ECCB2022

21st European Conference on Computational Biology

Planetary Health and Biodiversity

BOOK OF ABSTRACTS

12-21 September 2022
Sitges, Barcelona

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Spanish National Bioinformatics Institute

ELIXIR SPAIN

eccb2022.org

Table of content

Highlight Papers

Data

- PharmacoDB 2.0: improving scalability and transparency of in vitro pharmacogenomics analysis
- Polympact: exploring functional relations among common human genetic variants
- Scaling up oligogenic diseases research with OLIDA: the Oligogenic Diseases Database
- The AIMe registry for artificial intelligence in biomedical research

Genes

- Analysis of eukaryotic lincRNA sequences indicates signatures of hindered translation linked to selection pressure

Genomes

- A global metagenomic map of urban microbiomes and antimicrobial resistance
- Higher order genetic interactions switch cancer genes from two-hit to one-hit drivers
- plotsr: Visualising structural similarities and rearrangements between multiple genomes
- Revisiting genetic artifacts on DNA methylation microarrays exposes novel biological implications
- Towards a metagenomics machine learning interpretable model for understanding the transition from adenoma to colorectal cancer
- Unlocking capacities of genomics for the COVID-19 response and future pandemics

Proteins

- AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models
- Missense variants in human ACE2 strongly affect binding to SARS- CoV-2 Spike providing a mechanism for ACE2 mediated genetic risk in Covid-19: A case study in affinity predictions of interface variants
- Online predictions for protein biophysical features and their conservation
- PDBe-KB: collaboratively defining the biological context of structural data
- The clinical importance of tandem exon duplication-derived substitutions
- TITAN: T-cell receptor specificity prediction with bimodal attention networks

Systems

- Comprehensive prediction of robust synthetic lethality between paralog pairs in cancer cell lines
- Interpretable systems biomarkers predict response to immune- checkpoint inhibitors

Data

- A cloud-based resource to manage, access and explore omics datasets in multiuser environments
- Global biodata resources: challenges to long-term sustainability of a crucial data infrastructure
- Introducing X-Omics, the central multi-omics data integration and AI modeling platform for biomarker data at Merck
- Open Targets: A Platform for Biological Data Integration
- Scalable In-memory paradigm for genomics data processing

Genes

- Bioinformatics methods for the analysis of rare-disease patient data – applications for target discovery and obtaining phenotype associations
- Power analysis of cell-type deconvolution across human tissues

Genomes

- ChiTaH: a fast and accurate tool for identifying known human chimeric sequences from high-throughput sequencing data
- METALoci, identification of spatial enhancer hubs.
- Whole-genome sequencing analysis of food enzyme products reveals contaminations with genetically modified microorganism of related origin

Proteins

- A near-full compression of SARS-CoV-2 peptidome using UNIQmin

Systems

- Cancer patient stratification and molecular mechanism identification using patient clinotypes and transcriptomics embeddings
- Evaluation of Machine Learning Strategies for Imaging Confirmed Prostate Cancer Recurrence Prediction on Electronic Health Records
- AI microbiome-based recommendation system for improving soil health with bio-stimulants

Climate crisis and health

- Climate-sensitive disease outbreaks in the aftermath of extreme climatic events
- Infectious disease decision-support tools to enhance resilience in climate change hotspots
- Real-time Genomics for One Health
- The Catalan Initiative for the Earth Biogenome Project

ELIXIR talks

- AHoJ: rapid, tailored search and retrieval of apo and holo protein structures
- Current activities of the ELIXIR Machine Learning Focus Group
- Data security entry considerations for post covid data
- Genome-wide metabolic annotation for Methanocaldococcus (Methanococcus) jannaschii, the first member of the Archaea to be sequenced a quarter of a century ago
- Open-source genome-scale metabolic models: why and how
- Rare disease specific FAIR Maturity Indicators
- Using the IDP-KG to enable IDPcentral
- ELIXIR Europe: overview and opportunities
- Research data management (RDM) in ELIXIR and insight into the RDM Toolkit
- Creating paths for the development and application of bioinformatics in Mexico

Institutional talks

- Spanish Supercomputing Network (RES)
- Ersilia, a hub of open-source AI/ML models for drug discovery and global health
- The Bioinfo4Women Programme: towards gender equity and diversity in science

Sponsored talks

- Building Biomedical Knowledge Graphs for In-Silico Drug Discovery

Proceedings

Data

Detecting DNA of novel fungal pathogens using ResNets and a curated fungi-hosts data collection

Jakub M Bartoszewicz (Hasso Plattner Institute), Ferdous Nasri (Hasso Plattner Institute), Melania Nowicka (Hasso Plattner Institute) and Bernhard Y Renard (Hasso Plattner Institute).

Abstract*:*

Emerging pathogens are a growing threat, but large data collections and approaches for predicting the risk associated with novel agents are limited to bacteria and viruses. Pathogenic fungi, which also pose a constant threat to public health, remain understudied. We present a curated collection of fungal host range data, comprising records on human, animal and plant pathogens, as well as other plant-associated fungi, linked to publicly available genomes. The neural networks trained using our data collection enable accurate detection of novel fungal pathogens. We develop learned, numerical representations of the collected genomes, visualize the landscape of fungal pathogenicity and train multi-class models predicting if next-generation sequencing reads originate from novel fungal, bacterial or viral threats. A curated set of over 1,400 genomes with host and pathogenicity metadata supports training of machine learning models and sequence comparison, not limited to the pathogen detection task.

Proceedings

Data

Exploiting Pretrained Biochemical Language Models for Targeted Drug Design

Gökçe Uludoğan (Bogazici University), Arzucan Özgür (Bogazici University), Elif Özkırımlı (Roche AG), Kutlu Ülgen (Bogazici University) and Nilgün Karalı (Istanbul University).

Abstract*:*

The development of novel compounds targeting proteins of interest is crucial in the pharmaceutical industry. Target-specific molecule generation has been viewed as a translation between the protein language and the chemical language. Such a model is limited by the availability of interacting protein-ligand pairs. To address this, we propose exploiting pretrained biochemical language models utilizing large amounts of unlabeled sequences to initialize targeted models. We investigate two warm start strategies: (i) a one-stage strategy where the initialized model is trained on targeted molecule generation (ii) a two-stage strategy containing a pre-finetuning on molecular generation followed by target-specific training. The results show that the warm-started models perform better than a baseline trained from scratch. The proposed warm-start strategies achieve similar results to each other with respect to widely used metrics. However, docking evaluation of the generated compounds for novel proteins suggests that the one-stage strategy generalizes better than the two-stage strategy.

Proceedings

Data

Linking cells across single-cell modalities by synergistic matching of neighborhood structure

Borislav Hristov (University of Washington), Jeffrey Bilmes (University of Washington) and William Noble (University of Washington).

Abstract*:*

Integration of disjoint single-cell multi-omics data has been a particularly challenging problem because the different measurements typically do not share any features and there is no correspondence information available. We present an approach, Synmatch, which tackles the problem in unsupervised fashion by exploiting the neighborhood structure in each modality to find a matching of the cells from two different multi-omics datasets via discrete optimization. The key idea behind Synmatch is that the same cell, when measured in two different modalities, is likely to have similar sets of neighboring cells in the two spaces. We use this intuition to formulate the matching problem as a supermodular optimization over the neighborhood structure of the two modalities, and we solve the problem using a fast greedy heuristic that offers good theoretical guarantees.

Proceedings

Data

SimBu: Bias-aware simulation of bulk RNA-seq data with variable cell type composition

Alexander Dietrich (Chair of Experimental Bioinformatics, Technical University of Munich, 85354 Freising, Germany), Gregor Sturm (Biocenter, Institute of Bioinformatics, Medical University of Innsbruck, 6020 Innsbruck, Austria), Lorenzo Merotto (Institute of Molecular Biology, University of Innsbruck, 6020 Innsbruck, Austria), Federico Marini (Institute of Medical Biostatistics, Epidemiology and Informatics, Johannes Gutenberg University Mainz, Mainz, Germany), Francesca Finotello (Institute of Molecular Biology, University of Innsbruck, 6020 Innsbruck, Austria) and Markus List (Chair of Experimental Bioinformatics, Technical University of Munich, 85354 Freising, Germany).

Abstract*:*

As tissues are typically composed of various cell types, deconvolution tools have been developed to computationally infer their cellular composition from bulk RNA sequencing (RNA-seq) data. To assess deconvolution performance, gold-standard datasets are indispensable; experimental techniques can provide these, but cannot be systematically applied to the cell types and tissues profiled with transcriptomics. The simulation of 'pseudo-bulk' data using single-cell RNA-seq (scRNA-seq) expression profiles in pre-defined proportions offers a scalable alternative, making it feasible to create in silico gold-standard datasets. However, at present, no simulation software for generating pseudo-bulk RNA-seq data exists.

We developed SimBu, an R package capable of simulating pseudo-bulk samples and designed to test specific features of deconvolution methods. A unique feature of SimBu is the modelling of cell-type-specific total mRNA bias using scaling factors. We show that SimBu can generate realistic pseudo-bulk data and illustrate the impact of mRNA bias on the evaluation of deconvolution tools.

Proceedings

Data

This is GlycoQL

Catherine Hayes (University of Geneva), Vincenzo Daponte (University of Geneva), Julien Mariethoz (SIB Swiss Intitute of Bioinformatics) and Frederique Lisacek (SIB Swiss Intitute of Bioinformatics).

Abstract*:*

Glycosylation is a frequent posttranslational modification that impacts most cell surface or secreted protein structure and function. We have previously developed a tree-based ontology to represent, match and compare glycan structures. We now introduce GlycoQL, a query language for the actual implementation of a glycan (sub)structure search. The aim is to provide the means to consistently perform this type of search on a glycan structure triple store without expert knowledge of SPARQL. The methodology is described and illustrated with a use-case focused on SARS-CoV-2 spike protein glycosylation. We show how to enhance site annotation with federated queries involving UniProt and GlyConnect, our glycoprotein database.

Proceedings

Genes

Efficient Permutation-based Genome-wide Association Studies for Normal and Skewed Phenotypic Distributions

Maura John (TUM Campus Straubing for Biotechnology and Sustainability & Weihenstephan Triesdorf University of Applied Sciences), Markus Ankenbrand (Center for Computational and Theoretical Biology, University of Würzburg), Carolin Artmann (Center for Computational and Theoretical Biology, University of Würzburg), Jan Freudenthal (Center for Computational and Theoretical Biology, University of Würzburg), Arthur Korte (Center for Computational and Theoretical Biology, University of Würzburg) and Dominik Grimm (TUM Campus Straubing for Biotechnology and Sustainability & Weihenstephan Triesdorf University of Applied Sciences).

Abstract*:*

Genome-wide Association Studies (GWAS) are an integral tool for studying the architecture of complex genotype and phenotype relationships. Linear Mixed Models (LMMs) are commonly used to detect associations between genetic markers and a trait of interest. Assumptions of LMMs include a normal distribution of the residuals and that the genetic markers are independent and identically distributed. However, both assumptions are often violated in real data. Permutation-based methods can help to overcome some of these limitations and provide more realistic thresholds for the discovery of true associations. Still, in practice they are rarely implemented due to the high computational complexity. We propose permGWAS, an efficient LMM reformulation based on 4D-tensors that can provide permutation-based significance thresholds. We show that our method outperforms current state-of-the-art LMMs with respect to runtime and that permutation-based thresholds have lower false discovery rates for skewed phenotypes compared to the commonly used Bonferroni threshold.

Proceedings

Genes

Improved NSGA-II algorithms for multi-objective biomarker discovery

Luca Cattelani (University of Eastern Finland) and Vittorio Fortino (University of Eastern Finland (Kuopio campus)).

Abstract*:*

In modern translational research, the development of biomarkers heavily relies on use of omics technologies. A major problem in biomarker development is the discovery phase, which often lead to biomarkers that appear to predict patient diagnosis or prognosis but show up only by chance and not because of any biological connection to the disease. Here, we aim to showcase novel multi-objective genetic algorithms (NSGA2-CH and NSGA2-CHS) to address biomarker discovery as combinatorial optimization problem and find subsets of biomarkers exhibiting different trade-offs between accuracy and set size. Benchmark studies were conducted to find biomarkers predictive of cancer subtype from large-scale transcriptomics datasets. The TCGA-BRCA dataset was used for the training and validation steps, while the SCAN-B dataset was used as independent test set. NSGA2-CH and NSGA2-CHS outperformed alternative methods in most tests. The novel techniques can lead to simpler models, achieving high test accuracy with few selected biomarkers.

Genes

NSF4SL: negative-sample-free contrastive learning for ranking synthetic lethal partner genes in human cancers

Shike Wang (ShanghaiTech University), Yimiao Feng (ShanghaiTech University), Xin Liu (ShanghaiTech University), Yong Liu (Nanyang Technological University), Min Wu (I2R) and Jie Zheng (ShanghaiTech University).

Abstract*:*

Known as the Achilles' heel of cancer, synthetic lethality (SL) points to a gold mine of anti-cancer drug targets. Due to high cost of web-lab screening and availability of identified SL gene pairs, supervised machine learning for SL prediction has been popular. However, most methods formulate SL prediction as binary classification, and hence are limited by the lack of high-quality non-SL data. We propose NSF4SL, a negative-sample-free SL prediction model based on self-supervised contrastive learning. NSF4SL prioritizes SL partners for a given gene without any negative data and significantly outperforms all baselines which require negative samples. To the best of our knowledge, this is the first attempt to formulate SL prediction as a gene ranking problem, which is more practical than the conventional formulation as binary classification. NSF4SL is the first contrastive learning method for SL prediction, and its outstanding generalizability provides an avenue for discovery of novel SLs.

Genomes

3CAC: improving the classification of phages and plasmids in metagenomic assemblies using assembly graphs

Lianrong Pu (Tel Aviv University) and Ron Shamir (Tel Aviv University).

Abstract*:*

Viruses and plasmids are part of microbial communities and play a major role in disease and in antibiotic resistance. In metagenome sequence assembly, identifying virus and plasmid contigs is a hard task, since they tend to form shorter contigs and are overwhelmed by a larger mass of bacterial contigs. Here we present 3CAC, a new classifier that builds on existing classifiers and exploits the structure of the assembly graph for the classification of contigs into bacterial, viral, plasmidic, and unknown contigs. In simulated and real metagenomes of short and long reads, 3CAC outperformed the state-of-the-art algorithms.

Genomes

CRISPRtracrRNA: Robust approach for CRISPR tracrRNA detection

Alexander Mitrofanov (University of Freiburg), Marcus Ziemann (University of Freiburg), Omer Alkhnbashi (University of Freiburg), Wolfgang Hess (University of Freiburg) and Rolf Backfen (University of Freiburg).

Abstract*:*

The CRISPR-Cas9 system is a type-II system that has rapidly become the most versatile tool for genome engineering. It consists of two components, the Cas9 effector-protein, and a single guide RNA that combines the spacer with the tracrRNA, a trans-activating small RNA required for both crRNA maturation and interference. While there are well-established methods for screening Cas-effector proteins and CRISPR-arrays, the detection of tracrRNA remains the bottleneck in detecting Class-II systems. We introduce new pipeline CRISPRtracrRNA for screening and evaluation of tracrRNA candidates. This pipeline combines evidence from different components of the Cas9-sgRNA-complex. The core is a newly developed structural model via covariance models from a sequence-structure-alignment of experimentally validated tracrRNAs. As additional evidence, we determine the terminator signal (required for the tracrRNA transcription) and the RNA-RNA interaction between the CRISPR array repeat and the 5'-part of the tracrRNA. As additional evidence, we detect the cassette containing Cas9 and Cas12.

Genomes

DeepZF: Improved DNA-binding prediction of C2H2-zinc-finger proteins by deep transfer learning

Sofia Aizenshtein-Gazit (Ben-Gurion University) and Yaron Orenstein (Ben-Gurion University).

Abstract*:*

Cys2His2 zinc-finger (C2H2-ZF) proteins are the largest class of human transcription factors and hence play central roles in gene regulation and cell function. C2H2-ZF proteins are characterized by a DNA-binding domain containing multiple ZFs. A subset of the ZFs bind diverse DNA triplets. Despite their central roles, little is known about which of their ZFs are binding and how the DNA-binding preferences are encoded in the amino acid sequence of each ZF.

We present DeepZF, a two-step deep-learning-based pipeline for predicting binding ZFs and their DNA-binding preferences, given only the amino acid sequence of a C2H2-ZF protein. To the best of our knowledge, DeepZF includes the first ZF-binding classifier. Moreover, our predicting DNA-binding preferences model is the first to utilize deep learning for the task. DeepZF achieved a 0.42 average Pearson correlation in motif similarity, outperforming extant methods.

Interpretability techniques, show that DeepZF inferred biologically relevant binding principles.

Proceedings

Genomes

Discovering Significant Evolutionary Trajectories in Cancer Phylogenies

Leonardo Pellegrina (University of Padova) and Fabio Vandin (University of Padova).

Abstract*:*

Tumors evolve through complex processes driven by the accumulation genomic alterations, leading to substantial intra-tumor heterogeneity. While recent advances in single-cell and multi-region sequencing enable to accurately reconstruct such evolutionary processes and to describe the clonal architecture of tumors, the identification of significantly conserved evolutionary trajectories remains a challenging problem.

We present a new algorithm, MASTRO, to discover significantly conserved tumor evolutionary trajectories. MASTRO discovers all conserved trajectories from a collection of phylogenetic trees describing the evolution of a cohort of tumors. MASTRO identifies trajectories encoding complex interactions among genetic alterations, and assesses their statistical significance with a conditional statistical test that evaluates the coherence in the order in which alterations are observed.

When applied to data from non-small-cell lung cancer bulk sequencing and acute myeloid leukemia data from single-cell panel sequencing, MASTRO discovers significant trajectories that both recapitulate known important progression patterns and suggest new potentially interesting evolutionary trajectories.

Genomes

Improving Bacterial Genome Assembly Using a Test of Strand Orientation

Grant Greenberg (University of Illinois at Urbana-Champaign) and Ilan Shomorony (University of Illinois at Urbana-Champaign).

Abstract*:*

Genome assembly is is often confounded by the presence of reverse-complemented repeats, which can lead to incorrect inversions of large segments of the genome. To resolve such repeats, we propose a statistical test based on tetranucleotide frequency (TNF), which determines whether two segments from the same genome are on the same or in opposite strands. For most finished bacterial genomes, the test partitions the genome into two segments of equal length, corresponding to the segments between the DNA replication origin and terminus. We show in several cases where this balanced partition does not occur, the test identifies a potential inverted misassembly which is validated by the presence of a reverse-complemented repeat. After inverting the sequence between the repeat, the balance of the misassembled genome is restored. Furthermore, we show that this TNF-based orientation test can be useful in resolving repeat nodes in assembly graphs, by comparing the orientation of contigs before and after the repeat.

Proceedings

Genomes

SALAI-Net: Species-Agnostic Local Ancestry Inference Network

Benet Oriol Sabat (UPC), Daniel Mas Montserrat (Stanford University), Xavier Giro-i-Nieto (Universitat Politecnica de Catalunya) and Alexander Ioannidis (Stanford University).

Abstract*:*

Local Ancestry Inference (LAI) is the high resolution prediction of ancestry categories across a DNA sequence. LAI is becoming increasingly important in the study of human history, and precision medicine applications, including genome-wide association studies (GWAS) and polygenic risk scores (PRS). Most of modern LAI models do not generalize well between species, chromosomes, or even ancestry groups, requiring re-training for each different setting. Furthermore, such methods can lack interpretability, which is a paramount element in biomedical applications. We present SALAI-Net, a portable statistical LAI method that can be applied on any set of species and ancestries (species-agnostic), requiring only haplotype data and no other biological parameters, leading to an interpretable and fast technique. SALAI-Net outperforms previous methods in accuracy while generalizing between different settings, species, and datasets. Moreover, it is up to two orders of magnitude faster and uses considerably less RAM memory than competing methods. github.com/AI-sandbox/SALAI-Net.

Proteins

APPRIS Principal Isoforms and MANE Select Transcripts Define Reference Splice Variants

Fernando Pozo Ocampo (Spanish National Cancer Research Centre (CNIO)), Laura Martinez Gomez (Centro Nacional de Investigaciones Oncológicas), Jose Manuel Rodriguez (CNIC), Jesús Vázquez (CNIC) and Michael Tress (Spanish National Cancer Research Centre).

Abstract*:*

Ensembl/GENCODE and RefSeq have collaborated to produce MANE, a new reference transcript set for the human genome [1]. MANE defines a single reference isoform per coding gene, and these splice variants are described as a "high-value set of transcripts and corresponding proteins".

But does a single reference transcript per gene make sense from a biological and clinical point of view, or is the idea of a single reference isoform per gene over-simplifying and dangerous, as some believe?

Here we compared MANE Select variants against other reference transcript prediction methods, using data from large-scale proteomics experiments and human genetic variation studies. We find overwhelming support for a single main protein isoform in most coding genes, and that MANE Select is as effective as APPRIS principal isoforms at identifying this main isoform.

1. Morales et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature. 2022.

Proteins

Cross-Modality and Self-Supervised Protein Embedding for Compound-Protein Affinity and Contact Prediction

Yuning You (Texas A&M University) and Yang Shen (Texas A&M University).

Abstract*:*

Rational drug discovery is seeing a new wave of deep learning models with increasing accuracy, while increasingly demanding more explainability. We focus on explainable prediction of compound-protein interactions where intermolecular contact prediction underlies simultaneous affinity prediction. Current methods are challenged in explainability by the availability of protein structures and in accuracy by the availability of compound-protein affinity labels. We introduce a multi-modality and self-supervised learning framework to address the two challenges. Specifically, thanks to the recent breakthroughs of protein structure prediction, we consider protein data as available in both modalities of 1D amino-acid sequences and predicted 2D contact maps and we introduce cross-modality embedding schemes. Moreover, we adopt self-supervised learning, without the need of experimental affinity labels, to pre-train the embeddings. Our results indicate that our framework could improve the accuracy, the explanation, and the generalizability of affinity prediction especially for unseen proteins.

Proteins

DistilProtBert: A distilled protein language model used to distinguish between real proteins and their randomly shuffled counterparts

Yaron Geffen (Bar-Ilan university), Yanay Ofran (Bar-Ilan university) and Ron Unger (Bar-Ilan university).

Abstract*:*

Recently, deep learning NLP models were applied successfully to analyze protein sequences. A major drawback of these models is their size and amount of computational resources they require. Here, we adapted the concept of knowledge distillation to the problem of protein sequence analysis, by developing DistilProtBert, a distilled version of ProtBert. We reduced the size of the model and running time by 50%, and the computational resources needed for pretraining by 98% relative to ProtBert. Using two published tasks, we showed that the performance of DistilProtBert approaches that of ProtBert. We next tested the ability of DistilProtBert to distinguish between real and random protein sequences. The task is highly challenging if the composition is maintained on the level of singlet, doublet and triplet amino acids. Here, we show that DistilProtBert preforms very well on the human proteome singlet, doublet and triplet-shuffled versions with AUC of 0.92, 0.91, and 0.87 respectively.

Proteins

Group-walk, a rigorous approach to group-wise false discovery rate analysis by target-decoy competition

Jack Freestone (University of Sydney), Temana Short (University of Sydney), William Stafford Noble (University of Washington) and Uri Keich (The University of Sydney).

Abstract*:*

Target-decoy competition (TDC) is commonly used for false discovery rate (FDR) control when analyzing tandem mass spectrometry data.

This type of competition-based FDR control has recently gained significant popularity after Barber and Candes used it in controlling the FDR in feature selection in linear regression.

The effectiveness of TDC depends on whether the data is homogeneous, which is often not the case: the data frequently consists of groups with different score profiles or different proportions of true nulls. In such cases, applying TDC while ignoring the group structure typically yields imbalanced lists of discoveries, whereas, as we show, applying TDC separately to each group does not rigorously control the FDR.

Our Group-walk procedure, derived from the recently developed AdaPT, a general framework for controlling the FDR with side-information, rigorously controls the FDR in the competition-based setting while taking into account a given group structure.

Proteins

Insights into performance evaluation of compound-protein interaction prediction methods

Adiba Yaseen (Pakistan Institute of Engineering and Applied Sciences), Imran Amin (National Institute for Biotechnology and Genetic Engineering), Naeem Akhter (Pakistan Institute of Engineering and Applied Sciences), Asa Ben-Hur (Colorado State University) and Fayyaz Ul Amir Afsar Minhas (University of Warwick).

Abstract*:*

What are the limitations of existing machine learning based compound-protein interaction models? How can we analyze the true generalization performance of such models?

In this work, we systematically analyze the impact of several factors affecting generalization performance of CPI predictors that are overlooked in existing work. Using both state of the art approaches by other researchers as well as a simple kernel-based baseline, we have found that effective assessment of generalization performance of CPI predictors requires careful control over similarity between training and test examples. We show that, under stringent perfor-mance assessment protocols, a simple kernel based approach can exceed the predictive performance of existing state of the art methods. We also show how negative examples can be generated for effective training of such predictors. We propose several strategies that can lead to effective target compound screening for drug repurposing and discovery of putative chemical ligands of proteins. Such insights can be useful for the community to design effective CPI predictors.

Systems

Design centering enables robustness screening of pattern formation models

Anastasia Solomatina (TU Dresden, Faculty of Computer Science; Max Planck Institute of Molecular Cell Biology and Genetics), Alice Cezanne (Max Planck Institute of Molecular Cell Biology and Genetics), Yannis Kalaidzidis (Max Planck Institute of Molecular Cell Biology and Genetics), Marino Zerial (Max Planck Institute of Molecular Cell Biology and Genetics) and Ivo F. Sbalzarini (TU Dresden, Faculty of Computer Science; Max Planck Institute of Molecular Cell Biology and Genetics;).

Abstract*:*

We present a computational workflow to screen reaction-diffusion models for their capacity to robustly form Turing patterns anywhere in their parameter space. Reaction-diffusion models are frequently used to describe spatiotemporal processes in biochemical systems, including intra-cellular and tissue-scale pattern formation. Globally characterizing the potential behavior of such models is difficult if they contain many unknown parameters. We propose a screening framework based on the Lp-Adaptation algorithm and the principle of robustness-based model selection. Lp-Adaptation is a statistical method for approximate design centering and robustness estimation in high-dimensional parameter spaces with a computational cost that scales polynomially with dimension. We leverage Lp-Adaptation to globally characterize the capability (or incapability) of reaction-diffusion models to undergo pattern-forming instabilities, and we quantify the robustness with which they do so. We then apply this to the small GTPase Rab5, identifying novel hypothetical molecular mechanisms that drive pattern formation in cellular membrane compartments.

Proceedings

Systems

DrDimont: Explainable drug response prediction from differential analysis of multi-omics networks

Pauline Hiort (Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam), Julian Hugo (Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam), Justus Zeinert (Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam), Nataniel Müller (Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam), Spoorthi Kashyap (Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam), Jagath C. Rajapakse (School of Computer Science and Engineering, Nanyang Technological University), Francisco Azuaje (Genomics England), Bernhard Y. Renard (Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam) and Katharina Baum (Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam).

Abstract:

While it is well established that drugs affect and help patients differently, personalized drug response predictions remain challenging – especially solutions that leverage multiple omics datasets.

We present DrDimont, Drug response prediction from Differential analysis of multi-omics networks. It allows for comparative conclusions between two biological conditions and translates them into differential drug response predictions. DrDimont establishes condition-specific, multi-layer molecular networks, and provides a novel path-based integration step. DrDimont's predictions are fully explainable, they can be directly associated to molecular differences.

We predict differential drug response using transcriptomics, proteomics, phosphosite, and metabolomics measurements and contrast estrogen receptor positive and receptor negative breast cancer patients. DrDimont performs better than drug prediction based on differential protein expression or PageRank, and we find proteomic and phosphosite layers to carry most information.

DrDimont is available as ready-to-use R package that enables versatile network-based integration of multi-omics data.

Systems

GNN-SubNet: disease subnetwork detection with explainable Graph Neural Networks

Bastian Pfeifer (Medical University Graz), Anna Saranti (Medical University Graz) and Andreas Holzinger (Medical University Graz).

Abstract*:*

The tremendous success of graphical neural networks (GNNs) already had a major impact on systems biology research. For example, GNNs are currently used for drug target recognition in protein-drug interaction networks as well as cancer gene discovery and more. Important aspects whose practical relevance is often underestimated are comprehensibility, interpretability, and explainability.

In this work, we present a novel graph-based deep learning framework for disease subnetwork detection via explainable GNNs. Each patient is represented by the topology of a protein-protein network (PPI), and the nodes are enriched with multi-omics features from gene expression and DNA methylation. In addition, we propose a modification of the GNNexplainer that provides model-wide explanations for improved disease subnetwork detection.

The proposed methods and tools are implemented in the GNN-SubNet Python program, which we have made freely available on our GitHub for the international research community (https://github.com/pievos101/GNN-SubNet).

Systems

MERRIN: MEtabolic Regulation Rule INference from time series data

Kerian Thuillier (CNRS), Caroline Baroukh (INRAE), Alexander Bockmayr (Frein Berlin University), Ludovic Cottret (INRAE), Loïc Paulevé (CNRS/LaBRI, Bordeaux, France) and Anne Siegel (CNRS).

Abstract*:*

Many techniques have been developed to infer Boolean regulations from a prior knowledge network and experimental data. Existing methods are able to reverse-engineer Boolean regulations for transcriptional and signaling networks, but they fail to infer regulations that control metabolic networks.

We present a novel approach to infer Boolean rules for metabolic regulation from time series data and a prior knowledge network.

Our method is based on a combination of answer set programming and linear programming.

By solving both combinatorial and linear arithmetic constraints we generate candidate Boolean regulations that can reproduce the given data when coupled to the metabolic network.

We evaluate our approach on a core regulated metabolic network and show how the quality of the predictions depends on the available kinetic, fluxomics or proteomics time series data.

Systems

PiLSL: pairwise interaction learning-based graph neural network for synthetic lethality prediction in human cancers

Xin Liu (ShanghaiTech University), Jiale Yu (ShanghaiTech University), Siyu Tao (ShanghaiTech University), Beiyuan Yang (ShanghaiTech University), Shike Wang (ShanghaiTech University), Lin Wang (ShanghaiTech University), Fang Bai (ShanghaiTech University) and Jie Zheng (ShanghaiTech University).

Abstract*:*

Synthetic lethality (SL) is a type of genetic interaction in which the simultaneous inactivation of two genes leads to cell death, while the inactivation of a single gene does not affect the cell viability. It can effectively expand the range of anti-cancer therapeutic targets. Existing machine learning methods for SL prediction tend to learn the representations of single genes, but ignore the learning of pairwise interaction between two genes. These models are mostly not interpretable in terms of SL mechanisms. We propose a novel pairwise interaction learning-based graph neural network for SL prediction named PiLSL. It learns the representation of pairwise interaction between genes from enclosing subgraphs of a knowledge graph and multi-omics data. Extensive experimental results demonstrate that PiLSL outperforms all baselines and generalizes well under three realistic scenarios. Besides, PiLSL can reveal SL mechanisms via the weighted paths in the enclosing graphs by attention mechanism.

Systems

Small compound-based direct cell conversion with combinatorial optimization of pathway regulations

Toru Nakamura (Kyushu Institute of Technology), Michio Iwata (Kyushu Institute of Technology), Momoko Hamano (Kyushu Institute of Technology), Ryohei Eguchi (Kyushu Institute of Technology), Jun-Ichi Takeshita (National Institute of Advanced Industrial Science and Technology (AIST)) and Yoshihiro Yamanishi (Kyushu Institute of Technology).

Abstract:

Direct cell conversion, direct reprogramming (DR), is an innovative technology that directly converts source cells to target cells without bypassing induced pluripotent stem cells. The use of small compounds (e.g., drugs) for DR can help avoid carcinogenic risk; however, experimentally identifying small compounds remains challenging. In this paper, we present a new computational method, COMPRENDRE (combinatorial optimization of pathway regulations for direct reprograming), to elucidate the mechanism of small compound-based DR and predict new combinations of small compounds for DR. We developed a variant of a simulated annealing algorithm to identify the best set of compounds that can regulate DR-related pathways. Consequently, the proposed method enabled to predict new DR-inducing candidate combinations with fewer compounds and to successfully reproduce experimentally verified compounds inducing the direct conversion from fibroblasts to neurons or cardiomyocytes. The proposed method is expected to be useful for practical applications in regenerative medicine.

Highlight Papers

Data

Discovering cell types using manifold learning and enhanced visualization of single-cell RNA-Seq data

Akram Vasighizaker (University of Windsor), Saiteja Danda (University of Windsor) and Luis Rueda (University of Windsor).

Abstract*:*

We introduce a method that is used to identify representative clusters of different cell types by combining non-linear dimensionality reduction techniques and clustering algorithms on single-cell RNA-Seq data. Identifying relevant disease modules such as target cell types is a significant step for studying diseases. We assess the impact of different dimensionality reduction techniques combined with the clustering of thirteen publicly available scRNA-seq datasets of different tissues, sizes, and technologies. We further performed gene set enrichment analysis to evaluate the proposed method's performance. As such, our results show that modified locally linear embedding combined with independent component analysis yields overall the best performance relative to the existing unsupervised methods across different datasets.

Highlight Papers

Data

Flimma: Federated and privacy-aware medical differential gene expression analysis

Olga Zolotareva (University of Hamburg), Reza Nasirigerdeh (Technical University of Munich), Julian Matschinske (University of Hamburg), Reihaneh Torkzadehmahani (Technical University of Munich), Mohammad Bakhtiari (Chair of Computational system biology at University of Hamburg), Tobias Frisch (University of Southern Denmark), Julian Alexander Späth (University of Hamburg), David B. Blumenthal (Friedrich-Alexander-Universität), Amir Abbasinejad (Technical University of Munich), Paolo Tier

Abstract:

Aggregating transcriptomics data across hospitals can increase sensitivity and robustness of differential expression analyses, yielding deeper clinical insights. Data exchange is often restricted by privacy legislation and the accuracy of meta-analysis drops if class labels are inhomogeneously distributed among cohorts. To address these challenges, Flimma (https://featurecloud.ai/flimma/) implements the state-of-the-art workflow limma voom in a federated manner, i.e., patient data never leaves their source sites. Flimma's results are identical to those generated by limma voom on aggregated datasets, even in imbalanced scenarios, in contrast to meta-analyses. Flimma is available in FeatureCloud, an app store for privacy-by-design federated data analysis tools, to facilitate international collaborations while fully respecting patients' privacy by utilizing additive secret sharing for secure aggregation. In the talk, we will also introduce popular bioinformatics federated apps created by our group for biomedical scientists: sPLINK for GWAS and federated random forest.

Highlight Papers

Data

Marker-based annotation and integration of large scale single-cell transcriptomics data on a laptop

Sikander Hayat (Institute of Experimental Medicine and Systems Biology, Uniklinik RWTH Aachen, Germany), Yang Xu (UT-ORNL Graduate School of Genome Science and Technology, University of Tennessee, Knoxville, TN 37996, USA), Simon Baumgart (Novo Nordisk, Data Mining and Bioinformatics, Copenhagen, Denmark), Christian Stegmann (Pre-clinical Research, Vifor Pharma Group, Switzerland), Rafael Kramann (Institute of Experimental Medicine and Systems Biology, Uniklinik RWTH Aachen, Germany) and Rachel Patton McCord (Department of Biochemistry and Cellular and Molecular Biology, University of Tennessee, Knoxville, TN 37996, USA).

Abstract*:*

Single-cell transcriptomics data from millions of cells across tissues and conditions are available. Fast and efficient computational tools are needed to standardize and integrate these datasets. We present MACA (Xu et al., Bioinformatics, 2021) and MASI (Xu et al., bioRvix, 2022; under-review: Nature Communications) to automatically annotate and integrate millions of cells without using specialized hardware. As expensive and dedicated hardware isn't readily available to all researchers around the world, we believe cheap tools are essential to democratize single-cell analyses and enable researchers without access to dedicated hardware to work on large datasets. Our tools can annotate and integrate multiple datasets by mapping cell-type labels from reference data. They outperform existing supervised/semi-supervised methods based on speed, integration quality, and cell-type annotation. Using perturbation and developmental lineage datasets, we demonstrate that MASI preserves the underlying biological signal. Finally, we present three large case-studies in Human Kidney, Heart and Covid datasets.

Highlight Papers

Data

Orchestrating and sharing large multimodal data for transparent and reproducible research

Anthony Mammoliti (Princess Margaret Cancer Centre, University Health Network; Department of Medical Biophysics, University of Toronto), Petr Smirnov (Princess Margaret Cancer Centre, University Health Network; Department of Medical Biophysics, University of Toronto), Minoru Nakano (Princess Margaret Cancer Centre, University Health Network), Zhaleh Safikhani (Princess Margaret Cancer Centre, University Health Network; Department of Medical Biophysics, University of Toronto), Christopher Eeles (Princess Margaret Cancer Centre, University Health Network), Heewon Seo (Princess Margaret Cancer Centre, University Health Network; Department of Medical Biophysics, University of Toronto), Sisira Kadambat Nair (Princess Margaret Cancer Centre, University Health Network), Arvind S Mer (Princess Margaret Cancer Centre, University Health Network; Department of Medical Biophysics, University of Toronto), Ian Smith (Princess Margaret Cancer Centre, University Health Network; Department of Medical Biophysics, University of Toronto), Chantal Ho (Princess Margaret Cancer Centre, University Health Network), Gangesh Beri (Princess Margaret Cancer Centre, University Health Network), Rebecca Kusko (Immuneering Corporation, Cambridge, MA), Eva Lin (Department of Discovery Oncology, Genentech Inc, South San Francisco), Yihong Yu (Department of Discovery Oncology, Genentech Inc, South San Francisco), Scott Martin (Department of Discovery Oncology, Genentech Inc, South San Francisco), Marc Hafner (Department of Discovery Oncology, Department of Oncology Bioinformatics, Genentech Inc, South San Francisco) and Benjamin Haibe-Kains (Princess Margaret Cancer Centre, University Health Network; Department of Medical Biophysics, University of Toronto).

Abstract*:*

Reproducibility is essential to open science, as there is limited relevance for findings that can not be reproduced by independent research groups, regardless of its validity.

ORCESTRA is a cloud-based platform which provides a flexible framework for the reproducible processing of multimodal biomedical data. It enables processing of clinical, genomic and perturbation profiles of cancer samples through automated processing pipelines that are user-customizable. ORCESTRA creates integrated and fully documented data objects with persistent identifiers (DOI) and manages multiple dataset versions, which can be shared for future studies. By associating multimodal biomedical data with version-controlled data processing scripts, the platform aims to resolve the ever-increasing challenges of processing and sharing the biomedical data with the community in a FAIR (findable, accessible, interoperable, and reusable) manner. This presentation provides a technical overview of the platform, current limitations and ongoing improvements.

Highlight Papers

Data

PharmacoDB 2.0: improving scalability and transparency of in vitro pharmacogenomics analysis

Nikta Feizi (Princess Margaret Cancer Centre, University Health Network), Sisira Kadambat Nair (Princess Margaret Cancer Centre, University Health Network), Petr Smirnov (University of Toronto), Gangesh Beri (Princess Margaret Cancer Centre, University Health Network), Christopher Eeles (Princess Margaret Cancer Centre, University Health Network), Parinaz Nasr Esfahani (Princess Margaret Cancer Centre, University Health Network), Minoru Nakano (Princess Margaret Cancer Centre, University Health Network), Denis Tkachuk (Princess Margaret Cancer Centre, University Health Network), Anthony Mammoliti (Princess Margaret Cancer Centre, University Health Network), Evgeniya Goroberts (Princess Margaret Cancer Centre, University Health Network), Arvind Singh Mer (Princess Margaret Cancer Centre, University Health Network), Eva Lin (Department of Discovery Oncology, Genentech Inc), Yihong Yu (Department of Discovery Oncology, Genentech Inc), Scott Martin (Department of Discovery Oncology, Genentech Inc), Marc Hafner (Department of Oncology Bioinformatics, Genentech Inc) and Benjamin Haibe-Kains (Princess Margaret Cancer Centre, University Health Network).

Abstract*:*

In vitro pharmacogenomics combines high-throughput molecular and drug sensitivity screening of cancer models to investigate the determinants of therapy response. In PharmacoDB, we create a database integrating the 10 largest studies, encompassing 56K compounds tested across 1,756 cell lines, in over 6.3 Million dose-response experiments. In this presentation, we detail how the online web interface, together with the companion PharmacoGx R package, allows researchers to explore and leverage this resource. We give examples from literature of how the data has been used by others, as well as discuss our own work, using this resource to conduct a statistical meta-analysis to discover robust expression biomarkers for drug response. We present our experience validating findings from these in vitro datasets in retrospective clinical data, including the identification of a single gene marker competitive with complex models in predicting response to neoadjuvant Paclitaxel therapy in breast cancer patients.

Data

Polympact: exploring functional relations among common human genetic variants

Samuel Valentini (University of Trento, CIBIO), Francesco Gandolfi (University of Trento, CIBIO), Mattia Carolo (University of Trento, CIBIO), Davide Dalfovo (University of Trento, CIBIO), Lara Pozza (University of Trento, CIBIO) and Alessandro Romanel (University of Trento, CIBIO).

Abstract*:*

In the last years, Genome-Wide Associations Studies were able to identify associations between genetic variants and complex diseases. However, the mechanistic biological links explaining these associations are still mostly unknown.

This talk will introduce Polympact, a new web-based tool for analyzing the effects of common variants and their putative interactions.

Polympact is a web resource that characterizes over 18 million variants by combining their functional elements landscape, their impact on transcription factor binding motifs and their effects on the transcript levels of protein-coding genes.

Polympact introduces a new framework that combines clustering analysis, a new similarity network model aiming to identify putative variant-variant interactions and a variant-gene network model helpful to analyze in detail relations among variants and genes.

Polympact was successfully used to identify putative interactions among GWAS variants both in Breast Cancer and Alzheimer's disease.

Highlight Papers

Data

Scaling up oligogenic diseases research with OLIDA: the Oligogenic Diseases Database

Barbara Gravel (Université Libre de Bruxelles), Charlotte Nachtegael (Université Libre de Bruxelles), Arnau Dillen (Vrije Universiteit Brussel), Guillaume Smits (Hôpital Erasme), Ann Nowé (Vrije Universiteit Brussel), Sofia Papadimitriou (Université Libre de Bruxelles) and Tom Lenaerts (Université Libre de Bruxelles).

Abstract*:*

Understanding the relationship between genetic variants and disease remains an important challenge in human genetics. Although a large amount of data on single variants causing monogenic diseases has been made available in multiple databases, leading to numerous predictive computational tools, data linking combinations of variants in multiple genes with disease is still limited. We hereby present the Oligogenic Disease Database (OLIDA), the now largest curated repository of literature data on variant combinations underlying oligogenic diseases, which aims to enable novel developments in this area. OLIDA assigns a confidence score to each combination, which is based on standardized criteria assessing the genetic and functional evidence that supports the pathogenic link between a variant combination and a disease. With the creation of such scores, we aim to provide a high-quality benchmark data set that can further aid the research on oligogenic diseases and the development of improved computational tools in the field.

Highlight Papers

Data

The AIMe registry for artificial intelligence in biomedical research

Julian Matschinske (University of Hamburg), Nicolas Alcaraz (University of Copenhagen), Arriel Benis (Holon Institute of Technology), Gerda Cristal Villalba Silva (Federal University of Rio Grande do Sul), Martin Golebiewski (HITS gGmbH), Dominik G. Grimm (Technical University of Munich), Lukas Heumos (Technical University of Munich), Tim Kacprowski (Technical University of Braunschweig), Olga Lazareva (Technical University of Munich), Markus List (Technical University of Munich), Zakaria Louadi (Technical University of Munich), Josch Pauling (Technical University of Munich), Nico Pfeifer (University of Tübingen), Richard Röttger (University of Southern Denmark), Veit Schwämmle (University of Southern Denmark), Kristel Van Steen (University of Liège), Gregor Sturm (Medical University of Innsbruck), Alberto Traverso (Maastricht University), Martiela Vaz de Freitas (Federal University of Rio Grande do Sul), Leonard Wee (Maastricht University), Nina Wenke (University of Hamburg), Massimiliano Zanin (Instituto de Física Interdisciplinar y Sistemas Complejos), Olga Zolotareva (University of Hamburg), Jan Baumbach (University of Hamburg) and David B. Blumenthal (FAU Erlangen-Nuremberg).

Abstract*:*

We present the AIMe registry, a community-driven reporting platform for AI in biomedicine enhancing the accessibility, reproducibility and usability of biomedical AI models. Plenty of examples demonstrate how important thorough and complete information about methods and data are in the context of biomedical research, and how misleading purported findings can be otherwise. An extensive albeit flexible questionnaire (available at https://aime-registry.org) can significantly improve the documentation and reproducibility of research in this area. In the presentation, we show how it has been used so far and briefly introduce the newest AIMe 2022 standard, which will be published by the time of the presentation, and which will include new aspects such as privacy and explainability of AI in biomedicine.

Highlight Papers

Genes

Analysis of eukaryotic lincRNA sequences indicates signatures of hindered translation linked to selection pressure

Anneke Bruemmer (University of Lausanne), Rene Dreos (University of Lausanne), Ana Claudia Marques (University of Lausanne) and Sven Bergmann (University of Lausanne).

Abstract*:*

Although long intergenic noncoding RNAs (lincRNAs) are classified as noncoding, most lincRNAs contain open reading frames (ORFs), and it remains unclear why cytoplasmic lincRNAs are not or very inefficiently translated. Here, we analyzed signatures of hindered translation in lincRNA sequences from five eukaryotes. In species under stronger evolutionary selection, we detected significantly shorter ORFs, a suboptimal sequence context around start codons for translation initiation, and trinucleotides ("codons") corresponding to less abundant tRNAs than for neutrally evolving control sequences, likely impeding translation elongation. For human lincRNAs, we detected signatures for cell-type-specific hindrance of translation. In particular, codons in abundant cytoplasmic lincRNAs corresponded to lower expressed tRNAs than control codons, in three out of five human cell lines, in agreement with reduced ribosome binding to lincRNAs in these cell lines. The identified sequence signatures may improve predicting peptide-coding and genuine noncoding lincRNAs, in a cell-type-specific manner.

Highlight Papers

Genomes

A global metagenomic map of urban microbiomes and antimicrobial resistance

Alina Frolova (The Institute of Molecular Biology and Genetics of NASU, Kyiv, Ukraine), David Danko (Weill Cornell Medicine, New York, NY, USA), Daniela Bezdan (Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany), Christopher E. Mason (Weill Cornell Medicine, New York, NY, USA) and The International Consortium Metasub (The International MetaSUB Consortium).

Abstract*:*

We present a global atlas of 4,728 metagenomic samples from mass-transit systems in 60 cities over 3 years, representing the first systematic, worldwide catalog of the urban microbial ecosystem. This atlas provides an annotated, geospatial profile of microbial strains, functional characteristics, antimicrobial resistance (AMR) markers, and genetic elements, including 10,928 viruses, 1,302 bacteria, 2 archaea, and 838,532 CRISPR arrays not found in reference databases. We identified 4,246 known species of urban microorganisms and a consistent set of 31 species found in 97% of samples that were distinct from human commensal organisms. Profiles of AMR genes varied widely in type and density across cities. Cities showed distinct microbial taxonomic signatures that were driven by climate and geographic differences. These results constitute a high-resolution global metagenomic atlas that enables discovery of organisms and genes, highlights potential public health and forensic applications, and provides a culture-independent view of AMR burden in cities.

Highlight Papers

Genomes

Higher order genetic interactions switch cancer genes from two-hit to one-hit drivers

Solip Park (CNIO), Fran Supek (IRB) and Ben Lehner (CRG).

Abstract*:*

The classic two-hit model posits that both alleles of a tumor suppressor gene (TSG) must be inactivated to cause cancer. In contrast, for some oncogenes and haploinsufficient TSGs, a single genetic alteration can suffice to increase tumor fitness. Here, by quantifying the interactions between mutations and copy number alterations (CNAs) across 10,000 tumors, we show that many cancer genes actually switch between acting as one-hit or two-hit drivers.

Third order genetic interactions identify the causes of some of these switches in dominance and dosage sensitivity as mutations in other genes in the same biological pathway. The correct genetic model for a gene thus depends on the other mutations in a genome, with a second hit in the same gene or an alteration in a different gene in the same pathway sometimes representing alternative evolutionary paths to cancer.

Highlight Papers

Genomes

plotsr: Visualising structural similarities and rearrangements between multiple genomes

Manish Goel (Ludwig-Maximilians-Universität München) and Korbinian Schneeberger (Ludwig-Maximilians-Universität München).

Abstract*:*

Third-generation sequencing has revolutionized genome assembly resulting in a sharp increase in availability of high-quality genomes. This opens new avenues of comparative genome analyses and calls for novel visualization methods. We have developed plotsr, a new tool for generating publication-quality visualizations of the differences between multiple genomes. plotsr can generate visualizations in two modes: 1) stacked mode: for visualizing synteny and structural rearrangements between homologous chromosomes, and 2) itx mode: for visualizing inter-chromosomal rearrangements. Additionally, plotsr uses synteny between homologous chromosomes to allow for zoom-in visualizations without the need for computationally expensive all-versus-all genome alignments. plotsr can augment the visualisation by marking points of interest (e.g. genes or genomic markers) and plotting histogram tracks for the distribution of genomic features (e.g. SNP density, GC content). It can visualize genomic differences identified by different methods. plotsr is computationally efficient and can visualize multiple human genomes in under a minute

Highlight Papers

Genomes

Revisiting genetic artifacts on DNA methylation microarrays exposes novel biological implications

Benjamin Planterose Jiménez (Erasmus MC, University Medical Center Rotterdam), Manfred Kayser (Erasmus MC, University Medical Center Rotterdam) and Athina Vidaki (Erasmus MC, University Medical Center Rotterdam).

Abstract*:*

DNA methylation is the most studied epigenetic biomarker in human health and disease. Its epigenome-wide quantification has been standardized in association studies via microarrays. Nonetheless, current DNA methylation microarrays cannot fully consider the vast human genetic diversity leading to genetic artifacts. Typically, microarray probes predicted to be affected by underlying genetic variants are excluded based on lists compiled in early studies. Such lists, however, ignore the intricacies of the assay, remain empirically unvalidated and thus, suffer from undetermined false positive/negative levels. Towards benchmarking current practice, we perform a thorough characterization of genetic artifacts on DNA methylation microarrays. We unexpectedly uncover unaccounted factors and several unreported interactions between artifacts or with X-inactivation, imprinting and tissue-specific regulation. Additionally, a comethylation-based approach is proposed to distinguish artifacts from genuine epigenetic variation. Overall, we provide a framework for effective genetic artifact management, with implications for research on the genetics of DNA methylation.

Genomes

Towards a metagenomics machine learning interpretable model for understanding the transition from adenoma to colorectal cancer

Carlos Sánchez Casimiro-Soriguer (Clinical Bioinformatic Area, Fundacion Progreso y Salud), Carlos Loucera (Clinical Bioinformatic Area, Fundacion Progreso y Salud), María Peña-chilet (Clinical Bioinformatic Area, Fundacion Progreso y Salud) and Joaquin Dopazo (Clinical Bioinformatic Area, Fundacion Progreso y Salud).

Abstract*:*

Gut microbiome is gaining interest because of its links with several diseases, including colorectal cancer (CRC), as well as the possibility of being used to obtain non-intrusive predictive disease biomarkers. Here we performed a meta-analysis of 1042 fecal metagenomic samples from seven publicly available studies. We used an interpretable machine learning approach based on functional

profiles, instead of the conventional taxonomic profiles, to produce a highly accurate predictor of CRC with better precision than those of previous proposals. Moreover, this approach is also able to discriminate samples with adenoma, which makes this approach very promising for CRC prevention by detecting early stages in which intervention is easier and more effective.

In addition, interpretable machine learning methods allow extracting features relevant for the classification, which reveals basic molecular mechanisms accounting for the changes undergone by the microbiome functional landscape in the transition from healthy gut to adenoma and CRC conditions.

Highlight Papers

Genomes

Unlocking capacities of genomics for the COVID-19 response and future pandemics

Serghei Mangul (University of Southern California), Karishma Chhugani (University of Southern California), Sergey Knyazev (University of California, Los Angeles), Varuni Sarwal (University of California, Los Angeles), Ram Ayyala (University of Southern California), Angela Lu (University of Southern California) and Adam Smith (University of Southern California).

Abstract*:*

During the COVID-19 pandemic, genomics and bioinformatics have emerged as essential public health tools. The genomic data acquired using these methods have supported the global health response, facilitated development of testing methods, and allowed timely tracking of novel SARS-CoV-2 variants. Yet the virtually unlimited potential for rapid generation and analysis of genomic data is also coupled with unique technical, scientific, and organizational challenges. Here, we present the application of genomic and computational methods for the efficient data driven COVID-19 response, advantages of democratization of viral sequencing around the world, and challenges associated with viral genome data collection and processing.

Highlight Papers

Proteins

AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models

Mihaly Varadi (European Bioinformatics Institute), Sameer Velankar (European Bioinformatics Institute), Stephen Anyango (European Bioinformatics Institute), Mandar Deshpande (European Bioinformatics Institute), Sreenath Nair (European Bioinformatics Institute), Cindy Natassia (European Bioinformatics Institute), Galabina Yordanova (European Bioinformatics Institute), David Yuan (European Bioinformatics Institute), Oana Stroe (European Bioinformatics Institute), Gemma Wood (European Bioinformatics Institute), Agata Laydon (DeepMind), Augustin Zidek (DeepMind), Tim Green (DeepMind), Kathryn Tunyasuvunakool (DeepMind), Stig Petersen (DeepMind), John Jumper (DeepMind), Ellen Clancy (DeepMind), Richard Green (DeepMind), Ankur Vora (DeepMind), Mira Lutfi (DeepMind), Michael Figurnov (DeepMind), Andrew Cowie (DeepMind), Nicole Hobbs (DeepMind), Pushmeet Kohli (DeepMind), Gerard Kleywegt (European Bioinformatics Institute), Ewan Birney (European Bioinformatics Institute) and Demis Hassabis (DeepMind).

Abstract*:*

The AlphaFold Protein Structure Database (AlphaFold DB, https://alphafold.ebi.ac.uk) is an openly accessible, extensive database of high-accuracy protein-structure predictions. Powered by AlphaFold v2.0 of DeepMind, it has enabled an unprecedented expansion of the structural coverage of the known protein-sequence space. AlphaFold DB provides programmatic access to and interactive visualization of predicted atomic coordinates, per-residue and pairwise model-confidence estimates and predicted aligned errors. The current version of AlphaFold DB contains over 1 Million predicted structures across 37 model-organism proteomes, which will soon be expanded to cover most of the (over 100 million) representative sequences from the UniRef90 data set. The presentation will cover an overview of AlphaFold, the data types and access supported by AlphaFold DB, and the future perspectives of this data resource.

Highlight Papers

Proteins

Missense variants in human ACE2 strongly affect binding to SARS-CoV-2 Spike providing a mechanism for ACE2 mediated genetic risk in Covid-19: A case study in affinity predictions of interface variants

Stuart A. MacGowan (Division of Computational Biology, School of Life Sciences, University of Dundee), Michael I. Barton (Sir William Dunn School of Pathology, University of Oxford), Mikhail Kutuzov (Sir William Dunn School of Pathology, University of Oxford), Omer Dushek (Sir William Dunn School of Pathology, University of Oxford), P. Anton van der Merwe (Sir William Dunn School of Pathology, University of Oxford) and Geoffrey J. Barton (Division of Computational Biology, School of Life Sciences, University of Dundee).

Abstract*:*

SARS-CoV-2 infection manifests a range of clinical presentations from mild illness to life-threatening disease. As a mediator of viral entry, ACE2 is an a priori candidate genetic risk factor. The affinity of SARS-CoV-2 Spike for ACE2 is a key parameter influencing host-range and tropism and so we determined the affinities of several reported ACE2 population variants experimentally and predicted the effects of many more. We found ACE2 alleles that strongly inhibited binding to Spike and some with moderately increased affinity. Comparison to recent infectivity studies indicates that the affinity ranges of ACE2 variants can protect cells from infection and so some almost certainly confer resistance to carriers; this is now being tested with clinical data. We will also highlight the strengths and weaknesses of current generation predictors, and present new results on the interplay between ACE2 variants and different SARS-CoV-2 strains.

Proteins

Online predictions for protein biophysical features and their conservation

Wim Vranken (Interuniversity Institute of Bioinformatics in Brussels, Vrije Universiteit Brussel), Luciano Kagami (Interuniversity Institute of Bioinformatics in Brussels) and Adrian Diaz (Interuniversity Institute of Bioinformatics in Brussels, Vrije Universiteit Brussel).

Abstract*:*

Our understanding of how proteins operate and how evolution shapes them is mainly based on their overall fold and their amino acid sequence. With the direct relation between these now largely solved by AlphaFold2 and RosettaFold, the challenge of how to define highly dynamic and/or structurally ambiguous behavior in proteins remains. We provide integrated online protein sequence-based predictions to identify biophysical features of proteins that are not readily captured by protein fold or amino acid sequence, such as backbone dynamics or early folding regions. These predictions capture 'emergent' properties of proteins, i.e. the inherent biophysical propensities encoded in their sequence, rather than context-dependent behaviour (e.g. final folded state). This concept is extended to include multiple sequence alignments, so enabling exploration of the 'biophysical variation' in homologous proteins. In this way, biophysical limits can be defined for functionally relevant protein behaviour, with unusual residues flagged by a Gaussian mixture model analysis.

Highlight Papers

Proteins

PDBe-KB: collaboratively defining the biological context of structural data

Preeti Choudhary (EMBL-EBI).

Abstract*:*

The Protein Data Bank in Europe – Knowledge Base (PDBe-KB, https://pdbe-kb.org) is a collaborative resource between world-leading specialist data resources striving towards a two-fold goal: (i) to increase the visibility and reduce the fragmentation of annotations contributed by specialist data resources, and to make these data more findable, accessible, interoperable and reusable (FAIR) and (ii) to place macromolecular structure data in their biological context, thus facilitating their use by the broader scientific community in fundamental and applied research. Since we described PDBe-KB in 2019, there have been significant improvements in the variety of available annotation data sets and user functionality. Here we highlight all these additional annotations and new features such as a bulk download data service and a novel superposition service that generates clusters of superposed protein chains weekly for the whole PDB archive.

Proteins

The clinical importance of tandem exon duplication-derived substitutions

Laura Martinez Gomez (Centro Nacional de Investigaciones Oncológicas), Fernando Pozo Ocampo (Spanish National Cancer Research Centre (CNIO)), Thomas Walsh (Spanish National Cancer Research Centre (CNIO)), Federico Abascal (Wellcome Trust Sanger Institute) and Michael Tress (Spanish National Cancer Research Centre (CNIO)).

Abstract*:*

For this paper we manually curated a set of 236 tandem exon duplication-derived substitution events. These alternative splice events involve the use of one of two duplicated adjacent exons, generating two highly similar proteins.

We find that more than 90% of these tandem exon duplication-derived substitution events are conserved in at least one fish species, they are detected in proteomics experiments in much higher numbers, and they have proportionally 27 times more clinically important mutations than any other type of splice event.

Tandem exon duplication-derived substitutions clearly have important functional roles in the cell and the cross-species conservation suggests that they been a relevant factor in metazoan evolution. Curiously, despite their obvious importance, homologous exons are often left out of alternative splicing analyses. We hope that the annotation of a complete set of homologous substitutions for the human genome will inspire research into these important, highly conserved splice events.

Highlight Papers

Proteins

TITAN: T-cell receptor specificity prediction with bimodal attention networks

Anna Weber (IBM Research Zurich, ETH Zurich), Jannis Born (IBM Research Zurich, ETH Zurich) and Maria Rodriguez Martinez (IBM Research Zurich).

Abstract*:*

Reliable prediction of T cell receptor (TCR) specificity and understanding the mechanisms underlying the TCR-pMHC interaction is both a daunting and highly relevant challenge. To train our model TITAN (Tcr epITope bimodal Attention Networks) on this task, we leveraged machine learning techniques from transfer learning to interpretability to achieve a state-of-the-art prediction performance while also giving insights into the decision process of the model. In follow-up work we dove even deeper into explaining TCR specificity prediction models with a pipeline designed to extract binding and non-binding rules from any sequence-based model.

Systems

Comprehensive prediction of robust synthetic lethality between paralog pairs in cancer cell lines

Barbara De Kegel (University College Dublin), Niall Quinn (University College Dublin), Nicola Thompson (Wellcome Sanger Institute), David Adams (Wellcome Sanger Institute) and Colm Ryan (University College Dublin, Ireland).

Abstract*:*

Pairs of paralogs may share common functionality and, hence, display synthetic lethal interactions. As the majority of human genes have an identifiable paralog, exploiting synthetic lethality between paralogs may be a broadly applicable approach for targeting gene loss in cancer. However, only a biased subset of human paralog pairs has been tested for synthetic lethality to date. Here, by analyzing genome-wide CRISPR screens and molecular profiles of over 700 cancer cell lines, we identify features predictive of synthetic lethality between paralogs, including shared protein-protein interactions and evolutionary conservation. We develop a machine-learning classifier based on these features to predict which paralog pairs are most likely to be synthetic lethal and to explain why. We show that our classifier accurately predicts the results of combinatorial CRISPR screens in cancer cell lines and furthermore can distinguish pairs that are synthetic lethal in multiple cell lines from those that are cell-line specific.

Highlight Papers

Systems

Interpretable systems biomarkers predict response to immune-checkpoint inhibitors

Óscar Lapuente-Santana (Eindhoven University of Technology), Maisa van Genderen (Eindhoven University of Technology), Peter Hilbers (Eindhoven University of Technology), Francesca Finotello (University of Innsbruck) and Federica Eduati (Eindhoven University of Technology).

Abstract:

The response of patients to immunotherapy with immune checkpoint blockers is still poorly understood. Tumours are multicellular systems with several intra- and inter-cellular regulatory interactions, calling for new holistic approaches to quantitatively characterise the tumour microenvironment (TME) for stratification of patients for immunotherapy.

Here, we show how bulk RNA-seq data complemented by prior knowledge provides a high-level mechanistic representation of the multifaceted TME: cell type abundances, pathways, transcription factors and cytokines activity scores, quantification of ligand-receptor and cell-cell interactions. Using multi-task machine learning to learn associations between these features and different hallmarks of immune response, we identified biomarkers that are shown to be predictive of immunotherapy response in different cancer types. Additionally, we show preliminary results on how spatial information from tissue-slide imaging can improve therapy response prediction.

Data

A cloud based resource to manage, access and explore omics datasets in multiuser environments

Mario Looso (Max Planck Institute for Heart and Lung Research, Bioinformatics Core), Philipp Goymann (Max Planck Institute for Heart and Lung Research, Bioinformatics Core), Franz Ziegengeist (Max Planck Institute for Heart and Lung Research, Bioinformatics Core), Carsten Kuenne (Max Planck Institute for Heart and Lung Research, Bioinformatics Core), Daniel Spothelfer (Max Planck Institute for Heart and Lung Research, Bioinformatics Core), Noah Knoppik (Max Planck Institute for Heart and Lung Research, Bioinformatics Core) and Jasmin Walter (Max Planck Institute for Heart and Lung Research, Bioinformatics Core).

Abstract*:*

Omics-based screens supported by bioinformatics evolved to a de facto standard in many fields of research, diagnostics and industry. Central requirements defined by omics facilities relate to i) data storage, ii) standardized data analysis and data formats, iii) webapps for data exploration, and iv) on-demand scaling of computational resources. We found a combination of i) cloud computing, ii) application virtualization, iii) file based structures, iv) git based maintenance, and v) a custom made operator software optimal to fulfill these needs. Here we introduce the BCU repository, a flexible and easy to migrate data warehouse. It provides a web based front-end, and functionality to generate and search omics related metadata. Further it links omics data with explorative applications on demand via an inhouse implemented cloud operator. User attributes such as group assignments, integrated from remote user management tools, allow for sharing and collaborative work.

Applications

Data

Global biodata resources: challenges to long-term sustainability of a crucial data infrastructure

Guy Cochrane (Global Biodata Coalition) and Chuck Cook (Global Biodata Coalition).

Abstract*:*

Research in the life sciences is data-driven and critically dependent on data integration and analysis enabled by open-access biodata resources that collectively form the single most critical infrastructure for life sciences research. These resources are used globally, but funded by national and charitable funding bodies, typically in short-term funding cycles that do not provide the sustained long-term support necessary to ensure a healthy infrastructure. The Global Biodata Coalition (GBC) was formed and is supported by research funders to aid funders in collaborating to find more efficient and long-term solutions to support this infrastructure. Current GBC activities include undertaking an inventory of biodata resources around the globe that will provide the first ever overview of these resources as an infrastructure, and identification of the first set of Global Core Biodata Resources, which are those resources that provide the global foundations for data sharing and access across the life sciences.

Applications

Data

Introducing X-Omics, the central multi-omics data integration and AI modeling platform for biomarker data at Merck

Sven-Eric Schelhorn (Merck Healthcare KGaA).

Abstract*:*

Precision medicine at pharmaceutical organizations is driven by large interoperable datasets, high-performance computing, innovative statistical modeling, and cross-functional collaboration including also academic partners. Over the past few decades, the complexity, diversity, and amount of data produced by the biopharmaceutical industry have dramatically increased, making challenges of data management and analysis core topics in R&D. Scalable and FAIR Systems capable of overcoming these data-related and organizational complexities are critical for advancing the research and development of novel personalized therapeutics.

We introduce the X-Omics platform, a cloud-based, validated bioinformatics system that is the central locus for integration, multi-omics processing, and AI analysis of pre-clinical and clinical biomarker data at Merck Healthcare KGaA, Germany. The system centralizes access to clinical, genomics, histopathology, and real-world evidence data, enabling FAIRification and facilitating collaboration by a wide range of geographically dispersed users, including academic collaborators who can now work side-by-side on patient data with internal analysts.

Applications

Data

Open Targets: A Platform for Biological Data Integration

Irene Lopez Santiago (Open Targets).

Abstract*:*

The Open Targets Platform integrates 22 publicly available datasets as well as data generated within the Open Targets consortium to facilitate the identification and prioritisation of potential therapeutic targets. The Platform is regularly updated with new datasets and data types to establish target-disease associations and provide additional context to build therapeutic hypotheses, highlighting the fundamental challenge of designing optimised, well-integrated pipelines to present the data in a unified and intuitive way.

This talk will present Open Targets' approach to integrating and standardising diverse biological datasets. This approach is exemplified by the inclusion of a recently developed Natural Language Processing method developed within the Open Targets partnership to classify the reasons for which clinical trials have stopped.

Applications

Data

Scalable In-memory paradigm for genomics data processing

Ritesh Krishna (IBM Research), Vadim Elisseev (IBM Research) and Laura-Jayne Gardiner (IBM Research).

Abstract*:*

Accelerated data processing for genomics is a key capability for a wide variety of industrially relevant problems. Traditional bioinformatics workflows, whether running on in-house HPC clusters or cloud, are I/O intensive and introduce significant delays in overall processing of datasets. Here we present a proof of concept in-memory paradigm for developing and orchestrating bioinformatics workflows. We propose an architecture comprising of generally available industry-standard components like MPI, distributed in-memory key-value storage etc. that can minimize the I/O delays associated with traditional workflow designs and contribute towards improved and flexible handling of omics datasets. We show how the in-memory paradigm can inspire alternate algorithm designs offering significant speedups and possibilities to introduce event-driven behaviour in bioinformatics workflows. We demonstrate the proof of concept through an end-to-end bioinformatics pipeline to understand antimicrobial resistance profile of large microbiome sample.

Applications

Genes

Bioinformatics methods for the analysis of rare-disease patient data – applications for target discovery and obtaining phenotype associations

Elena Rojano (IBIMA; Universidad de Málaga), James Perkins (Universidad de Málaga; CIBERER; IBIMA), Fernando Moreno Jabato (Universidad de Málaga), José Córdoba-Caballero (Universidad de Málaga), Elena Diaz-Santiago (Universidad de Málaga), Federico García-Criado (Universidad de Málaga), Florencio Pazos (Centro Nacional de Biotecnología (CNB-CSIC)), Pedro Seoane (Universidad de Málaga - CIBERER) and Juan Ranea (Universidad de Málaga; CIBERER; INB ELIXIR; IBIMA).

Abstract*:*

We present methodologies aimed at the analysis of high-throughput disease-related data. They have been developed in collaboration with both basic and clinical groups within the Spanish biomedical research network for rare-diseases (CIBERER), and include methods for the analysis of gene-expression that have potential for target identification.

We also include methods to associate pathological phenotypes with genes and functional systems and to identify modules of phenotypes based on their tendency to co-occur across multiple patients, or to co-occur within diseases. We have applied them to large cohorts of rare and undiagnosed disease patients and to disease databases to analyse phenotypes related to neuromuscular diseases. These methods have clear applications outside of academia for improving diagnosis, and can be applied to the large scale patient cohort data being generated at the population level.

Applications

Genes

Power analysis of cell-type deconvolution across human tissues

Anna Vathrakokoili Pournara (EMBL-EBI), Zhichao Miao (EMBL-EBI) and Irene Papatheodorou (EMBL-EBI).

Abstract*:*

Cell-type deconvolution methods aim to infer cell-type heterogeneity and the cell abundances from bulk transcriptomic data. Since a plethora of methods have been developed, there is an urgent need for guidance on method selection. At the same time there is pressing interest to achieve decomposition of database-level transcriptomic data of different tissues, conditions and species. Here, we propose a multi-level assessment of 28 available deconvolution methods, leveraging 44 single-cell RNA-sequencing, from 7 tissues. We suggest a comprehensive simulation framework to evaluate deconvolution across a wide range of scenarios and we provide useful guidelines for method selection. We show that regression-based deconvolution methods are performing well but their performance is highly dependent on the reference selection and the tissue type. Lastly, we provide a modularised benchmarking pipeline that will speed up the evaluation of newly published methods and we showcase its applicability in the large-scale decomposition of healthy and cancer data.

Applications

Genomes

ChiTaH: a fast and accurate tool for identifying known human chimeric sequences from high-throughput sequencing data

Milana Frenkel-Morgenstern (Bar Ilan University), Rajesh Detroja (Bar Ilan University) and Sumit Mukherjee (Bar-Ilan University).

Abstract*:*

Fusion genes or chimeras typically comprise sequences from two different genes. The chimeric RNAs of such joined sequences often serve as cancer drivers. Identifying such driver fusions in a given cancer or complex disease is important for diagnosis and treatment. The advent of next-generation sequencing technologies, such as DNA-Seq or RNA-Seq, together with the development of suitable computational tools, has made the global identification of chimeras in tumors possible. However, the testing of over 20 computational methods showed these to be limited in terms of chimera prediction sensitivity, specificity, and accurate quantification of junction reads. These shortcomings motivated us to develop the first 'reference-based' approach termed ChiTaH (Chimeric Transcripts from High-throughput sequencing data). ChiTaH uses 43,466 non-redundant known human chimeras as a reference database to map sequencing reads and to accurately identify chimeric reads. We benchmarked ChiTaH and four other methods to identify human chimeras, leveraging both simulated and real sequencing datasets. ChiTaH was found to be the most accurate and fastest method for identifying known human chimeras from simulated and sequencing datasets. Moreover, especially ChiTaH uncovered heterogeneity of the BCR-ABL1 chimera in both bulk and single-cells of the K-562 cell line, which was confirmed experimentally.

Applications

Genomes

METALoci, identification of spatial enhancer hubs.

Marc A. Marti-Renom (CNAG-CRG), Irene Mota Gomez-Argente (MDC), Juan Antonio Rodriguez Perez (CNAG-CRG) and Dario Lupiañez (MDC).

Abstract*:*

We will introduce METALoci, a new computational framework to identify three-dimensional (3D) hubs of genomic signal. METALoci, which uses as input Hi-C interaction matrices and ChIP-seq profiles, makes use of autocorrelation analysis to identify whether the signal of interest is enriched in spatial compartments on the genome. We have applied METALoci to the discovery of genes associated with sex determination in mouse.

Applications

Genomes

Whole-genome sequencing analysis of food enzyme products reveals contaminations with genetically modified microorganism of related origin

Jolien D'Aes (Sciensano), Marie-Alice Fraiture (Sciensano), Bert Bogaerts (Sciensano), Sigrid C.J. De Keersmaecker (Sciensano), Nancy H.C. Roosens (Sciensano) and Kevin Vanneste (Sciensano).

Abstract*:*

Despite their presence being unauthorized on the European market, contaminations with genetically modified microorganisms (GMM) have repeatedly been reported in diverse commercial microbial fermentation produce types. Several of these contaminations are related to a GMM Bacillus velezensis used to synthesize a food enzyme protease, for which genomic characterization remains currently incomplete, and it is unknown whether these contaminations have a common origin.

In this study, GMM B. velezensis isolates from multiple food enzyme products were characterized by short- and long-read whole-genome sequencing (WGS), demonstrating that they harbor a free recombinant pUB110-derived plasmid carrying antimicrobial resistance genes. Additionally, single-nucleotide polymorphism (SNP) and whole-genome based comparative analyses showed that the isolates likely originate from the same parental GM strain.

This study highlights the added value of a hybrid WGS approach for accurate genomic characterization of GMM (e.g. genomic location of the transgenic construct), and of SNP-based phylogenomic analysis for source-tracking of GMM.

Applications

Proteins

A near-full compression of SARS-CoV-2 peptidome using UNIQmin

Li Chuin Chong (Bezmialem Vakıf Universty/TWINCORE GmbH) and Asif M. Khan (Perdana University/Bezmialem Vakif University).

Abstract*:*

The unprecedented increase in SARS-CoV-2 sequence data limits the application of alignment-dependent approaches to study viral diversity. Herein, we applied our recently published UNIQmin, an alignment-free tool to study the protein sequence diversity of SARS-CoV-2. Only less than 0.5% of the reported SARS-CoV-2 protein sequences were required to represent the inherent viral peptidome diversity, which only increased to a mere ~2% at the family rank. This is expected to remain relatively the same even with further increases in the sequence data. The findings have important implications in the design of vaccines, drugs, and diagnostics, whereby the number of sequences required for consideration of such studies is drastically reduced, short-circuiting the discovery process, while still providing for a systematics evaluation and coverage of the pathogen diversity.

Applications

Systems

Cancer patient stratification and molecular mechanism identification using patient clinotypes and transcriptomics embeddings

Zongliang Yue (University of Alabama at Birmingham), Samuel Bharti (University of Alabama at Birmingham), Eric Gong (University of Alabama at Birmingham), Radomir Slominski (University of Alabama at Birmingham), Thanh Nguyen (University of Alabama at Birmingham), Lara Lanov (University of Alabama at Birmingham), Christopher Willey (University of Alabama at Birmingham) and Jake Chen (University of Alabama at Birmingham).

Abstract*:*

Cancer patient stratification and molecular mechanism identification are essential in risk prediction and personalized therapy. We introduce an application integrated with our previous published tools, the metadata annotation tool called "Statistical Enrichment Analysis of Samples (SEAS)" and the functional genomics downstream analysis tool called "PAGER Web APP". The application enables the stratification of the cancer patients associated with molecular subtypes joined with clinotypes using the clinical feature weighted functional genomics embeddings and allows users to identify sub-cohorts in densMAP. In the functional genomics downstream analysis, the application identifies molecular mechanisms driving the clinical feature and systematically reviews the critical insights of the pathway crosstalk and gene mechanisms among those molecular subtypes. The application provides a visual exploration of the sub-cohort's gene panels in each enriched pathway coupling with gene networks, and a comprehensive review of genes by topology-based and transcriptomics-based prioritization.

Applications

Systems

Evaluation of Machine Learning Strategies for Imaging Confirmed Prostate Cancer Recurrence Prediction on Electronic Health Records

Jacqueline Beinecke (Institute for Medical Informatics at the University Medical Center Goettingen), Patrick Anders (Department of Nuclear Medicine, University Hospital Marburg), Tino Schurrat (Department of Nuclear Medicine, University Hospital Marburg), Dominik Heider (Department of mathematics and computer science at the Philipps University Marburg), Markus Luster (Department of Nuclear Medicine, University Hospital Marburg), Damiano Librizzi (Department of Nuclear Medicine, University Hospital Marburg) and Anne-Christin Hauschild (Institute for Medical Informatics at the University Medical Center Goettingen).

Abstract*:*

Prostate cancer is the second most commonly diagnosed cancer worldwide and one of the most leading causes of death in Western countries. After definitive primary treatment in up to 30% of all patients the cancer recurs. In the last years, Ga-68-PSMA PET/CT has become the primary method for additional diagnostics in recurring patients. While it performs great, it is an expensive and invasive examination. Therefore, we employed modern state-of-the-art multivariate machine learning (ML) methods on electronic health records of prostate cancer patients to improve the prediction of imaging confirmed prostate cancer recurrence (IPCR). This study demonstrates the potential of combining multitudes of parameters with multivariate ML models to improve identification of recurring patients compared to the current focus on the main screening parameter (PSA). Ultimately, the developed prediction tool can indicate patients with early-stage recurrence, detectable by Ga-68-PSMA PET/CT, and thereby pave the way for optimized early imaging and treatment.

Applications

Climate Crisis and Health

AI microbiome-based recommendation system for improving soil health with bio-stimulants

Beatriz García-Jiménez (Biome Makers Inc.), Sam Röttjers (Biome Makers Inc.), Diego Rodríguez-de-Prado (Biome Makers Inc.) and Alberto Acedo (Biome Makers Inc.).

Abstract*:*

Due to the impacts of climate change, we need to triple food reserves to prevent food insecurity, under limitations in highly-contaminant fertilizers.

We present a promising and actionable computational application to take care of the soil health: a microbiome-based system that recommends bio-stimulants, for improving the desired or identified deficiency. Similarly to how Netflix or Spotify recommends movies or music, our Artificial Intelligence system recommends biological products.

Preliminarily, our recommendation system reached an accuracy of 0.65 on average, from 0.81 to 0.52 in different indexes of interest in agriculture. The bio-stimulant would be recommended with high confidence to improve hormone properties or potassium mobilization in greater than 40% or 50%, respectively, of the tentative sampled locations of interest.

We will extend the system to recommend the best sustainable intervention to improve soil health, powered by the biological knowledge behind the Artificial Intelligence analysis of the microbiome.

Climate-sensitive disease outbreaks in the aftermath of extreme climatic events

Tilly Alcayna (London School of Hygiene and Tropical Medicine)

Abstract*:*

Outbreaks of climate-sensitive infectious diseases (CSID) in the aftermath of extreme climatic events, such as floods, droughts, tropical cyclones, and heatwaves, are of high public health concern. Recent advances in forecasting of extreme climatic events have prompted a growing interest in the development of prediction models to anticipate CSID risk, yet the evidence base linking extreme climate events to CSID outbreaks to date has not been collated and synthesized. This talk will draw on a recent scoping review which attempted to identify potential hydrometeorological drivers of outbreaks that could be used to inform trigger design for CSID prediction models for anticipatory public health action. We found higher evidence and higher agreement on the links between extreme climatic events and water-borne diseases than for vector-borne diseases. In addition, we found a substantial lack of evidence on the links between extreme climatic events and underlying vulnerability and exposure factors.

Infectious disease decision-support tools to enhance resilience in climate change hotspots

Rachel Lowe (Barcelona Supercomputing Center)

Abstract*:*

Extreme climatic events, environmental degradation, unplanned urbanization, and socio-economic inequalities exacerbate the risk of infectious disease emergence, spread and transmission. For example, mosquito-borne diseases, such as dengue and malaria, are highly sensitive to climate variability and climate change. A warming climate can lengthen the transmission season and alter the geographical range, potentially bringing diseases to regions which lack either population immunity or strong public health infrastructure. More frequent extreme weather events, such as storms, floods, and droughts, also affect the timing and intensity of vector and water-borne disease outbreaks. Despite the health threats of rapid environmental change, we lack the evidence-base to understand and predict the impacts of extreme events and landscape changes on disease risk, leaving communities vulnerable to increasing health threats. This talk will focus on the present and future risks of emerging infectious diseases and describe the partnerships and tools required to build climate resilience in climate change hotspots, including cities, the rainforest, highland areas, and small islands, to improve preparedness and response to emerging infectious disease threats and assist public health services adapt to climate change.

Climate crisis & Health

Real-time Genomics for One Health

Lara Urban (Helmholtz Munich).

Abstract*:*

While recent heatwaves have firmly established the link between climate change and human health, the impact of climate change on health is even more severe from a One Health perspective, which affirms that the health of humans, animals, and the environment are inextricably linked. In order to understand One Health in a global setting, we require fast and cost-efficient technologies that allow for point-of-care assessments of the interdependency between human and environmental health. In this talk, Lara will showcase three examples where she and her team have leveraged real-time genomics through portable nanopore sequencing in the realms of biodiversity, infectious disease, and air quality research: They used real-time genomics to rapidly monitor a recurring fungal infection in the critically endangered kākāpō parrot on a remote island in New Zealand. Within the PuntSeq initiative, they established real-time genomics for freshwater monitoring, with a focus on zoonotic disease. Finally, Lara's team is currently exploring nanopore sequencing for describing bioaerosols and airborne pathogens.

The Catalan Initiative for the Earth Biogenome Project

Montserrat Corominas (Universitat de Barcelona)

Abstract*:*

One of the most important scientific and social challenges is increasing our understanding of the Earth's biodiversity for managing resources responsibly. Global biodiversity is not fully characterized and it is increasingly under threat from climate change, habitat destruction, exploitation and human related activities. Genomics can contribute to characterize known species, discovering new ones and understanding evolution, speciation or the fate of endangered species. The Earth Biogenome Project (EBP) aims to sequence the genomes of all eukaryotes. The Catalan Initiative for the Earth Biogenome Project (CBP) is an EBP affiliated project that aims to sequence the genome of species that live in the Catalan territories, which are part of the Mediterranean basin. The CBP is a networked organization under the umbrella of the Institute for Catalan Studies (IEC), the academy tasked with the promotion of science and culture in these territories, which has provided seed funding for the project.

ELIXIR talks

AHoJ: rapid, tailored search and retrieval of apo and holo protein structures

Christos P. Feidakis (Department of Cell Biology, Faculty of Science, Charles University), Radoslav Krivak (Department of Software Engineering, Faculty of Mathematics and Physics, Charles University), David Hoksza (Department of Software Engineering, Faculty of Mathematics and Physics, Charles University) and Marian Novotný (Department of Cell Biology, Faculty of Science, Charles University).

Abstract*:*

Understanding the mechanism of action of a protein or designing better ligands for it, often requires access to a bound (holo) and an unbound (apo) state of the protein. Resources for the quick and easy retrieval of such conformations are severely limited. Apo-Holo Juxtaposition (AHoJ), is a web application for retrieving apo-holo structure pairs for user-defined ligands. Given a query structure and one or more user-specified ligands, it retrieves all other structures of the same protein that feature the same binding sites(s), aligns them, and examines the superimposed binding sites to determine whether each structure is apo or holo, in reference to the query. The resulting superimposed datasets of apo-holo pairs can be visualized and downloaded for further analysis. AHoJ accepts multiple input queries, allowing the creation of customized apo-holo datasets.

ELIXIR talks

Current activities of the ELIXIR Machine Learning Focus Group

Fotis Psomopoulos (CERTH), Emidio Capriotti (University of Bologna), Núria Queralt Rosinach (Leiden University), Mlfg Tasks Participants (ELIXIR), Leyla Jael Castro (ZB MED – Information Centre for Life Sciences) and Silvio Tosatto (University of Padova).

Abstract*:*

Machine Learning (ML) has emerged as a discipline that enables computers to assist humans in making sense of large and complex data sets. With the drop in the cost of high-throughput technologies, large amounts of omics data are being generated and made accessible to researchers. Analyzing these complex high-volume data is not trivial, and the use of classical statistics cannot explore their full potential. Machine Learning can thus be very useful in mining large omics datasets to uncover new insights that can consequently lead to the advancement of Life Sciences.

The ELIXIR Machine Learning Focus Group was initiated in October 2019, in order to capture the emerging need in Machine Learning expertise across the network. In addition to producing the DOME Recommendations, a set of community-wide recommendations for reporting supervised machine learning–based analyses applied to biological studies, the Focus Group is currently working on three main activities; (1) using the DOME recommendations to annotate relevant literature in order to gain insights into the level of adherence to DOME, (2) evaluating the gold standard datasets widely used in ML process in order to define and describe the aspects of a gold standard with particular focus on human data, and (3) reviewing the efforts around synthetic data in order to establish a set of best practices for their use and application in ML.

ELIXIR talks

Data security entry considerations for post covid data.

Dr Shane Lawrence (University of Cambridge Senstechse).

Abstract*:*

In the post covid environment there has been much welcome cooperation on the part of patients who have cooperated with investigative studies.

The main purpose on such studies has been to provide an insight into such specifics as the viral surface proteins,both the 'spike proteins' and others.

While such insights are important and should be widely accessed there should still be caution as to the total accessibility of such data.

ELIXIR talks

Genome-wide metabolic annotation for Methanocaldococcus (Methanococcus) jannaschii, the first member of the Archaea to be sequenced a quarter of a century ago

Ismini Baltsavia (UoC & CERTH), George Stamoulos (CSD-AUTH), Konstantinos Tziavaras (CPERI-CERTH), Alexandros Dermaris (CSD-AUTH), Ioannis Iliopoulos (UoC), Ron Caspi (SRI), Peter D. Karp (SRI), Nikos C. Kyrpides (JGI) and Christos A. Ouzounis (CSD-AUTH).

Abstract*:*
Methanocaldococcus jannaschii, the first archaeon for which the whole genome was sequenced (Bult et al. 1996), is an autotrophic hyperthermophile obligate anaerobic methanogen (Jeanthon et al. 1998). Physiologically, M. jannaschii has the ability to grow at extreme pressure and temperature conditions (Jones et al. 1983). Metabolic reconstruction for this organism is significant because it can help us understand the properties of methanogens as well as monitor re-annotation efforts (Ouzounis & Karp 2002). A metabolic reconstruction presented previously assigned enzymatic activity to 436 out of 1792 gene products (Tsoka et al. 2004). While some of these annotations have been incorporated into the public databases, currently a reconstruction is available only as a Tier 3 Pathway Genome Database (PGDB) at BioCyc.org (Karp et al. 2021). Here, we present the outcome of a massive, year-long, multi-level, collective and genome-wide curation for M. jannaschii (accession number NC_000909.1), paving the way towards a publicly available Tier 2 PGDB. Using methods reported elsewhere and datasets MjCyc-2005 (Tsoka et al. 2004) and PGDB-2020 (Karp et al. 2021), we examined each protein-coding gene with metabolic capacity potential. Of the 1851 genes in M. jannaschii, there are 1249 entries for mRNA genes with functional descriptions. For those, we have managed to assign 600 enzymatic functions, providing a comprehensive and up-to-date metabolic view, aiming at precision rather than coverage. We hope that the updated reconstruction will be useful to the wider community for research and development efforts in methanogen biology.

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NS, Venter JC (1996) Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. Science 273:1058-73. doi: 10.1126/science.273.5278.1058, PMID: 8688087.

Jeanthon C, L'Haridon S, Reysenbach AL, Vernet M, Messner P, Sleytr UB, Prieur D (1998) Methanococcus infernus sp. nov., a novel hyperthermophilic lithotrophic methanogen isolated from a deep-sea hydrothermal vent. Int J Syst Bacteriol. 48:913-9. doi: 10.1099/00207713-48-3-913, PMID: 9734046.

Jones WJ, Leigh JA, Mayer F, Woese CR, Wolfe RS (1983) Methanococcus jannaschii sp. nov., an extremely thermophilic methanogen from a submarine hydrothermal vent. Arch Microbiol. 136:254-261.

Karp PD, Midford PE, Billington R, Kothari A, Krummenacker M, Latendresse M, Ong WK, Subhraveti P, Caspi R, Fulcher C, Keseler IM, Paley SM (2021) Pathway Tools version 23.0 update: software for pathway/genome informatics and systems biology. Brief Bioinform. 22(1):109-126. doi: 10.1093/bib/bbz104, PMID: 31813964.

Ouzounis CA, Karp PD (2002) The past, present and future of genome-wide re-annotation. Genome Biol. 3(2):COMMENT2001. doi: 10.1186/gb-2002-3-2-comment2001, PMID: 11864365.

Tsoka S, Simon D, Ouzounis CA (2004) Automated metabolic reconstruction for Methanococcus jannaschii. Archaea 1(4):223-9. doi: 10.1155/2004/324925, PMID: 15810431.

ELIXIR talks

Open-source genome-scale metabolic models: why and how

Mihail Anton (ELIXIR Systems Biology Community, National Bioinformatics Infrastructure Sweden, Chalmers University of Technology), Barbara Szomolay (ELIXIR Systems Biology and Single Cell Communities, School of Medicine, Cardiff University) and Vitor A P Martins dos Santos (ELIXIR Systems Biology and Microbial Biotech. Communities, Laboratory of Bioprocess Engineering, Wageningen University).

Abstract*:*

Models in systems biology are being used as tools to understand biological processes by facilitating data interpretation, analysis and prediction. Iterative curation is paramount to render such models useful, in particular to improve model-based predictions and generate actionable knowledge.
Genome-scale, constraint-based metabolic models (GEMs) often require expert curations that involve scientific programming and reviews. Additionally, as GEMs are usually composed of thousands of variables, they also benefit from routinely being verified by validation tools such as MEMOTE (doi:10.1038/s41587-020-0446-y) and FROG (www.ebi.ac.uk/biomodels/curation/fbc).
In the recently formed ELIXIR Systems Biology Community (elixir-europe.org/focus-groups/systems-biology), we aim to align the already well-developed community standards to increase the FAIRness of the models. Moreover, we aim to facilitate the interlinking of models, workflows and data to lower the barrier to wider uptake of systems biology approaches and methods. Thus, the Community is designing a process to continuously develop models according to principles common in software engineering to make models easy to use in notebook-driven workflows.
Here, we will introduce techniques that can readily be applied when working on GEMs, and which are highly relevant for community-driven curations. These techniques include distributed version control, public issues and reviews, and automated workflows, such as sanity checks and model validation. Moreover, we will demonstrate how such techniques are the expression of open science, and how they go hand in hand with common toolboxes and workflows.

ELIXIR talks

Rare disease specific FAIR Maturity Indicators

Núria Queralt Rosinach (Leiden University Medical Center), Rajaram Kaliyaperumal (Leiden University Medical Center), Annika Jacobsen (Leiden University Medical Center), Mark Wilkinson (Centro de Biotecnología y Genómica de Plantas (CBGP, UPM-INIA) Universidad Politécnica de Madrid), Marc Hanauer (Inserm) and Marco Roos (Leiden University Medical Centre).

Abstract*:*

The Rare Disease (RD) community shows a strong interest in making their health data resources FAIR for accelerating research. These resources have great value to advancing diagnosis, prognosis to new therapeutic avenues, and understanding of the mechanisms underpinning rare disease conditions. Increasing the ability to efficiently and computationally find, access, interoperate and reuse data relevant for rare diseases across institutes and countries has been widely acknowledged, as well as the importance of community standards via which this can be achieved. The 15 FAIR Guiding Principles for scientific data management and stewardship as well as specific standards such as the Human Phenotype Ontology and the Orphanet Rare Disease Ontology have received the status of 'recognized resource' by the International Rare Disease Research Consortium (IRDiRC).

The ELIXIR Rare Disease community has subsequently made supporting the RD community in their FAIRification efforts a strategic goal. Within that a specific objective is to enable stewards to measure FAIR maturity in terms of the use of community standards while going through the steps of FAIRification. This applies to multiple types of resources relevant for rare diseases. In this talk we will present our latest outputs, prototype tests and services on developing FAIR infrastructure for FAIR evaluation specific to the RD community done in close collaboration with the EJP RD programme as a key FAIR application driver for the RD community.

ELIXIR talks

Using the IDP-KG to enable IDPcentral

Alasdair Gray (Heriot-Watt University), Ivan Mičetić (University of Padua) and Damiano Piovesan (University of Padua).

Abstract*:*

There are many data sources containing overlapping information about Intrinsically Disordered Proteins (IDPs). The ELIXIR IDP community (https://elixir-europe.org/communities/intrinsically-disordered-proteins) aims to have a central registry (IDPcentral – https://idpcentral.org/) that aggregates core data about disordered proteins. Traditional ETL approaches for populating IDPcentral would require the API and data model of each data source to be wrapped and then transformed into a common model.
The IDP Knowledge Graph (IDP-KG – https://alasdairgray.github.io/IDP-KG/) has been constructed by harvesting and aggregating Bioschemas markup available from the webpages of three IDP data sources; DisProt, MobiDB, and PED. Overlapping protein entries have been consolidated into a single node within the graph, resulting in the IDP-KG containing 2,701 disordered proteins together with their annotations. A REST API has been provided over the IDP-KG to enable the IDPcentral web application; providing methods to search for and retrieve core information about disordered proteins. The IDP-KG contains provenance data enabling links to the original sources of the data, and therefore retrieval of richer content that goes beyond the scope of a central registry. Queries over the IDP-KG endpoint have enabled insights into the state of data about IDPs across the three data sources. For example, it has been discovered that there are 663 distinct proteins predicted to be disordered that have not been manually curated. We plan to expand the number of community sources that the IDP-KG consumes from to increase the relevance to the IDP community.

ELIXIR LATAM

ELIXIR Europe: overview and opportunities

Andrew Smith (ELIXIR).

Abstract*:*

ELIXIR Europe coordinates bioinformatics services across its Members. Most of these services are available to anyone in the world to access, free of charge. This talk will introduce Latin American participants to ELIXIR, highlighting the activities they can engage in and the services they can benefit from.

ELIXIR LATAM

Research data management (RDM) in ELIXIR and insight into the RDM Toolkit

Frederik Coppens (Ghent University) and Carole Goble (The University of Manchester).

Abstract*:*

The ELIXIR RDMkit (https://rdmkit.elixir-europe.org) is a toolkit built by the biosciences community for the biosciences community to provide the RDM information they need. It is a framework for advice and best practice for RDM and acts as a "hub" of RDM information, with links to registries for tools, training materials, standards, and databases, and to services that offer deeper knowledge for DMP planning and FAIRification practices.

Launched in March 2021, over 120 contributors have provided nearly 100 pages of content and links to 300+ tools. Content covers the data lifecycle and specialised domains in biology, national considerations and examples of "tool assemblies" that have been developed to support RDM.

In this talk we will present the RDMkit - it's aims and context; it's content, community management and how folks can contribute; and our future plans and potential prospects for international cooperation.

ELIXIR LATAM

Creating paths for the development and application of bioinformatics in Mexico

Irma Martínez-Flores, Shirley Alquicira-Hernández and Alejandra Eugenia Medina-Rivera

Abstract*:*

Since 2006, the Center for Genomic Sciences-UNAM has carried out actions to improve bioinformatics skills, mainly for students but also for researchers, technicians, and the general public, through workshops, seminars, and courses, among others, promoting bioinformatics in Mexico and Latin America. To achieve this, several efforts were made, first at the local level (2006), then at the national level (2010), and as of 2012, at the international level, but all focused on a Spanish-speaking public.

To integrate all these initiatives, in 2018, the National Bioinformatics Network (RMB) was born as an organization that promotes the organized collaboration of all groups. Currently, the RMB has 761 members; we have collaboration alliances with nine national and international entities; and we get the support of several organizations such as Code for Science & Society, among others. We carry out permanent events and activities for the benefit of the members.

Institutional talks

Ersilia, a hub of open-source AI/ML models for drug discovery and global health

Gemma Turon (Ersilia Open Source Initiative) and Miquel Duran-Frigola (Ersilia Open Source Initiative).

Abstract*:*

The Ersilia Open Source Initiative (EOSI) is a non-profit organisation whose mission is to strengthen the research capacity in low- and middle-income countries. In particular, EOSI is focused on disseminating and deploying artificial intelligence and machine learning (AI/ML) tools for drug discovery as a means to minimise the cost and number of laboratory experiments. The main asset of EOSI is the Ersilia Model Hub, a free, online, open-source platform where scientists can browse through a catalogue of AI/ML models, select the ones that are relevant to their research and run predictions easily. We gather, in a single resource, two classes of models. On the one hand, we collect models developed by third parties and available in scientific publications. On the other hand, we develop models in-house and in collaboration with research groups that operate in the so-called Global South. In this presentation, we will explain how the Ersilia Model Hub is being deployed in the form of a fully-functional, comprehensive virtual screening cascade that is coupled with medicinal chemistry, parasitology and ADME experimental pipelines.

Institutional talks

Spanish Supercomputing Network (RES)

Alberto Antonio Gomez (Barcelona Supercomputing Center).

Abstract*:*

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 2022, the RES comprises 16 supercomputers, 14 institutions and 9 data management centres, 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 theprogress of excellent science and innovation in Spain.

Institutional talks

The Bioinfo4Women Programme: towards gender equity and diversity in science

Alba Jene-Sanz (Barcelona Supercomputing Center), María José Rementería (Barcelona Supercomputing Center), Eva Alloza (Barcelona Supercomputing Center, Spanish National Bioinformatics Institute (INB/ELIXIR-ES)) and Àtia Cortés (Barcelona Supercomputing Center).

Abstract*:*

The Bioinfo4Women programme (B4W) is an initiative that started in 2018 to promote the research done by women in computational biology, and it supports researchers by promoting the exchange of knowledge and experience of outstanding women researchers through activities such as seminars, conferences, training and mentorships. B4W has particular focus on the areas of personalised medicine, bioinformatics and HPC, and ultimately aims at building a more collaborative, supportive, and equal scientific community.

B4W kicked off a pilot mentoring programme in June 2022 to connect international, accomplished scientists as Mentors with junior researcher Mentees, providing role models for young researchers. Here we report on its implementation and on the feedbacks received from the training sessions and initial Mentor-Mentee meetings. The aim is to complete the pilot by May 2023, set the ground for more ambitious implementations, and expand towards providing guidance to postdoctoral researchers in their transition to independent researcher roles.

Sponsored talks

Building Biomedical Knowledge Graphs for In-Silico Drug Discovery

Tomas Sabat (Vaticle) and Wejdan Ismail (Vaticle).

Abstract*:*

The rapid development and spread of analytical tools in the biomedical sciences has produced a variety of information about all sorts of biological components and their functions. Though important individually, their biological characteristics need to be understood in relation to the interactions they have with other biological components, which requires the integration of vast amounts of complex, semantically-rich, heterogenous data.

Traditional systems are inadequate at accurately modelling and handling data at this scale and complexity, making solutions that speed up the integration and querying of such data a necessity.

In this talk, we present various approaches being used in organisations to build biomedical computational pipelines to address these problems using artificial intelligence and TypeDB. In particular, we discuss how to create an accurate and scalable semantic representation of molecular level data by presenting examples from drug discovery, precision medicine and competitive intelligence.

Technical Secretariat

Pl. Europa, 17-19 1st floor
08908 L'Hospitalet de Llobregat Barcelona, Spain
Ph: +34 93 882 38 78
eccb2022@bcocongresos.com