug discovery as well as clinical research.

### Biography

Bissan Al-Lazikani is Head of Data Science at the Institute of Cancer Research. There she leads the Big Data efforts to tackle key problems in Cancer drug discovery and Cancer therapy.
Bissan led the development of integrative computational approaches to inform drug discovery that are now internationally adopted and provided to the community via the canSAR knowledgebase. She applies data science and machine learning approaches to the discovery of novel therapies and pharmacological and radiation to adapting and individualising therapy to patients.
Bissan has a B.Sc (Hons) in Molecular Biology from University College London, an M.Sc in Computer Science from Imperial College and a PhD in Computational Biology from Cambridge University.
Bissan has worked on drug discovery and personalised medicine both in academia and industry.

## Londa Schiebinger

*Stanford University, United States.*

### Gendered Innovations in Biomedicine, Machine Learning, and Robotics

**Tuesday, 8 September | 16:30 - 17:20 (CEST)**

Chaired by **Niklas Blomberg**
*ELIXIR, United Kingdom.*

**Abstract**

How can we harness the creative power of gender analysis for discovery and innovation? In this talk I identify three strategic approaches to gender in research, policy, and practice: 1) "Fix the Numbers " focuses on increasing women's participation; 2) "Fix the Institutions" promotes gender equality in careers through structural change in research organizations; and 3) "Fix the Knowledge" or "Gendered Innovations" stimulates excellence in science and technology by integrating sex, gender, and intersectional analysis into research. This talk focuses on the third approach. I will discuss several case studies, including basic biomedical research, health & medicine, machine learning, and robotics. To match the global reach of science and technology, Gendered Innovations was developed through a collaboration of over 200 experts from across the United States, Europe, Canada, and Asia. Major funders include the European Commission, the U.S. National Science Foundation, and Stanford University. See AI can be Sexist and Racist—It's Time to Make it Fair Nature, 559.7714 (2018), 324-326; and Sex and Gender Analysis Improves Science and Engineering Nature, 575.7781 (2019), 137-146. For late-breaking news on Gendered Innovations, join our listserv or follow us on Twitter @GenderStanford

**Biography**

Londa Schiebinger is the John L. Hinds Professor of History of Science at Stanford University, and Director of EU/US Gendered Innovations in Science, Health & Medicine, Engineering, and Environment. She is a leading international expert on gender in science and technology and has addressed the United Nations on that topic. Schiebinger received her Ph.D. from Harvard University, is an elected member of the American Academy of Arts and Sciences, and the recipient of numerous prizes and awards, including the prestigious Alexander von Humboldt Research Prize and Guggenheim Fellowship.

Her global project, Gendered Innovations, harnesses the creative power of sex, gender, and intersectional analysis to enhance excellence and reproducibility in science and technology. See AI can be Sexist and Racist—It's Time to Make it Fair Nature, 559.7714 (2018), 324-326; Sex and Gender Analysis Improves Science and Engineering Nature, 575.7781 (2019), 137-146. For late-breaking news on Gendered Innovations, sign up here: https://mailman.stanford.edu/mailman/listinfo/genderedinnovations

# Abstracts

## Proceeding Talks

These are the 41 accepted proceedings, selected from a total of 203 submissions after a peer-review process and reviewed by the topic Chair and co-Chairs. The proceedings are divided in 5 topics: **Genomes**, **Systems**, **Data, Genes and Proteins,** and they will be published in a special issue of the **OUP Bioinformatics Journal**.

### Topic: GENOME

Chaired by **Stephane Rombauts** in Parallel Track #1
*Center For Plant System Biology, VIB - UGENT, Belgium.*
Chaired by **Toni Gabaldón** in Parallel Track #4 and Track #7
*Barcelona Supercomputing Center (BSC), Spain.*

### A general near-exact k-mer counting method with low memory consumption enables *de novo* assembly of 106× human sequence data in 2.7 hours

**Christina Huan Shi**[1] and Kevin Y. Yip[1,2,3]

[1] *Department of Computer Science and Engineering*
[2] *Hong Kong Bioinformatics Centre*
[3] *Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong SAR*

#### Abstract

**Motivation:** In *de novo* sequence assembly, a standard pre-processing step is k-mer counting, which computes the number of occurrences of every length-k sub-sequence in the sequencing reads. Sequencing errors can produce many k-mers that do not appear in the genome, leading to the need for an excessive amount of memory during counting. This issue is particularly serious when the genome to be assembled is large, the sequencing depth is high, or when the memory available is limited.

**Results:** Here we propose a fast near-exact k-mer counting method, CQF-deNoise, which has a module for dynamically removing noisy false k-mers. It automatically determines the suitable time and number of rounds of noise removal according to a user-specified wrong removal rate. We tested CQF-deNoise comprehensively using data generated from a diverse set of genomes with various data properties, and found that the memory consumed was almost constant regardless of the sequencing errors while the noise removal procedure had minimal effects on counting accuracy. Compared with four state-of-the-art k-mer counting methods, CQF-deNoise consistently performed the best in terms of memory usage, consuming 49-76% less memory than the second best method. When counting the k-mers from a human data set with around 60× coverage, the peak memory usage of CQF-deNoise was only 10.9GB (gigabytes) for k=28 and 21.5GB for k=55. De novo assembly of 106× human sequencing data using CQF-deNoise for k-mer counting required only 2.7 hours and 90GB peak memory.

## DeepSELEX: Inferring DNA-binding preferences from HT-SELEX data using multi-class CNNs

**Maor Asif**[1] and Yaron Orenstein[1]

[1] *School of Electrical and Computer Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel.*

### Abstract

**Motivation:** Transcription factor DNA-binding is a central mechanism in gene regulation. Biologists would like to know where and when these factors bind DNA. Hence, they require accurate DNA-binding models to enable binding prediction to any DNA sequence. Recent technological advancements measure the binding of a single transcription factor to thousands of DNA sequences. One of the prevailing techniques, highthroughput SELEX, measures protein-DNA binding by high-throughput sequencing over several cycles of enrichment. Unfortunately, current computational methods to infer the binding preferences from highthroughput SELEX data do not exploit the richness of these data, and are underusing the most advanced computational technique, deep neural networks.

**Results:** To better characterize the binding preferences of transcription factors from these experimental data, we developed DeepSELEX, a new algorithm to infer intrinsic DNA-binding preferences using deep neural networks. DeepSELEX takes advantage of the richness of high-throughput sequencing data and learns the DNA-binding preferences by observing the changes in DNA sequences through the experimental cycles. DeepSELEX outperforms extant methods for the task of DNA-binding inference from high-throughput SELEX data in binding prediction in vitro and is on par with the state of the art in in vivo binding prediction. Analysis of model parameters reveals it learns biologically relevant features that shed light on transcription factors' binding mechanism.

**Contact:** yaronore@bgu.ac.il

### Joint epitope selection and spacer design for string-of-beads vaccines

**Emilio Dorigatti**[1,2] and Benjamin Schubert[2,3]

[1] *Faculty of Mathematics, Informatics and Statistics, Ludwig Maximilian Universität, München, Germany.*
[2] *Institute of Computational Biology, Helmholtz Zentrum München - German Research Center for Environmental Health, Germany.*
[3] *Department of Mathematics, Technical University of Munich, Germany.*

#### Abstract

**Motivation:** Conceptually, epitope-based vaccine design poses two distinct problems: (1) selecting the best epitopes to elicit the strongest possible immune response, and (2) arranging and linking them through short spacer sequences to string-of-beads vaccines, so that their recovery likelihood during antigen processing is maximized. Current state-of-the-art approaches solve this design problem sequentially. Consequently, such approaches are unable to capture the inter-dependencies between the two design steps, usually emphasizing theoretical immunogenicity over correct vaccine processing, thus resulting in vaccines with less effective immunogencity *in vivo*.

**Results:** In this work, we present a computational approach based on linear programming, called JessEV, that solves both design steps simultaneously, allowing to weigh the selection of a set of epitopes that have great immunogenic potential against their assembly into a string-of-beads construct that provides a high chance of recovery. We conducted Monte Carlo cleavage simulations to show that a fixed set of epitopes often cannot be assembled adequately, whereas selecting epitopes to accommodate proper cleavage requirements substantially improves their recovery probability and thus the effective immunogenicity, pathogen, and population coverage of the resulting vaccines by at least two fold.

**Contact:** edo@stat.uni-muenchen.de

## panRGP: a pangenome-based method to predict genomic islands and explore their diversity

**Adelme Bazin**[1], Guillaume Gautreau[1], Claudine Médigue[1], David Vallenet [1], Alexandra Calteau[1]

*[1] LABGeM, Génomique Métabolique, CEA, Genoscope, Institut François Jacob, Université d'Évry, Université Paris-Saclay, CNRS, Evry, France*

### Abstract

**Motivation:** Horizontal gene transfer (HGT) is a major source of variability in prokaryotic genomes. Regions of Genome Plasticity (RGPs) are clusters of genes located in highly variable genomic regions. Most of them arise from HGT and correspond to Genomic Islands (GIs). The study of those regions at the species level has become increasingly difficult with the data deluge of genomes. To date no methods are available to identify GIs using hundreds of genomes to explore their diversity.

**Results:** We present here the panRGP method that predicts RGPs using pangenome graphs made of all available genomes for a given species. It allows the study of thousands of genomes in order to access the diversity of RGPs and to predict spots of insertions. It gave the best predictions when benchmarked along other GI detection tools against a reference dataset. In addition, we illustrated its use on Metagenome Assembled Genomes (MAGs) by redefining the borders of the leuX tRNA hotspot, a well studied spot of insertion in Escherichia coli. panRPG is a scalable and reliable tool to predict GIs and spots making it an ideal approach for large comparative studies.

**Availability:** The methods presented in the current work are available through the following software: https://github.com/labgem/PPanGGOLiN. Detailed results and scripts to compute the benchmark metrics are available at https://github.com/axbazin/panrgp_supdata.

**Contact:** vallenet@genoscope.cns.fr and acalteau@genoscope.cns.fr

## Graph Convolutional Networks for Epigenetic State Prediction Using Both Sequence and 3D Genome Data

**Jack Lanchantin**[1] and Yanjun Qi[1]

[1] *Department of Computer Science, University of Virginia, Charlottesville VA, USA.*

### Abstract

**Motivation:** Predictive models of DNA epigenetic state such as transcription factor binding are essential for understanding regulatory processes and developing gene therapies. It is known that the 3D genome, or spatial structure of DNA, is highly influential in the epigenetic state. Deep neural networks have achieved state of the art performance on epigenetic state prediction by using short windows of DNA sequences independently. These methods, however, ignore the long-range dependencies when predicting the epigenetic states because modeling the 3D genome is challenging.

**Results:** In this work, we introduce ChromeGCN, a graph convolutional network for epigenetic state prediction by fusing both local sequence and long-range 3D genome information. By incorporating the 3D genome, we relax the independent and identically distributed (i.i.d.) assumption of local windows for a better representation of DNA. ChromeGCN explicitly incorporates known long-range interactions into the modeling, allowing us to identify and interpret those important long-range dependencies in influencing epigenetic states. We show experimentally that by fusing sequential and 3D genome data using ChromeGCN, we get a significant improvement over the state-of-the-art deep learning methods as indicated by three metrics. Importantly, we show that ChromeGCN is particularly useful for identifying epigenetic effects in those DNA windows that have a high degree of interactions with other DNA windows.

**Availability:** https://github.com/QData/ChromeGCN
**Contact:** yanjun@virginia.edu

## Finding Orthologous Gene Blocks in Bacteria: The Computational Hardness of the Problem and Novel Methods to Address it

**Huy N Nguyen**[1,3], Alexey Markin[3], Iddo Friedberg[1,2], Oliver Eulenstein[2,3]

[1] *Dpt. of Veterinary Microbiology and Preventive Medicine, Iowa State University, Ames, IA, USA.*
[2] *Program in Bioinformatics and Computational Biology, Iowa State University, Ames, IA, USA.*
[3] *Dpt. of Computer Science, Iowa State University, Ames, IA, USA.*

### Abstract

**Motivation:** The evolution of complexity is one of the most fascinating and challenging problems in modern biology, and tracing the evolution of complex traits is an open problem. In bacteria, operons and gene blocks provide a model of tractable evolutionary complexity at the genomic level. Gene blocks are structures of co-located genes with related functions, and operons are gene blocks whose genes are co-transcribed on a single mRNA molecule. The genes in operons and gene blocks typically work together in the same system or molecular complex. Previously we proposed a method that explains the evolution of orthologous gene blocks (orthoblocks) as a combination of a small set of events that take place in vertical evolution from common ancestors. A heuristic method was proposed to solve this problem. However, no study was done to identify the complexity of the problem.

**Results:** Here we establish that finding the homologous gene block problem is NP-hard and APX-hard. We have developed a greedy algorithm that runs in polynomial time and guarantees an O(ln n) approximation. In addition, we formalize our problem as an integer linear program problem and solve it using the PuLP package and the standard CPLEX algorithm. Our exploration of several candidate operons reveals that our new method provides more optimal results than the results from the heuristic approach, and is significantly faster.

**Availability:** The software accompanying this paper is available x under the GPLv3 license on: https://github.com/nguyenngochuy91/Relevant-Operon

## PathFinder: Bayesian inference of clone migration histories in cancer

Sudhir Kumar[1,2,3], Antonia Chroni[1,2,], Koichiro Tamura[4,5], Maxwell Sanderford[1,2], Olumide Oladeinde[1,2], Vivian Aly[1,2], Tracy Vu[1,2] **Sayaka Miura**[1,2,]

[1]*Institute for Genomics and Evolutionary Medicine.*
[2]*Department of Biology, Temple University, Philadelphia, PA, USA.*
[3]*Center for Excellence in Genome Medicine and Research, King Abdulaziz University, Saudi Arabia.*
[4]*Research Center for Genomics and Bioinformatics, Tokyo Metropolitan University, Hachioji, Tokyo, Japan.*
[5]*Department of Biological Sciences, Tokyo Metropolitan University, Hachioji, Tokyo, Japan.*

### Abstract

**Summary**: Metastases form by dispersal of cancer cells to secondary tissues. They cause a vast majority of cancer morbidity and mortality. Metastatic clones are not medically detected or visible until later stages of cancer development. Thus, clone phylogenies within patients provide a means of tracing the otherwise inaccessible dynamic history of migrations of cancer cells. Here we present a new Bayesian approach, *PathFinder*, for reconstructing the routes of cancer cell migrations. *PathFinder* uses the clone phylogeny and the numbers of mutational differences among clones, along with the information on the presence and absence of observed clones in different primary and metastatic tumors. In the analysis of simulated datasets, *PathFinder* performed well in reconstructing migrations from the primary tumor to new metastases as well as between metastases. However, it was much more challenging to trace migrations from metastases back to primary tumors. We found that a vast majority of errors can be corrected by sampling more clones per tumor and by increasing the number of genetic variants assayed. We also identified situations in which phylogenetic approaches alone are not sufficient to reconstruct migration routes.

**Conclusions**: We anticipate that the use of *PathFinder* will enable a more reliable inference of migration histories, along with their posterior probabilities, which is required to assess the relative preponderance of seeding of new metastasis by clones from primary tumors and/or existing metastases.

**Availability:** PathFinder is available on the web at
**Contact:** s.kumar@temple.edu

### Detecting Evolutionary Patterns of Cancers using Consensus Trees

**Sarah Christensen**[1], Juho Kim[2], Nicholas Chia[3,4], Oluwasanmi Koyejo[1], Mohammed El-Kebir[1]

[1] *Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA.*
[2] *Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL, USA.*
[3] *Microbiome Program, Center for Individualized Medicine, Mayo Clinic, Rochester, MN, USA.*
[4] *Division of Surgical Research, Department of Surgery, Mayo Clinic, Rochester, MN, USA.*

#### Abstract

**Motivation:** While each cancer is the result of an isolated evolutionary process, there are repeated patterns in tumorigenesis defined by recurrent driver mutations and their temporal ordering. Such repeated evolutionary trajectories hold the potential to improve stratification of cancer patients into subtypes with distinct survival and therapy response profiles. However, current cancer phylogeny methods infer large solution spaces of plausible evolutionary histories from the same sequencing data, obfuscating repeated evolutionary patterns.

**Results:** To simultaneously resolve ambiguities in sequencing data and identify cancer subtypes, we propose to leverage common patterns of evolution found in patient cohorts. We first formulate the Multiple Choice Consensus Tree problem, which seeks to select a tumor tree for each patient and assign patients into clusters in such a way that maximizes consistency within each cluster of patient trees. We prove that this problem is NP-hard and develop a heuristic algorithm, RECAP, to solve this problem in practice. Finally, on simulated data, we show RECAP outperforms existing methods that do not account for patient subtypes. We then use RECAP to resolve ambiguities in patient trees and find repeated evolutionary trajectories in lung and breast cancer cohorts.

**Availability:** https://github.com/elkebir-group/RECAP
**Contact:** melkebir@illinois.edu

## svMIL: Predicting the pathogenic effect of TAD boundary-disrupting somatic structural variants through multiple instance learning

**Marleen M. Nieboer**[1] and Jeroen de Ridder[1]

[1] *Center for Molecular Medicine, Oncode Institute, University Medical Center Utrecht, Utrecht, 3584 CG, The Netherlands.*

### Abstract

**Motivation:** Despite the fact that structural variants (SVs) play an important role in cancer, methods to predict their effect, especially for SVs in non-coding regions, are lacking, leaving them often overlooked in the clinic. Non-coding SVs may disrupt the boundaries of Topologically Associated Domains (TADs), thereby affecting interactions between genes and regulatory elements such as enhancers. However, it is not known when such alterations are pathogenic. Although machine learning techniques are a promising solution to answer this question, representing the large number of interactions that an SV can disrupt in a single feature matrix is not trivial.

**Results:** We introduce svMIL: a method to predict pathogenic TAD boundary-disrupting SV effects based on multiple instance learning, which circumvents the need for a traditional feature matrix by grouping SVs into bags that can contain any number of disruptions. We demonstrate that svMIL can predict SV pathogenicity, measured through same-sample gene expression aberration, for various cancer types. In addition, our approach reveals that somatic pathogenic SVs alter different regulatory interactions than somatic non-pathogenic SVs and germline SVs.

**Availability:** All code for svMIL is publicly available on GitHub: https://github.com/UMCUGenetics/svMIL

**Contact:** J.deRidder-4@umcutrecht.nl

## Exploring Chromatin Conformation and Gene Co-Expression through Graph Embedding

Marco Varrone[1], **Luca Nanni**[1], Giovanni Ciriello[2,3], Stefano Ceri[1]

[1] *Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy.*
[2] *Department of Computational Biology, University of Lausanne, Lausanne, Lausanne, Switzerland.*
[3] *Swiss Institute of Bioinformatics, Lausanne, Lausanne, Switzerland.*

### Abstract

**Motivation:** The relationship between gene coexpression and chromatin conformation is of great biological interest. Thanks to high-throughput chromosome conformation capture technologies (Hi-C), researchers are gaining insights on the tri-dimensional organization of the genome. Given the high complexity of Hi-C data and the difficult definition of gene coexpression networks, the development of proper computational tools to investigate such relationship is rapidly gaining the interest of researchers. One of the most fascinating questions in this context is how chromatin topology correlates with gene coexpression and which physical interaction patterns are most predictive of coexpression relationships.

**Results:** To address these questions, we developed a computational framework for the prediction of coexpression networks from chromatin conformation data. We first define a gene chromatin interaction network where each gene is associated to its physical interaction profile; then we apply two graph embedding techniques to extract a low-dimensional vector representation of each gene from the interaction network; finally, we train a classifier on gene embedding pairs to predict if they are coexpressed.

Both graph embedding techniques outperform previous methods based on manually designed topological features, highlighting the need for more advanced strategies to encode chromatin information. We also establish that the most recent technique, based on random walks, is superior.

Overall, our results demonstrate that chromatin conformation and gene regulation share a non-linear relationship and that gene topological embeddings encode relevant information, which could be used also for downstream analysis.

**Availability:** The source code for the analysis is available at:
https://github.com/marcovarrone/geneexpression-chromatin
**Contact:** luca.nanni@polimi.it

**Topic: SYSTEMS**

Chaired by **Patrick Aloy** in Parallel Track #2 and Track #5
*Institute for Research in Biomedicine (IRB Barcelona), Spain.*

### A Neuro-Evolution Approach to Infer a Boolean Network from Time-Series Gene Expressions

**Shohag Barman**[1] and Yung-Keun Kwon[2]

*[1] Department of Computer Science, American International University-Bangladesh (AIUB), Dhaka, Bangladesh.*
*[2] School of IT Convergence, University of Ulsan, 93 Daehak-ro, Nam-gu, Ulsan, Republic of Korea.*

#### Abstract

**Motivation:** In systems biology, it is challenging to accurately infer a regulatory network from time-series gene expression data, and a variety of methods have been proposed.
Most of them were computationally inefficient in inferring very large networks, though, because of the increasing number of candidate regulatory genes.
Although a recent approach called GABNI was presented to resolve this problem using a genetic algorithm, there is room for performance improvement because it employed a limited representation model of regulatory functions.

**Results:** In this regard, we devised a novel genetic algorithm combined with a neural network for the Boolean network inference, where a neural network is used to represent the regulatory function instead of an incomplete Boolean truth table used in the GABNI.
In addition, our new method extended the range of the time-step lag parameter value between the regulatory and the target genes for more flexible representation of the regulatory function.
Extensive simulations with the gene expression datasets of the artificial and real networks were conducted to compare our method with five well-known existing methods including GABNI.
Our proposed method significantly outperformed them in terms of both structural and dynamics accuracy.

**Conclusion:** Our method can be a promising tool to infer a large-scale Boolean regulatory network from time-series gene expression data.

**Availability:** The source code is freely available at https://github.com/kwon-uou/NNBNI.
**Contact:** kwonyk@ulsan.ac.kr

### Feasible-Metabolic-Pathway-Exploration Technique using Chemical Latent Space

**Taiki Fuji**[1], Shiori Nakazawa[1], Kiyoto Ito[1]

[1] *Center for Exploratory Research, Research and Development Group, Hitachi, Ltd. 1-280, Higashi-Koigakubo, Kokubunji-shi, Tokyo, Japan*

#### Abstract

**Motivation:** Exploring metabolic pathways is one of the key techniques for developing highly productive microbes for the bioproduction of chemical compounds. To explore feasible pathways, not only examining a combination of well-known enzymatic reactions but also finding potential enzymatic reactions that can catalyze the desired structural changes are necessary. To achieve this, most conventional techniques use manually predefined-reaction rules, however, they cannot sufficiently find potential reactions because the conventional rules cannot comprehensively express structural changes before and after enzymatic reactions. Evaluating the feasibility of the explored pathways is another challenge because there is no way to validate the reaction possibility of unknown enzymatic reactions by these rules. Therefore, a technique for comprehensively capturing the structural changes in enzymatic reactions and a technique for evaluating the pathway feasibility are still necessary to explore feasible metabolic pathways.

**Results:** We developed a feasible-pathway-exploration technique using chemical latent space obtained from a deep generative model for compound structures. With this technique, an enzymatic reaction is regarded as a difference vector between the main substrate and the main product in chemical latent space acquired from the generative model. Features of the enzymatic reaction are embedded into the fixed-dimensional vector, and it is possible to express structural changes of enzymatic reactions comprehensively. The technique also involves differential-evolution-based reaction selection to design feasible candidate pathways and pathway scoring using neural-network-based reaction-possibility prediction. The proposed technique was applied to the non-registered pathways relevant to the production of 2-butanone, and successfully explored feasible pathways that include such reactions.

**Contact:** taiki.fuji.mn@hitachi.com

## Multi-view dual graph attention convolutional network for microbe-drug association prediction

Yahui Long[1,2], Min Wu[3], Yong Liu[4], Chee Keong Kwoh[2], Jiawei Luo[1], Xiaoli Li[3]

[1] *College of Computer Science and Electronic Engineering, Hunan University, Changsha, China.*
[2] *School of Computer Science and Engineering, Nanyang Technological University, Singapore, Singapore.*
[3] *Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore.*
[4] *Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly (LILY), Nanyang Technological University, Singapore, Singapore.*

### Abstract

### Motivation

Human microbes get closely involved in an extensive variety of complex human diseases and become new drug targets. In silico methods for identifying potential microbe-drug associations provide an effective complement to conventional experimental methods, which can not only benefit screening candidate compounds for drug development, but also facilitate novel knowledge discovery for understanding microbe-drug interaction mechanisms. On the other hand, the recent increased availability of accumulated biomedical data for microbes and drugs provides a great opportunity for a machine learning approach to predict microbe-drug associations. We are thus highly motivated to integrate these data sources to improve prediction accuracy. In addition, it is extremely challenging to predict interactions for new drugs or new microbes, which have no existing microbe-drug associations.

### Results

In this work, we leverage various sources of biomedical information and construct multiple networks (graphs) for microbes and drugs. Then, we develop a novel ensemble framework of graph attention networks with a hierarchical attention mechanism for microbe-drug association prediction from the constructed multiple microbe-drug graphs, denoted as EGATMDA. In particular, for each input graph, we design a graph convolutional network with node-level attention to learn embeddings for nodes (i.e., microbes and drugs). To effectively aggregate node embeddings from multiple input graphs, we implement graph-level attention to learn the importance of different input graphs. Experimental results under different cross-validation settings (e.g., the setting for predicting associations for new drugs) showed that our proposed method outperformed seven state-of-the-art methods. Case studies on predicted microbe-drug associations further demonstrated the effectiveness of our proposed EGATMDA method.

**Availability:** Python codes and dataset are available at: https://github.com/longyahui/EGATMDA

**Contact:** luojiawei@hnu.edu.cn and xlli@i2r.a-star.edu.sg

## A Siamese Neural Network model for the prioritization of metabolic disorders by integrating real and simulated data

**Gian Marco Messa**[1], Francesco Napolitano[1], Sarah H. Elsea[2],Diego di Bernardo[3,4], Xin Gao[1]

*[1] Computational Bioscience Research Center (CBRC), Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia.*
*[2] Department of Molecular and Human Genetics, One Baylor Plaza, Baylor College of Medicine, Houston, TX USA.*
*[3] Telethon Institute of Genetics and Medicine (TIGEM), Pozzuoli (NA) Italy.*
*[4] Department of Chemical, Materials and Industrial Production Engineering, University of Naples Federico II, Naples, Italy.*

### Abstract

**Motivation:** Untargeted metabolomic approaches hold a great promise as a diagnostic tool for inborn errors of metabolisms (IEMs) in the near future. However, the complexity of the involved data makes its application difficult and time consuming. Computational approaches, such as metabolic network simulations and machine learning, could significantly help to exploit metabolomic data to aid the diagnostic process. While the former suffers from limited predictive accuracy, the latter is normally able to generalize only to IEMs for which sufficient data are available. Here we propose a hybrid approach that exploits the best of both worlds by building a mapping between simulated and real metabolic data through a novel method based on Siamese neural networks.

**Results:** The proposed SNN model is able to perform disease prioritization for the metabolic profiles of IEM patients even for diseases that it was not trained to identify. To the best of our knowledge, this has not been attempted before. The developed model is able to significantly outperform a baseline model that relies on metabolic simulations only. The prioritization performances demonstrate the feasibility of the method, suggesting that the integration of metabolic models and data could significantly aid the IEM diagnosis process in the near future.

**Availability:** Datasets used in this study are publicly available from the cited sources. The trained models are publicly available online (Messa et al., 2020).

**Contact:** xin.gao@kaust.edu.sa

## Padhoc: A computational pipeline for Pathway Reconstruction On the Fly

**Salvador Casaní-Galdón**[1], Cecile Pereira[2,3], Ana Conesa[2]

[1] *Biobam Bioinformatics S.L., Valencia, Spain.*
[2] *Department of Microbiology and Cell Science, Institute for Food and Agricultural Sciences, Genetics Institute, University of Florida, Gainesville, Florida, USA.*
[3] *EURA NOVA, Marseille, France.*

### Abstract

**Motivation:** Molecular pathway databases represent cellular processes in a structured and standardized way. These databases support the community-wide utilization of pathway information in biological research and the computational analysis of high-throughput biochemical data. Although pathway databases are critical in genomics research, the fast progress of biomedical sciences prevents databases from staying up-to-date. Moreover, the compartmentalization of cellular reactions into defined pathways reflects arbitrary choices that might not always be aligned with the needs of the researcher. Today, no tool exists that allow the easy creation of user-defined pathway representations.

**Results:** Here we present Padhoc, a pipeline for pathway ad hoc reconstruction. Based on a set of user-provided keywords, Padhoc combines natural language processing, database knowledge extraction, orthology search and powerful graph algorithms to create navigable pathways tailored to the user's needs. We validate Padhoc with a set of well-established E. coli pathways and demonstrate usability to create not-yet-available pathways in model (human) and non-model (sweet orange) organisms.

**Availability:** Padhoc is freely available at https://github.com/ConesaLab/padhoc
**Contact:** aconesa@ufl.edu

## Probabilistic Graphlets Capture Biological Function in Probabilistic Molecular Networks

**Sergio Doria-Belenguer**[1,2], Markus K. Youssef[1,2], René Böttcher[1], Noël Malod-Dognin[1,3], Nataša Pr žulj[1,3,4]

*[1] Barcelona Supercomputing Center, Barcelona, Spain.*
*[2] Universitat Politècnica de Catalunya (UPC), Barcelona, Spain.*
*[3] Department of Computer Science, University College London, London, United Kingdom*
*[4] ICREA, Barcelona, Spain.*

### Abstract

**Motivation:** Molecular interactions have been successfully modelled and analysed as networks, where nodes represent molecules and edges represent the interactions between them. These networks revealed that molecules with similar local network structure also have similar biological functions. The most sensitive measures of network structure are based on graphlets. However, graphlet-based methods thus far are only applicable to unweighted networks, whereas real-world molecular networks may have weighted edges that can represent the probability of an interaction occurring in the cell. This information is commonly discarded when applying thresholds to generate unweighted networks, which may lead to information loss.

**Results:** We introduce probabilistic graphlets as a tool for analyzing the local wiring patterns of probabilistic networks. To assess their performance compared to unweighted graphlets, we generate synthetic networks based on different well-known random network models and edge probability distributions and demonstrate that probabilistic graphlets outperform their unweighted counterparts in distinguishing network structures. Then we model different real-world molecular interaction networks as weighted graphs with probabilities as weights on edges and we analyse them with our new weighted graphlets-based methods. We show that due to their probabilistic nature, probabilistic graphlet-based methods more robustly capture biological information in these data, while simultaneously showing a higher sensitivity to identify condition-specific functions compared to their unweighted graphlet-based method counterparts.

**Availability:** Our implementation of probabilistic graphlets is available at https://github.com/Serdobe/Probabilistic_Graphlets
**Contact:** natasha@bsc.es

### FBA reveals guanylate kinase as a potential target for antiviral therapies against SARS-CoV-2

**Alina Renz**[1,2], Lina Widerspick[1], Andreas Dräger[1,2,3]

*[1] Computational Systems Biology of Infection and Antimicrobial-Resistant Pathogens, Institute for Bioinformatics and Medical Informatics (IBMI), University of Tübingen, Tübingen, Germany.*
*[2] Department of Computer Science, University of Tübingen, Tübingen, Germany.*
*[3] German Center for Infection Research (DZIF), partner site Tübingen, Germany.*

#### Abstract

**Motivation:** The novel coronavirus (SARS-CoV-2) currently spreads worldwide, causing the disease COVID-19. The number of infections increases daily, without any approved antiviral therapy. The recently released viral nucleotide sequence enables the identification of therapeutic targets, e.g., by analyzing integrated human-virus metabolic models. Investigations of changed metabolic processes after virus infections and the effect of knock-outs on the host and the virus can reveal new potential targets.

**Results:** We generated an integrated host-virus genome-scale metabolic model of human alveolar macrophages and SARS-CoV-2. Analyses of stoichiometric and metabolic changes between uninfected and infected host cells using flux balance analysis (FBA) highlighted the different requirements of host and virus. Consequently, alterations in the metabolism can have different effects on host and virus, leading to potential antiviral targets. One of these potential targets is guanylate kinase (GK1). In FBA analyses, the knock-out of the guanylate kinase decreased the growth of the virus to zero, while not affecting the host. As GK1 inhibitors are described in the literature, its potential therapeutic effect for SARS-CoV-2 infections needs to be verified in in-vitro experiments.

**Availability:** The computational model is accessible at https://identifiers.org/biomodels.db/MODEL2003020001

**Contact:** renz@informatik.uni-tuebingen.de; draeger@informatik.uni-tuebingen.de

### Inferring Signaling Pathways with Probabilistic Programming

**David Merrell**[1,2] and Anthony Gitter[1,2,3]

[1] *Department of Computer Sciences, University of Wisconsin–Madison, USA.*
[2] *Morgridge Institute for Research, Madison, Wisconsin, USA*
[3] *Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, USA.*

### Abstract

**Motivation:** Cells regulate themselves via dizzyingly complex biochemical processes called signaling pathways. These are usually depicted as a network, where nodes represent proteins and edges indicate their influence on each other. In order to understand diseases and therapies at the cellular level, it is crucial to have an accurate understanding of the signaling pathways at work. Since signaling pathways can be modified by disease, the ability to infer signaling pathways from condition- or patient-specific data is highly valuable. A variety of techniques exist for inferring signaling pathways. We build on past works that formulate signaling pathway inference as a Dynamic Bayesian Network structure estimation problem on phosphoproteomic time course data. We take a Bayesian approach, using Markov Chain Monte Carlo to estimate a posterior distribution over possible Dynamic Bayesian Network structures. Our primary contributions are primo a novel proposal distribution that efficiently samples sparse graphs and secundo the relaxation of common restrictive modeling assumptions.

**Results:** We implement our method, named Ourmethodfull, in Julia using the Gen probabilistic programming language. Probabilistic programming is a powerful methodology for building statistical models. The resulting code is modular, extensible, and legible. The Gen language, in particular, allows us to customize our inference procedure for biological graphs and ensure efficient sampling. We evaluate our algorithm on simulated data and the HPN-DREAM pathway reconstruction challenge, comparing our performance against a variety of baseline methods. Our results demonstrate the vast potential for probabilistic programming, and Gen specifically, for biological network inference.

**Availability:** Find the full codebase at https://github.com/gitter-lab/ssps
**Contact:** gitter@biostat.wisc.edu

## Topic: DATA

Chaired by **Josep Lluís Gelpí** in Parallel Track #3 and Track #6
*Barcelona Supercomputing Center (BSC), Spain.*
Chaired by **Toni Gabaldón** in Parallel Track #7
*Barcelona Supercomputing Center (BSC), Spain.*

### Dementia Key Gene Identification with MultiLayered SNP-Gene-Disease Network

**Dong-gi Lee**[1], Myungjun Kim[1], Sang Joon Son[2], Chang Hyung Hong[2], Hyunjung Shin[1]

*[1] Department of Industrial Engineering, Ajou University, 206 Worldcup-ro, Yeongtong-gu, Suwon 16499, Republic of Korea.*
*[2] Department of Psychiatry, Ajou University School of Medicine, 206 Worldcup-ro, Yeongtonggu, Suwon 16499, Republic of Korea.*

#### Abstract

**Motivation:** Recently, various approaches for diagnosing and treating dementia have received significant attention, especially in identifying key genes that are crucial for dementia. If the mutations of such key genes could be tracked, it would be possible to predict the time of onset of dementia and significantly aid in developing drugs to treat dementia. However, gene finding involves tremendous cost, time, and effort. To alleviate these problems, research on utilizing computational biology to decrease the search space of candidate genes is actively conducted.

**Methods:** In this study, we propose a framework in which diseases, genes, and single nucleotide polymorphisms (SNPs) are represented by a layered network, and key genes are predicted by a machine learning algorithm. The algorithm utilizes a network-based semi-supervised learning model that can be applied to layered data structures.

**Results:** The proposed method was applied to a dataset extracted from public databases related to diseases and genes with data collected from 186 patients. A portion of key genes obtained using the proposed method was verified in silico through PubMed literature, and the remaining genes were left as possible candidate genes.

**Availability:** The code for the framework will be available at http://www.alphaminers.net/
**Contact:** shin@ajou.ac.kr

## Enhancing statistical power in temporal biomarker discovery through representative shapelet mining

**Thomas Gumbsch**[1,2], Christian Bock[1,2], Michael Moor[1,2], Bastian Rieck[1,2], Karsten Borgwardt[1,2]

[1] *Department of Biosystems Science and Engineering, ETH Zurich, Basel, Switzerland.*
[2] *SIB Swiss Institute of Bioinformatics.*

### Abstract

**Motivation:** Temporal biomarker discovery in longitudinal data is based on detecting reoccurring trajectories, so-called shapelets. The search for shapelets requires considering all subsequences in the data. While the accompanying issue of multiple testing has been mitigated in previous work, the redundancy and overlap of the detected shapelets results in an a priori unbounded number of highly similar and structurally meaningless shapelets. As a consequence, current temporal biomarker discovery methods are impractical and underpowered.

**Results:** We find that the pre- or post-processing of shapelets does not sufficiently increase the power and practical utility. Consequently, we present a novel method for temporal biomarker discovery: Statistically Significant Submodular Subset Shapelet Mining (S5M) that retrieves short subsequences that are 1. occurring in the data, 2. are statistically significantly associated with the phenotype, and 3. are of manageable quantity while maximizing structural diversity. Structural diversity is achieved by pruning nonrepresentative shapelets via submodular optimization. This increases the statistical power and utility of S5M compared to state-of-the-art approaches on simulated and real-world data sets. For patients admitted to the intensive care unit (ICU) showing signs of severe organ failure, we find temporal patterns in the sequential organ failure assessment score that are associated with in-ICU mortality.

**Availability:** S5M is an option in the python package of S3M: github.com/BorgwardtLab/S3M
**Contact:** thomas.gumbsch@bsse.ethz.ch, karsten.borgwardt@bsse.ethz.ch

## SCHNEL: Scalable clustering of high dimensional single-cell data

**Tamim Abdelaal**[1,2], Paul de Raadt[2], Boudewijn P.F. Lelieveldt[1,2], Marcel J.T. Reinders[1,2,3], Ahmed Mahfouz[1,2,3]

[1] *Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands.*
[2] *Leiden Computational Biology Center, Leiden University Medical Center, Leiden, The Netherlands.*
[3] *Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands.*

### Abstract

**Motivation:** Single cell data measures multiple cellular markers at the single-cell level for thousands to millions of cells. Identification of distinct cell populations is a key step for further biological understanding, usually performed by clustering this data. Dimensionality reduction based clustering tools are either not scalable to large datasets containing millions of cells, or not fully automated requiring an initial manual estimation of the number of clusters. Graph clustering tools provide automated and reliable clustering for single cell data, but suffer heavily from scalability to large datasets.

**Results:** We developed SCHNEL, a scalable, reliable and automated clustering tool for high-dimensional single-cell data. SCHNEL transforms large high-dimensional data to a hierarchy of datasets containing subsets of data points following the original data manifold. The novel approach of SCHNEL combines this hierarchical representation of the data with graph clustering, making graph clustering scalable to millions of cells. Using seven different cytometry datasets, SCHNEL outperformed three popular clustering tools for cytometry data, and was able to produce meaningful clustering results for datasets of 3.5 and 17.2 million cells within workable timeframes. In addition, we show that SCHNEL is a general clustering tool by applying it to single-cell RNA sequencing data, as well as a popular machine learning benchmark dataset MNIST.

**Availability and Implementation:** Implementation is available on GitHub : https://github.com/biovault/SCHNELpy
**Contact:** a.mahfouz@lumc.nl

### FastSK: Fast Sequence Analysis with Gapped String Kernels

Derrick Blakely[1], Eamon Collins[1], Ritambhara Singh[2], Andrew Norton[1], Jack Lanchantin[1], **Yanjun Qi**[1]

[1] *Department of Computer Science, University of Virginia, Charlottesville VA USA.*
[2] *Computer science department and Center for Computational Molecular Biology, Brown University, USA.*

### Abstract

**Motivation:** Gapped k-mer kernels with Support Vector Machines (gkm-SVMs) have achieved strong predictive performance on regulatory DNA sequences on modestly-sized training sets. However, existing gkm-SVM algorithms suffer from slow kernel computation time, as they depend exponentially on the subsequence feature-length, number of mismatch positions, and the task's alphabet size.

**Results:** In this work, we introduce a fast and scalable algorithm for calculating gapped k-mer string kernels. Our method, named FastSK , uses a simplified kernel formulation that decomposes the kernel calculation into a set of independent counting operations over the possible mismatch positions. This simplified decomposition allows us to devise a fast Monte Carlo approximation that rapidly converges. FastSK can scale to much greater feature lengths, allows us to consider more mismatches, and is performant on a variety of sequence analysis tasks. On multiple DNA transcription factor binding site (TFBS) prediction datasets, FastSK consistently matches or outperforms the state-of-the-art gkmSVM2.0 algorithms in AUC, while achieving average speedups in kernel computation of ~ 100× and speedups of ~ 800× for large feature lengths. We further show that FastSK outperforms character-level recurrent and convolutional neural networks while achieving low variance. We then extend FastSK to 7 English language medical named entity recognition datasets and 10 protein remote homology detection datasets. FastSK consistently matches or outperforms these baselines.

**Availability:** Our algorithm is available as a Python package and as C++ source code.
**Supplementary:** Install with the command make or pip install from: https://github.com/Qdata/FastSK/ .
**Contact:** yanjun@virginia.edu

## Batch Equalization with a Generative Adversarial Network

**Wesley Wei Qian**[1], Cassandra Xia[2], Subhashini Venugopalan[2], Arunachalam Narayanaswamy[2], Michelle Dimon[2], George W. Ashdown[3], Jake Baum[3], Jian Peng[1], Michael Ando[2]

[1] *Department of Computer Science, University of Illinois at Urbana-Champaign.*
[2] *Google Research.*
[3] *Department of Life Sciences, Imperial College London.*

### Abstract

**Motivation:** Advances in automation and imaging have made it possible to capture a large image dataset that spans multiple experimental batches of data. However, accurate biological comparison across the batches is challenged by batch-to-batch variation (i.e., batch effect) due to uncontrollable experimental noise (e.g., varying stain intensity or cell density). Previous approaches to minimize the batch effect have commonly focused on normalizing the low-dimensional image measurements such as an embedding generated by a neural network. However, normalization of the embedding could suffer from over-correction and alter true biological features (e.g., cell size) due to our limited ability to interpret the effect of the normalization on the embedding space. While techniques like flat-field correction can be applied to normalize the image values directly, they are limited transformations that handle only simple artifacts due to batch effect.

**Results:** We present a neural network based batch equalization method that can transfer images from one batch to another while preserving the biological phenotype. The equalization method is trained as a generative adversarial network (GAN), using the StarGAN architecture that has shown considerable ability in style transfer. After incorporating new objectives that disentangle batch effect from biological features, we show that the equalized images have less batch information and preserve the biological information. We also demonstrate that the same model training parameters can generalize to two dramatically different types of cells, indicating this approach could be broadly applicable.

## Supervised learning on phylogenetically distributed data

**Elliot Layne**[1], Erika Dort[2], Richard Hamelin[2], Yue Li[1], Mathieu Blanchette[1]

[1] *School of Computer Science, McGill, Montreal, Canada.*
[2] *Department of Forestry and Conservation Sciences, University of British Columbia, Vancouver, Canada.*

### Abstract

**Motivation:** The ability to develop robust machine-learning models is considered imperative to the adoption of ML techniques in biology and medicine fields. This challenge is particularly acute when data available for training is not independent and identically distributed, in which case trained models are vulnerable to out-of-distribution generalization problems. Of particular interest are problems where data corresponds to observations made on phylogenetically related samples (e.g. antibiotic resistance data).

**Results:** We introduce DendroNet, a new approach to train neural networks in the context of evolutionary data. DendroNet explicitly accounts for the relatedness of the training/testing data, while allowing the model to evolve along the branches of the phylogenetic tree, hence accommodating potential changes in the rules that relate genotypes to phenotypes. Using simulated data, we demonstrate that DendroNet produces models that can be significantly better than non-phylogenetically aware approaches. DendroNet also outperforms other approaches at two biological tasks of significant practical importance: antiobiotic resistance prediction in bacteria, and trophic level prediction in fungi.

**Availability:** https://github.com/BlanchetteLab/DendroNet
**Contact:** elliot.layne@mail.mcgill.ca, blanchem@cs.mcgill.ca

### Using a GTR+Γ substitution model for dating sequence divergence when stationarity and time-reversibility assumptions are violated

**Jose Barba-Montoya**[1,2], Qiqing Tao[1,2], Sudhir Kumar[1,2,3]

[1] *Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA.*
[2] *Department of Biology, Temple University, Philadelphia, PA, USA.*
[3] *Center for Excellence in Genome Medicine and Research, King Abdulaziz University, Jeddah, Saudi Arabia.*

#### Abstract

**Motivation:** As the number and diversity of species and genes grow in contemporary datasets, two common assumptions made in all molecular dating methods, namely the time-reversibility and stationarity of the substitution process, become untenable. No software tools for molecular dating allow researchers to relax these two assumptions in their data analyses. Frequently the same General Time Reversible (GTR) model across lineages along with a gamma (+Γ) distributed rates across sites is used in relaxed clock analyses, which assumes time-reversibility and stationarity of the substitution process. Many reports have quantified the impact of violations of these underlying assumptions on molecular phylogeny, but none have systematically analyzed their impact on divergence time estimates.

**Results:** We quantified the bias on time estimates that resulted from using the GTR+Γ model for the analysis of computer-simulated nucleotide sequence alignments that were evolved with non-stationary (NS) and non-reversible (NR) substitution models. We tested Bayesian and RelTime approaches that do not require a molecular clock for estimating divergence times. Divergence times obtained using a GTR+Γ model differed only slightly (~3% on average) from the expected times for NR datasets, but the difference was larger for NS datasets (~10% on average). The use of only a few calibrations reduced these biases considerably (~5%). Confidence and credibility intervals from GTR+Γ analysis usually contained correct times. Therefore, the bias introduced by the use of the GTR+Γ model to analyze datasets, in which the time-reversibility and stationarity assumptions are violated, is likely not large and can be reduced by applying multiple calibrations.

**Availability:** All datasets are deposited in Figshare: https://doi.org/10.6084/m9.figshare.12594638.
**Contact:** s.kumar@temple.edu

## Matrix (Factorization) Reloaded: Flexible Methods for Imputing Genetic Interactions with Cross-Species and Side Information

**Jason Fan**[1], Xuan Cindy Li[2], Mark Crovella[3], Mark D.M. Leiserson[1]

*[1] Department of Computer Science and Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, USA.*
*[2] Program in Computational Biology, Bioinformatics, and Genomics, University of Maryland, College Park, MD, USA.*
*[3] Department of Computer Science, Boston University, MA, USA.*

### Abstract

**Motivation:** Mapping genetic interactions (GIs) can reveal important insights into cellular function, and has potential translational applications. There has been great progress in developing high-throughput experimental systems for measuring GIs (e.g. with double knockouts) as well as in defining computational methods for inferring (imputing) unknown interactions. However, existing computational methods for imputation have largely been developed for and applied in baker's yeast, even as experimental systems have begun to allow measurements in other contexts. Importantly, existing methods face a number of limitations in requiring specific side information and with respect to computational cost. Further, few have addressed how GIs can be imputed when data is scarce.
**Results:** In this paper we address these limitations by presenting a new imputation framework, called Extensible Matrix Factorization (EMF). EMF is a framework of composable models that flexibly exploit cross-species information in the form of GI data across multiple species, and arbitrary side information in the form of kernels (e.g. from protein-protein interaction networks). We perform a rigorous set of experiments on these models in matched GI datasets from baker's and fission yeast. These include the first such experiments on genome-scale GI datasets in multiple species in the same study. We find that EMF models that exploit side and cross-species information improve imputation, especially in data-scarce settings. Further, we show that EMF outperforms the state-of-the-art deep learning method, even when using strictly less data, and incurs orders of magnitude less computational cost.

**Availability:** Implementations of models and experiments are available at: github.com/lrgr/emf
**Contact:** mdml@umd.edu

## The Effect of Kinship in Re-identification Attacks Against Genomic Data Sharing Beacons

**Kerem Ayoz**[1], Miray Aysen[1], Erman Ayday[1,2], A. Ercument Cicek[1,3]

[1] *Computer Engineering Department, Bilkent University, Ankara, Turkey.*
[2] *Computer and Data Sciences Department, Case Western Reserve University, Cleveland, OH* ¡.
[3] *Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA*¡.

### Abstract

**Motivation:** Big data era in genomics promises a breakthrough in medicine, but sharing data in a private manner limits the pace of field. Widely accepted "genomic data-sharing beacon" protocol provides a standardized and secure interface for querying the genomic datasets. The data is only shared if the desired information (e.g., a certain variant) exists in the dataset. Various studies showed that beacons are vulnerable to re-identification (or membership inference) attacks. As beacons are generally associated with sensitive phenotype information, re-identification creates a significant risk for the participants. Unfortunately, proposed countermeasures against such attacks have failed to be effective, as they do not consider the utility of beacon protocol.

**Results:** In this study, for the first time, we analyze the mitigation effect of the kinship relationships among beacon participants against re-identification attacks. We argue that having multiple family members in a beacon can garble the information for attacks since a substantial number of variants are shared among kinrelated people. Using family genomes from HapMap and synthetically-generated datasets, we show that having one of the parents of a victim in the beacon causes (i) significant decrease in the power of attacks and (ii) substantial increase in the number of queries needed to confirm an individual's beacon membership. We also show how the protection effect attenuates when more distant relatives, such as grandparents are included alongside the victim. Furthermore, we quantify the utility loss due adding relatives and show that it is smaller compared to flipping based techniques.

## DeepCDR: a hybrid graph convolutional network for predicting cancer drug response

**Qiao Liu**[1], Zhiqiang Hu[2], Rui Jiang[1], Mu Zhou[3]

*[1] MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for Synthetic & Systems Biology, BNRIST; Department of Automation, Tsinghua University, Beijing, China.*
*[2] SenseTime Research, Shanghai, China.*
*[3] SenseBrain Research, CA, USA.*

### Abstract

**Motivation:** Accurate prediction of cancer drug response (CDR) is challenging due to the uncertainty of drug efficacy and heterogeneity of cancer patients. Strong evidences have implicated the high dependence of CDR on tumor genomic and transcriptomic profiles of individual patients. Precise identification of CDR is crucial in both guiding anti-cancer drug design and understanding cancer biology.
**Results:** In this study, we present DeepCDR which integrates multi-omics profiles of cancer cells and explores intrinsic chemical structures of drugs for predicting cancer drug response. Specifically, DeepCDR is a hybrid graph convolutional network consisting of a uniform graph convolutional network (UGCN) and multiple subnetworks. Unlike prior studies modeling hand-crafted features of drugs, DeepCDR automatically learns the latent representation of topological structures among atoms and bonds of drugs. Extensive experiments showed that DeepCDR outperformed state-of-the-art methods in both classification and regression settings under various data settings. We also evaluated the contribution of different types of omics profiles for assessing drug response. Furthermore, we provided an exploratory strategy for identifying potential cancer-associated genes concerning specific cancer types. Our results highlighted the predictive power of DeepCDR and its potential translational value in guiding disease-specific drug design.

**Availability:** DeepCDR is freely available at https://github.com/kimmo1019/DeepCDR
**Contact:** ruijiang@tsinghua.edu.cn; muzhou@sensebrain.site

## SCIM: Universal Single-Cell Matching with Unpaired Feature Sets

Stefan G. Stark[1,2], **Joanna Ficek**[1,2,3], Francesco Locatello[1,4,5], Ximena Bonilla[1,2], Stephane Chevrier[6], Franziska Singer[2,7],Gunnar Ratsch[1,2,5,8,9] , Kjong-Van Lehmann[1,2]

[1] *Department of Computer Science, ETH Zurich, Universitätstrasse 6, Zurich, Switzerland.*
[2] *Swiss Institute of Bioinformatics, Quartier Sorge - Batiment Amphipole, Lausanne, Switzerland.*
[3] *Life Science Zurich Graduate School, PhD Program Molecular & Translational Biomedicine, Zurich, Switzerland.*
[4] *Max Planck Institute for Intelligent Systems, Empirical Inference Department, Tuebingen, Germany.*
[5] *Center for Learning Systems, ETH Zurich, Switzerland*
[6] *Department of Quantitative Biomedicine, University of Zurich,Zurich, Switzerland*
[7] *Nexus Personalized Health Technologies, ETH Zurich, Zurich, Switzerland.*
[8] *University Hospital Zurich, Zurich, Switzerland*
[9] *Department of Biology, ETH Zurich, Zurich, Switzerland.*

### Abstract

**Motivation:** Recent technological advances have led to an increase in the production and availability
of single-cell data. The ability to integrate a set of multi-technology measurements would allow the
identification of biologically or clinically meaningful observations through the unification of the perspectives afforded by each technology. In most cases, however, profiling technologies consume the used cells and thus pairwise correspondences between datasets are lost. Due to the sheer size single-cell datasets can acquire, scalable algorithms that are able to universally match single-cell measurements carried out in one cell to its corresponding sibling in another technology are needed.

**Results:** We propose Single-Cell data Integration via Matching (SCIM), a scalable approach to recover
such correspondences in two or more technologies. SCIM assumes that cells share a common (low-dimensional) underlying structure and that the underlying cell distribution is approximately constant across technologies. It constructs a technology-invariant latent space using an autoencoder framework with an adversarial objective. Multi-modal datasets are integrated by pairing cells across technologies using a bipartite matching scheme that operates on the low-dimensional latent representations. We evaluate SCIM on a simulated cellular branching process and show that the cell-to-cell matches derived by SCIM reflect the same pseudotime on the simulated dataset. Moreover, we apply our method to two real-world scenarios, a melanoma tumor sample and a human bone marrow sample, where we pair cells from a scRNA dataset to their sibling cells in a CyTOF dataset achieving 90% and 78% cell-matching accuracy for each one of the samples respectively.

**Availability:** https://github.com/ratschlab/scim
**Contact:** Gunnar.Ratsch@ratschlab; Kjong.Lehmann@inf.ethz.ch

## Topic: GENES

Chaired by **Ana Conesa** in Parallel Track #8
*University of Florida, United States.*
Chaired by **Artemis Hatzigeorgiou** in Parallel Track #9
*University of Thessaly, Hellenic Pasteur Institute, Greece.*

### An efficient framework to identify key miRNA–mRNA regulatory modules in cancer

**Milad Mokhtaridoost**[1] and Mehmet Gönen[2,3,4]

*[1] Graduate School of Science and Engineering.*
*[2] Department of Industrial Engineering, College of Engineering.*
*[3] School of Medicine, Koç University, Istanbul, Turkey.*
*[4] Department of Biomedical Engineering, School of Medicine, Oregon Health & Science University, Portland, OR, USA.*

### Abstract

**Motivation:** Micro RNAs (miRNAs) are known as the important components of RNA silencing and posttranscriptional gene regulation, and they interact with messenger RNAs (mRNAs) either by degradation or by translational repression. miRNA alterations have a significant impact on the formation and progression of human cancers. Accordingly, it is important to establish computational methods with high predictive performance to identify cancer-specific miRNA–mRNA regulatory modules.

**Results:** We presented a two-step framework to model miRNA–mRNA relationships and identify cancerspecific modules between miRNA and mRNA from their matched expression profiles of more than 9000 primary tumors. We first estimated the regulatory matrix between miRNA and mRNA expression profiles by solving multiple linear programming problems. We then formulated a unified regularized factor regression (RFR) model that simultaneously estimates the effective number of modules (i.e. latent factors) and extracts modules by decomposing regulatory matrix into two low-rank matrices. Our RFR model groups correlated miRNAs together and correlated mRNAs together, and also controls sparsity levels of both matrices. These attributes lead to interpretable results with high predictive performance. We applied our method on a very comprehensive data collection including 32 TCGA cancer types. To find the biological relevance of our approach, we performed functional gene set enrichment and survival analyses. A large portion of the identified modules are significantly enriched in Hallmark, PID and KEGG pathways/gene sets. To validate the identified modules, we also performed literature validation as well as validation using experimentally supported miRTarBase database.

**Availability and implementation:** Our implementation of proposed two-step RFR algorithm in R is available at https://github.com/MiladMokhtaridoost/2sRFR together with the scripts that replicate the reported experiments.
**Contact:** mehmetgonen@ku.edu.tr

## RAINFOREST: A random forest approach to predict treatment benefit in data from (failed) clinical drug trials

**Joske Ubels**[1,2,3,4], Tilman Schaefers[1,4], Cornelis Punt[5], Henk-Jan Guchelaar[6] and Jeroen de Ridder[1,4]

[1] *Center for Molecular Medicine, UMC Utrecht, Utrecht, The Netherlands.*
[2] *Erasmus MC Cancer Institute, ErasmusMC, Rotterdam, The Netherlands.*
[3] *SkylineDx, Rotterdam, The Netherlands.*
[4] *Oncode Institute, Utrecht, The Netherlands.*
[5] *Department of Medical Oncology, Amsterdam University Medical Center, University of Amsterdam, The Netherlands.*
[6] *Department of Clinical Pharmacy and Toxicology, Leiden University Medical Center, Leiden, The Netherlands.*

### Abstract

**Motivation:** When phase III clinical drug trials fail their end-point, enormous resources are wasted. Moreover, even if a clinical trial demonstrates a significant benefit, the observed effects are often small and may not outweigh the side effects of the drug. Therefore, there is a great clinical need for methods to identify genetic markers that can identify subgroups of patients which are likely to benefit from treatment as this may i) rescue failed clinical trials and/or ii) identify subgroups of patients which benefit more than the population as a whole. When single genetic biomarkers cannot be found, machine learning approaches that find multivariate signatures are required. For SNP profiles this is extremely challenging owing to the high dimensionality of the data. Here we introduce RAINFOREST (tReAtment benefIt prediction using raNdom FOREST), which can predict treatment benefit from patient SNP profiles obtained in a clinical trial setting.

**Results:** We demonstrate the performance of RAINFOREST on the CAIRO2 dataset, a phase III clinical trial which tested the addition of cetuximab treatment for metastatic colorectal cancer and concluded there was no benefit. However, we find that RAINFOREST is able to identify a subgroup comprising 27.7% of the patients that do benefit, with a hazard ratio of 0.69 ($p = 0.04$) in favor of cetuximab. The method is not specific to colorectal cancer and could aid in reanalysis of clinical trial data and provide a more personalized approach to cancer treatment, also when there is no clear link between a single variant and treatment benefit.

**Availability:** The R code used to produce the results in this paper can be found at github.com/jubels/RAINFOREST. A more configurable, user-friendly Python implementation of RAINFOREST is also provided. Due to restrictions based on privacy regulations and informed consent of participants, phenotype and genotype data of the CAIRO2 trial cannot be made freely available in a public repository. Data from this study can be obtained upon request. Requests should be directed towards Prof. dr. H.J. Guchelaar (h.j.guchelaar@lumc.nl).
**Contact:** j.deridder-4@umcutrecht.nl

## Conditional out-of-distribution generation for un-paired data using transfer VAE

**Mohammad Lotfollahi**[1,2], Mohsen Naghipourfar[1,4], Fabian J. Theis[1,2,3], F. Alexander Wolf[1]

*[1] Institute of Computational Biology, Helmholtz Center Munich, Neuherberg, Germany.*
*[2] School of Life Sciences Weihenstephan, Technical University of Munich, Munich, Germany.*
*[3] Department of Mathematics, Technische Universität München, Munich, Germany.*
*[4] Department of Computer Engineering, Sharif University of Technology, Tehran, Iran.*

### Abstract

**Motivation:** While generative models have shown great success in sampling high-dimensional samples conditional on low-dimensional descriptors (stroke thickness in MNIST, hair color in CelebA, speaker identity in WaveNet), their generation OOD poses fundamental problems due to the difficulty of learning compact joint distribution across conditions. The canonical example of the conditional variational autoencoder (CVAE), for instance, does not explicitly relate conditions during training and, hence, has no explicit incentive of learning such a compact representation.

**Results:** We overcome the limitation of the CVAE by matching distributions across conditions using maximum mean discrepancy (MMD) in the decoder layer that follows the bottleneck. This introduces a strong regularization both for reconstructing samples within the same condition and for transforming samples across conditions, resulting in much improved generalization. As this amounts to solving a styletransfer problem, we refer to the model as transfer VAE (trVAE). Benchmarking trVAE on high-dimensional image and single-cell RNA-seq, we demonstrate higher robustness and higher accuracy than existing approaches.We also show qualitatively improved predictions by tackling previously problematic minority classes and multiple conditions in the context of cellular perturbation response to treatment and disease based on high-dimensional single-cell gene expression data. For generic tasks, we improve Pearson correlations of high-dimensional estimated means and variances with their ground truths from 0.89 to 0.97 and 0.75 to 0.87, respectively. We further demonstrate that trVAE learns cell-type-specific responses after perturbation and improves the prediction of most cell-type-specific genes by 65%.

**Availability:** The trVAE implementation is available via github.com/theislab/trvae. The results of this paper can be reproduced via github.com/theislab/trvae_reproducibility.

### MirCure: A tool for quality control, filter, and curation of microRNAs of animals and plants

**Guillem Ylla**[1,2], Tianyuan Liu[1], and Ana Conesa[1]

[1]Microbiology and Cell Science Department, University of Florida, Gainesville, FL, USA.
[2]*Current affiliation:* Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA.

#### Abstract

**Motivation:** microRNAs (miRNAs) are essential components of gene expression regulation at the post-transcriptional level. miRNAs have a well-defined molecular structure and this has facilitated the development of computational and high-throughput approaches to predict miRNAs genes. However, due to their short size, miRNAs have often been incorrectly annotated in both plants and animals. Consequently, published miRNA annotations and miRNA databases are enriched for false miRNAs, jeopardizing their utility as molecular information resources. To address this problem, we developed MirCure, a new software for quality control, filtering, and curation of miRNA candidates. MirCure is an easy-to-use tool with a graphical interface that allows both scoring of microRNA reliability and browsing of supporting evidence by manual curators.

**Results:** Given a list of miRNA candidates, MirCure evaluates a number of miRNA-specific features based on gene expression, biogenesis, and conservation data, and generates a score that can be used to discard poorly supported miRNA annotations. MirCure can also curate and adjust the annotation of the 5p and 3p arms based on user-provided small RNA-seq data. We evaluated MirCure on a set of manually curated animal and plant microRNAs and demonstrated great accuracy. Moreover, we show that MirCure can be used to revisit previous bona fide miRNAs annotations to improve microRNA databases.

**Availability:** The MirCure software and all the additional scripts used in this project are publicly available at https://github.com/ConesaLab/MirCure. A Docker image of MirCure is available at https://hub.docker.com/u/conesalab.
**Contact:** aconesa@ufl.edu

## Adversarial Deconfounding Autoencoder for Learning Robust Gene Expression Embeddings

**Ayse B. Dincer**[1], Joseph D. Janizek[1, 2], Su-In Lee[1]

[1] *Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA.*
[2] *Medical Scientist Training Program, University of Washington, Seattle, WA, USA.*

### Abstract

**Motivation:** Increasing number of gene expression profiles has enabled the use of complex models, such as deep unsupervised neural networks, to extract a latent space from these profiles. However, expression profiles, especially when collected in large numbers, inherently contain variations introduced by technical artifacts (e.g., batch effects) and uninteresting biological variables (e.g., age) in addition to the true signals of interest. These sources of variations, called confounders, produce embeddings that fail to transfer to different domains, i.e., an embedding learned from one dataset with a specific confounder distribution does not generalize to different distributions. To remedy this problem, we attempt to disentangle confounders from true signals to generate biologically informative embeddings.

**Results:** In this paper, we introduce the AD-AE (Adversarial Deconfounding AutoEncoder) approach to deconfounding gene expression latent spaces. The AD-AE model consists of two neural networks: (i) an autoencoder to generate an embedding that can reconstruct original measurements, and (ii) an adversary trained to predict the confounder from that embedding. We jointly train the networks to generate embeddings that can encode as much information as possible without encoding any confounding signal. By applying AD-AE to two distinct gene expression datasets, we show that our model can (1) generate embeddings that do not encode confounder information, (2) conserve the biological signals present in the original space, and (3) generalize successfully across different confounder domains. We demonstrate that AD-AE outperforms standard autoencoder and other deconfounding approaches.

**Availability:** Our code and data are available at https://gitlab.cs.washington.edu/abdincer/ad-ae
**Contact:** abdincer@cs.washington.edu; suinlee@cs.washington.edu

## DriverGroup: A novel method for identifying driver gene groups

**Vu VH Pham** [1] , Lin Liu [1] , Cameron P Bracken [2,3] , Gregory J Goodall [2,3] , Jiuyong Li [1] and Thuc D Le[1]

*[1] UniSA STEM, University of South Australia, Mawson Lakes, SA, Australia.*
*[2] Centre for Cancer Biology, an alliance of SA Pathology and University of South Australia, Adelaide, SA, Australia.*
*[3] Department of Medicine, The University of Adelaide, Adelaide, SA, Australia.*

### Abstract

**Motivation:** Identifying cancer driver genes is a key task in cancer informatics. Most existing methods are focused on individual cancer drivers which regulate biological processes leading to cancer. However, the effect of a single gene may not be sufficient to drive cancer progression. Here, we hypothesise that there are driver gene groups that work in concert to regulate cancer and we develop a novel computational method to detect those driver gene groups.

**Results:** We develop a novel method named DriverGroup to detect driver gene groups by using gene expression and gene interaction data. The proposed method has three stages: (1) Constructing the gene network, (2) Discovering critical nodes of the constructed network, and (3) Identifying driver gene groups based on the discovered critical nodes. Before evaluating the performance of DriverGroup in detecting cancer driver groups, we firstly assess its performance in detecting the influence of gene groups, a key step of DriverGroup. The application of DriverGroup to DREAM4 data demonstrates that it is more effective than other methods in detecting the regulation of gene groups. We then apply DriverGroup to the BRCA dataset to identify driver groups for breast cancer. The identified driver groups are promising as several group members are confirmed to be related to cancer in literature. We further use the predicted driver groups in survival analysis and the results show that the survival curves of patient subpopulations classified using the predicted driver groups are significantly differentiated, indicating the usefulness of DriverGroup.

**Availability and implementation:** DriverGroup is available at https://github.com/pvvhoang/DriverGroup
**Contact:** Thuc.Le@unisa.edu.au

## Topic: PROTEINS

Chaired by **Ana Conesa** in Parallel Track #8
*University of Florida, United States.*
Chaired by **Mark Wass** in Parallel Track #10
*University Of Kent, United Kingdom.*

### CLPred: A sequence-based protein crystallization predictor using BLSTM neural network

**Wenjing Xuan**[1,2], Ning Liu[1], Neng Huang[1], Yaohang Li[3,], Jianxin Wang[1,2]
*[1] School of Computer Science and Engineering, Central South University, Changsha, China.*
*[2] Hunan Provincial Key Lab on Bioinformatics, Central South University, Changsha, China.*
*[3] Department of Computer Science, Old Dominion University, Norfolk, Virginia, United States.*

### Abstract

**Motivation:** Determining the structures of proteins is a critical step to understand their biological functions. Crystallography-based X-ray diffraction technique is the main method for experimental protein structure determination. However, the underlying crystallization process, which needs multiple time-consuming and costly experimental steps, has a high attrition rate. To overcome this issue, a series of *in-silico* methods have been developed with the primary aim of selecting the protein sequences that are promising to be crystallized. However, the predictive performance of the current methods is modest.

**Results:** We propose a deep learning model, so-called CLPred, which uses a bidirectional recurrent neural network with long short-term memory (BLSTM) to capture the long-range interaction patterns between *k*-mers amino acids to predict protein crystallizability. Using sequence only information, CLPred outperforms the existing deep-learning predictors and a vast majority of sequence-based diffraction-quality crystals predictors on three independent test sets. The results highlight the effectiveness of BLSTM in capturing non-local, long-range inter-peptide interaction patterns to distinguish proteins that can result in diffraction-quality crystals from those that cannot. CLPred has been steadily improved over the previous window-based neural networks, which is able to predict crystallization propensity with high accuracy. CLPred can also be improved significantly if it incorporates additional features from pre-extracted evolutional, structural, and physicochemical characteristics. The correctness of CLPred predictions is further validated by the case studies of Sox transcription factor family member proteins and Zika virus non-structural proteins.

**Availability:** https://github.com/xuanwenjing/CLPred
**Contact:** jxwang@mail.csu.edu.cn; yaohang@cs.odu.edu

## Geometricus Represents Protein Structures as Shape-mers Derived from Moment Invariants

**Janani Durairaj**[1], Mehmet Akdel[1], Dick de Ridder[1], Aalt DJ van Dijk[1,2]

[1] *Bioinformatics Group, Department of Plant Sciences, Wageningen University and Research.*
[2] *Mathematical and Statistical Methods - Biometris, Department of Plant Sciences, Wageningen University and Research.*

### Abstract

**Motivation:** As the number of experimentally solved protein structures rises, it becomes increasingly appealing to use structural information for predictive tasks involving proteins. Due to the large variation in protein sizes, folds, and topologies, an attractive approach is to embed protein structures into fixed-length vectors, which can be used in machine learning algorithms aimed at predicting and understanding functional and physical properties. Many existing embedding approaches are alignment-based, which is both time-consuming and ineffective for distantly related proteins. On the other hand, library- or model-based approaches depend on a small library of fragments or require the use of a trained model, both of which may not generalize well.

**Results:** We present Geometricus, a novel and universally applicable approach to embedding proteins in a fixed-dimensional space. The approach is fast, accurate, and interpretable. Geometricus uses a set of 3D moment invariants to discretize fragments of protein structures into shape-mers, which are then counted to describe the full structure as a vector of counts. We demonstrate the applicability of this approach in various tasks, ranging from fast structure similarity search, unsupervised clustering, and structure classification across proteins from different superfamilies as well as within the same family.

**Availability:** Python code available at https://git.wur.nl/durai001/geometricus
**Contact:** aaltjan.vandijk@wur.nl, janani.durairaj@wur.nl

## GRaSP: a graph-based residue neighborhood strategy to predict binding sites

Charles A. Santana[1,2], **Sabrina de A. Silveira**[3,5], João P. A. Moraes[4], Sandro C. Izidoro[4], Raquel C. de Melo-Minardi[1,2], António J. M. Ribeiro[5], Jonathan D. Tyzack[5], Neera Borkakoti[5], Janet M. Thornton[5]

[1] *Department of Biochemistry and Immunology, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.*
[2] *Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.*
[3] *Department of Computer Science, Universidade Federal de Viçosa, Viçosa, Brazil.*
[4] *Institute of Technological Sciences (ICT), Advanced Campus at Itabira, Universidade Federal de Itajubá, Itabira, Brazil.*
[5] *European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.*

### Abstract

**Motivation:** The discovery of protein-ligand binding sites is a major step for elucidating protein function and for investigating new functional roles. Detecting protein-ligand binding sites experimentally is time-consuming and expensive. Thus a variety of in silico methods to detect and predict binding sites were proposed as they can be scalable, fast and present low cost.

**Results:** We proposed GRaSP, a novel residue centric and scalable method to predict ligand binding site residues. It is based on a supervised learning strategy that models the residue environment as a graph at the atomic level. Results show that GRaSP made compatible or superior predictions when compared with methods described in the literature. GRaSP outperformed 6 other residue-centric methods, including the one considered as state-of-the-art. Also, our method achieved better results than the method from CAMEO independent assessment. GRaSP ranked second when compared with 5 state-of-the-art pocketcentric methods, which we consider a significant result, as it was not devised to predict pockets. Finally, our method proved scalable as it took 10 to 20 seconds on average to predict the binding site for a protein complex whereas the state-of-the art residue-centric method takes 2 to 5 hours on average.

**Availability and implementation:** The source code and datasets are available at https://github.com/charles-abreu/GRaSP
**Contact:** sabrina@ufv.br

## PROBselect: accurate prediction of protein-binding residues from proteins sequences via dynamic predictor selection

Fuhao Zhang[1], Wenbo Shi[1], Jian Zhang[2,3], Min Zeng[1], Min Li[1], **Lukasz Kurgan**[3]
*[1] Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, China.*
*[2] School of Computer and Information Technology, Xinyang Normal University, Xinyang, China.*
*[3] Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA.*

### Abstract

**Motivation:** Knowledge of protein-binding residues (PBRs) improves our understanding of protein-protein interactions, contributes to the prediction of protein functions, and facilitates protein-protein docking calculations. While many sequence-based predictors of PBRs were published, they offer modest levels of predictive performance and most of them cross-predict residues that interact with other partners. One unexplored option to improve the predictive quality is to design consensus predictors that combine results produced by multiple methods.

**Results:** We empirically investigate predictive performance of a representative set of nine predictors of PBRs. We report substantial differences in predictive quality when these methods are used to predict individual proteins, which contrast with the dataset-level benchmarks that are currently used to assess and compare these methods. Our analysis provides new insights for the cross-prediction concern, dissects complementarity between predictors, and demonstrates that predictive performance of the top methods depends on unique characteristics of the input protein sequence. Using these insights, we developed PROBselect, first-of-its-kind consensus predictor of PBRs. Our design is based on the dynamic predictor selection at the protein level, where the selection relies on regression-based models that accurately estimate predictive performance of selected predictors directly from the sequence. Empirical assessment using a low-similarity test dataset shows that PROBselect provides significantly improved predictive quality when compared with the current predictors and conventional consensuses that combine residue-level predictions. Moreover, PROBselect informs the users about the expected predictive quality for the prediction generated from a given input protein.

**Availability:** PROBselect is available at http://bioinformatics.csu.edu.cn/PROBselect/home/index
**Contact:** lkurgan@vcu.edu

## New mixture models for decoy-free false discovery rate estimation in mass-spectrometry proteomics

**Yisu Peng**[1], Shantanu Jain[1], Yong Fuga Li[2], Michal Greguš[3,4], Alexander R. Ivanov[3,4], Olga Vitek [1,4], Predrag Radivojac[1]

[1] *Khoury College of Computer Sciences, Northeastern University, Boston, Massachusetts, USA.*
[2] *Illumina Inc., San Diego, California, USA.*
[3] *Department of Chemistry and Chemical Biology, Northeastern University, Boston, Massachusetts, USA.*
[4] *Barnett Institute of Chemical and Biological Analysis, Northeastern University, Boston, Massachusetts, USA.*

### Abstract

**Motivation:** Accurate estimation of false discovery rate (FDR) of spectral identification is a central problem in mass spectrometry-based proteomics. Over the past two decades, target-decoy approaches (TDAs) and decoy-free approaches (DFAs), have been widely used to estimate FDR. TDAs use a database of decoy species to faithfully model score distributions of incorrect peptide-spectrum matches (PSMs). DFAs, on the other hand, fit two-component mixture models to learn the parameters of correct and incorrect PSM score distributions. While conceptually straightforward, both approaches lead to problems in practice, particularly in experiments that push instrumentation to the limit and generate low fragmentation-efficiency and low signal-to-noise-ratio spectra.

**Results:** We introduce a new decoy-free framework for FDR estimation that generalizes present DFAs while exploiting more search data in a manner similar to TDAs. Our approach relies on multi-component mixtures, in which score distributions corresponding to the correct PSMs, best incorrect PSMs, and second-best incorrect PSMs are modeled by the skew normal family. We derive EM algorithms to estimate parameters of these distributions from the scores of best and second-best PSMs associated with each experimental spectrum. We evaluate our models on multiple proteomics datasets and a HeLa cell digest case study consisting of more than a million spectra in total. We provide evidence of improved performance over existing DFAs and improved stability and speed over TDAs without any performance degradation. We propose that the new strategy has the potential to extend beyond peptide identification and reduce the need for TDA on all analytical platforms.

**Availability:** https://github.com/shawn-peng/FDR-estimation

## APOD: accurate sequence-based predictor of disordered flexible linkers

**Zhenling Peng**[1], Qian Xing[1], Lukasz Kurgan[2]
[1]*Center for Applied Mathematics, Tianjin University, Tianjin, China.*
[2]*Department of Computer Science, Virginia Commonwealth University, Richmond, VA, USA.*

### Abstract

**Motivation:** Disordered flexible linkers (DFL) are abundant and functionally important intrinsically disordered regions that connect protein domains and structural elements within domains and which facilitate disorder-based allosteric regulation. While computational estimates suggest that thousands of proteins have DFLs, they were annotated experimentally in less than 200 proteins. This substantial annotation gap can be reduced with the help of accurate computational predictors. The sole predictor of DFLs, DFLpred, trades-off accuracy for shorter runtime by excluding relevant but computationally-costly predictive inputs. Moreover, it relies on the local/window-based information while lacking to consider useful protein-level characteristics.

**Results:** We conceptualize, design and test APOD (Accurate Predictor Of DFLs), the first highly accurate predictor that utilizes both local and protein-level inputs that quantify propensity for disorder, sequence composition, sequence conservation and selected putative structural properties. Consequently, APOD offers significantly more accurate predictions when compared with its faster predecessor, DFLpred, and several other alternative ways to predict DFLs. These improvements stem from the use of a more comprehensive set of inputs that cover the protein-level information and the application of a more sophisticated predictive model, a well-parametrized Support Vector Machine. APOD achieves AUC = 0.82 (28% improvement over DFLpred) and MCC = 0.42 (180% increase over DFLpred) when tested on an independent/low-similarity test dataset. Consequently, APOD is a suitable choice for accurate and small-scale prediction of DFLs.

**Availability:** https://yanglab.nankai.edu.cn/APOD/
**Contact:** zhenling@tju.edu.cn or lkurgan@vcu.edu

## A Glimpse into Global Bioinformatics Communities: Latin America

This session has been organised in collaboration with the **Iberoamerican Society for Bioinformatics (SolBio)**.

Session Chaired by **Javier De Las Rivas**

*Spanish National Research Council (CSIC), University of Salamanca (USAL).*

## The Brazilian Initiative on Precision Medicine: Strategies and Findings

**Benilton Carvalho**
*University of Campinas, Brazil.*

### Abstract

The Brazilian Initiative on Precision Medicine (BIPMed) was created by the joint effort of five Research Innovation and Dissemination Centers (RIDCs) supported by FAPESP in Brazil. It aims to assist in the consolidation of the Precision Medicine infrastructure in Brazil by bringing together different stakeholders (decision makers, researchers, medical doctors, patients). To date, BIPMed has genotyped and sequenced several hundreds of samples, in addition to setting up computational solutions to assist in data processing. In this talk, we will present these solutions and findings derived from these data.

## Building city-scale genomic cartographies for improved response to emerging infectious diseases

**Gregorio Iraola**
*Microbial Genomics Laboratory, Institut Pasteur de Montevideo, Uruguay.*

## Molecular Modeling of Ion Channels-Associated Diseases

**Wendy González Diaz**
*Center for Bioinformatics and Molecular Simulations (CBSM), University of Talca, Chile.*

### Abstract

There are many diseases related to ion channels. For example, the utilization of blockers acting simultaneously on Na V 1.5 and K V 1.5 channels or K V 1.5 and TASK-1 channels have been suggested as innovative strategies against atrial fibrillation (AF), the most common type of arrhythmia. One single ion channel can contain about 7,000 atoms and this kind of proteins have been the focus of computational approaches in my group to relate their three-dimensional structure to their physiological function. Usually, we embed the protein in a lipid membrane patch surrounded by water molecules and ions totaling around 100,000 atoms. Molecular dynamics simulations provide information on the kinetics of ion channels.

We have been using additional computational approaches - such a Pharmacophore-Based Virtual Screening or docking - for finding ion channels modulators and for elucidating the action mechanisms of small molecules. In this talk, I will present results of collaborations between my research group and others from Latin America and Germany related to ion channel structure-function. A key aspect of this common research is to validate our models experimentally, or to explain their experimental findings by means of molecular modeling.

## Logical modeling of dendritic cells in vitro differentiation from human monocytes unravels novel transcriptional regulatory interactions

Karen J. Nuñez-Reza[1], Aurélien Naldi[2], Arantza Sanchéz-Jiménez[1], Ana V. Leon-Apodaca[1], M. Angélica Santana[3], Morgane Thomas-Chollier[2], Denis Thieffry[2], **Alejandra Medina-Rivera**[1]

[1] *Laboratorio Internacional de Investigación sobre el Genoma Humano, Universidad Nacional Autónoma de México, Juriquilla, México.*
[2] *Computational Systems Biology team, Institut de Biologie de l'Ecole normale supérieure, Inserm, CNRS, Université* PSL, Paris, France.
[3] *Centro de Investigación en Dinámica Celular (IICBA), Universidad Autónoma del Estado de Morelos, Cuernavaca, México.*

### Abstract

Dendritic cells (DCs) are antigen-presenting cells commonly used to develop immunotherapies. In 1994, Sallusto and Lazavecchia described a protocol to obtain DCs (called moDCs) from monocytes using a culture medium containing IL4 and CSF2 (Sallusto and Lanzavecchia 1994), which are both required for moDC differentiation and maintenance. IL4 and CSF2 activate the signaling pathways JAK/STAT, NFKB, PI3K, and MAPK, which have been widely studied. The main transcription factors (TFs) targeted by these two signaling cascades in monocytes are known, but how genes specific for DCs are activated remains to be clarified.

In this work, we performed a comprehensive literature search to identify the TFs controlling the differentiation of moDCs from monocytes, and we integrated them together with IL4 and CSF2 signaling pathways in a logical model. To complete this model, we further considered experimentally validated and computationally predicted TF-DNA regulatory interactions inferred from the analysis of ChIP-seq (H3K4me1, H3K4me3, H3K27ac, H3K36me3, H3K9me3, H3K27me3) and RNA-seq data generated by the Blueprint consortium, for monocytes, as well as for monocyte-derived dendritic cells and macrophages. Interestingly, this analysis of Blueprint data enabled us to identify active and poised promoters, as well as the genes up and down-regulated in each cell type. Macrophages were included in our model for DC differentiation since they correspond to an alternative fate when monocytes are incubated with CSF2 but no IL4.

Built with the software GINsim (Naldi, Hernandez, Abou-Jaoudé, et al. 2018), the model was further analysed using the tools integrated into the CoLoMoTo toolbox (Naldi, Hernandez, Levy, et al. 2018). The model gives raise to four stable states, each corresponding to one combination of input cytokines (CSF2 and IL4), and that generate outputs that matches our four expected cellular states. We further challenged our model by performing *in silico* simulations of nine single gene mutants documented in the literature, which result in stable states matching the outcomes of the corresponding experiments.

In conclusion, our logical model analysis supports the roles of 96 novel regulatory interactions inferred from blueprint data, which should now be assessed experimentally.

## A Glimpse into Global Bioinformatics Communities: Europe

This session has been organised in collaboration with **ELIXIR** **(European Life Science Infrastructure for Biological Information).**

Session Chaired by **Jen Harrow**
*ELIXIR, United Kingdom.*

### 3D-Beacons: An integrative, distributed platform for FAIR access to experimental and predicted macromolecular structures

**Sameer Velanker**[1]**,** Mihaly Varadi[1] , Christine Orengo[2], Toirsten Schwede[3], Maria-Jesus Martin[4]

[1] *Protein Data Bank in Europe, United Kingdom.*
[2] *Genome3D, United Kingdom.*
[3] *SWISSMODEL, Switzerland.*
[4]*Uniprot, United Kingdom.*

### Abstract

The Protein Data Bank in Europe - Knowledge Base (PDBe-KB) is an open collaboration between PDBe and world-leading specialist data resources who contribute functional and biophysical annotations derived from molecular structure data in the Protein Data Bank (PDB). The goal of PDBe-KB is to place macromolecular structure data in their biological context by developing standardised data exchange formats and via integrating functional annotations from the contributing partner resources into a knowledge graph that can provide valuable biological insights.

Unfortunately, the overwhelming majority of the known protein sequence space is not represented in the PDB, with merely 0.027% of sequences being directly mapped to structures. However, rapid advances in structure modelling methods have resulted in tools that can provide high-quality models. 3D-Beacons is an integrative, distributed platform that aims to provide FAIR access to both experimental (PDB) and predicted (Genome3D, SWISS-MODEL) protein structures, combined with their functional annotations in a transparent way, such that data provenance is clear to the end-users.

## 3DBionotes-COVID19 Edition: bringing together structural and functional information on SARS-CoV-2 proteome

**Jose Ramon Macias**[1], Ruben Sanchez-Garcia[1], Pablo Conesa[1], Erney Ramírez-Aportela[1], Marta Martinez Gonzalez[1], Carlos Oscar Sanchez Sorzano[1], Jose Maria Carazo[1]

*[1]Spanish National Bioinformatics Institute, INB/ELIXIR-ES. Biocomputing Unit, CNB-CSIC. Instruct Image Processing Center, I2PC. Spain*

### Abstract

3DBionotes-WS, an ELIXIR Recommended Interoperability Resource, is a set of Web Services providing multiple annotations oriented to structural biology analysis, plus a website interface including a 3D viewer for macromolecular models and Cryo-EM derived maps. The display is fully interactive and includes functional, genomic, proteomic and structural feature annotations.

COVID-19 pandemic calls for high quality data suitable for the required structural analysis of SARS-CoV-2 and other related viruses or interacting human macromolecules. To fulfill this demand, we have set up an entry website aiming to collect and provide easy access to all available structural information related to the topic.

A schematic representation of the virus proteome at the top of the page is used as an index, leading the user to panels with information for every protein. Every panel has a set of sections containing small images representing different entries organised by source, PDB or EMDB. When available, the refined structures (PDB-Redo, Isolde) are provided too. Structures containing interacting proteins or protein binding ligands are collected in its own panel. A set of computational models are also provided, as well as related structures from SARS-CoV or other coronaviruses. When clicking in each structure image, the corresponding entry and its annotations, including automatically retrieved and manually curated, are loaded into the 3D viewer.

### COVoc: a COVID-19 ontology to support literature triage

**Déborah Caucheteur**[1], Paola Roncaglia[2], Julien Gobeill[1], Zoë May Pendlington[2], Luc Mottin[1], Nicolas Matentzoglu[2], David Osumi-Sutherland[2], Donat Agosti[3], Helen Parkinson[2], Patrick Ruch[1]

*[1]HES-SO/HEG, SIB, Switzerland.*
*[2]EMBL-EBI, United Kingdom.*
*[3]SIB, Platzi, Switzerland*

#### Abstract

We report on the design of an original resource: the COVID-19 Vocabulary. A Google Spreadsheet was made available on March 25, 2020 to collect feedback from a wide range of research communities. Over several hackathons, a relatively stable resource was obtained containing 9 sheets for as many unambiguous semantic types, corresponding to different aspects of the pandemics (host organisms, pathogenicity, gene and gene products, barrier gestures, treatments, etc). When possible, each of the 543 concepts - and 1625 synonyms - were then mapped to some pre-existing reference ontology, resulting in 2770 cross-references to core onto-terminological resources. We then used the COVoc concepts and synonyms to fully annotate MEDLINE and PMC, as well as the CORD-19 preprint collection, for half a billion annotations (N=695 665 604  by August 18, 2020), which allow researchers and biocurators to navigate from the literature to a wide range of ELIXIR core databases (e.g. UniProt, ChEBI). A triage service was built on top of these annotations to search for COVID-related information across the pre-defined aspects with daily updates. The ontological resource is browsable in OLS and available on GitHub. The curation-support literature triage demonstrator can be found here: http://candy.hesge.ch/CovidTriage/. All APIs and data are publicly available.

## The ELIXIR Tools Platform - solutions for COVID-19 research

**Bjoern Gruening**[1], Herve Menager[2], Jennifer Harrow[3], Salvador Capella-Gutierrez[4]

*[1]Uni-Freiburg, Germany.*
*[2]Institut Pasteur, France.*
*[3]ELIXIR, United Kingdom.*
*[4]Barcelona Supercomputing Center (BSC), Spain.*

### Abstract

The Tools Platform (https://elixir-europe.org/platforms/tools) is one of 5 ELIXIR platforms with the aim to improve software quality and documentation and to help life scientists to find, deploy and benchmark software tools, including workflows. In this talk we will demonstrate the recent achievements from the Tools Platform by highlighting how different aspects of the platform have contributed to COVID-19 research, recent applications and use cases. We will begin with a brief introduction to the ELIXIR Tools platform and update of the progress being made around the Tools Ecosystem. This ecosystem is an ongoing effort to federate the tools and services information into a centralized and transparent repository. This repository will be a cornerstone for the sustainability and interoperability between services, both within (bio.tools, BioContainers, OpenEBench, Bioconda) and outside (e.g. Workflow Hub, Debian Med, Galaxy, etc.) the Tools Platform. Participants will be walked through an end to end real-life example of how the tools platform and registries are utilized for COVID-19 research and how a distributed ELIXIR-wide compute network enabled researchers to conduct COVID-19 research at large scale.

## Nordic development on federating the EGA

Lucile Soler[1], Juha Törnroos[2], **Abdulrahman Azab**[3], Stefan Negru[2], Dmytro Titov[3], Johan Viklund[1], Joakim Bygdell[1], Jon Ander Novella[1], Dimitrios Bampalikis[1]

[1] *ELIXIR-SE, Sweden.*
[2] *ELIXIR-FI, Finland.*
[3] *ELIXIR-NO, Norway.*

### Abstract

Research in biomedical sciences aims ultimately at curing diseases and improving the quality of life. Successful research in the field requires the use of human data of various types and from various sources. To support these three ELIXIR nodes (Finland, Sweden, and Norway) from the Nordics, together with the Nordic e-Infrastructure Collaboration NeIC, are working together in NeIC Tryggve2 project. One of the project objectives is to establish a technological basis to operate sensitive data archives in the participating ELIXIR nodes. This will enable the archiving of sensitive human data inside the national boundaries and researchers to both archive and access their valuable data in a secure manner.

The aim is also to connect these sensitive data archives to the European Genome-Phenome Archive (EGA) and create a federated archive that will support the researchers all across Europe. The EGA is operated by the EMBL-EBI, UK, and the CRG, Spain. The core components support the base functionalities of 1) data upload, 2) file ingestion & archiving, and 3) controlled data access. In the federated EGA set-up, the user interfaces for metadata submission, metadata administration, persistent identifier generation, and the same dataset catalog will be provided by functionality in the Central EGA at EBI and CRG. The system is being built upon a container-based microservices architecture pattern, with a design that should be fairly agnostic of the underlying compute and storage infrastructure, be they virtual or bare-metal.

The current state of technological development is that there is a technological capability in all three Nordic countries to join the EGA federation. The development efforts are currently focused on stabilising the solution and the infrastructure in general. Also, the capability to operate the archives in a standalone manner, without the EGA, is being developed.

## FAIR data by design

**Flora D'Anna**[1], Vahid Kiani[1], Stuart Owen[2], Carole Goble[2], Frederik Coppens[1]

[1] *ELIXIR-BE, VIP, Belgium.*
[2] *ELIXIR-UK, University of Manchester, United Kingdom.*

### Abstract

Despite the general demand for research data that meet FAIR principles (Findability, Accessibility, Interoperability and Reusability), there are very few tools available that try to help researchers generate FAIR data by design. FAIRDOMHub/SEEK is a very flexible platform for sharing research projects; ISA tools and formats are a general purpose framework to collect and communicate complex metadata. Both systems are used by several scientific communities, within ELIXIR and beyond, as suitable systems for research data management. However, additional features should be implemented in order to allow researchers to make their data FAIR by design.

We will describe how we aim to provide support for researchers to publish FAIR data by collecting the necessary metadata as part of the research project, rather than an afterthought. Based on e.g. the ELIXIR Deposition Databases, researchers will be able to choose the appropriate metadata schema to describe study design and assays. By making this a step-wise process throughout the data life cycle, the quality of the metadata will be improved. Integration of the submission procedures to the commonly used deposition databases will allow for easy deposition of high quality data, contributing to Open and FAIR science. By adhering to standards such as ISA and the metadata checklists, we ensure interoperability and enable implementation of the same principles in other platforms.

## Meet our Sponsors

## Sponsored by European Bioinformatics Institute (EMBL-EBI)

Monday, 7 September | 13:10 - 13:30 (CEST)

### What's new at EMBL-EBI?

**Sarah Morgan** and **Amonida Zadissa**
*European Bioinformatics Institute (EMBL-EBI), United Kingdom.*

The European Bioinformatics Institute (EMBL-EBI) is a worldwide leader in bioinformatics. We make the world's public biological data freely available to the scientific community by providing data resources and tools. We also perform computational research, provide professional training in bioinformatics and collaborate with the private sector to enable innovation in drug discovery.
This year at ECCB2020, we will give you a glimpse of the latest updates from the EMBL-EBI data resources. We will update you on the progress of our new training portal where you can find a selection of our new look training courses. You will also hear about our new COVID-19 Data Portal, which lets you access and analyse COVID-19 related reference data and specialist datasets including sequences, expression data, proteins and structural features.

Don't forget to visit our virtual exhibition booth to hear more information about these features, and much more that EMBL-EBI has to offer, including our latest job vacancies and a demonstration of the new COVID-19 Data Portal.

## Sponsored by The Spanish Supercomputing Network (RES)

Tuesday, 8 September | 13:10 - 13:30 (CEST)

### The Spanish Supercomputing Network offers HPC resources to the scientific community

**Jordi Mas**
*The Spanish Supercomputing Network (RES), Spain.*

The Spanish Supercomputing Network (RES, Red Española de Supercomputación, www.res.es) was created in 2006, in response to the need of the Spanish scientific community for intensive calculation resources. The RES is a Unique Scientific and Technical Infrastructure (ICTS) distributed throughout Spain, which aims to support the development of top-quality cutting-edge research. In 2020, the RES comprises 13 supercomputers and 12 institutions, and it is coordinated by the Barcelona Supercomputing Center – Centro Nacional de Supercomputación (BSC-CNS).

The RES aims to manage high performance computing technologies to promote the progress of excellent science and innovation in Spain. It offers its resources through an open competitive access. Thus, the application procedure is the same for all the RES nodes and based on criteria of efficacy, efficiency and transparency. This common access guarantees optimal use of the resources available in the network (computing, storage, parallelization, data management, etc.). The RES also promotes and executes common interest actions for all its nodes. For instance, it promotes shared investment plans, training and dissemination, and joint participation in national and international calls and projects.

**Technical Secretariat**

Bco Congresos

Barceló GRUPO

Pl. Europa, 17-19 1st floor
08908 L'Hospitalet de Llobregat Barcelona, Spain
Ph: +34 93 882 38 78
eccb2020@bcocongresos.com